orban.tamas_188_24

MTA doktori értekezés

„Nem-kódoló” és invazív elemek az eukarióta genomokban: szerepük és orvosbiológiai hasznosításuk

Orbán Tamás István

Molekuláris Élettudományi Intézet

MTA
Kiváló Kutatóhely

Budapest

2024

TARTALOMJEGYZÉK

RÖVIDÍTÉSEK JEGYZÉKE

ACTC	szívspecifikus alfa-aktin
Adh	alkohol-dehidrogenáz
Ago	Argonaute fehérjecsalád
Amp	ampicillin
Aub	Aubergine (fehérje)
C14MC	a 14. kromoszómán elhelyezkedő humán miRNS klaszter
C19MC	a 19. kromoszómán elhelyezkedő humán miRNS klaszter
CAG	mesterséges hibrid promóter, a CMV vírus *immediate early* enhanszer, a csirke β-aktin promóter, és a nyúl β1-globin promóter szekvenciák fúziójából
CAPG	*actin filament capping protein, gelsolin like*
CDS	kódoló szekvencia (*coding sequence*)
ChIP	kromatin immunprecipitácó (*chromatin immunoprecipitation*)
CMV	citomegalovírus
CMV IE	citomegalovírus *immediate early* (enhanszer)
CPM	*count per million* (új generációs szekvenálás értékelésénél)
CRD	ciszteinben gazdag domén (*cystein-rich domain*)
CSB	*Cockayne Syndrome group B*
DFX	deferasirox
DGCR8	*DiGeorge Syndrome Critical Region 8*
DHS	DNáz I hiperszenzitív hely
dsRNS	duplaszálú RNS
EF1α	elongációs faktor 1α
(E)GFP	zöld fluoreszkáló fehérje ((*Enhanced*) *Green Fluorescent Protein*)
ENCODE	*Encyclopedia of DNA Elements*
ESC	embrionális őssejt (*embryonic stem cell*)
Exc	excízió (transzpozon kivágódás)
gDNS	genomiális DNS
HUES	humán embrionális őssejtvonal
hyPB	hiperaktív *piggyBac* transzpozáz
hyPBdel	első 100 aminosav hiányos hyPB mutáns
iPS	indukált pluripotens őssejt

LAM-PCR	*Linear-amplification mediated*-PCR
lncRNS	hosszú nem-kódoló (*long non-coding*) RNS
MEF	egér embrionális fibroblaszt (*mouse embryonic fibroblast*)
miRNS	mikroRNS
mPB	emlős kodon-optimalizált *piggyBac* transzpozáz
mPBdel	első 100 aminosav hiányos mPB mutáns
MSCL	mezenchimális őssejtvonal
Mtn	metallothionein
mutPB	katalitikusan inaktív *piggyBac* transzpozáz
NPC	neuronális progenitor sejt
NPPA	*natriuretic peptide precursor A*
NSD	*„nonstop" decay*
NTDR	N-terminális rendezetlen régió (*N-terminal disordered region*)
PAX6	*paired box gene 6*
PCR	polimeráz láncreakció
PB	*piggyBac* transzpozáz
PGBD	*piggyBac-derived* szekvencia
PGK	foszfoglicerát-kináz
piRNS	Piwi-asszociált kisRNS
Piwi	*P-element induced wimpy testis* (gén és fehérje)
PLN	*phospholamban*
pre-miRNS	prekurzor miRNS
pri-miRNS	elsődleges (*primary*) miRNS átirat
Puro	puromicin
RDRP	RNS-függő RNS polimeráz
RISC	RNS-indukálta csendesítési komplex (*RNA-Induced Silencing Complex*)
RITS	RNS-indukálta transzkripciós csendesítés (*RNA-Induced Transcriptional Silencing*)
RNAi	RNS interferencia (*RNA interference*)
RT	reverz transzkripció
qRT-PCR	kvantitatív, reverz-transzkripcióval egybekötött PCR
SB	*Sleeping Beauty* transzpozáz
SBmut	katalitikusan inaktív *Sleeping Beauty* transzpozáz
Seq	szekvenálás

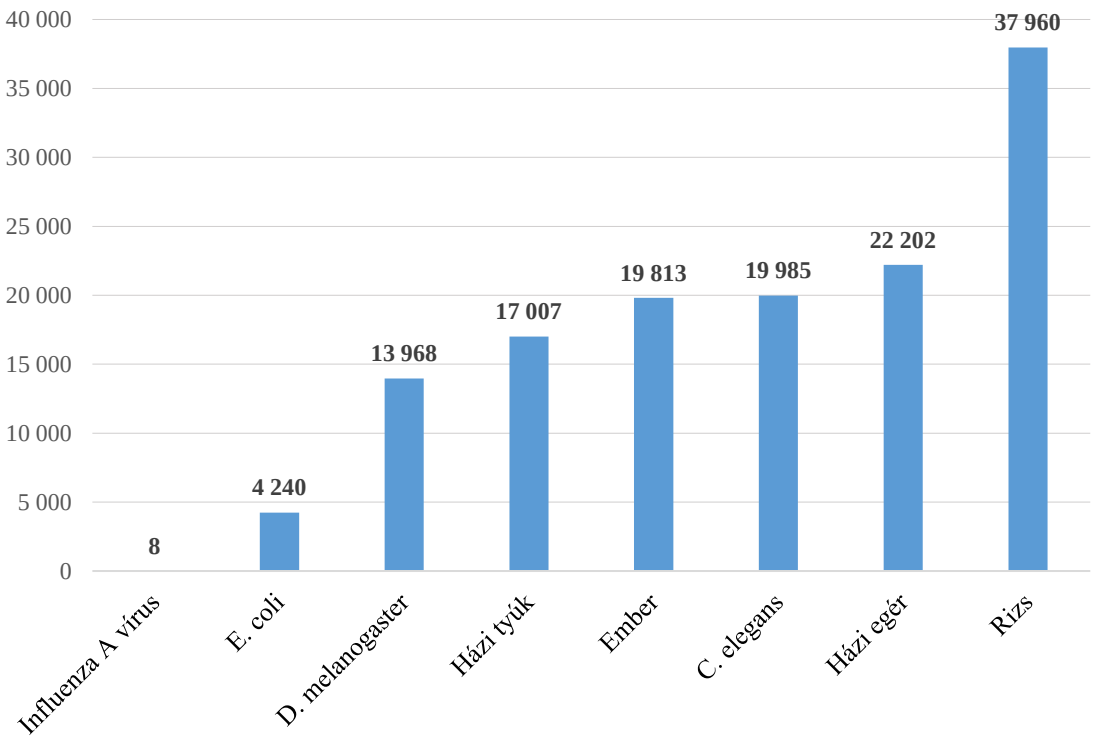
siRNS	kis interferáló RNS (*small interfering RNA*)
TIR	terminális fordított repetitív szekvencia (*terminal inverted repeat*) (LE – baloldali (*left end*), RE – jobboldali (*right end*))
TSD	célszekvencia duplikáció (*target site duplication*)
TSS	transzkripciós indítóhely (*transcription start site*)
UTR	nem transzlálódó régió (*untranslated region*)
VEGFA	*Vascular Endothelial Growth Factor A*

BEVEZETÉS, IRODALMI ÁTTEKINTÉS – A „NEM-KÓDOLÓ” GENOMRÓL RÖVIDEN

„Egy modern felfedezés azért felfedezés, mert ellentétben áll az összes ismert tudással. Ha nem áll ellentétben, akkor csak egy kis adalék.”

(Szent-Györgyi Albert)

A XX. század utolsó két évtizedében indított genom projektek azzal a reménnyel kecsegtettek, hogy egy élőlény örökítőanyagának, a dezoxiribonukleinsav (DNS) bázissorrendjének a meghatározásával egyértelműen fel lehet térképezni a működést biztosító géneket és szabályozó elemeket. Ez a kijelentés ugyan nyilván megállja a helyét, az eredmények azonban sok esetben sem a várt gének számában, sem ezek egy genomon belüli arányában nem egyeztek meg a korábbi predikciókkal és várakozásokkal. Ennek az egyik nyilvánvaló oka az volt, hogy a megelőző időszak kutatásai alapján géneknek legfőképpen a „kódoló” (értsd fehérjekódoló) szekvenciákat tekintették, bár az alaposabb definíciók megemlítették a már akkor is ismert RNS-kódoló géneket, mint például a riboszómális (rRNS) vagy a transzfer RNS (tRNS) géneket. A humán genom esetében például az akkor ismert fehérje repertoár alapján a gének számát 100 000 körülire becsülték, az emberi genomszekvencia első „piszkozata” azonban elég nagy csalódást okozott: mindösszesen ~30 000 fehérjekódoló gént sikerült azonosítani, és ami még megdöbbentőbb volt, hogy ezek a teljes genomnak csak egy kis töredékét, kevesebb mint 5%-át tették ki (Lander és mtsai, 2001; Venter és mtsai, 2001). Az ezt követő vizsgálatok és az egyre pontosabb annotálások a fehérjekódoló gének számát tovább csökkentették, és a legfrissebb adatok alapján ezek száma az emberi genomban kicsivel 20 000 alattira tehető (Nurk és mtsai, 2022). A számok több szempontból is zavarba ejtőek, hiszen az emberi gének száma így arányaiban nem sokkal múlja felül a kevésbé „komplexnek” gondolt modellorganizmusok, így például a gyümölcslégy (*Drosophila melanogaster)* vagy a házi tyúk (*Gallus gallus domesticus*) génjeinek a számát, nagyon hasonló a laboratóriumi fonalféreg (*Caenorhabditis elegans*) génjeinek a számához, illetve jócskán elmarad sok más eukarióta élőlény, így például a rizs (*Oryza sativa*), vagy akár a házi egér (*Mus musculus*) genomjában leírt gének számától (**1. ábra**). Ezek az eredmények nagyon sok kérdést vetettek fel, és rengeteg további kutatási irányt jelöltek ki, amelyek részben rávilágítottak a genomiális működés több, eddig ismeretlen aspektusára.



1. ábra: A fehérjekódoló gének száma néhány kiválasztott élőlény példáján. 2023. január 10-i állapot, forrás: http://ensemblgenomes.org/.

Az egyik legfontosabb kérdés a detektált és leírt fehérjék száma és az őket kódoló gének száma közötti jelentős különbség: milyen mechanizmusok teszik lehetővé, hogy egy gén egyszerre többféle, esetenként jelentős számú, és sokszor eltérő funkciójú fehérjét kódoljon? Az elmúlt évtizedek molekuláris biológiai kutatásai sokféle, ebben a folyamatban szerepet játszó mechanizmust tártak fel, amelyek közül genomi szinten kétségtelenül az alternatív splicing jelensége tűnik a leginkább meghatározónak (Baralle és Giudice, 2017; Shenasa és Hertel, 2019). Az eukarióta gének általában „hasított” gének, vagyis a genomban található szekvencia nem folyamatosan kódolja az információt (például a kódolt fehérje szekvenciáját), hanem a kódoló szakaszokat (exonok) nem-kódoló intronok szakítják meg. A DNS-ről képződő elsődleges átiratból tehát az intronok kivágódnak, és az exonok nélkülük illesztődnek össze: ez a folyamat a splicing, amelynek során létrejön a hírvivő RNS (*messenger* RNS, mRNS) végső szekvenciája, amely további folyamatok következtében módosul, és válik „érett”, transzláció képes mRNS-sé (Schmid és Jensen, 2010; Darnell, 2013). A valóság azonban ennél még bonyolultabb, ugyanis az exonok használata és összeillesztési sorrendje egy adott génen belül is változhat, és az alternatív splicing folyamata kapcsán így egy génről több, extrém esetben

akár 38 000 (!) különféle fehérjetermék is létrejöhet (Nilsen és Graveley, 2010). Természetesen a fehérjék sokfélesége egyéb folyamatok, például alternatív leolvasási keretek használatával, vagy különféle poszttranszlációs módosításokkal is növelhető, ami szintén magyarázatot adhat arra, hogy egy adott élőlény bonyolultságát a genom által kódolt gének számán túl más tényezők is befolyásolják. Ugyanakkor mivel a legfrissebb genom annotálási adatok alapján az emberi fehérjekódoló gének túlnyomó többségében (>95%) kimutattak többféle érett mRNS variánst, az alternatív splicing folyamata biztosan jelentős szerepet játszik a gének által kódolt fehérjék diverzitásában és evolúciójában (lásd például az *exon shuffling* jelenségét), így a „komplexitás" növelésének folyamatában (Patthy, 2021; Marasco és Kornblihtt, 2022).

A genom szerkezetével kapcsolatos problémák ugyanakkor továbbra is fennállnak: miért van az, hogy a fehérjekódoló gének szekvenciái az emberi genomnak megközelítőleg csak ~2%-át, az exonoknál általában jóval hosszabb intronokkal együtt is csak ~5%-át teszik ki? (Pennisi, 2005) Mi van a géneken „túl"? A genom nem-kódoló részei pusztán haszontalan „szemétnek", „*junk*" DNS-nek tekinthetők? A kérdés megválaszolására irányuló kutatások egyre több esetben igazolták, hogy a „nem-kódoló", vagyis a nem-fehérjekódoló régióknak is nagyon fontos szerepe van a genetikai anyag szabályozásában, a gének regulációjában. Az első erre utaló jelek megint csak a splicinghoz köthetők: kiderült, hogy léteznek nem-kódoló exonok is, amelyek gyakran a fehérjekódoló gének 5' végén találhatók, és amelyek szerepet játszhatnak a mRNS stabilitásában, vagy a transzláció hatékonyságának szabályozásában (Kimura és mtsai, 2006; Rojas-Duran és Gilbert, 2012; Sandor és mtsai, 2016). Mivel alternatív exonok például rRNS kódoló génekben is előfordulnak, egyre több kutatás vette célba a genom „nem-kódoló" régióit, és az intronokon, alternatív nem-kódoló exonokon kívül a „géneken túli régiók" is egyre inkább az érdeklődés középpontjába kerültek (Zerbino és mtsai, 2020; Gates és mtsai, 2021).

A több párhuzamos szálon elindult kutatások első igazán izgalmas, de – a genom projektek első szakaszának lezárultához hasonlóan – zavarba ejtő eredményeit minden kétséget kizáróan az ENCODE (***Enc****yclopedia* ***of*** ***D****NA* ***E****lements*) projekt eredményei szolgáltatták. 2012-ben egyszerre nagyjából 30 közlemény jelent meg rangos nemzetközi folyóiratokban (Consortium, 2012), amelyek a projekt különböző aspektusainak részleteit mutatták be. Nagyszámú sejtvonal igen sokféle vizsgálata alapján a projekt legfontosabb üzenete az volt, hogy az emberi genom minimum 80%-áról történik valamilyen sejttípusban génátírás – hozzátéve, hogy az ebből a szempontból „legaktívabb" sejtekben

is egyszerre maximum a genom ~57%-áról történik transzkripció (Djebali és mtsai, 2012). A kutatók ugyanakkor hangsúlyozták, hogy a nagyszámú átirat jelentős részéhez egyelőre nem tudnak „hagyományos" biokémiai funkciót társítani, de az eredmények tükrében mégis fontosnak tartották, hogy a gén fogalmán finomítsanak. A definíció szerint továbbra is a DNS egy adott szakaszát tekintjük génnek, amely komplementer RNS molekulákat, transzkriptum izoformákat kódol, de a továbbiakban magukat a transzkriptumokat kell az öröklődés alapvető egységeinek tekinteni. A gén fogalma ehhez kapcsolódva pedig egy magasabb rendű egységet takar: adott transzkriptumok hálózatát, amelyek együttesen alakítanak ki egy fenotípust (Djebali és mtsai, 2012). Az eredmények interpretálásának azonban jócskán akadtak kritikusai is: a szkeptikusok fő érvei azon alapultak, hogy jelenleg nagyon nehéz definiálni a „biológiai zaj" fogalmát, és ismert funkció hiányában egy DNS szakasz esetleges átíródása okán azt nem lehet feltétlenül biológiailag fontos működésként interpretálni (Graur és mtsai, 2013). Kétségtelen tény, hogy az ENCODE projekt a rengeteg adat mellett sok esetben nem szolgált magyarázattal új funkciókra, ugyanakkor a technológia robbanásszerű fejlődése kapcsán elérhetővé vált új metodikákkal igenis új utakat és irányokat szabott a molekuláris genetikai kutatásoknak.

Mi van tehát a géneken túl, mit találunk a nem-kódoló régiókban? A molekuláris biológia fejlődésére sokszor jellemző volt az, hogy a kutatók egy modellorganizmus egy szokatlanul hangsúlyos tulajdonságát vizsgálva új mechanizmusokat tártak fel, amelyek még akkor is hasznosnak és úttörőnek bizonyultak, ha a későbbiek során nem minden élőlényben, vagy nem olyan mértékben sikerült igazolni azok jelenlétét. Gyakorta előfordult persze az is, hogy elszigeteltnek tűnő jelenségek érdekes kuriózumként lettek interpretálva, amelyekről később kiderült, hogy nagyon fontos, eddig nem ismert vagy technikailag még nem vizsgálható, de általános mechanizmusokat takarnak. A számtalan példa közül talán elég csak említeni az 1977-ben felfedezett splicing folyamatát (Berget és mtsai, 1977; Chow és mtsai, 1977), vagy a mikroRNS-ek (miRNS-ek) 1993-as felfedezését (Lee és mtsai, 1993; Wightman és mtsai, 1993). Az elmúlt két évtizedben a technológiai forradalomnak köszönhetően egyre több ilyen eredményre támaszkodva egyre több, a genom projektek és az ENCODE projekt által felvetett kérdést sikerült megválaszolni, és egyre jobban megismerni a korábban a kozmológiai kutatások mintájára „sötét anyagnak" is elnevezett genomiális régiókat (Johnson és mtsai, 2005). Ma már tudjuk, hogy nagyon sok genom meghatározó részét, a repetitív szekvenciák egy jelentős hányadát mobilis genetikai elemek, a transzpozonok teszik ki. Ezek között

vannak a mai napig aktív molekuláris paraziták, vannak az evolúciós folyamatok során inaktívvá vált „halott” szekvenciák, és vannak olyan elemek, amelyek a gazdagenom szempontjából hasznosulva endogén funkciókra domesztikálódtak (Volff, 2006; Feschotte és Pritham, 2007; Jangam és mtsai, 2017; Cosby és mtsai, 2019). Az emberi genom esetében például a genetikai állomány megközelítőleg 45%-a tekinthető igazoltan transzpozon szekvenciának, ugyanakkor nagyon sok domesztikálódott transzpozáz (például a RAG rekombinázok, centroméra fehérjék, stb.) példáján látszik, hogy ebben a tekintetben egy kicsit kezd elmosódni a határ a fehérjekódoló és a „nem-kódoló” genomiális szakaszok között (Alzohairy és mtsai, 2013; Morales Poole és mtsai, 2017). A fehérjekódoló kapacitással önmagában nem rendelkező, nagyszámú Alu-szekvencia között is szép számmal akadnak olyan elemek, amelyek bizonyítottan fontos evolúciós szerephez jutottak az alternatív splicing és az *exon shuffling* folyamatában, így sokszor egyértelműen hasznos, funkcionális elemeknek tekinthetők (Hurst és Werren, 2001; Chuong és mtsai, 2017; Burns, 2020; Almeida és mtsai, 2022). Mindezek az eredmények azt mutatják, hogy a genomokat mindenképpen egy evolúciós szempontból dinamikus rendszerként kell kezelnünk, és ezt a szempontot a genomok funkcionális annotálása során is érvényesíteni kell.

A genomiális „sötét anyag” (Johnson és mtsai, 2005) feltérképezésének egyik legizgalmasabb új vetületét a nagyszámú RNS-kódoló gén jelenti. Ma már tudjuk, hogy a genomokban jelentős számban találunk szabályozó funkciójú, úgynevezett kisRNS géneket, amelyek az RNS interferencia különböző útvonalaihoz kapcsolódnak. A belső szabályozási hálózatot képviselő miRNS gének mellett ide tartoznak az invazív genetikai elemek (például transzpozonok vagy vírusok) elleni védekezésben szerepet játszó kis interferáló RNS-ek (*small interfering* RNS-ek, siRNS-ek), illetve a csíravonal sejtjeiben meghatározó genomiális védelmet ellátó Piwi-asszociált RNS-ek (piRNS-ek) (Ghildiyal és Zamore, 2009; Gebert és MacRae, 2019; Onishi és mtsai, 2021; Kong és mtsai, 2022; Wang és mtsai, 2022). Az utóbbi évek kutatásai egyre több nagyobb méretű (>200 nukleotid), hosszú nem-kódoló RNS-t (lncRNS-t) írtak le, amelyek a DNS hibajavító folyamatokban való részvételtől kezdve a sejtorganellumok membránokkal nem körülhatárolt, fázis-szeparációval mégis elválasztott régióinak „állványmolekuláin” át változatos funkciókon keresztül vesznek részt nagyon sok sejtélettani folyamatban (Ransohoff és mtsai, 2018). Az RNS-kódoló gének ilyen repertoárja véleményem szerint nemcsak egyfajta bizonyíték a molekuláris evolúció kezdetén létező „RNS-világra”, hanem arra is rávilágít, hogy a fehérjék mellett az RNS molekulák a mai napig aktív

részesei és alakítói a sejtek és a többsejtű élőlények világának, és meghatározó részét képezik az élőlények genomjainak is. Ennek alátámasztására az említett példákon túl elég csak elkezdeni katalogizálni a sejtekben működő ribonukleoprotein-komplexeket (riboszómák, spliceoszómák, telomeráz, RNáz P komplex, szignál felismerő részecskék, stb.), vagy számba venni a létező RNS vírusokat, amelyek egyik képviselője épp most kényszerítette térdre az emberi fajt is…

Kutatásaim során nagyon sokat foglalkoztam a nem-kódoló elemek genetikai szabályozásban betöltött szerepével. A PhD éveket követően is megmaradt az érdeklődésem a splicing folyamata iránt, és mindig lenyűgözött az, hogy az RNS molekulákhoz köthető szabályozási folyamatok, például az RNS interferencia megismerése egyben milyen fantasztikus eszköztárat is biztosít a genetikai vizsgálatok újszerűbb és gyorsabb tervezéséhez és elvégzéséhez. Posztdoktori kutatásaim második szakaszától kezdődően lehetőségem volt behatóbban foglalkozni a transzpozonokkal, megismerni az „önző" genetikai elemek működésének részleteit, és azt, hogy hogyan lehet a „megszelídített" transzpozonokat génbeviteli eljárásokban biztonságosan alkalmazni. A későbbiekben kutatócsoport vezetőként alkalmam nyílt olyan folyamatokat is behatóbban vizsgálni, ahol igyekeztünk megérteni, hogy a gazdagenom miként képes az invazív elemeket nemcsak hogy féken tartani, hanem azok működését saját szolgálatába állítva domesztikálni azokat. Azt gondolom, szerencsésnek mondhatom magam abban, hogy az elmúlt években a munkám során olyasmikkel foglalkozhattam, amelyek mindig is érdekeltek, és a fent említett mechanizmusok vizsgálatával lehetőségem nyílt molekuláris szinten megismerni az evolúciós folyamatokat és az élet működését. A jelen értekezésben bemutatom azt a két fő kutatási irányt, az RNS interferencia folyamatainak, illetve a DNS transzpozonok működésének a vizsgálatát, amely témakörök meghatározták az elmúlt időszakban végzett tudományos tevékenységemet. Arra törekedtem, hogy az egyes részek önmagukban is érthető és élvezhető egységek legyenek, ezért az adott területek irodalmi hátterét mindig a megfelelő fejezetek elején tekintem át. Az itt bemutatott munkákkal az izgalmas tudományos felfedezések ismertetésén túl azt is igyekszem demonstrálni, hogy a nem-kódoló elemekhez köthető szabályozási folyamatok megismeréséhez, és az ezen a tudáson alapuló orvosbiológiai hasznosításhoz néhány szeletkével talán nekem is sikerült hozzájárulnom.

CÉLKITŰZÉSEK

A PhD fokozat megszerzése óta eltelt időszakban kutatásaim a genomok „nem-kódolónak” vagy „invazívnak” nevezett régióira irányultak. Igyekeztem olyan tudományos kérdéseket feltenni, amelyekkel nemcsak a vizsgált RNS és DNS szekvenciák evolúciósan kialakult funkcióira derülhet fény, hanem esetleg lehetőség nyílik ezen elemek molekuláris genetikai vagy orvosbiológiai felhasználására is. A jelen disszertációban bemutatott vizsgálatokat alapvetően két témakörben folytattam, az ezekhez kapcsolódó kutatási célok pedig a következő pontokban fogalmazhatók meg:

1. Az RNS interferencia különböző folyamatainak a megértése, ezen belül:

❶ az siRNS útvonal által kiváltott RNS degradáció molekuláris mechanizmusának a feltárása;

❷ egy speciális miRNS csoport, a splicing folyamatával összekapcsolt mirtron útvonal feltérképezése emlős rendszerekben;

❸ a miRNS-ek érése során a karhasználat és az izomiR-ek képződésének vizsgálata;

❹ a klaszterekben elhelyezkedő miRNS-ek speciális érési folyamatainak a vizsgálata.

2. A DNS transzpozonok működésének és evolúciójának részletes vizsgálata, ezen belül:

❶ a *piggyBac* elemek működésének és domesztikációjának vizsgálata különböző genomokban;

❷ a DNS transzpozonokkal végrehajtott génbevitel molekuláris jellemzői a *Sleeping Beauty* és a *piggyBac* rendszerek példáján;

❸ transzgenikus sejt- és állatmodellek létrehozása és részletes karakterizálása DNS transzpozonok segítségével.

EREDMÉNYEK ÉS MEGVITATÁSUK

„There is no last word in science. It is a living conversation."
(Prof. Manuel Ares, Jr. – UC Santa Cruz)

1. Az RNS interferenciához köthető nem-kódoló elemek

Tudománytörténeti szempontból a különféle RNS interferencia (*RNA interference*, RNAi) útvonalak felfedezése egymástól függetlenül történt. Először növényekben figyeltek meg „szokatlan" géncsendesítési folyamatokat, amelyekről csak később igazolódott, hogy kisRNS-ekhez köthető szabályozásról van szó. A kutatók a különféle színanyagok szintéziséért felelős enzimek génjeinek bejuttatásával petúniákat szerettek volna színesebbé, vagy más színűvé változtatni, azonban a génbevitel hatására az endogén gének is csendesültek, ami végül színtelen virágok kialakulásához vezetett (Napoli és mtsai, 1990). Hasonló jelenséget gombákban is megfigyeltek (Romano és Macino, 1992), majd az első miRNS-ek leírását követően (Lee és mtsai, 1993; Wightman és mtsai, 1993) pár év múlva jelent meg az siRNS rendszert szisztematikusan bemutató, korszakalkotó közlemény: a duplaszálú RNS (dsRNS) molekulák kiváltotta géncsendesítést „RNS interferenciának" nevezték el, és ennek felfedezéséért később 2006-ban Andrew Z. Fire és Craig C. Mello megkapta az orvosi-élettani Nobel-díjat (Fire és mtsai, 1998). Néhány évnek azonban el kellett telnie ahhoz, hogy az előbbiekben röviden bemutatott kísérletek mögötti folyamatokról kiderüljön: lényegében egy hasonló, RNS-fehérje kapcsolatokon alapuló szabályozási rendszert képviselnek.

RNS interferencia alatt ma már több, kisRNS molekulákhoz köthető szabályozási folyamatot értünk. Ezek közös jellemzője, hogy olyan ribonukleoprotein-komplexek vesznek benne részt, amelyek rövid (20-35 nukleotid hosszúságú), egyszálú RNS molekulát tartalmaznak, és mindig megtalálható bennük az Argonaute fehérjecsalád egy képviselője (egy Ago vagy egy Piwi fehérje). Funkcionális szempontból ezek többségükben géncsendesítést végző szabályozó komplexek, hatásukat pedig transzkripcionális vagy poszttranszkripcionális szinten fejtik ki, így elnevezéseik is ehhez köthetők: a citoplazmában működő komplexek neve „RISC" (=***R****NA-****I****nduced* ***S****ilencing* ***C****omplex*), míg a sejtmagi komplexeké „RITS" (=***R****NA-****I****nduced* ***T****ranscriptional* ***S****ilencing*) komplex (Ghildiyal és Zamore, 2009). Az siRNS és a piRNS útvonalak közös jellemzője, hogy egyfajta molekuláris immunrendszerként működve invazív genetikai

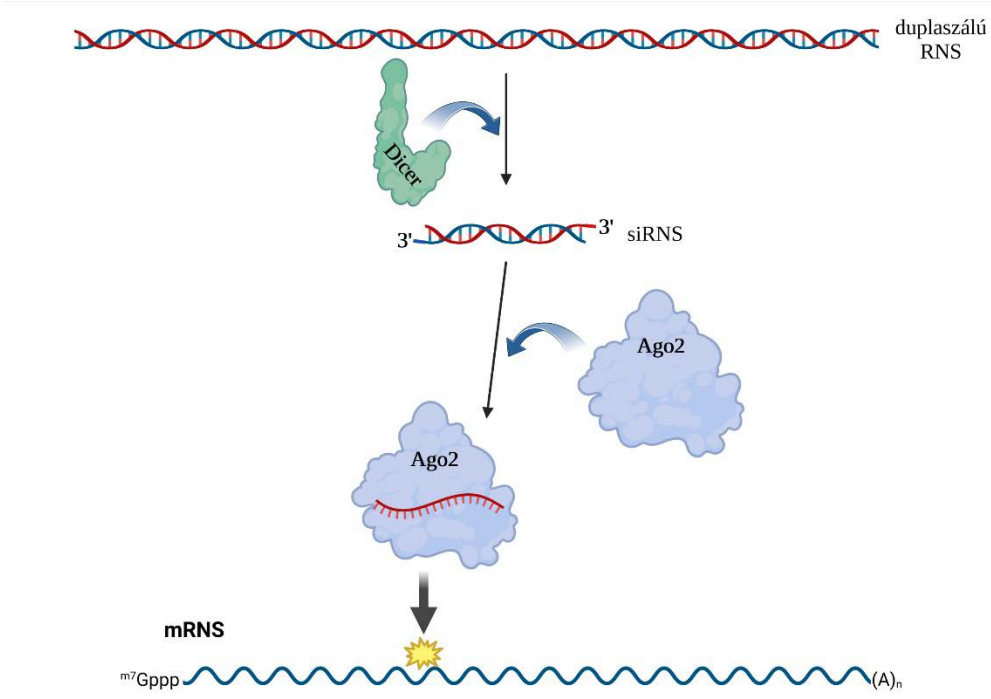
elemek, elsősorban vírusok vagy transzpozonok ellen védik a sejtek genetikai állományát. Míg az siRNS rendszer (vagy az ahhoz hasonló rendszerek) a legtöbb eukariótában megtalálhatók, addig a piRNS rendszer csak az állati szervezetekben fordul elő, és hangsúlyozottan expresszálódik a csíravonal sejtjeiben, őrizve a sejtek genetikai állományának stabilitását (Onishi és mtsai, 2021; Wang és mtsai, 2022). A miRNS rendszer ugyanakkor az evolúció során a gazdagenom saját génjeit reguláló endogén szabályozási hálózattá fejlődött ki, amely a legtöbb eukarióta sejtben előfordul, és ma már tudjuk, hogy komplexitásában összemérhető a transzkripciós faktorok alkotta szabályozási hálózattal (Chen és Rajewsky, 2007; Arora és mtsai, 2013; Gebert és MacRae, 2019; Trabucchi és Mategot, 2019).

Az RNAi folyamatok tehát a genom kódoló és nem-kódoló elemeinek finom összjátékán alapulnak: a célmolekulák kijelölését bázispárosodáson keresztül a kisRNS komponens végzi, míg a végső effektor funkció (például a cél-mRNS hasítása) viszont az Argonaute fehérje feladata. Az elmúlt években az RNAi útvonalak közül részletesen foglalkoztam az siRNS rendszer és a miRNS rendszer mechanizmusaival, különböző modellorganizmusokban vizsgálva a molekuláris működések sajátosságait. Az értekezés első fejezetében ebben a témában azokat a nemzetközi folyóiratokban közölt munkákat mutatom be, amelyekben meghatározó szerzőként vettem részt.

1.1. Az siRNS rendszer szerepe

Ahogy korábban már röviden említésre került, az siRNS rendszer egyfajta sejtszintű adaptív immunrendszerként működik: a sejtbe bejutó és aktiválódó invazív elemeket a citoplazmában megjelenő dsRNS molekulákon keresztül ismeri fel. Az ilyen molekulák vírusok vagy transzpozonok különféle életciklusaiban jelenhetnek meg, és a sejt számára idegen entitásnak számítanak. Amikor a 2000-es évek elején elkezdtünk ezzel az útvonallal foglalkozni, akkor a működés molekuláris mechanizmusának bizonyos részei még tisztázatlanok voltak (Dykxhoorn és mtsai, 2003). Annyit már lehetett tudni, hogy a dsRNS-t a citoplazmában a Dicer nevű RNáz fehérje ismeri fel, és egyfajta molekuláris „ollóként” a dsRNS-t feldarabolja 21 nukleotid hosszúságú siRNS molekulákra. Ezen kisebb vágási termékeket a pontos méretükön túl még az is jellemzi, hogy mindkét összetapadó szál 3’ végén egy 2 nukleotid hosszúságú, egyszálú túlnyúló véget tartalmaznak. Az siRNS-ek ezek után egy hosszabb érési folyamat kapcsán beépülnek az effektor komplexbe: a Dicer-ről leválva segédfehérjéken keresztül összekapcsolódnak

egy Argonaute fehérjével (például az Ago2-vel), majd további érési lépések során az egyik szál eliminálódik, és kialakul az aktív, Ago fehérjével asszociált, rövid egyszálú RNS-t tartalmazó RISC (**2. ábra**). A további hatásmechanizmusra is voltak már bizonyítékok: a RISC a benne található kisRNS segítségével letapogatja („szkenneli") a citoplazmában található RNS molekulákat, és ha tökéletes illeszkedést talál (megtalálja például a vírus RNS-ét), akkor megindítja a cél-RNS lebomlását. Indirekt kísérletes eredmények arra utaltak, hogy a degradáció egy endonukleolitikus hasítással veszi kezdetét (Elbashir és mtsai, 2001; Nykanen és mtsai, 2001; Martinez és mtsai, 2002; Martinez és Tuschl, 2004), de akkor még nem azonosították magát az endonukleázt (azóta már tudjuk, hogy az maga az Ago2 fehérje, (Liu és mtsai, 2004; Meister és mtsai, 2004; Rand és mtsai, 2004)), és nem volt ismert a degradáció pontos mechanizmusa sem. Mivel posztdoktori kutatásaim során RNS lebomlási folyamatokkal foglalkoztam, ezért szerettük volna alaposan feltérképezni ezt a mechanizmust, és azonosítani a degradációs útvonal molekuláris szereplőit.



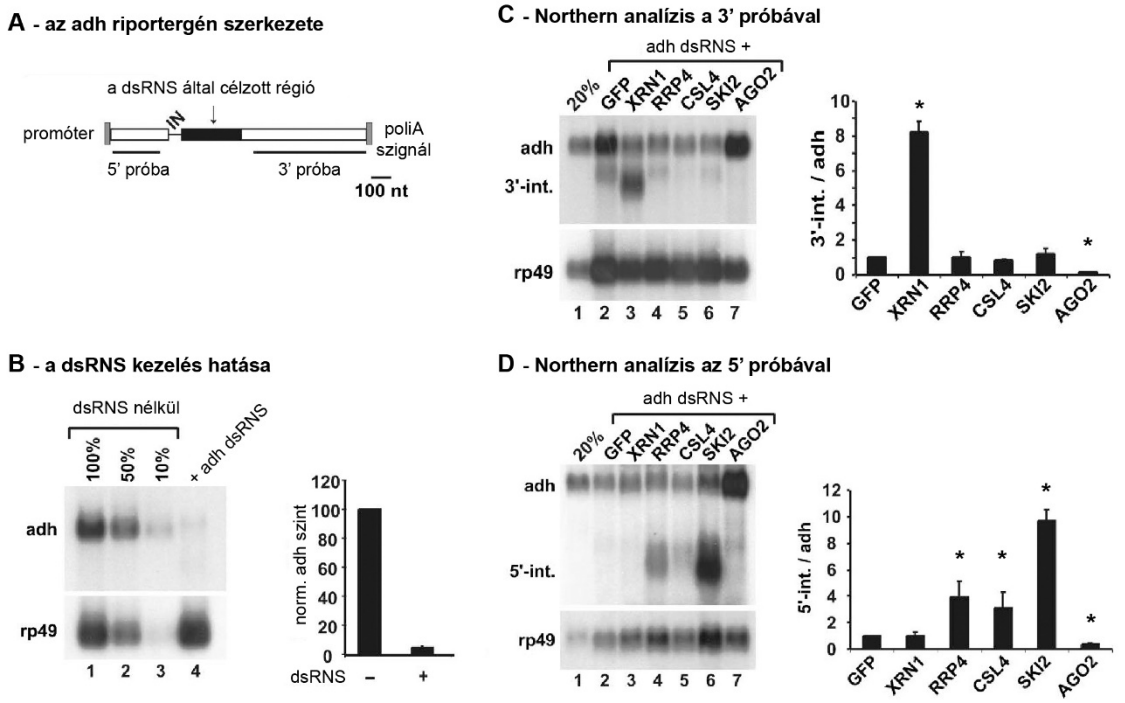
2. ábra: Az siRNS-ek képződésének egyszerűsített modellje. Az invazív elemek (például vírusok vagy transzpozonok) életciklusa során keletkező hosszú duplaszálú RNS-ek aktiválják az útvonalat. A Dicer enzim ezeket felismeri, és ~21 bázispár hosszú kis duplaszálú RNS-ekre (siRNS-ekre) hasítja. Az Ago2 tartalmú komplex (RISC) az egyik szálat eliminálja, majd a másik szál segítségével felismeri az invazív elem mRNS-ét, és elindítja annak lebomlását.
Az ábra a *biorender.com* alkalmazással készült.

1.1.1. A célzott RNS degradáció mechanizmusa

A kutatásokhoz a *Drosophila melanogaster* nagyon jó modellnek bizonyult: mivel ennek az élőlénynek nincs sejtes adaptív immunrendszere, ezért nagyon jól kifejezett, és a többi RNAi útvonaltól molekuláris szinten is nagyon jól elkülönült siRNS rendszerrel rendelkezik. A kísérletekhez létrehoztunk több S2 (Schneider) *Drosophila* sejtvonalat, amelyek különböző riporter konstrukciókat expresszáltak: voltak intront tartalmazó és nem tartalmazó riporterek, illetve ezeket konstitutív aktin promóterrel, vagy indukálható metallothionein (Mtn) promóterrel meghajtott konstrukciók. A kísérletek során az RNS interferenciát a riporter gének középső szekvenciájára specifikus, 600-700 bázispár hosszúságú dsRNS molekulák transzfektálásával váltottuk ki, majd a riporter konstrukciók 5’ és 3’ végeire specifikus próbákkal, *Northern* analízis segítségével vizsgáltuk a degradáció folyamatát (a próbák és a dsRNS között nem volt szekvenciális átfedés, **3A ábra**). A dsRNS transzfekcióját követően a ripotergének RNS-szintű expressziója jelentősen lecsökkent, konstrukciótól függően az eredeti szint 30-10%-a alá (**3B ábra**). Kérdés volt azonban, hogy az így detektált alacsony „*steady-state*” szintek a csökkent transzkripció, vagy a megnövekedett RNS lebomlás eredményei-e. Ennek eldöntésére a dsRNS-ek bejuttatását követően actinomycin D hozzáadásával gátoltuk a transzkripciót, majd meghatározott időintervallumok elteltével mintát vettünk a tenyészetből, és megmértük a riporter gének RNS-szintű expresszióját. A dsRNS hozzáadása nélkül a mért mRNS szintek hosszú ideig stabilak voltak, a féléletidejük 4 óránál is hosszabbnak bizonyult, ami *Drosophila* rendszerben nagyon stabilnak számít. Az RNAi kiváltását követően azonban az RNS szint nagyon hamar, már 15 perc elteltével a kiindulási szint kevesebb mint 20%-a alá csökkent. Ezzel a kísérlettel igazoltuk, hogy ebben a rendszerben a dsRNS által kiváltott hatás poszttranszkripcionális szinten, az RNS lebomlásán keresztül valósul meg. A következő kérdés tehát az volt, hogy milyen géntermékek vesznek részt a RISC által kijelölt cél-mRNS molekulák degradációjában.

A továbbiakban a kiválasztott és vizsgálandó fehérjék szintjének drasztikus csökkentését (de nem teljes eliminációját) szintén specifikus dsRNS hozzáadásával váltottuk ki (a „*knock out*” helyett a „*knock down*” módszerrel). Erre azért volt szükség, mert a legtöbb tesztelendő faktor általában esszenciális a sejt működése szempontjából, így a célzott inaktiváló mutációk helyett a gyorsabb és egyszerűbb géncsendesítést alkalmaztuk. Az RNAi szempontjából kulcsfontosságú kérdés az volt, hogy az akkor ismert négy *Drosophila* Argonaute fehérje közül ebben a folyamatban melyik vesz részt

a RISC komplex felépítésében. Vizsgálataink egyértelműen igazolták, hogy az RNS degradáció mechanizmusában az Ago2 játszik szerepet (**3C** és **3D ábrák**), míg az Ago1, a Piwi és az Aubergine (Aub) fehérjék csendesítése a riportergén expressziós szintjére nem volt hatással. Akkori kísérleteinkben azonban mi még nem tudtuk igazolni, hogy maga az Ago2 rendelkezik az endonukleáz („*slicer*”) aktivitással (Liu és mtsai, 2004; Meister és mtsai, 2004; Rand és mtsai, 2004). (Azóta már azt is lehet tudni, hogy az Ago1 fehérje *Drosophila*-ban a miRNS útvonal központi szereplője, míg a Piwi, az Aub, és a később felfedezett Ago3 pedig a piRNS útvonalban játszanak szerepet (Ghildiyal és Zamore, 2009).) A továbbiakban különböző RNS lebomlási útvonalak ismert komponenseit teszteltük: az 5’→3’ degradáció főszereplőjeként az Xrn1 exonukleázt (*Drosophila*-ban „*PacMan*” néven is ismert), míg a 3’→5’ degradációs folyamatok szereplői közül az exoszóma több komponensét (Rrp4, Csl4), illetve az útvonalban fontos szerepet játszó RNS-helikáz komplex, a Ski-komplex összetevőit (Ski2, Ski3, Ski8) vizsgáltuk. Az eredmények azt mutatták, hogy az Xrn1 csendesítése a dsRNS által megcélzott régiótól 3’ irányban elhelyezkedő degradációs intermediert stabilizálja, és a teljes hosszúságú riporter transzkriptummal szemben ennek az intermediernek a féléletidejét jelentősen megnöveli (**3C ábra**). Ugyanakkor, ha az exoszóma és a Ski-komplex komponenseit csendesítettük, akkor viszont az 5’ fragmentum stabilitása növekedett meg (**3D ábra**). Az eredmények arra utaltak, hogy a RISC által kiváltott endonukleázos hasítást követően a képződött fragmentumokat a sejt endogén RNS lebontó rendszerei eliminálják: az 5’ hasítási terméket az exoszóma, míg a 3’ terméket az Xrn1 bontja le.



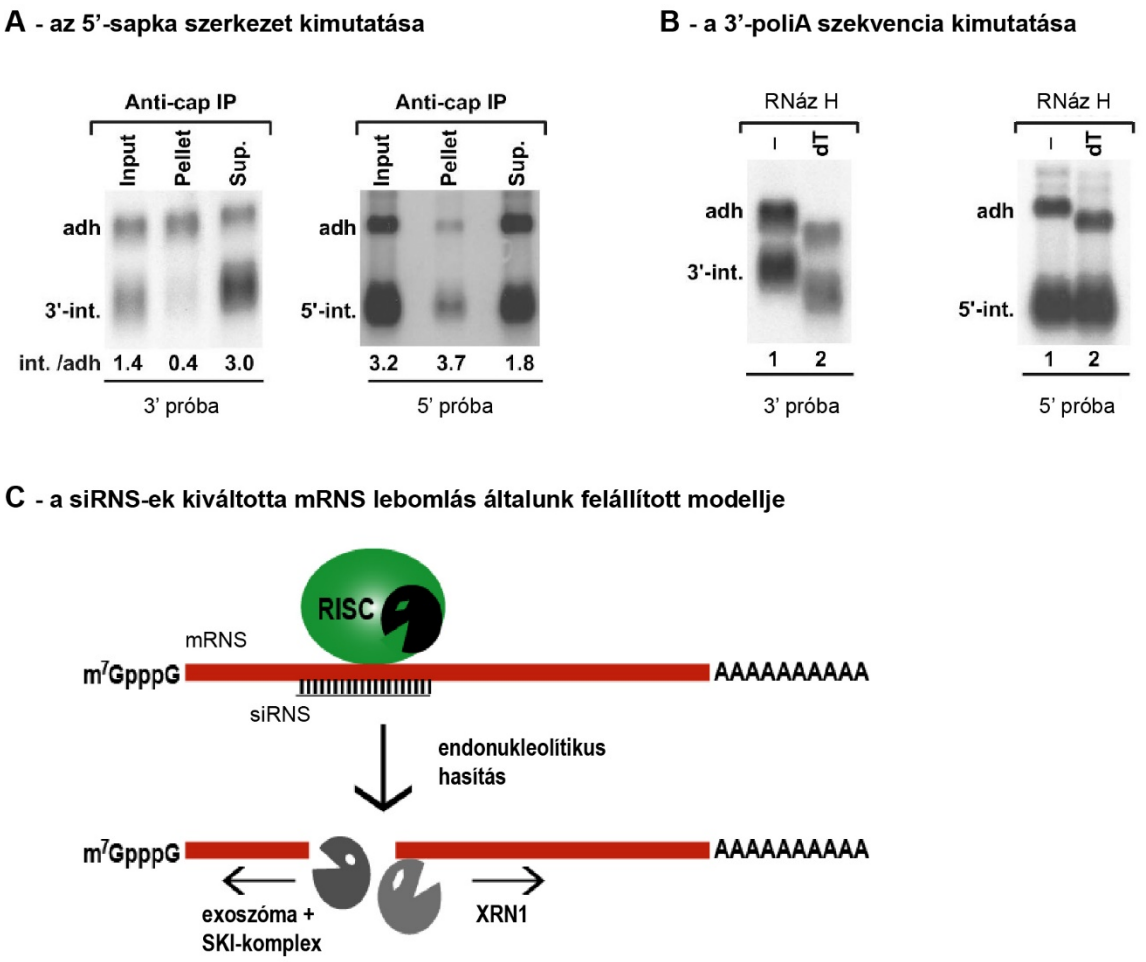
3. ábra: Az siRNS útvonal által kiváltott mRNS degradációban résztvevő faktorok tesztelése az adh riportergén példáján bemutatva. (**A**) Az alkohol-dehidrogenáz (adh) gén módosított, rövidebb szekvenciájából létrehozott riportergén szerkezete, IN: intron. (**B**) Az adh transzkriptum „*steady-state*" szintjének csökkenése a dsRNS kezelés hatására. A gél bal oldalán referenciaként egy kezeletlen minta hígítási sora van feltüntetve. (**C**) A *Northern* analízis során a 3' próbával a teljes hosszúságú transzkriptum mellett az esetlegesen megjelenő/stabilizálódó 3' degradációs intermedier (3'-int.) is kimutatható. A gél feletti génnevek mutatják, hogy az adott kísérletben a sejteket az adh riporterre specifikus dsRNS mellett milyen másik, a különféle degradációs útvonalakban szerepet játszó gének vagy a kontroll (zöld fluoreszkáló fehérje, ***G****reen* ***F****luorescent* ***P****rotein*, GFP) mRNS elleni dsRNS-sel kezeltük. A gél bal oldalán egy dsRNS kezelés nélküli sejttenyészet 20%-ra hígított RNS mintája szerepel referenciaként. A kísérletekből látszik, hogy a 3'-intermedier az XRN1 faktor hiányában stabilizálódik, vagyis eredetileg ez a fehérje felelős a lebontásáért. A jobb oldali grafikonon ezt mutatják a 3'-intermedier és a teljes hosszúságú transzkriptum arányának kvantitatív értékei is. (**D**) Az előzőhöz hasonló vizsgálatok szerepelnek az 5' próbával elvégezve. Az 5'-intermedier (5'-int.) lebontásában az exoszóma (pl. RRP4, CSL4) és a Ski-komplex (pl. SKI2) komponensei vesznek részt.

A grafikonokon legalább három független biológiai párhuzamos mérés átlagai szerepelnek, a hibazászlók a szórás (SD) értékeket mutatják. *: $p<0,05$; rp49: endogén kontrollként használt mRNS. Az ábra az (Orban és Izaurralde, 2005) közleményünk alapján készült.

Felmerült azonban a kérdés, hogy a mRNS-ek degradációjában szerepet játszó egyéb komponensek, például az 5’ „sapka” („5’-*cap*”) szerkezetet lebontó „*decapping*” enzimek, vagy a 3’ végen található poliA szekvenciát rövidítő deadenilázok nem vesznek-e részt a folyamatban. Ennek a kérdésnek a megválaszolását úgy végeztük el, hogy megvizsgáltuk a lebomlási intermedierek 5’ és 3’ végeinek szerkezetét. Az 5’ sapka kimutatásához specifikus antitestet használtunk: miután az exoszóma komponensek csendesítésével az 5’ intermediert stabilizáltuk, az antitesttel történt immunprecipitációt követően igazolni tudtuk, hogy az 5’ intermedieren még megtalálható az 5’-sapka szerkezet, így a lebontási folyamatokban a „*decapping*” enzimek nem vesznek részt (**4A ábra**). A 3’ fragmentum szerkezetét más módszerrel vizsgáltuk: az Xrn1 csendesítését követő stabilizáció után az izolált RNS mintákhoz rövid dT oligókat adtunk, majd a keveréket RNáz H enzimmel kezeltük, és azt követően *Northern* analízisnek vetettük alá. Az oligodT molekulák hibridizálódnak a hosszabb adenin tartalmú RNS szekvenciákhoz, a riporter konstrukciók esetén leginkább a 3’ végen található poliA-farokhoz. Az RNáz H enzim specifikus az RNS-DNS hibrid szekvenciákra, azokat lebontja, az egyszálú RNS régiókat viszont érintetlenül hagyja, így az emésztést követően a mRNS molekulák megrövidülnek. A kísérletek során pontosan ezt tapasztaltuk: a 3’ fragmentumok a kezelést követően lerövidültek, ami azt igazolta, hogy eredetileg tartalmaztak poliA szekvenciát, lebomlásukban tehát a deadenilázok nem vesznek részt (**4B ábra**).

Az előbbiekben leírt kísérletek kapcsán felállítottunk egy modellt, amely a RISC által kiváltott, endonukleázos hasítással kezdődő RNS degradáció általános modelljévé vált (Orban és Izaurralde, 2005). A komplexben található egyszálú kisRNS molekula bázispárosodáson keresztül kijelöli a lebontásra kijelölt cél-RNS molekulát, és az érintett szakaszon megtörténik a cukor-foszfát gerinc elhasítása. Az RNS lebomlása ezt követően a hasítási ponttól folytatódik: a 3’ intermedier lebontását az 5’→3’ exonukleáz aktivitással rendelkező Xrn1 fehérje végzi el, míg az 5’ intermedier lebomlásáért a 3’→5’ exonukleáz aktivitású exoszóma a felelős, amelynek a folyamathoz szüksége van az RNS helikáz aktivitással bíró Ski-komplexre is (**4C ábra**). A folyamat nem csak az siRNS rendszer sajátossága: növényekben, ahol az állati rendszerekkel ellentétben a miRNS-ek legtöbbször teljes komplementaritást mutatva kötődnek akár a cél-mRNS kódoló régiójához is (nemcsak a nem transzlálódó, úgynevezett UTR-hez (*untranslated region*)), szintén kimutatták a homológ Xrn4 szerepét a hasítás során képződő 3’ fragmentum lebontásában (Souret és mtsai, 2004). Ugyanez a kutatócsoport ugyan csak feltételezte az exoszóma szerepét az 5’ intermedier eliminálásában, mások későbbi kísérletei azonban

igazolták, hogy az általunk leírt degradációs mechanizmus valóban általánosnak bizonyult (Pillai, 2005; Zamore és Haley, 2005).



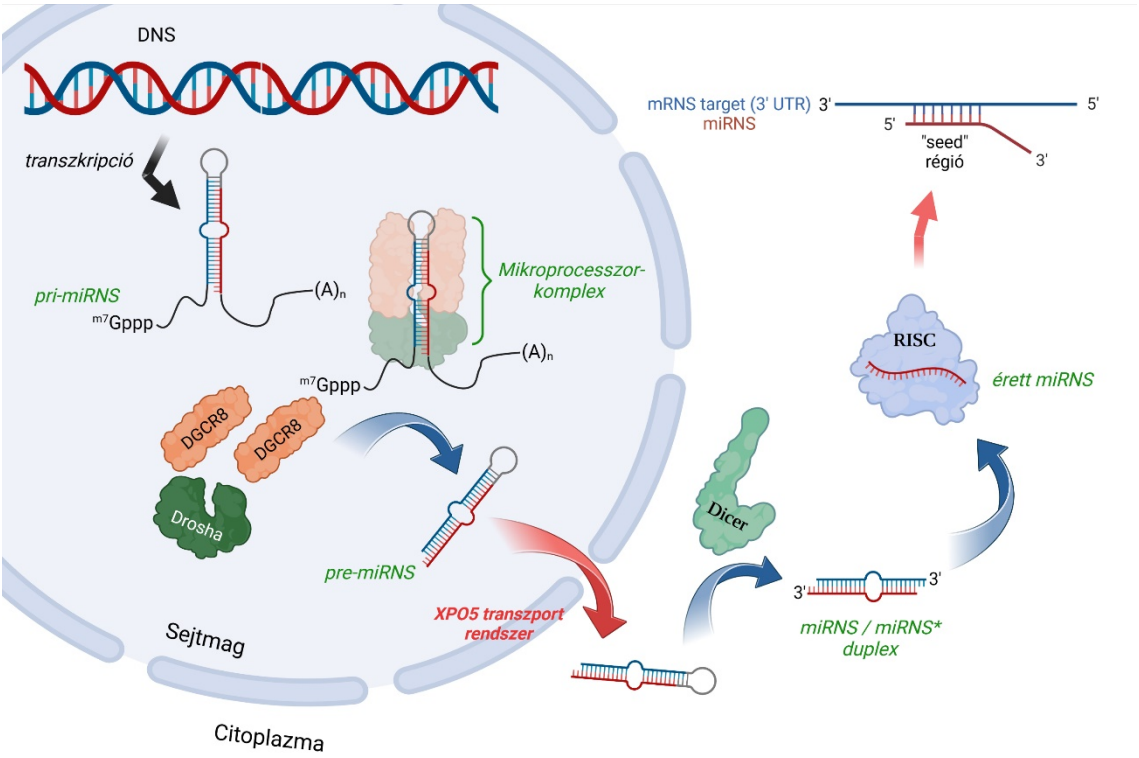
4. ábra: Az siRNS útvonalon inicializált mRNS degradáció kísérleteinken alapuló modellje. (**A**) Az 5'-sapka szerkezetre specifikus antitesttel végzett immunprecipitációs termékek vizsgálata *Northern* analízissel, a „*decapping*" enzimek hatásának kizárása. Az 5'-intermedier esetében a precipitátumban (pelletben) a teljes hosszúságú transzkriptumhoz mért arány (int. /adh) az inputhoz hasonló, tehát tartalmazza az 5'-sapkát, míg a negatív kontrollként használt 3'-intermedier fragmentum esetében ez nincs így. Sup.: felülúszó. (**B**) Az oligodT (dT) jelenlétében végzett RNáz H emésztés hatására a 3'-intermedier és a teljes hosszúságú transzkriptum (adh) esetében is detektálható rövidülés („*mobility shift*"), míg a negatív kontrollként vizsgált 5'-intermedier esetében nem. A 3'-intermedier tehát tartalmazza a 3'-poliA szekvenciát, így a deadeniláz enzimek szerepe a folyamatban kizárható. (**C**) A kísérleteink alapján felállított modell részletei.

Az ábra az (Orban és Izaurralde, 2005) közleményünk alapján készült.

A sejtben található RNS lebomlási útvonalak bonyolultságát mutatja, hogy a RISC aktivitása által képződött degradációs termékek más, úgynevezett „minőségbiztosítási rendszerek”, mint például a „*nonstop decay*” (NSD) szubsztrátjai is lehetnek (Siwaszek és mtsai, 2014). Itt elsősorban a hasítást követően a transzlációban megakadt, de még aktív riboszómákat tartalmazó 5’ intermedier mihamarabbi eltávolításáról van szó. Munkánk során mi is megvizsgáltuk ezt a lehetőséget, és sikerült igazolni, hogy a transzláció gátlása hatással van az 5’ fragmentumra, és kismértékben stabilizálja azt. Az NSD útvonal *Drosophila*-ban akkor ismert komponenseinek (pl. a Hbs1 fehérje) vizsgálata azonban nem mutatott egyértelmű összefüggést az NSD és az RNAi folyamatok között (Orban és Izaurralde, 2005). Évekkel később Silhavy Dániel kutatócsoportjával közös munkánk során növényi és emlős rendszerekben újra felmerült ez a kérdés, hogy van-e kapcsolat a két fenti RNS lebomlási útvonal között. Emlős modellrendszeren a *Drosophila*-hoz hasonlóan nekünk újfent nem sikerül kimutatni egyértelmű kapcsolatot, a növényi modellek ugyanakkor azt mutatták, hogy ott az interakció fontos szereppel bír. A növényekben az siRNS rendszer sokkal kiterjedtebb, és a vírusok elleni védekezésben fontos szerepe van az RNS-dependens RNS polimerázoknak (RDRP) is: az eredetileg kiváltott siRNS hasítást követően az 5’ hasítási termékek ezen enzimek szubsztrátjaként további dsRNS molekulákat generálnak, amelyek tovább erősítik az így kiváltott „immunválaszt”. A növényi miRNS-ek által célpontként kijelölt mRNS-ek esetében azonban ilyen felerősítésre nincs szükség, sőt a szabályozás szempontjából kifejezetten károsnak mondható. Úgy tűnik, hogy ennek kiküszöbölésére lép közbe az NSD útvonal, és a transzláció jeleit mutató 5’ hasítási termékeket mielőbb, még az RDRP-k aktivitása előtt eliminálja (Szadeczky-Kardoss és mtsai, 2018). Valószínűsíthető, hogy az RNAi és az NSD rendszerek ilyen szoros kapcsolata elsősorban olyan élőlényekben mutatható ki, amelyek RDRP géneket kódolnak – ez megmagyarázná, hogy miért nem olyan erős a két RNS rendszer közötti kapcsolat *Drosophila*-ban, illetve emlősökben. Felveti ugyanakkor annak a lehetőségét, hogy például *C. elegans* sejtekben a növényekéhez hasonló szoros összefüggést lehet-e kimutatni, amit a jövőben érdemes lenne megvizsgálni. A RISC által kiváltott endonukleázos hasítás, és a cél-RNS molekulák lebomlási folyamata azonban úgy tűnik, hogy az egész eukarióta világban hasonló módon zajlik le, és ennek a folyamatnak a feltérképezésében a korábban említett munkánkkal mi is tevékeny részt vettünk.

1.2. A miRNS rendszer szerepe

Amint az a korábbi fejezetekben már említésre került, az RNS interferencia ősi, adaptív védekezési rendszereiből az evolúció során valószínűleg több alkalommal is kialakult egy új, endogén szabályozási hálózat, a miRNS útvonal (Moran és mtsai, 2017). Ennek működése a genomban kódolt miRNS génekhez kötött: az ezekről átíródó elsődleges transzkriptumok (*primary*-miRNS-ek, pri-miRNS-ek) erőteljes másodlagos szerkezetű, hajtűszerűen összeálló szekvenciákat kódolnak, amelyek processzálását állati szervezetek esetében a sejtmagban egy ún. „Mikroprocesszor-komplex" indítja el (**5. ábra**). A komplex RNáz komponense a Drosha fehérje, amelyhez a működését segítő DGCR8 (***D****i****G****eorge Syndrome* ***C****ritical* ***R****egion* ***8***) fehérje két molekulája kapcsolódik (Nguyen és mtsai, 2015). Az aktív heterotrimer komplex által lehasított hajtűszerkezetek a pre-miRNS (*precursor*-miRNS) molekulák, amelyeket egy speciális, az Exportin-5 (XPO5) fehérjéhez kötött RNS export rendszer juttat ki a citoplazmába. A pre-miRNS-eket itt az siRNS útvonalból már megismert RNáz, a Dicer molekula ismeri fel, amely a prekurzorok hurok alakú végét (a *loop* szekvenciát) lehasítva egy rövid, kettős szálú kisRNS molekulát hoz létre, amelynek jellegzetessége, hogy a két szál gyakran nem teljesen komplementere egymásnak, hanem kisebb hurkokat vagy egymással nem komplementer (*mismatch*) szekvenciákat tartalmaz. A további érési folyamatok során egy Ago fehérje kapcsolódását, majd a kettős szál egyik szálának eliminálását követően kialakul az érett miRNS-t tartalmazó miRISC, amely effektor funkciója a célzott mRNS molekulákról történő transzláció gátlása, vagy gyakran ezzel párhuzamosan a mRNS degradációjának elindítása (Huntzinger és Izaurralde, 2011). Állati szervezetekben a miRISC-ek legtöbbször a mRNS molekulák 3'-UTR szekvenciáihoz kapcsolódnak, meglehetősen komplex módon: egy miRNS számos mRNS molekulát szabályozhat, illetve egy fehérjekódoló mRNS egyszerre többféle miRNS kötőhelyét is tartalmazhatja. Ezáltal egy roppant bonyolult szabályozási hálózat alakul ki, amelynek az eredménye egy kifinomult, poszttranszkripciós szabályozás, amely képes a transzkriptumok sokaságának finomhangolására (Gebert és MacRae, 2019).



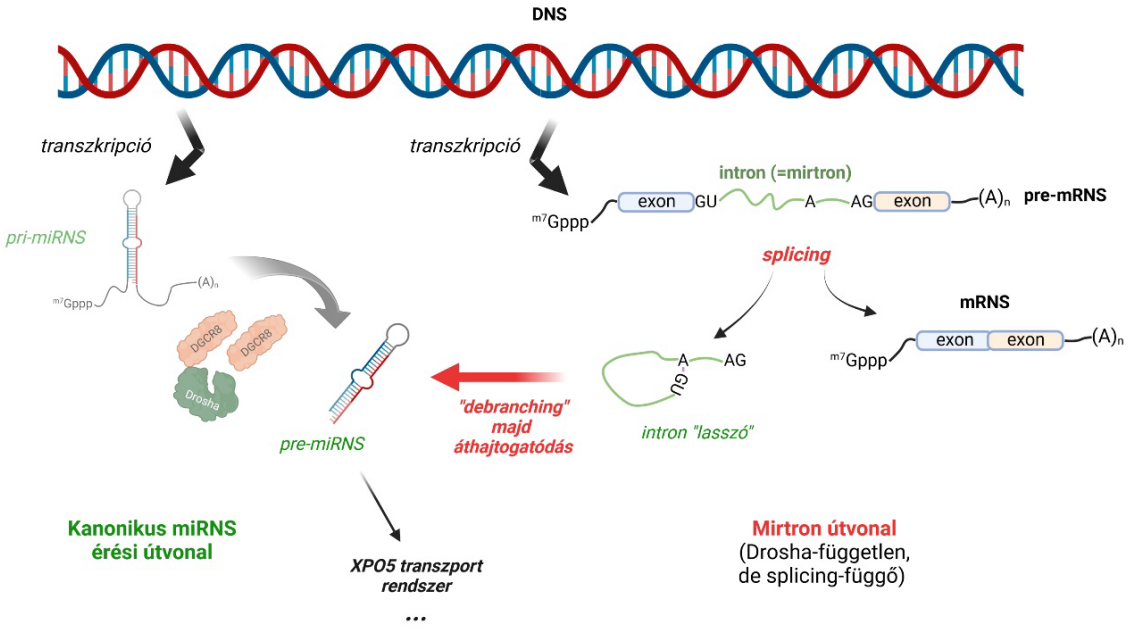
5. ábra: A miRNS molekulák érése állati szervezetekben – a „kanonikus" útvonal. A mechanizmus részletes magyarázatát lásd a szövegben. A miRNS-eknek általában nem a teljes szekvenciája mutat komplementaritást a cél-RNS szekvenciával: a gyakran csak a miRNS 5' részére (2-8. nukleotid) korlátozódó részt az irodalomban „*seed*" régiónak hívják. Az áttekinthetőség kedvéért az ábrán bizonyos komponensek, így a Dicer partnerfehérjéi, vagy a RISC-kel asszociált fehérjék nincsenek feltüntetve.
Az ábra a *biorender.com* alkalmazással készült.

A röviden felvázolt érési útvonal az úgynevezett „kanonikus" miRNS érés folyamata, amelynek azonban a sejtekben meglepően sok alternatív útvonala létezik (Winter és mtsai, 2009; Miyoshi és mtsai, 2010). Ezek alapvetően a két főbb lépést, a Drosha/DGCR8 processzálást, vagy a Dicer hasítás lépéseit kerülik ki, tovább bővítve a miRNS képződés szabályozásának repertoárját. Egy korábbi munkánkban az egyik legjelentősebb alternatív útvonal, a mirtron útvonal emlős sejtekben történő feltérképezésébe vágtunk bele, amelynek kapcsán a miRNS-ek érésének egyéb érdekes aspektusai is felszínre kerültek; ezeket az izgalmas kutatásokat mutatom be a következő fejezetekben.

1.2.1. Egy speciális miRNS érési útvonal: a mirtronok

A mirtronok olyan miRNS szekvenciák a genomban, ahol a pre-miRNS tulajdonképpen egy fehérjekódoló gén egy intronját jelenti. Az érésük a „gazda” mRNS splicing folyamatához kötött: a Drosha/DGCR8 komplex helyett a splicing hozza létre pre-miRNS-t, amely ezután bekapcsolódik a kanonikus érési útvonalba, és az Exportin-5 rendszer, majd a Dicer aktivitását követően a miRNS beépül a RISC-be (**6. ábra**). A mirtronokat, ezeket a Drosha-független úton érő miRNS molekulákat *Drosophila* és *C. elegans* rendszerekben fedezték fel (Okamura és mtsai, 2007; Ruby és mtsai, 2007): ezen organizmusok genomjára általában jellemző, hogy rövid intronokkal rendelkeznek, amelyek mérete összevethető a pre-miRNS-ek méretével, így evolúciósan is „jól kihasználhatók” a miRNS rendszer szempontjából. Több kísérletes eredmény is alátámasztja, hogy az alternatív miRNS érési útvonalaknak, így a mirtron útvonalnak is fontos sejtélettani szerepe van: vannak például olyan sejtdifferenciációs lépések, ahol a Drosha/DGCR8 komplex kikapcsolásával a sejt teljesen leállítja a miRNS-ek többségének képződését, a mirtron útvonalon ugyanakkor továbbra is keletkezhet néhány miRNS molekula, amelyeknek fontos szerepe van az adott folyamatok szabályozásában. Ugyanakkor míg az említett modellorganizmusok esetén a kutatók kísérletesen is igazolták több mirtron létezését, más genomok, így az emlős genomok esetében funkcionális tesztek nélkül, pusztán bioinformatikai predikciókra hagyatkozva biztosra vették a mirtonok előfordulását (Berezikov és mtsai, 2007). Ez utóbbi hozzáállás nem kis vitát váltott ki a tudományos közéletben, és az emlős mirtronok karakterizálásán keresztül ezekben a kutatásokban nekünk is jelentős szerep jutott.

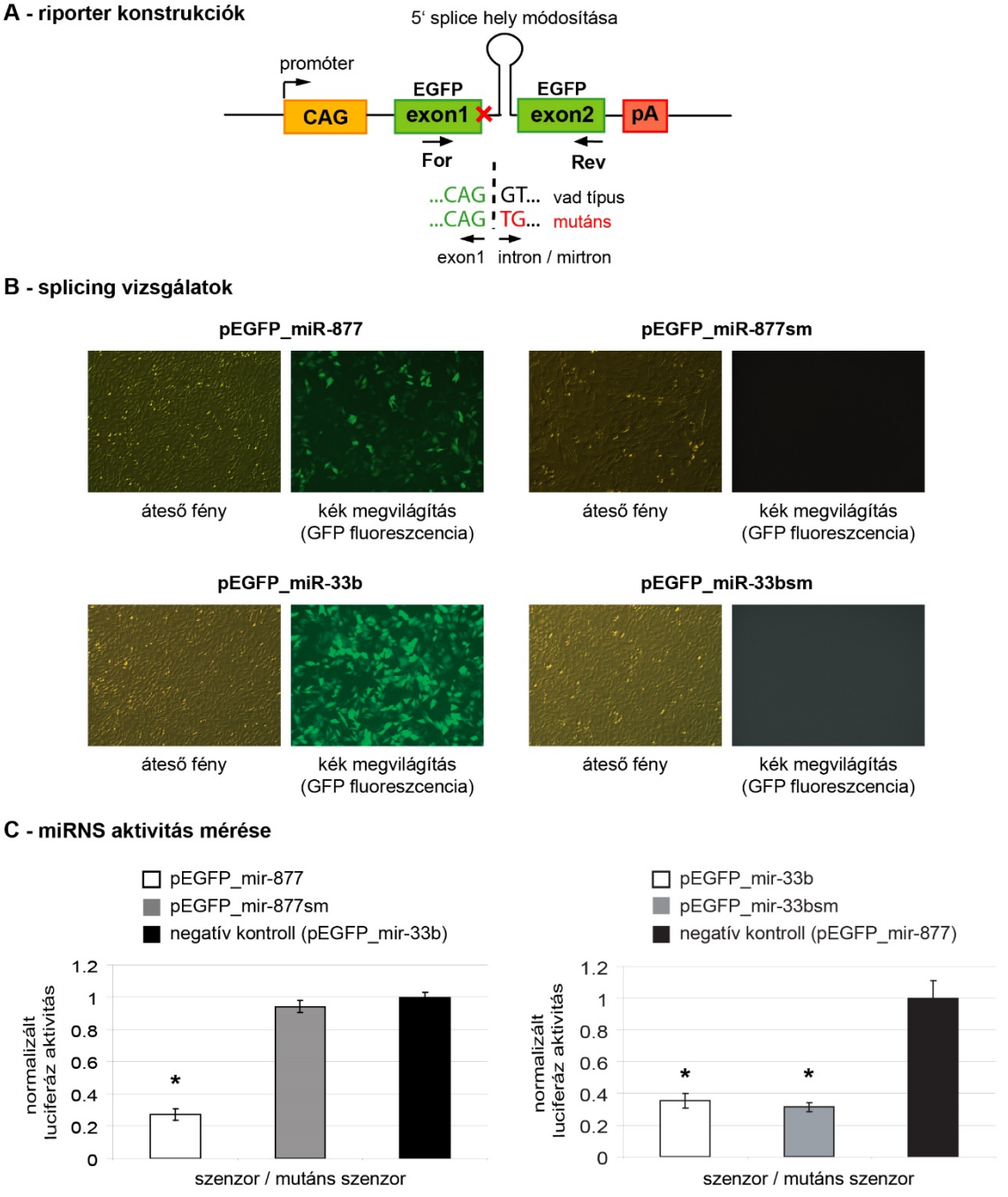
Az emlős genomokban, köztük az emberi genomban is, általában nagy méretű intronok találhatók, amelyek sok esetben akár nagyszámú pre-miRNS szekvenciát is tartalmazhatnak. Ezek átíródása történhet a gazda mRNS transzkripciójával szinkronban, de alternatív, intronban elhelyezkedő promóter és poliA-szignál szekvenciák segítségével attól függetlenül is. Kim és munkatársai már korábban kimutatták, hogy a nagy intronokban elhelyezkedő pre-miRNS-ek processzálódásához nem szükséges a splicing: ha ez utóbbi folyamatot gátolták, akkor az intronokat is tartalmazó elsődleges átiratból továbbra is képződtek miRNS-ek, hiszen a Mikroprocesszor-komplex felismerte a jellegzetes másodlagos szerkezetű szekvenciarészeket, és elindította miRNS érés folyamatát (Kim és Kim, 2007). Felmerült a kérdés, hogy például az emberi genomban ez miért ne következne be a rövid, mirtronként azonosított intronok esetében is?



6. ábra: A mirtron útvonal összehasonlítása a kanonikus miRNS érési útvonallal. A mirtronok tulajdonképpen rövid, a pre-miRNS-nek megfelelő mérettartományú intronok, amelyek legtöbbször fehérjekódoló gének intronjainak feleltethetők meg. Az ábrán az egyszerűség kedvéért a pre-mRNS-nek csak a mirtront tartalmazó régiója van részletesen ábrázolva. Az intron 5’ és 3’ végén a meghatározó nukleotid szekvenciák, illetve az intronon belül a splicing szempontjából kulcsfontosságú elágazási pont (*branch point*) adenin nukleotidja vannak feltüntetve. A lasszó formájú intron az átszerkesztését követően felveszi a pre-miRNS-re jellemző hajtű szerkezetet, majd az érési folyamat a kanonikus útvonal további lépéseivel folytatódik.
Az ábra a *biorender.com* alkalmazással készült.

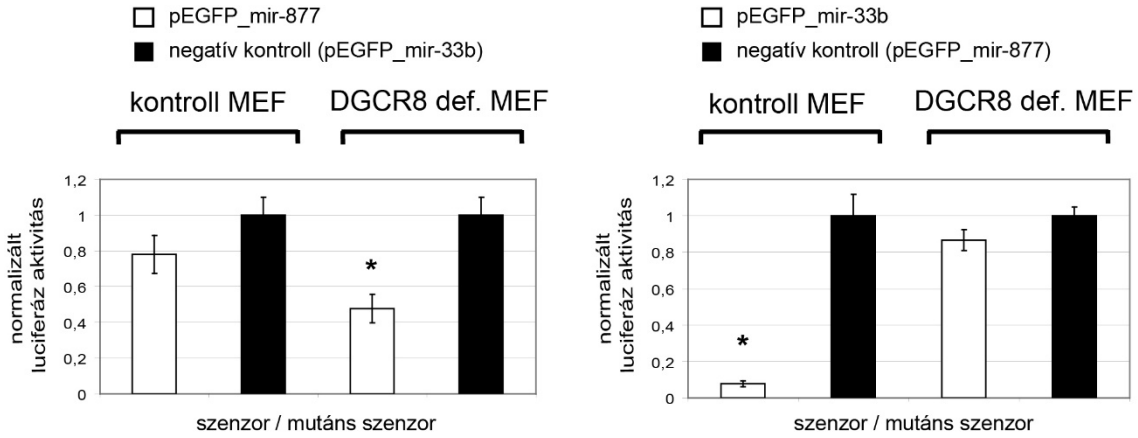
A kérdés eldöntésére kiválasztottunk néhány mirtronnak prediktált humán miRNS szekvenciát és megvizsgáltuk, hogy ezek érési folyamatai mennyiben kapcsolódnak a splicing folyamatához. Sikerült igazolnunk, hogy az eredeti szekvenciakörnyezeten túl, egy riportergén mesterséges intronjaként expresszálva is a prediktált mitronok többségéről érett, funkcionális miRNS molekulák képződnek különböző sejtvonalakban (lásd a hsa-miR-877 példáját a **7. ábrán**). Ezen miRNS-ek érése azonban szigorúan a splicing folyamatához kötődött, ugyanis az 5’ illesztési pontban (5’ *splicing site*) mutációt hordozó konstrukciókból nem képződtek miRNS-ek. Ezek alapján úgy tűnt, hogy valódi humán mirtronokat azonosítottunk, azonban egy bizonyításra még szükség volt: megvizsgáltuk a mirtronok, illetve kontrollként kanonikus miRNS-ek érését DGCR8-deficiens sejtekben. Ezeket a sejteket nem lehet hosszú ideig fenntartani, hiszen funkcionális Mikroprocesszor-komplex hiányában előbb-utóbb elpusztulnak,

ugyanakkor primer tenyészetek formájában vizsgálati célból rövidebb ideig fenntarthatók (Wang és mtsai, 2007; Babiarz és mtsai, 2008). Ilyen emlős sejtekben vizsgálva igazoltuk, hogy míg a kanonikus miRNS-t (pl. a hsa-miR-33b-t) kifejező konstrukciók itt nem működnek, az igazolt mirtron szekvenciákból érett, funkcionális miRNS-ek képződnek (**8. ábra**). Ezekkel a kísérletekkel bebizonyítottuk, hogy az emberi genomban is léteznek valódi mirtronok, amelyekről alternatív útvonalon, a Drosha/DGCR8 komplextől függetlenül, a splicing folyamatának felhasználásával indul el a lókuszon kódolt miRNS-ek érése. Ugyanakkor ezen munkáink fontos tanulsága volt, hogy nem minden rövid intronban elhelyezkedő miRNS viselkedik mirtronként (az általunk vizsgált hsa-miR-1233 például nem), és az adott lókusz vizsgálata mindenképpen biokémiai/molekuláris biológiai bizonyítást igényel (Schamberger és mtsai, 2012; Schamberger és Orban, 2014b). Hogy bizonyos rövid, mirtronszerű szekvenciák miért nem processzálódnak ezen az alternatív útvonalon, azt még további vizsgálatoknak kell feltárnia. Léteznek „hosszabb végű", úgynevezett *tailed*-mirtronok is, amelyeknél a pre-miRNS 5' és/vagy 3' vége az intronon belül, a splicing illesztési pontoktól távolabb esik, a splicing hiányában a Drosha/DGCR8 mégsem tudja elindítani ezek érési folyamatát (Li és mtsai, 2013; Kock és mtsai, 2015; Wen és mtsai, 2015). Valószínű, hogy a Mikroprocesszor-komplex és a splicing apparátus kölcsönhatásának bonyolultsága áll ennek a háttérében, és ez lehet a kulcsa a mirtron útvonal működésének. A kölcsönhatás pontos molekuláris mechanizmusait azonban még nem sikerült minden részletében feltárni, amely így bőven nyújt még a jövőben is kutatási irányokat a mirtron útvonal szabályozásával kapcsolatban.



7. ábra: Egy mirtron (hsa-miR-877) és egy kanonikus miRNS (hsa-miR-33b) összehasonlító vizsgálata HEK-293 sejtekben. (**A**) Az alkalmazott riporter konstrukciók szerkezete. Az EGFP (*Enhanced* GFP) kódoló szekvenciát két exonra bontottuk, közéjük pedig vagy egy mirtron, vagy egy hosszabb, kanonikus miRNS-t tartalmazó intront illesztettünk. (**B**) A vad típusú konstrukciók esetében a működő splicing indikátora az EGFP fluoreszcencia, amely a splicing mutánsok (sm) esetében nem detektálható. (**C**) Az érett, működő miRNS-t a luciferáz szenzor aktivitásának csökkenése jelzi. A mirtron esetében a splicing mutáns konstrukciókról nem képződik érett miRNS. Az adatok 3 független mérés átlagát (± SD) mutatják, *: $p<0,05$.

Az ábra a (Schamberger és mtsai, 2012) közleményünk alapján készült.

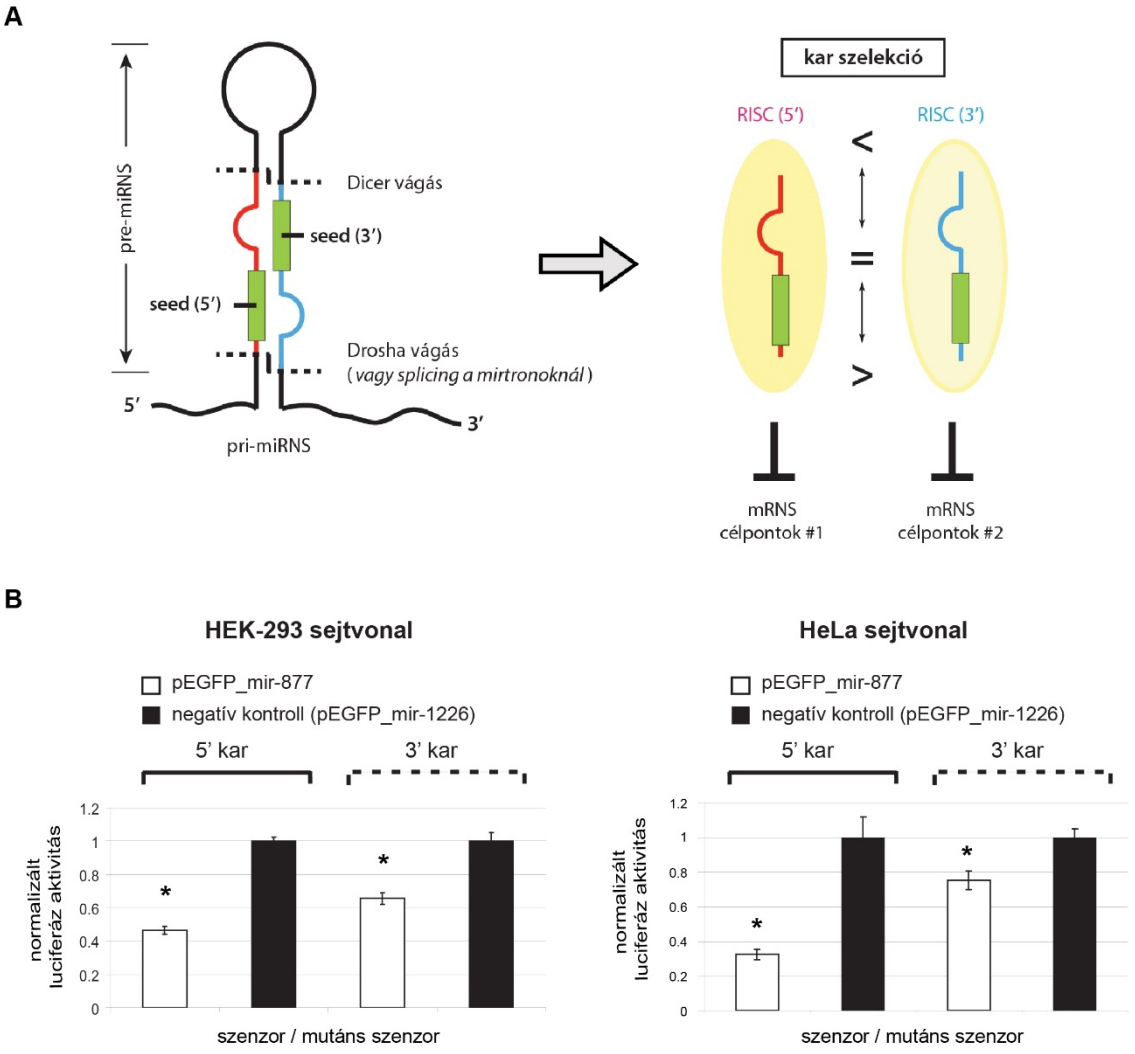


8. ábra: Egy mirtron (hsa-miR-877) és egy kanonikus miRNS (hsa-miR-33b) aktivitásának vizsgálata normál és DGCR8 deficiens sejtekben. A mirtron esetében az érett miRNS aktivitása a DGCR8 deficiens sejtekben is detektálható, míg a kanonikus miRNS esetében a kontroll sejtvonalban mért erős aktivitás a DGCR8 deficiens sejtvonalban majdnem teljesen eltűnik. Az adatok 3 független mérés átlagát (± SD) mutatják, *: $p<0,05$. MEF: egér embrionális fibroblaszt (*mouse embryonic fibroblast*).
Az ábra a (Schamberger és mtsai, 2012) közleményünk alapján készült.

1.2.2. Változatok egy lókuszról: miRNS karhasználat és a 3’ izomiR-ek

A humán mirtronok vizsgálata kapcsán merült fel először a kérdés, hogy a funkcionális tesztelés során egy adott lókusznál a pre-miRNS melyik karjából, az 5’ vagy a 3’ karból érő potenciális miRNS-t vizsgáljuk. A tudomány akkori állása szerint minden miRNS lókusz esetén dominánsan az egyik karból képződik funkcionális termék, a másik szál pedig a RISC érése során eliminálódik (Khvorova és mtsai, 2003; Schwarz és mtsai, 2003). Ez az elképzelés alapvetően egy, az siRNS-eket célzó vizsgálaton alapult, ahol erős összefüggést találtak a szálak beépülése, és szálak 3’ végének termodinamikai stabilitása között (Hutvagner, 2005). Kérdés volt persze, hogy ezt a modellt mennyiben lehet a miRNS-ek érésére is adaptálni, időközben ugyanis megjelentek olyan közlemények, amelyek adott miRNS lókuszok esetén szövetspecifikus karhasználatot igazoltak, illetve bizonyos esetekben felmerült az aktívan szabályozott „karváltás” jelensége is (Ruby és mtsai, 2006; Ro és mtsai, 2007; Grimson és mtsai, 2008; Kuchenbauer és mtsai, 2008; Morin és mtsai, 2008; Okamura és mtsai, 2008; Packer és mtsai, 2008; Yang és mtsai, 2011). Az eredmények arra utaltak, hogy a miRNS-ek

érésének ezen lépése is meglehetősen bonyolult szabályozás alatt áll, amely végső soron növeli a miRNS lókuszokról képződő szabályozó kisRNS-ek repertoárját (**9A ábra**). Ezekre az eredményekre alapozva úgy döntöttünk, hogy bár voltak predikciók a vizsgált mirtron eredetű miRNS-ek karhasználatára vonatkozóan, a funkcionális teszteket mi minden esetben az 5’ és a 3’ karra is egyaránt elvégeztük. Ez a megközelítés nagyon szerencsésnek bizonyult, ugyanis a korábbi vizsgálatok mellett, ahol az eredetileg prediktált 5’ kar működésére fókuszáltunk (lásd a **7.** és **8. ábrákat**), további kísérletekben a hsa-miR-877 esetében sikerült igazolnunk, hogy több sejttípusban mind az 5’, mind pedig a 3’ karból származó érett miRNS-ek egyszerre vannak jelen az effektor komplexekben (**9B ábra**). Kísérletes eredményeink így elsők között mutatták be egy miRNS lókuszon az aktív karhasználat szabályozásának több sejttípusban elvégzett szisztematikus vizsgálatát. A megfigyelésnek nagy jelentősége van a miRNS molekulák általi szabályozásban, hiszen tovább növeli a potenciális célpontok számát, ezáltal bővítve a hálózat regulációs képességét (**9A ábra**). Az eredmények a szakterületünkön nagy érdeklődést váltottak ki, és a közleményünket elfogadó folyóirat aktuális számában az erről szóló ábránk a címlapra került (Schamberger és mtsai, 2012). Az azóta eltelt időszakban számos eredmény, köztük nagyszámú kisRNS szekvenálási adat igazolta, hogy számos pre-miRNS esetén a karhasználat sejttípustól függően aktívan szabályozott, és sok esetben sikerült azonosítani az ebben szerepet játszó regulációs faktorokat is (Noland és Doudna, 2013; Winter és Diederichs, 2013; Suzuki és mtsai, 2015; Wilson és mtsai, 2015; Kim és mtsai, 2020). Az eredetileg feltételezett, a miRNS karok termodinamikai stabilitás-különbségén alapuló szelekciós modell kiegészült több aspektussal, így például az Ago fehérjéknek a miRNS-ek 5’ végén található nukleotidok felé mutatott preferenciájával, egyfajta „digitális alapú szelekciós kóddal” (Suzuki és mtsai, 2015). Úgy tűnik tehát, hogy az általunk a mirtron eredetű miRNS-eken megfigyelt jelenség sokkal általánosabb, és a miRNS-ek érési folyamatainak egy finoman szabályozott szegmenséről van szó.



9. ábra: A pre-miRNS karok közötti szelekció szerepének vizsgálata. (**A**) A pre-miRNS általános szerkezete és a karválasztás egyszerűsített modellje. A szaggatott vonalak a miRNS érés két meghatározó endonukleolitikus hasítási lépését, míg a zöld téglalapok az eltérő karokon található, szekvenciálisan különböző *seed* regiókat jelölik. Adott sejttípusokban eltérő arányban képződhetnek az 5' vagy 3' eredetű, érett miRNS-eket tartalmazó csendesítési komplexek (RISC-ek). (**B**) A mirtron eredetű hsa-miR-877 miRNS karhasználatának funkcionális vizsgálata luciferáz esszékkel, két különböző sejtvonalon. A sejtekben egyszerre vannak jelen az 5' vagy a 3' karból ért miRNS-t tartalmazó csendesítési komplexek, a funkcionális esszék alapján azonban az 5'-RISC tűnik dominánsabbnak. Az adatok 3 független mérés átlagát (± SD) mutatják, *: $p<0,05$.

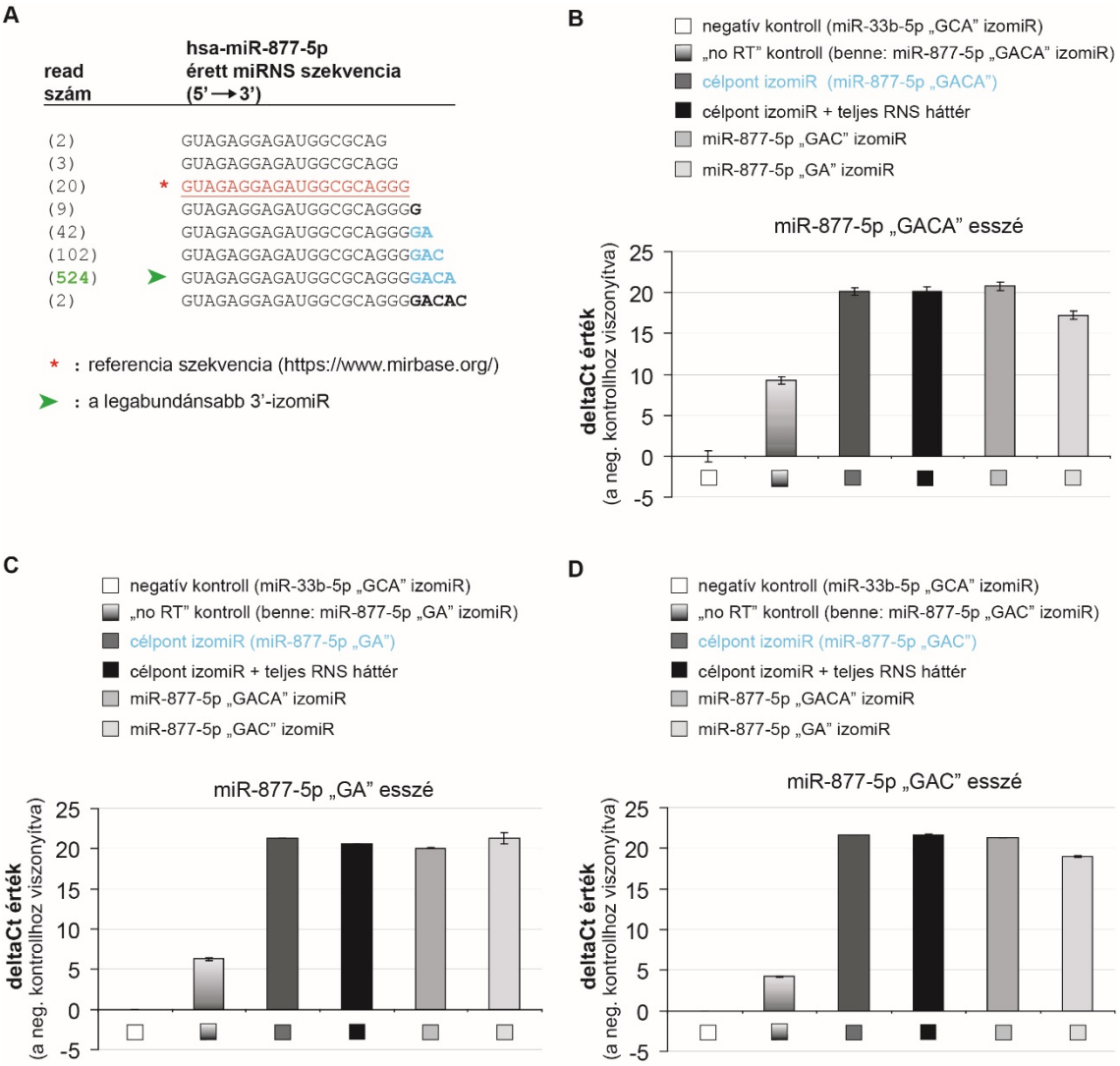
Az (**A**) ábra az (Orban, 2023) közleményem alapján, míg a (**B**) ábra a (Schamberger és mtsai, 2012) közleményünk alapján készült.

A miRNS-ek célzott kimutatása és szekvenciális validálása vezetett el minket egy másik izgalmas, és korábban kevéssé ismert területre, az érett miRNS molekulák végeinek szekvenciális heterogenitásához. A fehérjekódoló mRNS-ek esetében pontosan lehet tudni, hogy egy adott transzkriptumnak hol van az 5’ kezdete és a 3’ vége, amely információk biztonsággal felhasználhatók a kimutatáshoz tervezett próbák előállításánál. A miRNS-ek esetében az új generációs szekvenálási adatok rávilágítottak azonban arra, hogy ott az adott lókuszról sok esetben egy miRNS populáció képződik, amelyben 5’ és 3’ szekvencia variánsok, izomiR-ek vannak jelen (Burroughs és mtsai, 2011; Neilsen és mtsai, 2012; Zhou és mtsai, 2012). Ezek sok esetben 1-2 nukleotid különbséget jelentenek a fő izomiR-hez képest, de előfordulnak a végükön jóval nagyobb hosszúságban különböző variánsok is. Ma már lehet tudni, hogy az ilyen miRNS izoformák képződéséhez nagyon sok tényező járulhat hozzá, ilyen például a Drosha vagy a Dicer enzimek szekvenciális „lötyögése” vagy „promiszkuitása”, amikor a prekurzor molekulákból a kivágódás során hosszabb vagy rövidebb miRNS molekulák jönnek létre (Gu és mtsai, 2012; Starega-Roslan és mtsai, 2015a; Starega-Roslan és mtsai, 2015b). Előfordul azonban, hogy bizonyos enzimek utólag módosítják a miRNS szekvenciáját, és az eredeti prekurzor szekvenciától különböző, úgynevezett templát-független addíciók (*non-template additions*) keletkeznek (Ebhardt és mtsai, 2009; Heo és mtsai, 2009; Wyman és mtsai, 2011; Heo és mtsai, 2012). Funkcionális szempontból az izomiR-eknek azért is van jelentősége, mert a miRNS-ek a célpont felismerése során sokszor csak az 5’ végük egy részén, a 2-8. nukleotid által meghatározott magi régióban (*seed sequence*) mutatnak komplementaritást az mRNS szekvenciával, így az 5’ végükön hosszabb izomiR-eknél változhat a célpont specifitása. A templát-független addíciók leginkább a miRNS-ek 3’ végét érintik, és sok esetben bizonyítottan megváltoztatják a molekula stabilitását (Jones és mtsai, 2012; Westholm és mtsai, 2012; Scheer és mtsai, 2016).

Az általunk végzett vizsgálatokban az izomiR-ek jelenléte technikai szempontból jelentett kihívást: amíg a *Northern* analízis esetén (megfelelő expressziós szint mellett) az izomiR-ek egyértelműen azonosíthatók, addig az expresszió mérésére szintén általánosan elterjedt real-time PCR módszer esetében a szekvencia heterogenitása komolyan befolyásolhatja a megbízhatóságot. A kisRNS szekvenálási adatokat elemezve a hsa-miR-877-5p (az 5’ karról képződő érett miRNS) esetében például kiderült, hogy az adatbázisokban eredetileg szereplő izomiR (az akkor még „referencia” miRNS) csak nagyon alacsony szinten fejeződik ki, helyette viszont a lókuszról képződő miRNS populációban egy másik, a 3’ végén 4 nukleotiddal hosszabb izomiR a leggyakoribb (**10A**

ábra). Az eredeti referencia szekvenciára tervezett *real-time* PCR esszé a hosszabb izomiR-t nem mutatta ki, és csak egy egyedileg tervezett új esszével sikerült ez utóbbinak az expresszióját detektálni (**10B ábra**). A probléma ugyanakkor még összetettebb volt, mert a miRNS populációban több, hosszúságukban 1-2 nukleotidban különböző 3' izomiR volt jelen: ezek bármelyikére tervezett esszé keresztreakciót mutatott a többi izoformával, így ez a módszer ebben a hagyományos formájában nem volt alkalmas arra, hogy megbízhatóan és kvantitatív módon detektálja az izomiR-eket (**10C** és **10D ábrák**). További technikai nehézséget jelentett még, hogy az eredeti, „első generációs" stem-loop esszék DNS és RNS kontaminációra is meglehetősen érzékenyek voltak, így az izomiR-ek megbízható detektálása ezzel a módszerrel akkor emiatt is nehézségekbe ütközött (Schamberger és Orban, 2014a). Az általunk is leírt problémára azóta születtek áthidaló megoldások, így ma már több módszer is létezik, amelyek – még ha kicsit bonyolultabbak is az eredeti technológiánál – alkalmasak az izomiR-ek pontosabb mérésére (Honda és Kirino, 2015; Wang és mtsai, 2020; Franco és mtsai, 2022).

A 2014-es közleményünk megjelenése óta már jóval többet tudunk az izomiR-ek világáról. A 5' és a 3' szekvencia variánsok (és ezek kombinációi) mellett ma már ismerünk ún. polimorfikus izomiR-eket, amelyek hosszúsága ugyan megegyezik a referencia miRNS-ével, belső szekvenciája azonban egy vagy több nukleotidban eltér attól, például az RNS editálás jelenségének köszönhetően (Bofill-De Ros és mtsai, 2020; Tomasello és mtsai, 2021). A jelenség részletesebb megismerése ugyanakkor rávilágított a „nem-kódoló" gének képződésének sokszínűségére, amelyek populációs-, vagy akár egyedi-szintű különbségekkel lehetővé teszik a génexpressziós szabályozás nagyon összetett „finomhangolását" – a kódoló és a nem-kódoló célpontokat tekintve egyaránt (Orban, 2023).



10. ábra: A 3' izomiR-ek és detektálásuk problémái a hsa-miR-877-5p példáján. (**A**) A hsa-miR-877-5p miRNS populációban detektálható 3' izomiR-ek listája és mennyiségi eloszlása. A referencia szekvenciaként megjelölt kisRNS előfordul ugyan, de alacsony mennyiségben. A további vizsgálatok a három leggyakoribb izomiR-en történtek (ezek 3' „extra" szekvenciái kékkel kiemelve). Az eredmények a (Lee és mtsai, 2010) közleményhez tartozó szekvenálási adatokból származnak. (**B-D**) Az izomiR-ek detektálását az adott célpont szekvenciára tervezett, specifikus „stem-loop" TaqMan esszékkel végeztük, templátként szintetikus RNS oligókat használva. A negatív kontroll méréseknél 36-39 közötti Ct értékeket mértünk, míg a célpont RNS jelenléte reverz transzkripció (RT) hiányában is jelentős hátteret (deltaCt emelkedést) adott. Minden esszé erős keresztreakciót mutat valamennyi vizsgált izomiR jelenlétében. A grafikonok 3 mérés átlagát mutatják, a hibazászlók a szórásértékeket (SD) jelzik. A (**B-D**) ábrák a (Schamberger és Orban, 2014a) közleményünk alapján készültek.

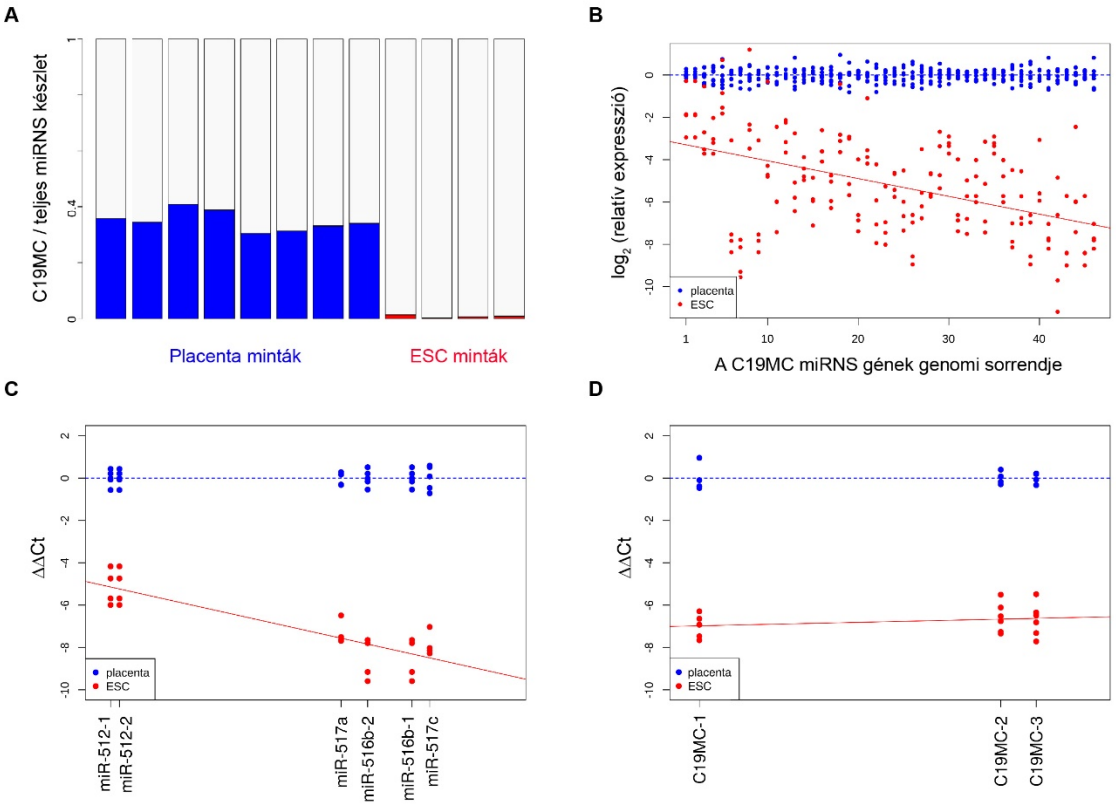
1.2.3. miRNS-ek klaszterekben

Mint minden genetikai elem, így a miRNS-ek kifejeződésének szabályozásában is kulcsfontosságú lépés a transzkripciós szintű reguláció. Ahogy korábban már említésre került, a miRNS-ek ezen a téren is meglehetősen nagy variabilitást mutatnak: vannak önálló transzkripciós egységgel rendelkező kisRNS gének; vannak olyanok, amelyek expressziója más gének működéséhez kötött (például a korábban bemutatott mirtronok); de gyakran előfordul, hogy több miRNS egy közös klasztert alkotva együttes szabályozás alatt áll (Altuvia és mtsai, 2005; Kim és Kim, 2007; Ramalingam és mtsai, 2014; Gebert és MacRae, 2019). Az ilyen klaszterekben legtöbbször kevés számú (<10) miRNS található, extrém esetekben azonban akár 40-50 miRNS lókusz is egy közös transzkripciós egységet alkothat, amelyekre a humán genom 14. és 19. kromoszómáján is találunk példákat (C14MC – *chromosome 14 miRNA cluster*, C19MC – *chromosome 19 miRNA cluster*). A klaszter csoportosulás evolúciós előnye, hogy nagyszámú regulátort lehet egyszerre „mozgósítani", ugyanakkor a génexpresszió további lépései, a poszttranszkripciós mechanizmusok továbbra is lehetővé teszik a közös klaszterből átíródott miRNS-ek individuális finomhangolását (Kim és mtsai, 2009; Michlewski és Caceres, 2019). A nagyobb miRNS klaszterek genetikai szempontból is nagyon összetettek: a C14MC esetében az 52 miRNS gén közé ékelődve találunk még egy kisebb C/D snoRNS (kis nukleoláris RNS) klasztert is, illetve mivel sok miRNS lókuszról az 5' és a 3' karról egyaránt képződik termék, így a klaszterről jóval nagyobb számú érett miRNS képződik (Morales-Prieto és mtsai, 2013). További érdekesség, hogy mindkét nagy humán klaszter esetében megfigyelhető a genetikai *imprinting* jelensége: a C14MC az anyai allélról, míg a C19MC az apától örökölt allélról íródik át (Malnou és mtsai, 2019).

Kutatásaink során a C19MC klaszter regulációjával foglalkoztunk részletesen, mert ez a 46 miRNS lókuszt tömörítő régió több szempontból is felkeltette az érdeklődésünket. A nagyméretű (>100 kilobázis szekvenciát lefedő) kromoszómarégió szövetspecifikus expressziót mutat: kifejeződik embrionális őssejtekben és trofoblaszt sejtekben, illetve jelentős mértékben placenta sejtekben, más szövetekben azonban (néhány tumor kivételével) nem lehet kimutatni a transzkripcióját (Bar és mtsai, 2008; Ren és mtsai, 2009; Flor és Bullerdiek, 2012). A klaszter érdekessége, hogy a miRNS lókuszok között rengeteg Alu szekvencia található: ezek a primáta-specifikus, nem-autonóm retrotranszpozonok valószínűleg fontos evolúciós szerepet játszottak a C19MC

kialakulásában, repetitív jellegüknél fogva azonban nagyon megnehezítik a régió expressziós (például PCR-alapú) vizsgálatát (Zhang és mtsai, 2008; Lehnert és mtsai, 2009). A lókuszról egy nagyon hosszú transzkriptum íródik át, amely komplex splicing mintázatot is mutat, ugyanakkor fehérjekódoló részeket a mai napig nem sikerült benne kimutatni, és úgy tűnik, hogy ez a régió „tisztán" nem-kódoló elemek, miRNS-ek és Alu szekvenciák forrása (Bortolin-Cavaille és mtsai, 2009).

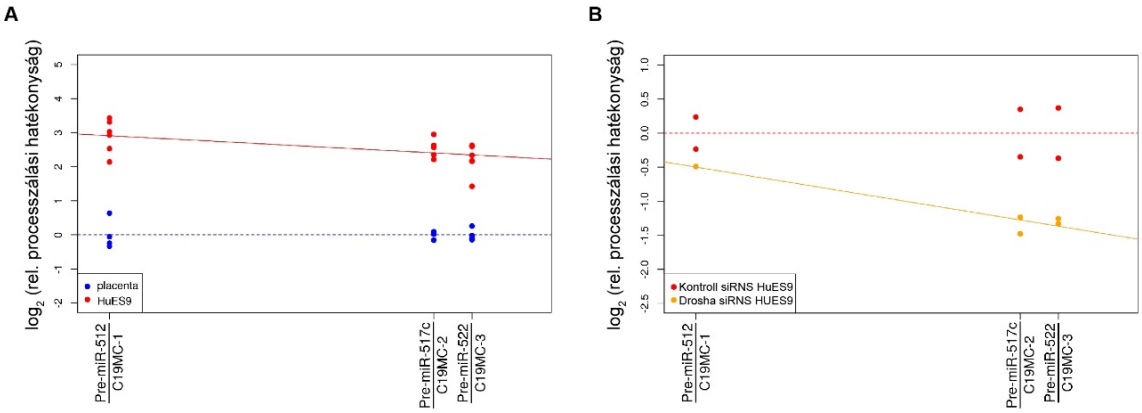
Vizsgálataink első izgalmas eredményét az embrionális őssejtek és a placentaszövet összehasonlító expressziós vizsgálatai hozták. Kiderült, hogy placentában jóval magasabb a C19MC expresszió, olyannyira, hogy a placentában található teljes miRNS mennyiség megközelítőleg 40%-a erről a klaszterről származik (**11A ábra**). Őssejtekben ugyanakkor az alacsony expresszió mellett szembetűnő volt, hogy a miRNS-ek szintje a klaszter 3' irányába egy erőteljes, pozíció függő fokozatos csökkenést mutat (**11B ábra**). Ezt a jelenséget több független, különböző adatbázisokban és közleményekben publikált adatsor analízise is megerősítette, illetve saját laboratóriumi mintáinkban egyedi miRNS-ek expressziós szintjeinek a vizsgálatával is igazoltuk (**11C ábra**). Mivel első közelítésben transzkripciós szintű szabályozásra gyanakodtunk, specifikus kvantitatív, reverz-transzkripcióval egybekötött *real-time* PCR (qRT-PCR) esszékkel megvizsgáltuk a klaszter különböző régióinak átíródási szintjeit, vagyis tulajdonképpen a pri-miRNS szinteket, itt azonban nem tudtuk detektálni a pozíció függő expressziócsökkenést (**11D ábra**). A transzkripció gátlása placenta eredetű sejtvonalban is hasonló eredményre vezetett: ebben a sejttípusban a miRNS expressziós profilok klaszteren belüli, 3' irányú csökkenését így sem lehetett előidézni. Úgy tűnt, hogy a jelenség mögött valamilyen poszttranszkripciós szabályozási mechanizmus állhat, ugyanakkor továbbra is zavarba ejtő volt, hogy őssejtekben a miRNS expressziós szintek miért mutatnak az egyedi szekvenciáktól függetlenül ennyire a genomi pozíciótól függő, fokozatos csökkenést.



11. ábra: A C19MC miRNS-ek genomi pozícióhoz kötött expressziós szintje a két vizsgált sejttípusban. (**A**) A C19MC-ről képződő miRNS-ek gyakorisága a sejt teljes expresszálódó miRNS készletében, 8 placenta és 4 embrionális őssejt (*embryonic stem cell*, ESC) minta kisRNS szekvenálási adatainak az elemzése alapján. (**B**) A C19MC 46 miRNS pozíciójában mért relatív expressziós szintek összehasonlítása placenta (n=8) és ESC (n=4) mintákban. Az adott pozíció pre-miRNS-ének 5' és 3' karjából érő miRNS-ek értékeit összevontuk, az expressziós szinteket pedig a placentában mért átlagokra normalizált, $\log_2$ transzformált CPM (*count per million*) értékekként ábrázoltuk. A piros vonal az ESC értékekre illesztett lineáris modellt mutatja (a regressziós β-koefficiens értéke -0,08401, p=3,5E-12). (**C**) A szekvenálási adatok validálása néhány kiválasztott miRNS pozícióban, qRT-PCR módszerrel, független placenta (n=5) és ESC (n=4) mintákon. A piros vonal az ESC értékekre illesztett lineáris modellt mutatja (a regressziós β-koefficiens értéke -0,08134, p=2,53E-6). (**D**) A pri-miRNS-ek szintjének mérése qRT-PCR módszerrel, 3 kiválasztott genomi pozícióban. Az ESC mintákban mért expressziós szintek itt nem mutatják a placentához viszonyított, genomi pozíciótól függő fokozatos csökkenést (az illesztett piros vonal esetében a β-koefficiens értéke 0,00928, p=0,321).

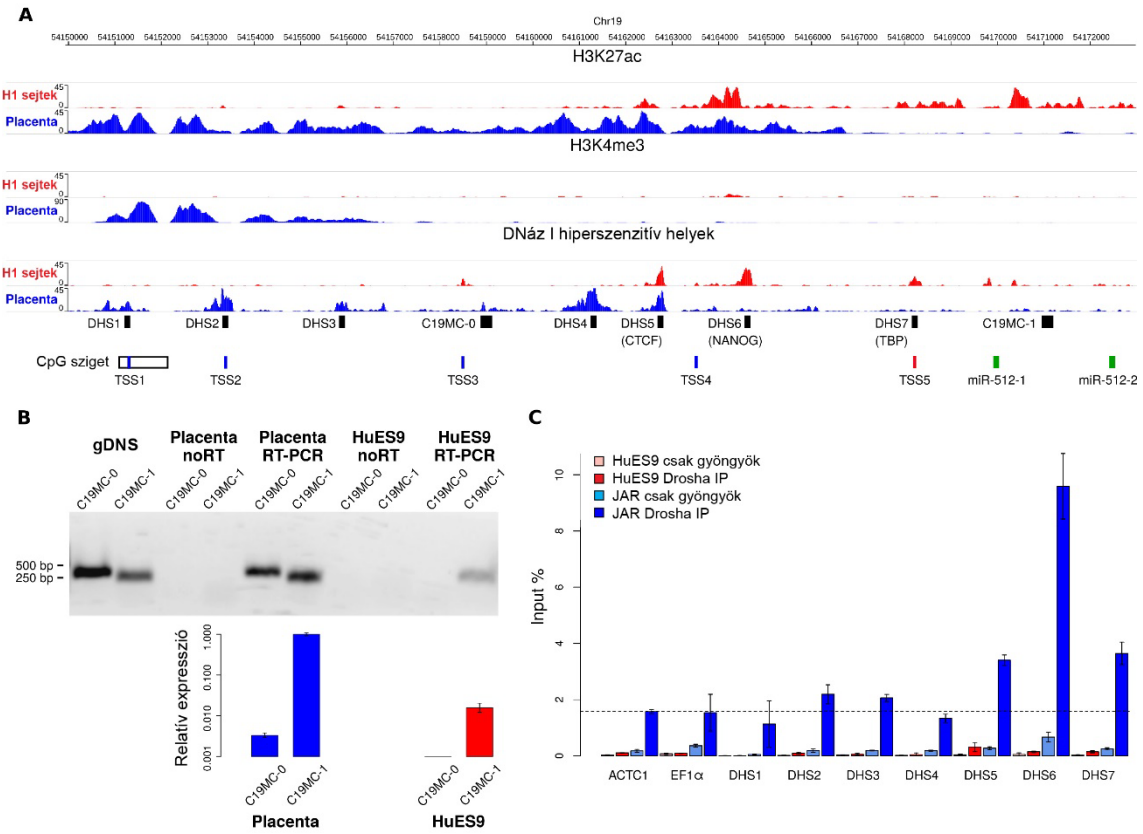
Az ábra a (Fothi és mtsai, 2021) közleményünk alapján készült.

A szokatlan jelenség további vizsgálata során először próbáltunk azonosítani olyan, elsősorban fehérje faktorokat, amelyek más miRNS klaszterek esetében igazoltan felelősek az egyedi miRNS szintek poszttranszkripciós különbségeiért (Michlewski és Caceres, 2019), ez azonban nem vezetett eredményre. Megvizsgáltuk ugyanakkor a C19MC különböző régióiban elhelyezkedő miRNS-ek processzálási hatékonyságát, vagyis a pri-miRNS → pre-miRNS átalakulás hatékonyságát, és azt tapasztaltuk, hogy őssejtekben ez kismértékű, de a klaszteren belül 3' irányban fokozatos csökkenést mutat (**12A ábra**). Az eredmények így a Mikroprocesszor-komplex valamilyen szabályozására utaltak, ezért kipróbáltuk, hogy a komplex fő effektorának, a Drosha enzimnek a mennyiségi csökkentése hasonló jelenséget eredményez-e. A placenta sejtvonalban (JAR) a specifikus siRNS kezelésnek minimális hatása volt, az őssejtekben ugyanakkor nagyon markáns hatást tapasztaltunk: a miRNS processzálás a klaszteren belül 3' irányban tovább csökkent, vagyis a pozícióhatás egyértelműen a Drosha (és így a Mikroprocesszor-komplex) lokális koncentrációjához köthető (**12B ábra**). A kérdés már csak az volt, hogy tudunk-e azonosítani bármilyen szövetspecifikus mechanizmust, amely a jelenség mögött állhat.

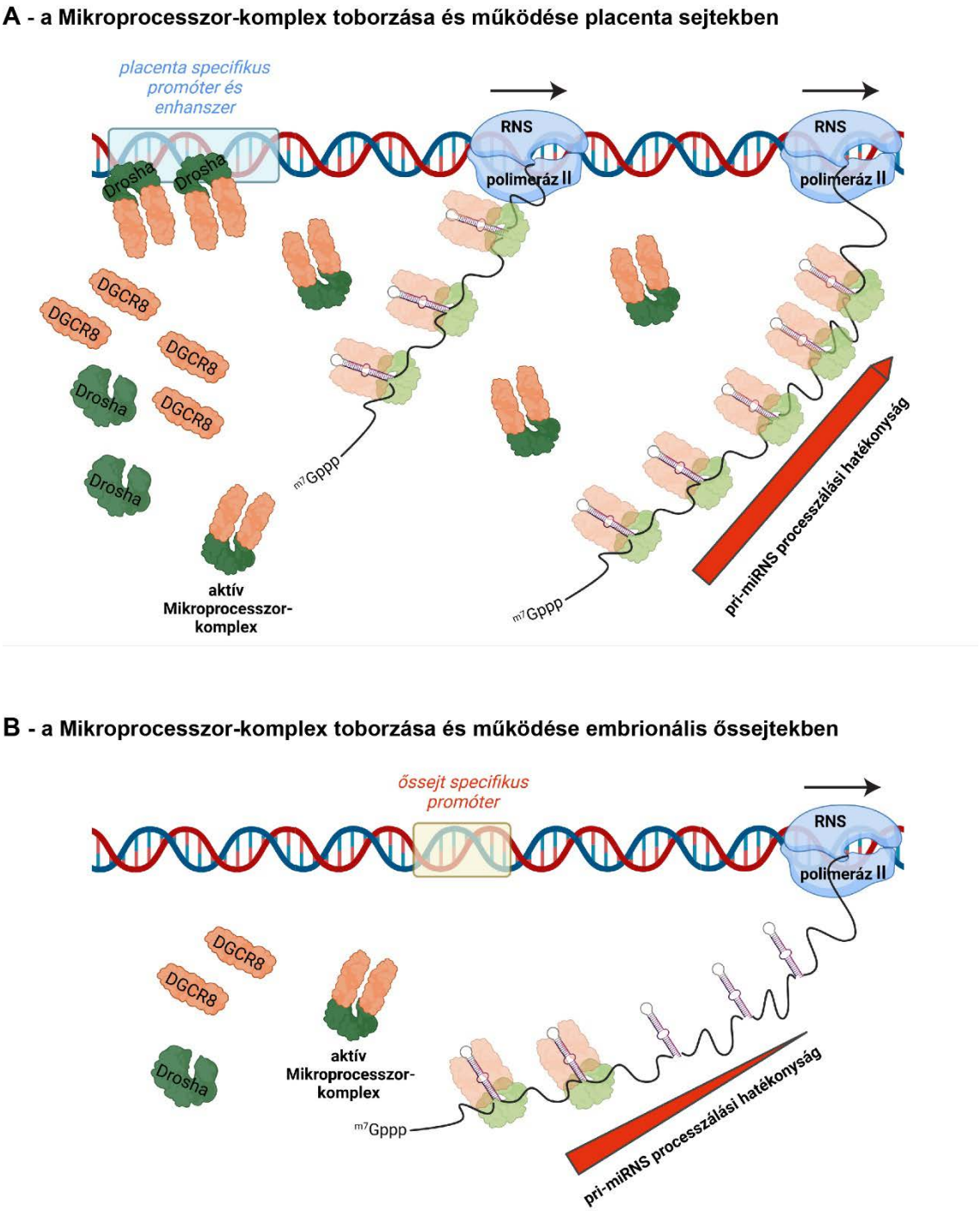


12. ábra: A C19MC belüli expressziós pozíció effektus a pri-miRNS → pre-miRNS processzáláshoz köthető. (**A**) A qRT-PCR módszerrel mért pri-miRNS vágási hatékonyság a HUES9 őssejtvonalon általában magasabb, mint a placenta sejtekben, viszont a klaszter 3' végi irányába csökkenést mutat (a piros illesztett egyenes esetén a a regressziós β-koefficiens értéke -0,014825, p=0,0184). (**B**) A Drosha siRNS-sel történő csendesítése őssejtekben pozíció függő módon, a klaszter 3' vége felé tovább csökkentette a pri-miRNS vágási hatékonyságot (a sárga illesztett egyenes esetén a a regressziós β-koefficiens értéke -0,02282, p=0,00131).
Az ábra a (Fothi és mtsai, 2021) közleményünk alapján készült.

A kérdés megválaszolásához végül egy akkor friss tudományos eredmény segített hozzá, amelyben egy kutatócsoport leírta, hogy a Mikroprocesszor-komplex bizonyos miRNS gének előtt elhelyezkedő, úgynevezett „szuper-enhanszer" régiókhoz kötődik, és így már a promóternél összeálló RNS polimeráz holoenzimhez kötődve, ko-transzkripcionális módon elősegítheti az elsődleges átiratokból a pre-miRNS-ek kivágódását (Suzuki és mtsai, 2017). Hasonló mechanizmust feltételezve először az *NIH Roadmap Epigenomics Project* és az ENCODE nyilvánosan hozzáférhető epigenetikai vizsgálati adatait elemeztük, elsősorban különféle hiszton módosulások (H3K4me3 és H3K27ac), DNáz I hiperszenzitív (DHS) régiók, és ChIP-seq (***C**hromatin **I**mmuno**p**recipitation-**seq**uencing)* módszerrel meghatározott transzkripciós faktor kötőhelyekre fókuszálva. Ezen epigenetikai markerek segítségével a C19MC-től 5' irányban először azonosítottunk egy placenta-specifikus promóter és enhanszer régiót (**13A ábra**), és igazoltuk a placentában átíródó, hosszabb transzkriptum (tk. pri-miRNS) jelenlétét (**13B ábra**). Ezt követően a Drosha fehérje elleni specifikus antitest használatával, ChIP-qPCR módszerrel bizonyítottuk, hogy placentában a Drosha hatékonyan képes kötődni ehhez a régióhoz, méréseink alapján elsősorban 3 enhanszer szekvenciához (**13C ábra**, DHS5-7 régiók). Az eredmények alapján azt feltételezzük, hogy a vizsgált miRNS-klaszter placenta-specifikus promótere nemcsak a magasabb szintű transzkripcióhoz, hanem a Mikroprocesszor-komplex hatékony „toborzásán" keresztül a nagyszámú miRNS hatékony éréséhez is jelentősen hozzájárul (**14. ábra**) (Fothi és mtsai, 2021). A mechanizmus részleteivel kapcsolatban természetesen még akadnak tisztázandó kérdések, például a komplex másik tagjának, a DGCR8 fehérjének a szerepe a Drosha „toborzásában" (Ree és mtsai, 2022). Ugyanakkor más miRNS-klaszterek expressziós mintázatait elemezve úgy tűnik, hogy hasonló, „szuper-enhanszer" mediált mechanizmusok ott is szerepet játszhatnak, és az általunk felállított modell választ adhat az ott tapasztalt szövetspecifikus miRNS-mintázatok kialakulására (Fothi és mtsai, 2021).



13. ábra: Alternatív promóterhasználat placentában és embrionális őssejtekben (H1 és HuES9 sejtekben) a C19MC-től 5' irányban található 22 kb hosszú DNS szegmensen. (**A**) H3K4me3 hiszton módosulások kijelölnek két feltételezett szövetspecifikus promótert, míg a H3K27ac hiszton módosulások és a DNáz I hiperszenzitív (DHS) helyek több aktív enhanszer régiót. Az 5' promóter a placentában aktív, 4 igazolt transzkripciós indítóhellyel (*transcription start sites*, TSS, kékkel jelölve), míg a 3' promóter őssejtekben aktív, 1 transzkripciós indítóhellyel (pirossal jelölve). A TSS-eket a (Noguchi és mtsai, 2017) közlemény „CAGE-seq" adatai alapján azonosítottuk; a TSS1 egy hosszabb CpG sziget része, amely a placentán kívül egyéb szövetekben erősen metilált (Noguer-Dance és mtsai, 2010). A fekete téglalapok a PCR amplikonokat jelölik, köztük az általunk feltételezett 7 enhanszert (DHS1-7), és a két promóterhez tartozó pri-miRNS régiókat (C19MC-0 és C19MC-1). Zárójelben a „ChIP-seq" módszerrel azonosított transzkripciós faktor kötőhelyeket, zöld téglalappal pedig a C19MC első 2 miRNS lókuszát jelöltük. (**B**) Az ábrán az alternatív promóterekhez tartozó pri-miRNS-ek kimutatása (felső gélkép, PCR kontrollként genomiális DNS-t (gDNS) használtunk), és qRT-PCR-rel történő kvantitálása látható (a PolIIA endogén kontrollhoz viszonyítva). (**C**) ChIP-qPCR módszerrel azonosított 3 Drosha kötőhely (DHS5-7) placenta sejtekben. Háttérnek a két független promóter (ACTC – szívspecifikus alfa-aktin és EF1α – elongációs faktor 1α) értékeit állítottuk be (szaggatott vonal). A grafikonokon a mérések átlagát ± szórás értékeket ábrázoltuk; JAR: placenta eredetű sejtvonal. Az ábra a (Fothi és mtsai, 2021) közleményünk alapján készült.



14. ábra: A C19MC miRNS-ek feldolgozási mechanizmusának az eredményeink alapján felállított modellje. A placenta-specifikus promóter (**A**) a robosztus transzkripciós aktivitás mellett hatékonyan képes a Drosha fehérjén keresztül toborozni a Mikroprocesszor-komplexeket. A komplex lokális koncentrációja ezáltal megnő, így a klaszter teljes hosszában képes a pre-miRNS-ek hatékony felismerésére és kivágására. Ezzel szemben az embrionális őssejtekben működő promóter (**B**) mind transzkripcionálisan, mind a Mikroprocesszor-komplex toborzásában kevésbé aktív, ezáltal a pre-miRNS-ek processzálása a klaszter 3' vége felé fokozatosan csökken. Az ábra a *biorender.com* alkalmazással készült.

A C19MC placenta-specifikus magas expressziója kapcsán adódik a kérdés, hogy vajon a kóros placenta működések hátterében nem állhat-e az előbbiekben ismertetett szabályozási mechanizmus zavara. A Semmelweis Egyetem I. számú Szülészeti és Nőgyógyászati Klinikával folytatott együttműködésünk keretében lehetőségünk nyílt a kérdés vizsgálatára: normális és preeklampsziás placenta minták, valamint a hozzájuk tartozó vérminták miRNS mintázatát vizsgálva elemeztük azok miRNS mintázatát, köztük a C19MC-ről származó miRNS-ek expressziós szintjét is. Sajnos a minták között nagy volt a variabilitás, és a klaszterről származó miRNS-ek esetében nem tudtunk szignifikáns különbséget kimutatni a normális és a kóros placenta minták között. A mások által is potenciális biomarkerként azonosított, nem a C19MC-ről képződő hsa-miR-210 esetében viszont mi is találtunk összefüggést a preeklampsziával kapcsolatban, és a betegség egyik lehetséges etiológiai hátttereként azonosított hipoxiát sejtes modellekben (HTR8 és JAR sejtvonalak) vizsgálva eredményeink megerősítették a vizsgált miRNS expressziójának emelkedését oxigénhiányos állapotban. A vérminták vizsgálata azt mutatta, hogy preeklampsziában a hsa-miR-210 szintje jelentősen megemelkedik, de a vérben ez nem az extracelluláris vezikulákban (exoszómákban), hanem az Ago-fehérjékhez kötötten, az úgynevezett „szabad frakcióban” volt kimutatható. Sejtes modellekben azt is sikerült igazolni, hogy a miRNS-ek sejtekből történő szekréciója szigorúan szabályozott: csak bizonyos miRNS-eket érint, és sejttípustól, illetve a terhesség korától függően változik, hogy exoszómák formájában, vagy Ago-kötött állapotban történik (Biro és mtsai, 2019). További vizsgálatok szükségesek a szekretált miRNS-ek pontos funkciójának felderítéséhez, ugyanakkor ezek a kutatások a biomarkerek azonosításán túl közelebb vihetnek minket a kóros placentaműködésre visszavezethető betegségek modern, genetikai alapú diagnózisához, és az ezeken alapuló jövőbeli gyógyításához is.

1.3. Kitekintés – az RNS-ek csodálatos világa

Kétség sem fér hozzá, hogy az RNS interferencia, illetve az ebbe a tág fogalomkörbe tartozó változatos, a kisRNS-fehérje komplexeken alapuló regulációs folyamatok nagyon fontos szereppel bírnak az eukarióta sejtek endogén szabályozásában. A molekuláris biológia dinamikusan fejlődő eszköztára segítségével egyre többet tárhatunk fel ezen folyamatok részleteiről, és az elmúlt évtizedek tapasztalatai alapján elkezdtünk egy kicsit más szemmel tekinteni a ribonukleoprotein-komplexek izgalmas világára. A miRNS-ek szempontjából nagyon fontos tanulság, hogy a lókuszok tekintetében már itt sem igaz a még korábban, a molekuláris biológia hajnalán megfogalmazott „egy gén – egy géntermék” hipotézis (eredetileg „egy gén – egy enzim” hipotézis, (Beadle és Tatum, 1941)). A miRNS gének esetén a „termék” kifejezetten egy komplex kisRNS populáció, amely például a pre-miRNS karválasztásán vagy az izomiR-ek képződésén keresztül egy nagyon összetett, poszttranszkripciós szabályozási hálózat létrejöttét eredményezi (Orban, 2023). Ennek a dinamikusan finomhangolt hálózatnak a jelentőségét leginkább talán akkor vesszük észre, ha az valami miatt zavart szenved, és különféle betegségek, például neurodegeneratív szindrómák vagy rákos elváltozások kialakulásához vezet (Salem és mtsai, 2016; Telonis és mtsai, 2017; Lu és mtsai, 2020; Xiong és mtsai, 2020; Nikolova és mtsai, 2021; Zelli és mtsai, 2021; Scheper és mtsai, 2022). A nemrégiben a világon végigsöpört COVID-19 pandémia egyik érdekes tudományos „melléktermé ke”, hogy a miRNS szabályozás populációk illetve egyedek közötti, vagy akár nemek közötti különbségei hozzájárulhatnak az emberek közötti immunitásbeli különbségekhez, így a fertőzések elleni védekezésben tapasztalható heterogenitáshoz (Loher és mtsai, 2014; Rotival és mtsai, 2020). Az RNS vírusok és az új generációs vakcinák mellett azt hiszem, hogy ez a fontos orvosbiológiai aspektus is ráirányítja a figyelmet az RNS-alapú kutatások fontosságára, és örvendetes módon ehhez a 2023-ban Karikó Katalinnak és Drew Weissmannak odaítélt élettani és orvostudományi Nobel-díj is jelentősen hozzájárult.

Az elmondottak alapján azt gondolom, hogy a sejtjeinkben még ma is jelenlevő „RNS világ” nem pusztán evolúciós múltunk relikviája, hanem a mai napig a földi élet aktív formálója. Én pedig személy szerint örülök, hogy az elmúlt években a nem-kódoló RNS-ekre irányuló kutatásaim kapcsán betekintést nyerhettem ebbe a csodálatos „RNS világba”.

„If you know you are on the right track, if you have this inner knowledge, then nobody can turn you off... no matter what they say."
(Barbara McClintock)

2. Mobilis genetikai elemek – a kódoló és a nem kódoló elemek határán

A XX. század elején az öröklődés Mendel által felismert szabályainak „újrafelfedezése" kétségkívül világszerte óriási lendületet adott a genetika fejlődésének. Az elképzelés, hogy az egyedek tulajdonságait az örökítőanyag diszkrét egységei, a gének határozzák meg, egy nagyon használható, a korabeli keresztezési kísérleteket többségében jól magyarázó modellnek bizonyult – annak ellenére, hogy az öröklődés anyagi alapjai még jó ideig tisztázatlanok maradtak (Keynes és Cox, 2008; Barton, 2022). Volt azonban egy kutató, aki már az 1940-es években elkezdte feszegetni a genetika frissen megszilárdult határait: ő volt Barbara McClintock, aki kukoricán végzett keresztezési eredményeinek egy részét sehogy se tudta a Mendel-féle génmodell segítségével megmagyarázni. Egy új ötlettől vezérelve azonban feltételezte, hogy bizonyos szabályozó gének elmozdulhatnak eredeti pozíciójukból, és egy új régióba átkerülve befolyásolhatják a környező gének működését (McClintock, 1950). Az új, mobilitáson alapuló modelljét a kortársak azonban nem fogadták el, és néhány közleményt követően (McClintock, 1953) McClintock sem publikált jó ideig ebben a témában. A következő évtizedekben azonban sorra kerültek elő bizonyítékok mobilis genetikai elemekről, „transzpozonokról": először baktériumokban írtak le „inszerciós szekvenciákat" (Shapiro, 1969), majd egyre több növényi és állati genomban derült fény ezek jelenlétére, ill. kiderült, hogy az emberi genomban előforduló repetitív elemek jó része transzpozon szekvenciákból áll (Flavell és mtsai, 1974; Baldari és Amaldi, 1976; Crain és mtsai, 1976; Picard és mtsai, 1978; Rubin és mtsai, 1982; Rubin és Spradling, 1982; Venturini és mtsai, 1987). A tudományos közösség számára is nyilvánvalóvá vált, hogy a genomok többségében léteznek mobilis genetikai elemek, és ezek az eredmények Barbara McClintock számára is meghozták az elismerést: számos tudományos díj mellett a mobilis genetikai elemek felfedezéséért 1983-ban megkapta az élettani és orvostudományi Nobel-díjat. Tudománytörténeti és egyben tudománypolitikai érdekesség, hogy máig ő az egyetlen nő, aki ebben a kategóriában osztatlanul egymaga kapta meg a díjat (https://www.nobelprize.org/prizes/facts/facts-on-the-nobel-prize-in-physiology-or-medicine/).

No de mik is azok a mobilis genetikai elemek, vagy más néven transzpozonok? A népszerűsítő irodalomban gyakran „ugráló gének”-ként hivatkoznak rájuk, ami szemléletes, ugyanakkor nem teljesen pontos meghatározás. Vannak ugyanis olyan (jellemzően nem-autonóm, lásd később) transzpozonok, amelyek még a legfrissebb ENCODE definíciók alapján sem tekinthetők géneknek. Transzpozonok alatt olyan, prokariótákban és eukariótákban egyaránt előforduló DNS szekvenciákat értünk, amelyek képesek áthelyeződni egy DNS lókuszról egy másikra[1]. A transzpozíció történhet egy DNS molekulán (például egy kromoszómán vagy egy plazmidon) belül, vagy különböző, akár független eredetű DNS molekulák között (például egy prokarióta plazmidról egy eukarióta genomrégióba). A transzpozonokat többféle szempont alapján lehet csoportosítani (Wicker és mtsai, 2007): a legelterjedtebb, és egyben a modern transzpozon osztályozás alapját is képező szempont a transzpozícióban résztvevő intermedier alapján történő csoportosítás, így megkülönböztetve az **RNS** (más néven Retro-, vagy *Class I*) **transzpozonokat** a **DNS** (más néven *Class II*) **transzpozonoktól**. Egy másik szempont a transzpozíció mechanizmusát veszi alapul, megkülönböztetve a **konzervatív** („*cut-and-paste*”) mechanizmust mutató transzpozonokat a **replikatív** („*copy-and-paste*”) transzpozonoktól. Ez utóbbi felosztás az intermediereken alapuló csoportosítással nagyrészben, de nem teljesen átfedő kategorizálás: az RNS transzpozonok mobilizálódása replikatív mechanizmussal, a DNS transzpozonok többségénél pedig konzervatív mechanizmussal történik, de például a DNS transzpozonokhoz tartozó Helitronok transzpozíciója replikatív mechanizmussal megy végbe. Végül fontos megkülönböztetnünk **autonóm** és **nem-autonóm** transzpozonokat: az előbbieknél a transzpozon tartalmaz minden, a transzpozícióhoz szükséges kódoló és nem-kódoló szekvenciát, míg az utóbbi csoportba tartozók hiányosak, és a transzpozíciójuk csak más aktív transzpozonok jelenlétében történhet.

A transzpozonok tehát a szekvenciájukat tekintve összetett entitások: egy autonóm transzpozon mindig tartalmaz fehérjekódoló géneket, amelyek a transzpozíció mechanizmusában kulcsszerepet töltenek be. Az egyszerű felépítésű DNS transzpozonok esetében ez sokszor egyetlenegy fehérje, a transzpozáz kódolását jelenti, amely gyakran más makromolekuláris partnerekkel (például fehérjékkel) való kölcsönhatás nélkül, akár *in vitro* is képes a célszekvencia hasítására (Claeys Bouuaert és mtsai, 2013). A

[1] A kezdetben felfedezett prokarióta mobilis elemeket „pusztán” inszerciós szekvenciáknak hívták, és csak az általuk közrefogott szekvenciákkal együtt nevezték őket transzpozonoknak (Kleckner, 1981), ez a nevezéktan azonban mára már idejétmúlt (Harmer és mtsai, 2020).

transzpozonoknak ugyanakkor nagyon fontos részét képezik nem-kódoló szekvenciák is: ezek legtöbbször a transzpozont határoló, repetitív szekvenciák, amelyek a transzpozíciót végző fehérjék célszekvenciái, és amelyek az egyes transzpozon csoportokra specifikusak (Wicker és mtsai, 2007). A nem-autonóm elemek az esetek többségében transzpozonok nem-kódoló szegmenseinek a maradványai, ugyanakkor ilyenek más mechanizmusokkal is létrejöhetnek (például az Alu elemek, lásd később). A transzpozonok eredetére vonatkozóan többféle elgondolás létezik, az azonban biztos, hogy az egész élővilágban elterjedtek (Feschotte és Pritham, 2007; Cosby és mtsai, 2019; Wells és Feschotte, 2020). Felvetődik viszont a kérdés, hogy mitől ilyen sikeresek, és a genomok szempontjából alapvetően invazív tulajdonságuk ellenére miért voltak képesek évmilliókon keresztül megmaradni, sőt továbbterjedni a genomokban? Kutatócsoportommal az elmúlt évtizedben ennek a kérdéskörnek nagyon sok aspektusával foglalkoztunk, és a következő fejezetekben ezeknek a kutatási irányoknak a főbb eredményeit fogom ismertetni.

2.1. Transzpozonok a genomban: paraziták vagy domesztikált rezidensek?

Egyes genomokban a transzpozonok meglepően nagy arányban vannak jelen: az emberi genom legalább 45%-a transzpozon eredetű, a kukorica genomjában ez a szám 85% (Wells és Feschotte, 2020)! Nagy kérdés persze, hogy ezeknek vajon mekkora hányada aktív transzpozon, de ezzel együtt mégis hogyan lehetséges az, hogy ilyen nagyszámú invazív elem mellett a genom látszólag jól működik? A kérdés azért is zavarba ejtő, hiszen kezdetben a transzpozonokra alapvetően mint genomi parazitákra gondoltak: ők a tipikus megtestesítői a Richard Dawkins által elnevezett „önző gének”-nek (Dawkins, 2021). Kétségkívül igaz, hogy a transzpozíció önmagában is mutációs teherrel jár, illetve a nagyszámú transzpozon szekvencia a genomban képes nem kívánt genomi átrendeződéseket („illegitim” rekombinációkat) indukálni, amelyek negatív hatására már csak az emberi betegségek tekintetében is számos példát ismerünk (Hancks és Kazazian, 2016; Burns, 2020). Természetesen véletlen mutációk kapcsán a transzpozonok is inaktiválódhatnak, de ez önmagában nem jelent hatékony védelmet a paraziták ellen. Az elmúlt évtizedekben ugyanakkor nagyon sok olyan mechanizmust is megismertünk, amelyeken keresztül a genomok aktívan védekeznek a transzpozonok ellen. Ilyenek a már korábban ismertetett, az RNAi folyamatok közé tartozó siRNS és piRNS útvonalak (Ghildiyal és Zamore, 2009; Onishi és mtsai, 2021; Kong és mtsai, 2022; Wang és mtsai, 2022), illetve néhány, transzpozon-specifikus miRNS által kiváltott szabályozási útvonal

(Pedersen és Zisoulis, 2016), de ide sorolhatunk nagyszámú, a transzpozonok átíródását gátló DNS metilációs vagy hiszton módosítással járó folyamatot, amelyek a transzpozonok promóter működését célzottan inaktiválják (Cosby és mtsai, 2019; Almeida és mtsai, 2022). Ez természetesen egy folyamatos evolúciós versengés, amelynek az eredményeképpen azt várnánk, hogy a transzpozonok számát és arányát tekintve egy adott genomban hosszú távon beáll egy egyensúly – érdekes azonban, hogy a transzpozon tartalom tekintetében milyen nagy különbségek vannak az egyes genomok között. Ha csak a transzpozíció mechanizmusát nézzük, akkor pedig azt várnánk, hogy alapvetően a replikatív mechanizmusú, elsősorban tehát a retrotranszpozonok teszik ki a genomban található transzpozonok nagy részét. Több faj esetében ez megállja a helyét: két élesztőfaj, a *Saccharomyces cerevisiae* és a *Schizosaccharomyces pombe* genomjában kizárólag, míg az egér és az ember genomjában túlnyomórészt retrotranszpozonokat találunk. Sok genom vizsgálata azonban nem támasztotta alá ezt a várakozást: szép számmal akadnak olyan genomok is, ahol arányaiban jóval több (a rizs vagy a *Caenorhabditis elegans*), vagy kizárólag csak konzervatív módon replikálódó DNS transzpozonok fordulnak elő, mint például a *Trichomonas vaginalis* élősködő genomjában (Biemont és Vieira, 2006; Feschotte és Pritham, 2007; Wells és Feschotte, 2020). Úgy tűnik tehát, hogy a különböző mechanizmusok genomonként jelentősen eltérőek lehetnek, és ebből a szempontból minden genomot egyedileg kell vizsgálni.

Az első eredmények, amelyek a transzpozonok kizárólagos negatív szerepét kétségbe vonták, a baktériumok vizsgálatával kezdődtek. Kiderült, hogy a baktériumok antibiotikum rezisztenciájáért felelős géneket akár transzpozonok is mozgathatják, amelyek így a saját mobilizációjuk „melléktermékeként” terjeszthetik a rezisztenciát a populációban (Firth és Skurray, 1998). Ez mindenképp egy előnyös tulajdonság, hiszen az adott környezetben a gazdaszervezet hozzájut a túléléshez szükséges eszköztárhoz. A későbbiekben pedig zavarba ejtően komplex kölcsönhatásokat írtak le a transzpozonok és gazdagenomjaik között. Meglepő felfedezés volt például az ún. processzált pszeudogének jelenléte az emlős genomokban, amelyek túlnyomó többsége a L1 típusú retrotranszpozonok „promiszkuitása” kapcsán jött létre: a transzpozon által kódolt ORF1 és ORF2 fehérjék kötődhetnek a citoplazmában nagyszámban jelenlevő mRNS molekulák egyikéhez, és az L1 retrotranszpozon a reverz transzkripció során az erről képződött kópiát is képes integrálni a genomba (Kazazian, 2004; Kroutter és mtsai, 2009). A processzált pszeudogének jellemzője, hogy nem tartalmaznak intronokat, illetve a reverz transzkriptáz alacsonyabb hatékonysága miatt gyakran csak az eredeti mRNS

csonkolt, 3’ végi szekvenciáját tartalmazzák (Pavlicek és mtsai, 2002a; Pavlicek és mtsai, 2002b; Baertsch és mtsai, 2008). Bizonyos esetekben ugyanakkor megtörténhet, hogy teljes fehérjekódoló szekvenciák egy promóter mögé illesztődve az evolúció során a gazdagenom számára hasznos, aktív génekké alakulnak, amelyre a multigénes fehérjecsaládok esetében számos példát találunk (Nei és Rooney, 2005; Cheetham és mtsai, 2020). A LINE-1 (L1) típusú retrotranszpozonok „felelősek” továbbá a primáta genomokban a nem-autonóm Alu-szekvenciák elterjedéséért: ezek a szignálfelismerő részecskében található 7SL RNS-sel rokonságot mutató szekvenciák a humán genomban például megközelítőleg 1,1 millió kópiában találhatók meg, ami a genomunk ~11%-át teszi ki (Lander és mtsai, 2001; Venter és mtsai, 2001; Kazazian, 2004). Ezek a rövid (~300 bp hosszúságú) nem-kódoló elemek ugyanakkor fontos RNS processzálási szignálokat, például splicing donor és akceptor szekvenciákat hordoznak, amelyek az *exon shuffling* folyamatán keresztül bizonyítottan meghatározó szerepet játszanak a fehérjék evolúciójában (Long és mtsai, 2013; Cosby és mtsai, 2021). Érdekes kérdés, hogy a primátáknak az emlősökön belüli látványos (és számunkra előnyös) evolúciós sikere mennyiben oka vagy következménye az Alu-elemek aktivitásának, de kétségtelen tény, hogy ezek a retrotranszpozonok nagyon is aktív formálói például az emberi genomnak (Schrider és mtsai, 2013; Cheetham és mtsai, 2020).

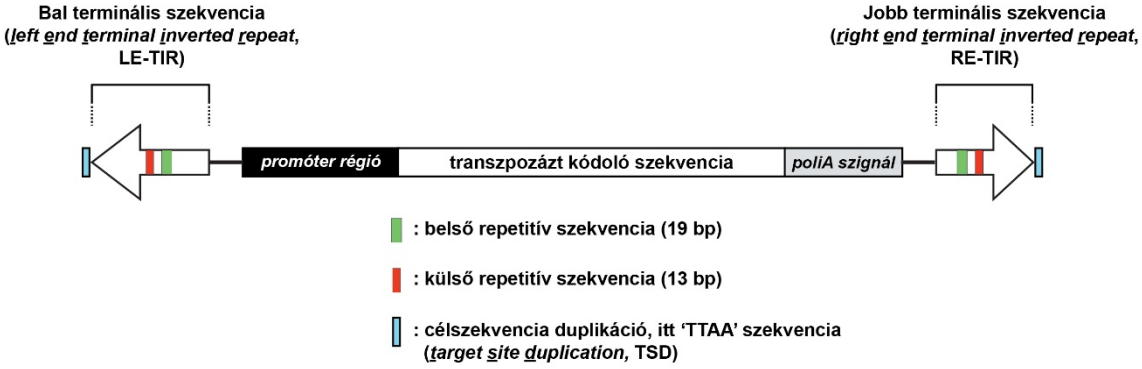
A transzpozonok aktív, genomot formáló hatásának kétségkívül a legizgalmasabb példájára az emlős idegrendszer fejlődésének tanulmányozása kapcsán bukkantak. Kiderült, hogy a neuronális progenitor sejtekben hirtelen megnő a LINE-1 retrotranszpozonok aktivitása, amely a potenciális genotoxikus hatás (az inszerciós mutagenezis) mellett néha előnyökkel is járhat: előfordulhat, hogy egy, a neuronális fejlődésben fontos gén (például egy neurotranszmittert kódoló gén) közelébe integrálódva serkenti annak expresszióját. Mivel a neuronális progenitorok nagyszámú osztódása kapcsán ebben az időszakban a sejtek között egy klonális szelekció zajlik, a negatív hatással bíró sejtek kiszelektálódnak, míg az esetlegesen jobban működő idegsejtek az erősebb sejt-sejt közötti kapcsolataik révén nagyobb eséllyel lesznek részesei a kialakuló idegpályáknak (Muotri és mtsai, 2005; Singer és mtsai, 2010). A transzpozonok hatása itt komolyan elgondolkodtató: egy olyan mechanizmusról van szó, amely az idegrendszerben szomatikus mozaicizmust hoz létre, amelynek eredményeképpen lehetőség van egyedi szinten egy jobban működő idegrendszer felépítésére, és ezáltal az egyed nagyobb esélyű túlélésére. A jelenséget eredetileg egerek idegrendszerét vizsgálva írták le, de van bizonyíték rá emberi neuronok esetében is (Coufal és mtsai, 2009). A

bemutatott folyamat a genotípus-fenotípus összefüggések komplexitása mellett további érdekes filozófiai/etikai kérdéseket is felvet: mivel egy sztochasztikus jelenségről van szó, ez tovább csökkenti a tökéletes klónok létrehozásának az esélyét. Ugyanis még ha egy adott genomot tartalmazó egyed azonos környezeti körülmények között fejlődik is, a fent leírt véletlenszerű folyamatok miatt tulajdonságait tekintve valószínűleg teljesen egyedi lesz, és nem teljesen identikus a genetikai klónjával. Nem mintha ehhez kapcsolódóan bárki is etikusnak tartaná például tökéletes „Einstein-klónok" létrehozását, de a természet – úgy tűnik, például a transzpozonokon keresztül – ennek a lehetőségét igencsak korlátozta…

A transzpozonok pozitív módon kétségkívül a szekvenciális repertoárjukkal járulnak hozzá leginkább az élőlények genetikai változatosságához. A fent említett védekező mechanizmusok, illetve a véletlenszerű mutációk révén a mobilis elemeknek sok inaktív kópiája halmozódhat fel a genomokban. Érdekes jelenség az ún. genetikai domesztikáció (vagy ko-optálás), amikor a parazitából „rezidens" lesz, vagyis a transzpozon szekvenciák a genom saját endogén működésében hasznosulnak (Sinzelle és mtsai, 2009; Alzohairy és mtsai, 2013; Jangam és mtsai, 2017). Meglepően sok példát ismerünk arra, amikor fontos sejtélettani funkciókat transzpozáz eredetű fehérjék látnak el: ilyenek például az immunfolyamatokban az antitestek variabilitásának kialakításáért felelős RAG rekombinázok, amelyek a *Transib* szupercsaládba tartozó transzpozonok leszármazottjai; a CENP-B centroméra fehérjék, amelyek egy, a *TC1/mariner* szupercsalád *pogo* alcsoportjába tartozó transzpozonból eredeztethetők; vagy például a *Drosophila melanogaster* telomeráz fehérjéi, amelyek nem-LTR típusú retrotranszpozonokra vezethetők vissza (Pardue és DeBaryshe, 2003; Feschotte és Pritham, 2007; Cosby és mtsai, 2019; Burns, 2020). A nem-kódoló régiók evolúciójának pedig különösen érdekes esetei, amikor miRNS kötőhelyeknek, vagy akár maguknak a miRNS-eknek a kialakulása repetitív transzpozon régiókra vezethető vissza (Piriyapongsa és mtsai, 2007; Qin és mtsai, 2015). Mindezen példák nagyban hozzájárultak ahhoz az árnyalt képhez, ahogy ma a transzpozonok szerepére tekintünk: a genomokra nehezedő mutációs teher mellett számtalan esetben igazolódott, hogy a transzpozonok pozitív, az élőlények szempontjából előnyös módon is befolyásolhatják a genetikai anyag evolúcióját. Az elmúlt évek kutatásai során az általam vezetett munkacsoportban a *piggyBac* transzpozonok példáján keresztül vizsgáltuk ezt az érdekes jelenséget, vagyis elsősorban az emberi genomban a transzpozonokhoz köthető aktív evolúciós folyamatokat.

2.1.1. piggyBac-szerű elemek a humán genomban

Az első *piggyBac* transzpozont a *Trichoplusia ni*[2] nevű bagolylepkében azonosították (Fraser és mtsai, 1983), később azonban a szupercsalád több képviselőjét az egysejtűektől a primátákig számos élőlényben megtalálták (Lander és mtsai, 2001; Pritham és mtsai, 2005; Bouallegue és mtsai, 2017). Az eredetileg leírt rovar transzpozon az egyszerű szerkezetű DNS transzpozonok közé tartozik: az 1,8 kb hosszúságú, egy 68 kDa méretű transzpozáz fehérjét kódoló szakaszt 200-300 bp méretű végszekvenciák határolják, amelyekben egy-egy 13 bp és 19 bp hosszúságú terminális fordított repetitív szekvenciát találunk (***T****erminal* ***I****nverted* ***R****epeat* – TIR). A konzervatív *„cut and paste”* mechanizmussal működő transzpozon célszekvenciája egy ’TTAA’ szekvencia, amely a transzpozíció során megduplázódik, és a már beintegrálódott transzpozont két oldalról ez a szekvencia határolja (***T****arget* ***S****ite* ***D****uplication* – TSD; részletek a **15. ábrán**).



15. ábra: A természetes piggyBac transzpozon szerkezete. A transzpozázt kódoló transzkripciós egységet határoló terminális szekvenciák ugyan eltérő hosszúságúak (a LE-TIR hosszabb), de mindkét TIR szekvenciában megtalálható egy belső és egy külső repetitív szekvenciaelem, ellentétes orientációban.

A kezdeti vizsgálatok során kiderült, hogy a transzpozon hatékonyan működik nemcsak rovarsejtekben, hanem más állati, illetve emberi eredetű sejtekben egyaránt (Cary és mtsai, 1989; Yusa, 2015). Ez a tulajdonság, valamint a *piggyBac* transzpozáz fehérje azon képessége, hogy „nyom nélkül” képes a transzpozon szubsztrátot kivágni a DNS-ből, vezetett ahhoz, hogy a *piggyBac* transzpozonos rendszert széles körben alkalmazzák génbeviteli eljárásokban, illetve génterápiás fejlesztések során egyaránt

[2] A hazánkban is előforduló lepkefaj hivatalos magyar neve vándor aranybagoly, de korábban u-betűs vagy v-betűs aranybagolyként is említették. (https://www.izeltlabuak.hu/)

(Ding és mtsai, 2005; Cadinanos és Bradley, 2007; Mitra és mtsai, 2008; Manuri és mtsai, 2010; Yusa és mtsai, 2011) (ill. lásd később, 2.2. fejezet). Ugyanakkor az ilyen irányú felhasználások kapcsán komoly problémát jelentett, hogy az emberi genomban több *piggyBac*-eredetű szekvencia is előfordul, amelyek esetleges aktivitásukkal komolyan befolyásolhatják a génbevitel hatékonyságát, illetve a létrehozott sejtek genetikai stabilitását (Sarkar és mtsai, 2003). Az emberi génállomány vizsgálata során arra derült fény, hogy a DNS transzpozon eredetű elemekből a retrotranszpozonokhoz képest jóval kevesebb található meg a humán genomban, és ezen belül is a *piggyBac* szupercsalád tagjai vannak a legkisebb számban reprezentálva. Az adatok alaposabb elemzése ugyanakkor azt mutatta, hogy a genetikai állományunkban még így is megközelítőleg 2000 darab *piggyBac*-eredetű szekvencia található, amely génterápiás szempontból egy komoly veszélyforrást jelent (Mandal és Kazazian, 2008). A részletesebb bioinformatikai elemzések és annotációs adatok alapján azt lehetett tudni, hogy az emberi génállományban 5 darab, a rovar *piggyBac* transzpozáz fehérjével rokon szekvenciát kódoló gén található, amelyeket ehhez kapcsolódóan neveztek el PGBD elemeknek (***p**iggy**B**ac-**d**erived sequences*, PGBD1-5) (Bouallegue és mtsai, 2017). A többi azonosítható *piggyBac*-eredetű szekvencia valószínűleg a TIR elemek maradványaiból származik, ugyanakkor az emberi genomban az „eredeti” transzpozonos konstellációt mutató, vagyis a transzpozázt kódoló szekvenciát mindkét oldalról TIR szekvenciák által határolt DNS szakaszokat nagyon keveset találunk. Ezek az adatok arra utaltak, hogy a *piggyBac* transzpozonok valaha ugyan aktívak voltak a hominida őseinkben, transzpozíciós képességüket azonban mostanra valószínűleg elvesztették (Sarkar és mtsai, 2003; Bouallegue és mtsai, 2017).

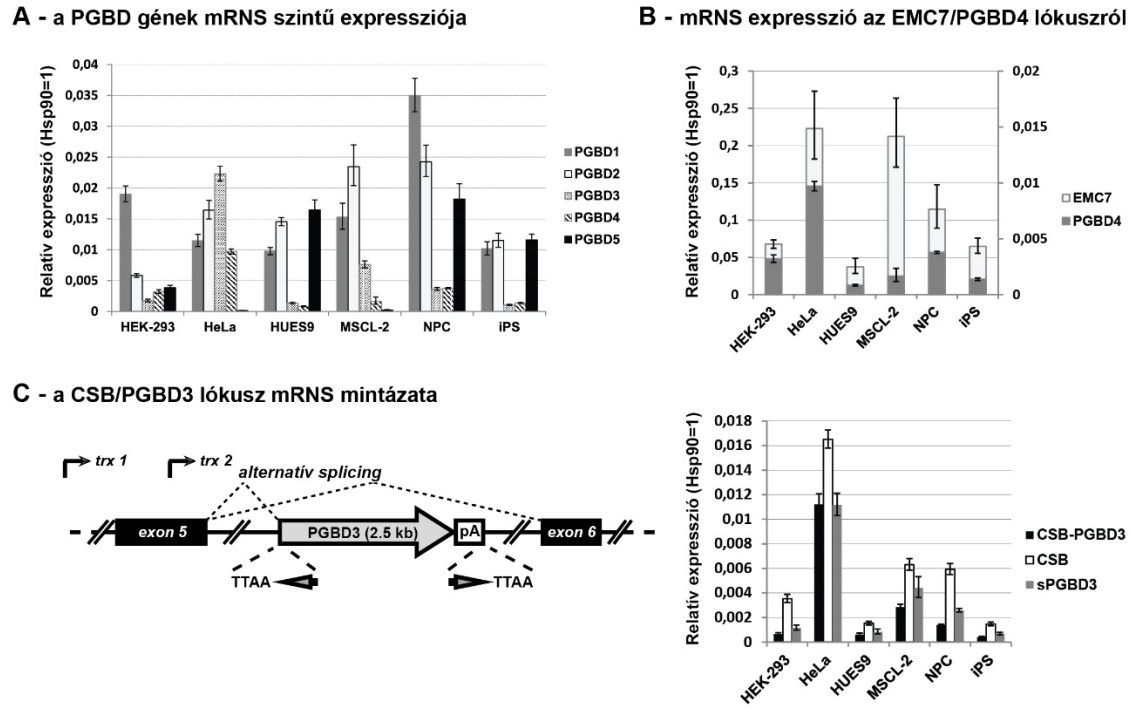
A génterápiás alkalmazások esetében ugyanakkor nem lehet valószínűségekre hagyatkozni, hiszen bármilyen alacsony szintű transzpozíciós aktivitás, vagy a rovar eredetű génbeviteli rendszerrel akár csak a legkisebb mértékű keresztreakció is a beteg számára komoly veszélyt jelenthet. A problémát tovább bonyolítja, hogy az elmúlt évek vizsgálatai kapcsán ebben a kérdésben a tudományos szakirodalomban sincsenek egységes álláspontok. A kezdeti bioinformatikai és kísérletes eredmények arra mutattak, hogy nagy valószínűséggel sem a humán PGBD fehérjék, sem a TIR szekvenciák nem képesek transzpozícióra, ugyanakkor a vizsgálatok csak néhány kiválasztott elem vizsgálatán alapultak (Pavelitz és mtsai, 2013; Saha és mtsai, 2015). Időközben megjelent azonban egy olyan közlemény, amelyben a szerzők állításuk szerint bizonyítják, hogy a PGBD5 képes a transzpozícióra, sőt képes mozgatni a rovar eredetű transzpozonos

szekvenciákat (Henssen és mtsai, 2015). Az eredmény azért is meglepő volt, mert az összes PGBD fehérje közül az emlős PGBD5 a legősibb szekvencia, amely az evolúciós törzsfán már a félgerinchúrosok őseiben megjelenhetett, és az evolúciós távolságból adódó diverzifikációt is figyelembe véve ennek a fehérjének az esetében a legkisebb a valószínűsége, hogy a szupercsaládon belül egy távoli rovar eredetű transzpozonnal keresztreagáljon (Pavelitz és mtsai, 2013). Az általam vezetett kutatócsoport is ebben az időszakban kezdett el behatóbban foglalkozni a *piggyBac* transzpozonok működésével és evolúciójával, és a témában dolgozó több kutatócsoporttal együtt mi is értetlenül álltunk a publikált eredmény előtt. A komolyabb problémát az jelentette, hogy sem mi, sem pedig más általunk ismert kutatócsoport nem tudta reprodukálni a PGBD5-tel kapcsolatos eredményeket. A kérdéskört tovább árnyalta, hogy a szokatlan eredményt bemutató eredeti kutatócsoport tagjai néhány későbbi publikációjukban már sokkal árnyaltabban nyilatkoztak a kérdésről, a PGBD5 transzpozíciós aktivitása helyett egyre inkább csak a fehérje által kiváltott DNS törésekről és rekombinációs eseményekről beszéltek (Henssen és mtsai, 2016; Henssen és mtsai, 2017a; Henssen és mtsai, 2017b), és felmerült a gyanú, hogy az eredeti kísérleti eredmények esetleg valamilyen technikai hibára vezethetők vissza. Ugyanakkor az eredeti eredmény óriási kockázatot jelentene a *piggyBac*-alapú génterápiás eljárásokra nézve (Ivics, 2016), így több laboratórium is alapos vizsgálatokba kezdett a PGBD5, illetve a többi PGBD fehérje feltételezett transzpozíciós aktivitása kapcsán. Ezekre alapozva kutatásaink során mi is azt tűztük ki célul, hogy az összes humán PGBD fehérje szisztematikus vizsgálatával igyekszünk feltérképezni a *piggyBac* transzpozonok evolúcióját az emberi genomban, és választ kapni arra, hogy ezek vajon jelenleg is aktív mobilis elemek, vagy inkább már domesztikálódott „rezidens” fehérjékként működnek.

A vizsgálatokat a humán PGBD gének expressziós mintázatainak analízisével kezdtük. Megmutattuk, hogy ezek a transzpozáz eredetű fehérjéket kódoló gének mRNS szinten magas és nagyon markáns, de egymástól eltérő szövetspecifikus expressziót mutatnak (**16A ábra**). A PGBD3 például egy természetes „géncsapdaként” a ***C**ockayne **S**yndrome group **B*** (CSB) gén 5. intronjában helyezkedik el, és a lókusz egy komplex alternatív splicing mintázatot mutat (Newman és mtsai, 2008; Bailey és mtsai, 2012; Gray és mtsai, 2012; Weiner és Gray, 2013). Az általunk erre a régióra speciálisan kifejlesztett *real-time* PCR esszé (Kolacsek és mtsai, 2017) segítségével igazoltuk, hogy a splicing variánsok expressziós szintje jellegzetes szövetspecifikus mintázatot mutat, amely aktív szabályozásra, és fontos endogén regulációs szerepre utal (**16C ábra**). Hasonlóan

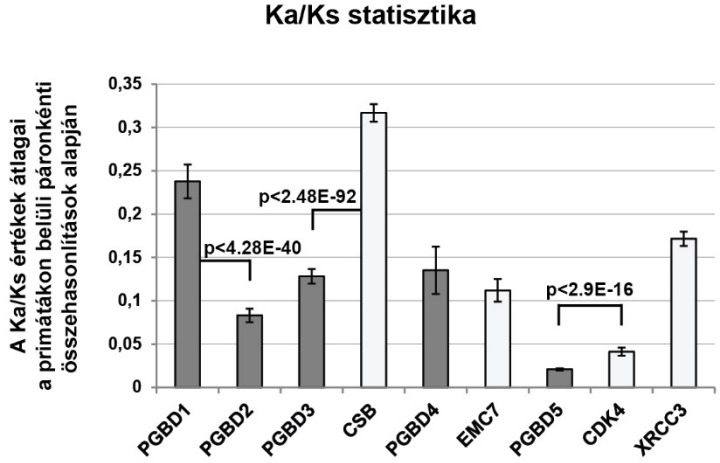
összetett szabályozást mutattunk ki a PGBD4 esetében is: a transzkripciót ebben az esetben egy bidirekcionális promóter irányítja, és a másik irányban átíródó EMC7 gén valamint a PGBD4 expressziós szintje egymástól független szövetspecifikus mintázatot mutat (**16B ábra**). Az expressziós mintázatok mellett a főemlősökön (primátákon) belül a PGBD fehérjék evolúciós konzerváltságát a Ka/Ks analízis segítségével is megvizsgáltuk, amely a fehérjekódoló szekvenciák páronkénti összehasonlításával, a nem-szinonim (aminosav cserével járó, 'Ka') és a szinonim (*silent*, 'Ks') változások arányából következtet az evolúciós divergenciára (Hurst, 2002). Az adatokat összehasonlítottuk más konzervált gének, például a sejtciklus gének, vagy a DNS javítórendszerek génjeinek értékeivel. A PGBD fehérjék Ka/Ks értékei általában alacsonynak, 0,3 alattinak bizonyultak, amely általánosságban is az erősen konzervált génekre jellemző tulajdonság (Hurst, 2002). A korábbi analízisekhez (Pavelitz és mtsai, 2013) hasonlóan a PGBD5 bizonyult a legkonzerváltabbnak, még a CDK4 sejtciklus génhez viszonyítva is, ugyanakkor például a PGBD2 is nagyon alacsony Ka/Ks értéket, így magas fokú evolúciós konzerváltságot mutat (**17. ábra**). Érdekes, hogy míg a közös promótert használó PGBD4 és EMC7 fehérjék esetében hasonló értékeket kaptunk, addig a PGBD3 viszonylag erős konzerváltsága mellett a „gazdagén" vizsgálatakor sokkal magasabb Ka/Ks értéket, vagyis alacsonyabb konzerváltságot tudtunk kimutatni, amely valószínűleg a génekre ható eltérő szelekciós nyomásból adódhat (**17. ábra**). Mind az expressziós értékek, mind pedig az evolúciós konzerváltsági adatok azt támasztották alá, hogy a PGBD géneknek fontos endogén funkcióik lehetnek, amelyek feltérképezése további vizsgálatokat igényel (Kolacsek és mtsai, 2022).

Az irodalmi adatok fényében egy átfogó analízissel szerettük volna tisztázni, hogy bármelyik *piggyBac*-eredetű humán szekvencia megőrizte-e a transzpozíciós képességét. Van ugyan példa olyan transzpozon eredetű génekre, amelyek a domesztikáció során megőrizték a mobilizációs képességüket (ilyen a THAP9 gén, (Majumdar és mtsai, 2013)), az esetek többségében azonban a ko-optált elemek elveszítik a transzpozíciós aktivitásukat (Kapitonov és Jurka, 2005; Cordaux és mtsai, 2006; Cosby és mtsai, 2019; Miskei és mtsai, 2021). A humán PGBD gének esetében ennek a vizsgálata nagy kihívást jelent, hiszen a transzpozáz fehérjék esetében nehéz, vagy sokszor nem is lehet azonosítani a feltételezett szubsztrátként funkcionáló TIR szekvenciákat.



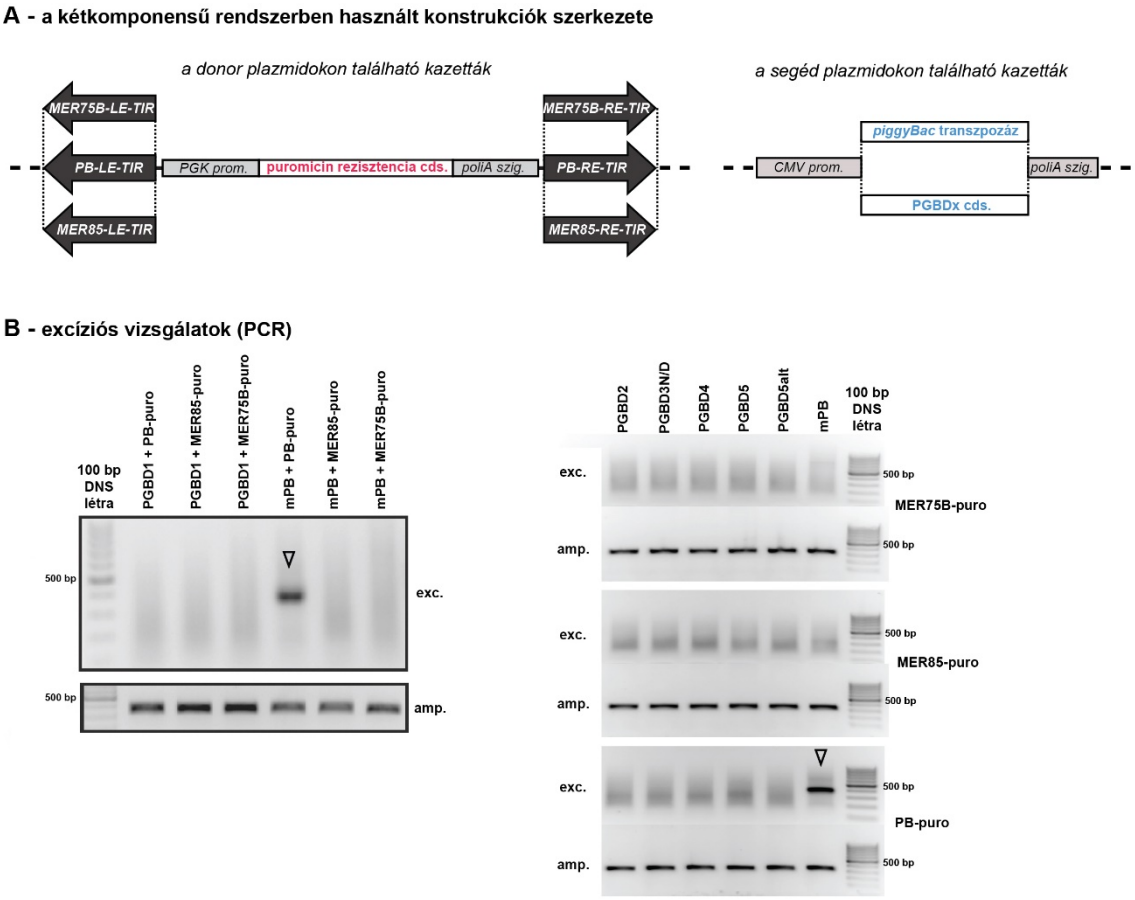
16. ábra: A humán PGBD gének expressziós mintázatai. (**A**) A humán PGBD gének expressziós mintázata különböző sejtvonalakon, *real-time* kvantitatív PCR módszerrel meghatározva. (**B**) A bidirekcionális promóterről átíródó PGBD4 és EMC7 gének relatív expressziója. (**C**) A CSB gén 5. intronjában elhelyezkedő humán PGBD3 szerkezete (bal oldali ábra), illetve a lókuszról átíródó splicing variánsok expressziós mintázata (jobb oldali grafikon). A 'trx 1' a CSB gén promóteréről történő transzkripciót, míg a 'trx 2' az 5. exonban elhelyezkedő, kriptikus promóterről történő átírás kezdetét jelöli. Az RNS-szintű méréseknél a Hsp90 mRNS expressziós szintje szolgált endogén kontrollként, a grafikonokon bemutatott értékek pedig legalább 3 mérés átlagát jelentik; a hibazászlók a 95%-os konfidencia intervallumot jelölik. A HEK-293 és a HeLa ismert sejtvonalak mellett az expressziós vizsgálatokat egy humán embrionális őssejtvonalon (HUES9), egy abból differenciáltatott mezenchimális őssejtvonalon (MSCL-2), őssejtekből differenciáltatott neuronális progenitor sejteken (NPC), valamint indukált pluripotens őssejteken (iPS) végeztük. Az őssejtvonalakból és az ezekből differenciáltatott sejtekből származó mintákat Dr. Apáti Ágota laboratóriuma bocsátotta a rendelkezésünkre.
Az ábra a (Kolacsek és mtsai, 2022) közleményünk alapján készült.

Az emberi genomban fellelhető szerkezeti együttállások kapcsán ugyanakkor komoly eséllyel feltételezhető, hogy a PGBD3-nak valószínűleg egy, a MER85 repetitív szekvenciákhoz, míg a PGBD4 esetében a MER75 szekvenciákhoz hasonló szubsztrátja lehetett (Lander és mtsai, 2001; Sarkar és mtsai, 2003); a PGBD3 esetében kísérletesen is kimutatták a fehérje kötődését ezen TIR elemekhez (Bailey és mtsai, 2012). Míg a



17. ábra: A humán PGBD gének evolúciós vizsgálata. A grafikon a PGBD gének primátákon belüli, páronkénti összehasonlítása alapján készült Ka/Ks analízisek eredményét mutatja. A sötét oszlopok az ortológ PGBD1-5 transzpozáz eredetű géneket, a világos oszlopok más endogén géneket jelölnek. A fontosabb kiemelt esetekben a statisztikai analízisek (kétoldali t-próbák) során kapott p értékeket is mutatjuk (magyarázat a szövegben). A grafikon értékei az adott ortológokon kapott számítások átlagát jelentik, a hibazászlók pedig a 95%-os konfidencia intervallumot jelölik.
Az ábra a (Kolacsek és mtsai, 2022) közleményünk alapján készült.

CSB gén 5. intronjában elhelyezkedő, MER85 szekvenciák által határolt PGBD3 kópia tűnik a leginkább hasonlatosnak egy „kanonikus" *piggyBac*-szerű DNS transzpozonhoz, a transzpozíciós aktivitás ellen szól az a tény, hogy az itt található, és egyébként aktívan átíródó PGBD3 fehérje épp a katalitikus domén kritikus DDD aminosavak mintázatában mutációt hordoz (D352N mutációt, vagyis a motívum DND) (Keith és mtsai, 2008b). A PGBD4 fehérje esetében a DDD motívum érintetlen, a többi PGBD fehérje esetében viszont vagy nehéz azonosítani a 3 kitüntetett aminosav pozíciót, vagy a feltételezett helyeken szintén más aminosava(ka)t találunk. A PGBD5 gén esetében ráadásul kiderült, hogy a Henssen és munkatársai által aktívnak mondott fehérje valójában egy hibás annotáció eredménye, és a legújabb humán genom adatbázisban más szekvencia szerepel (Kolacsek és mtsai, 2022). Mindezek alapján úgy határoztunk, hogy mind az 5 humán PGBD fehérjét, a PGBD5 esetén a korábban használt alternatív terméket („PGBD5alt") is kísérletesen teszteljük, és a rovar *piggyBac* TIR szekvenciák mellett a MER75 és a MER85 szekvenciákkal kombinálva, egy antibiotikum szelekción alapuló transzpozíciós esszében megvizsgáljuk.

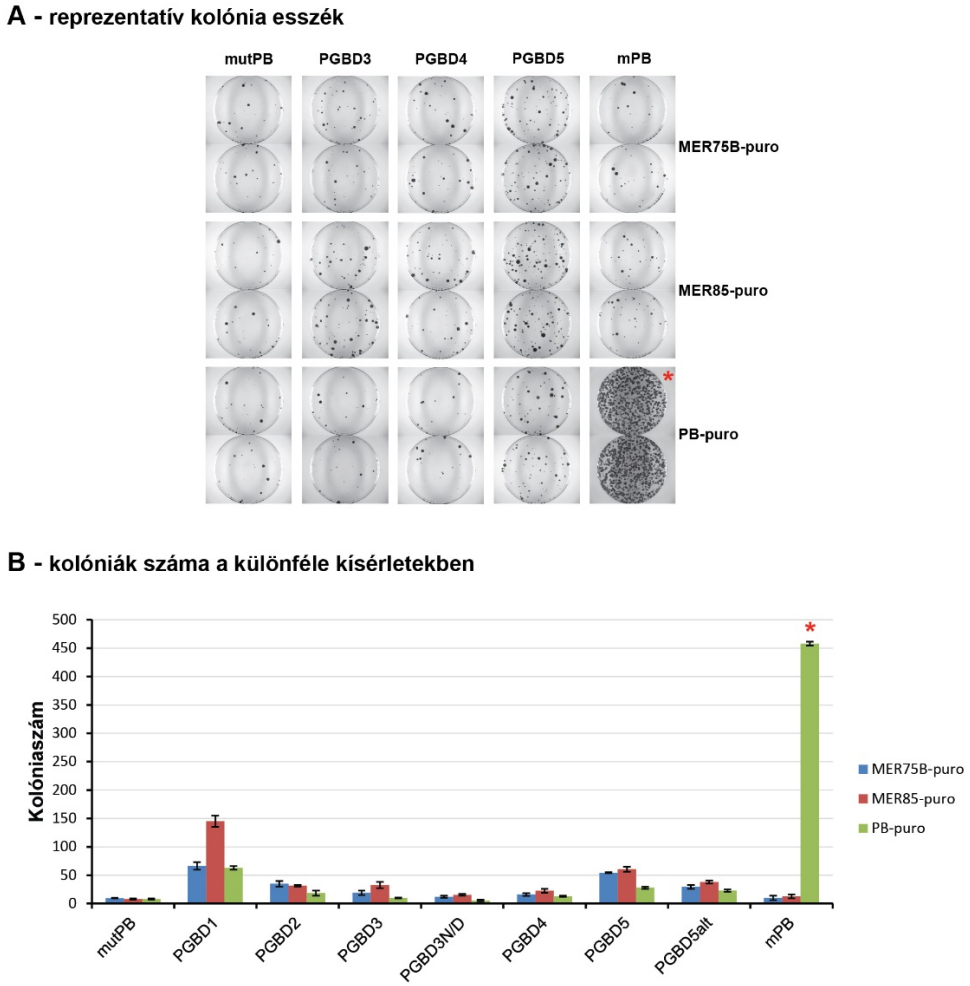


18. ábra: A kétkomponensű transzpozonos rendszer: a plazmidok szerkezete és az excízió vizsgálata. (**A**) A kísérletek során a donor plazmidokon a transzkripciós egységet két oldalról az ábrán jelölt TIR szekvenciapárok egyike határolja, a segéd plazmidokon pedig az aktív transzpozázt, vagy a tesztelendő transzpozáz eredetű fehérjéket („PGBDx") kifejező expressziós kazetták találhatók. PB: rovar *piggyBac*; PGK: foszfoglicerát-kináz promóter; CMV: citomegalovírus promóter; szig.: szignál szekvencia; cds.: kódoló szekvencia. (**B**) A transzpozon excíziót kimutató diagnosztikus PCR vizsgálatok reprezentatív eredményei. A kísérletekben használt transzpozázt vagy transzpozáz eredetű fehérjéket, illetve a velük együttesen expresszált transzpozon konstrukciókat a gélek felett (bal oldali panel), vagy a gélek felett és tőlük jobbra található feliratok mutatják (jobb oldali panel). A transzpozon kivágódása esetén egy 381 bp hosszúságú termék („exc.") jelenik meg, amely a kísérleteink során csak és kizárólag a rovar *piggyBac* transzpozáz és a rovar TIR szekvenciákkal határolt transzpozonok kombinációja esetében volt detektálható (nyilakkal jelölve). Az „amp." felirat a kontroll reakció termékét, az izolált plazmidok gerincén található ampicillin rezisztencia génre specifikus, 340 bp hosszúságú PCR terméket jelöli. mPB: a rovar eredetű *piggyBac* transzpozáz emlős kodon-optimalizált változata; PGBD3N/D: a DDD-motívumban helyreállított humán PGBD3 fehérje; PGBD5alt: egy korábbi hibás annotációból származó alternatív PGBD5 fehérje.

Az ábra a (Kolacsek és mtsai, 2022) és a (Rasko és mtsai, 2022) közleményeink alapján készült.

Kísérleteink során egy kétkomponensű rendszerben, egy „donor” és egy „segéd” („*helper*”) plazmid együttes transzfekcióját követően vizsgáltuk a transzpozíció jeleit a sejtekben. A donor plazmid tartalmazza a transzpozont, vagyis ebben az esetben a feltételezett TIR szekvenciák által közrefogott puromicin rezisztencia gént, míg a segéd plazmid a tesztelendő transzpozáz eredetű szekvenciákat kifejező vektort jelenti (**18A ábra**). A transzpozíció megtörténtét két ponton ellenőriztük: az első lépést jelentő transzpozon kivágódást (excíziót) egy diagnosztikus PCR segítségével, az excízión átesett, majd a sejt DNS javító apparátusa által újra cirkulárissá alakított plazmidok megfelelő régiójának amplifikálásával teszteltük. A transzpozíció befejező lépése pedig az integrációs lépés, amelyet vizsgálatainkban az antibiotikum rezisztenciát mutató sejtek számával, tulajdonképpen a transzgenikus rátával közelítettük. A vizsgálatokban pozitív kontrollként a rovar eredetű *piggyBac* transzpozáz saját szubsztrát TIR szekvenciájával kapott eredményeket, míg negatív kontrollnak a rovar eredetű transzpozáz katalitikusan inaktív (D268G) mutánsával kapott eredményeket használtuk. A kísérletsorozat eredményeképpen megállapítottuk, hogy egyik humán eredetű PGBD fehérje sem volt képes transzpozícióra, mert sem az excízió (**18B ábra**), sem a transzpozon integráció szintjén (**19. ábra**) nem detektáltunk a háttértől jelentősen különböző aktivitást egyik TIR szubsztrát esetében sem. A legerősebb bizonyíték az excízió teljes hiánya volt, hiszen a transzpozíció kezdeti lépése nélkül bármilyen magasabb kolóniaszám más jelenségekre, például random integrációkra, vagy szekvencia mediált illegitim rekombinációkra vezethető vissza (ahogy ez például az alkalmazott MER85 szekvencia esetében valószínűsíthető, lásd **19B ábra**). A szekvenciális adatok és a genomi szerkezet alapján leginkább a DDD-motívumban helyreállított PGBD3 fehérje és a MER85 kombinálásától vártunk pozitív eredményt, azonban ebben az esetben sem tudtunk transzpozíciót kimutatni. Eredményeink így arra utaltak, hogy az emberi genomban található *piggyBac*-eredetű fehérje szekvenciák elvesztették a mobilizációs képességüket, és nagy valószínűséggel ugyanez igaz a *piggyBac*-eredetű TIR szekvenciákra is (Kolacsek és mtsai, 2022). Időközben más laboratóriumokból is megjelentek a témához kapcsolódó közlemények, főleg a PGBD5 feltételezett aktivitásával kapcsolatban: az eredeti PGBD5-ös eredményeket közlő labor vezetője egy francia csoporttal együttműködve publikált olyan adatokat, amelyek a PGBD5 transzpozáz funkcióját hivatottak alátámasztani, bár a transzpozon excízióra vonatkozó egyértelmű eredményeket továbbra sem tudtak bemutatni (Helou és mtsai, 2021b). Megjelent viszont egy komolyabb közlemény egy, a *piggyBac* transzpozonokkal régóta foglalkozó kutatócsoporttól, és az ő adataik

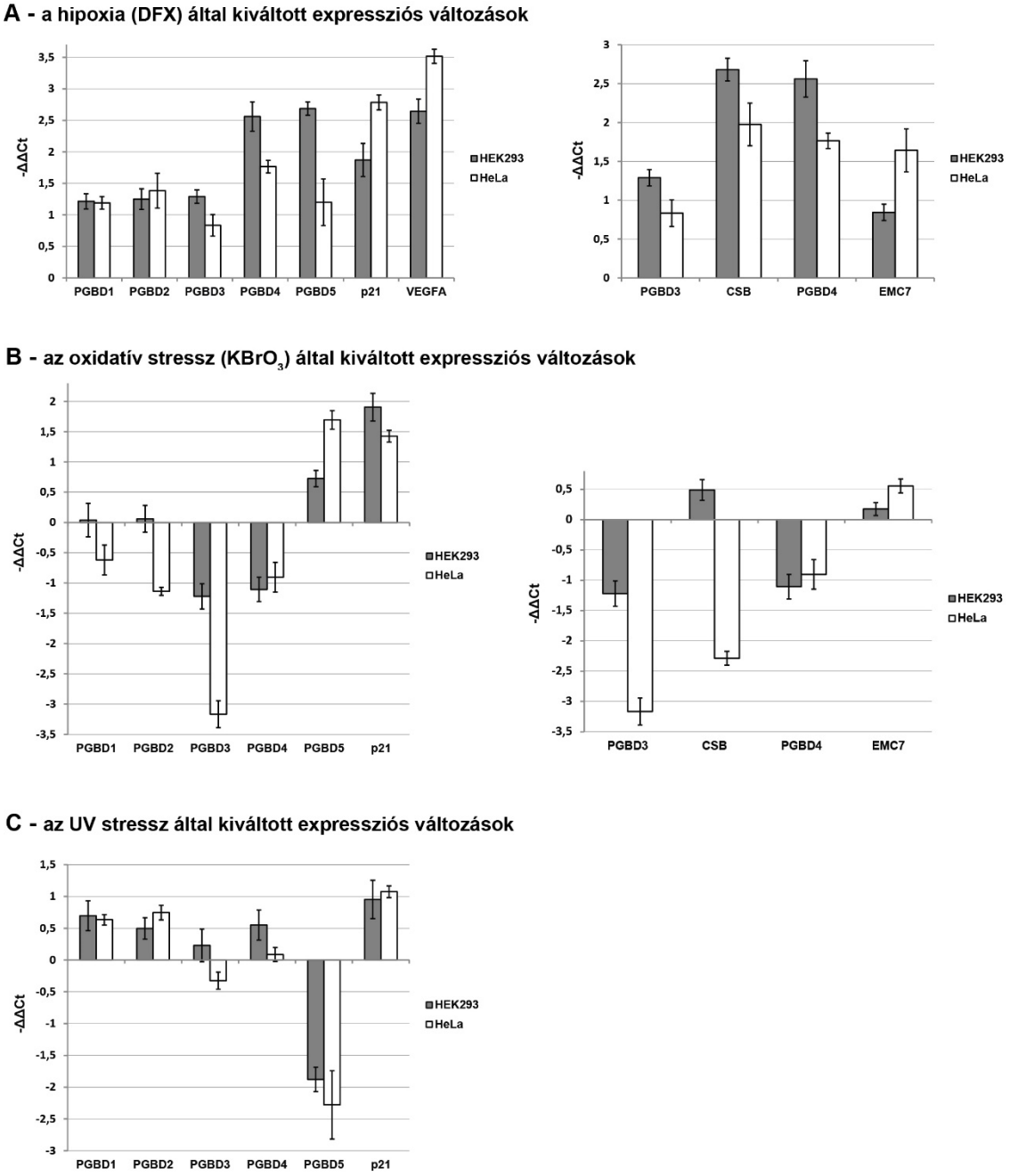
egyértelműen a mi eredményeinkkel vannak összhangban: a PGBD5 esetében nem tudtak kimutatni transzpozíciós aktivitást, és azt is megmutatták, hogy a jelenleg is aktív, de evolúciósan távoli csoportokban azonosított transzpozázok, mint a rovar eredetű *piggyBac* vagy a denevér eredetű *piggyBat* fehérjék sem tudják egymás szubsztrátjait mobilizálni, csak a saját TIR szekvenciájukat (Beckermann és mtsai, 2021). A tudományos vita egyelőre tehát még nem zárult le, de a szakmai közösség kutatóinak többsége egyetért az általunk is bizonyított eredményekkel, hogy a humán PGBD fehérjék elvesztették a transzpozíciós képességüket.



19. ábra: A kétkomponensű transzpozonos rendszer: a transzpozon integráció vizsgálata. (**A**) Az antibiotikumos szelekciót követően életben maradt és megfestett sejtkolóniák néhány példája (kísérletenként két technikai párhuzamossal). Az ábrák felett az alkalmazott transzpozáz eredetű fehérjéket, jobb oldalt pedig a transzpozonos konstrukciókat jelöltük. (**B**) A kísérletekben mért átlagos kolóniaszámok bemutatása (a hibazászlók a szórást jelölik). Excízió hiányában (lásd **18. ábra**) a PGBD1 és a PGBD5 esetében tapasztalt növekedés nem transzpozícióra vezethető vissza. mutPB: katalitikusan inaktív rovar *piggyBac* transzpozáz; *: pozitív kontroll kísérlet. Az ábra a (Kolacsek és mtsai, 2022) közleményünk alapján készült.

További vizsgálataink során arra fókuszáltunk, hogy bizonyítsuk a humán PGBD fehérjék domesztikációját, illetve feltérképezzük azok esetleges endogén funkcióit. Ehhez kapcsolódva elsőként különféle stresszhatásoknak kitett sejtekben vizsgáltuk a PGBD-k expresszióját. A deferasirox (DFX) vegyülettel kiváltott hipoxiás állapot, a $KBrO_3$-tal kiváltott oxidatív stressz, illetve az UV kezelés indukálta stresszfolyamat hatására az 5 PGBD fehérje esetében sokszor egymástól eltérő mintázatot detektáltunk, amelyek az egyes fehérjék esetében sejttípustól függően is különböztek (**20. ábra**). A legmarkánsabb eltérést a PGBD5 esetében tudtuk kimutatni, amely az oxidatív stressz és az UV kezelés hatására is a többi PGBD-vel ellentétes expressziós változást mutatott. Emellett megfigyeltük, hogy a PGBD3 irányultságot tekintve sokszor nem, de a változás mértékét tekintve jelentősen különbözött a többi PGBD expressziós mintázatától. Érdekes volt továbbá, hogy mind a PGBD3, mind pedig a PGBD4 esetében a genomi pozícióban található, korábban közösen reguláltnak vélt gének (a CSB, illetve az EMC7) expressziós mintázata a különböző stresszfolyamatokban sokszor egyaránt más mértékben és más irányban változott (**20A és 20B ábrák**). Mindezen eredmények arra utaltak, hogy a humán PGBD fehérjék egyértelműen endogén funkciókra domesztikálódtak, ugyanakkor az expressziós vizsgálatok rávilágítottak arra, hogy valószínűleg eltérő szereppel bírnak, így a működésük megértéséhez mindegyiküket egyedileg kell vizsgálni (Kolacsek és mtsai, 2022).

A kutatások következő izgalmas lépése a transzpozon eredetű PGBD fehérjék egyedi domesztikált funkcióinak a feltárása, és ezen – több laboratóriumhoz hasonlóan – jelenleg a saját kutatócsoportomban is aktívan dolgozunk. Ígéretes eredmények vannak azzal kapcsolatban, hogy a PGBD3, illetve a PGBD4 működése a splicing folyamatok szabályozásához kapcsolódik, de a végleges konklúziókhoz még további megerősítő vizsgálatokra van szükség. A humán PGBD1 funkciójának feltárásában azonban Izsvák Zsuzsanna laboratóriumával együttműködve nagyon izgalmas eredményre jutottunk. Ez a fehérje a többi PGBD-től eltérően az evolúció során egy fúziós fehérjévé vált, és az N-terminális régiójában tartalmaz egy SCAN és egy KRAB-like domént, amelyekről feltételeztük, hogy fehérje-fehérje interakciókban vesznek részt (Rasko és mtsai, 2022). Miután kimutattuk, hogy a fúziós PGBD1 sem mutat transzpozíciós aktivitást, többféle interakciós vizsgálatban igyekeztünk feltárni a fehérje funkcióját. Elsőként azt igazoltuk, hogy a kezdeti feltételezések ellenére a PGBD1 nem vesz részt közvetlenül a splicing szabályozásában, viszont továbbra is megőrizte DNS kötő képességét.



20. ábra: A humán PGBD gének expressziós változásai különféle stresszfolyamatokban. (**A**) A DFX kezelés hatására bekövetkező változások. (**B**) A $KBrO_3$ által indukált oxidatív stressz hatására bekövetkező változások. Az (**A**) és a (**B**) panelek jobboldali grafikonjain külön kiemeltük a PGBD3/CSB és a PGBD4/EMC7 közös lókuszok tagjainak változásait. (**C**) Az UV stressz hatása a PGBD gének expressziójára. A mRNS szintű méréseket HEK-293 és HeLa sejtvonalakon, RT-qPCR-t alkalmazva, a ΔΔCt módszerrel kvantitálva végeztük, a stresszhatás kontrolljaként a p21 és a VEGFA (***V****ascular* ***E****ndothelial* ***G****rowth* ***F****actor* ***A***) géneket vizsgáltuk. A feltüntetett értékek három független mérés átlagát jelentik, a hibazászlók a 95%-os konfidencia intervallumot jelölik.

Az ábra a (Kolacsek és mtsai, 2022) közleményünk alapján készült.

Egy hosszú kísérletsorozat kapcsán ugyanakkor kiderült, hogy a fehérje közvetlenül szabályozza a NEAT1 hosszú nem-kódoló RNS expresszióját, és kifejezetten erősen gátolja a NEAT1_2 (hosszabb) izoforma transzkripcióját. Ahogy ezt Izsvák Zsuzsanna laboratóriumában részletesen igazolták, a NEAT1 fontos szerepet játszik egy emlős-specifikus sejtmagi struktúra, a „*paraspeckle*" kialakításában és fenntartásában, különösen a neuronok esetében. Ha neuronális progenitor sejtekben gátoljuk a PGBD1 expresszióját, akkor az a NEAT1 RNS expressziójának és a *paraspeckle* struktúrák számának növekedésével jár, illetve serkenti a neuronális differenciációt (Rasko és mtsai, 2022). A PGBD1-nek így tehát fontos szerepe van neuronális fejlődésben, amit az is bizonyít, hogy a gén öröklött mutációit megtalálták például skizofréniában szenvedő betegek örökítőanyagában (Stefansson és mtsai, 2009; Yue és mtsai, 2011; Prata és mtsai, 2019). Ezek az érdekes és újszerű eredmények rávilágítottak arra, hogy az emberi evolúció során bizonyos transzpozonok a domesztikációjuk során jelentős mértékben hozzájárultak új endogén funkciók kialakulásához, amelyek akár új sejtstruktúrák kialakulásának a molekuláris hátterét is biztosíthatják, ahogy ezt a PGBD1 fehérje esetében igazoltuk (Rasko és mtsai, 2022). Az elmúlt évek eredményeinek fényében büszke vagyok arra, hogy a *piggyBac* transzpozonok vizsgálatán keresztül ebbe az izgalmas tudományterületbe az általam irányított kutatócsoport is bekapcsolódhatott.

2.1.2. piggyBac transzpozázok: rendezetlenség a végeken

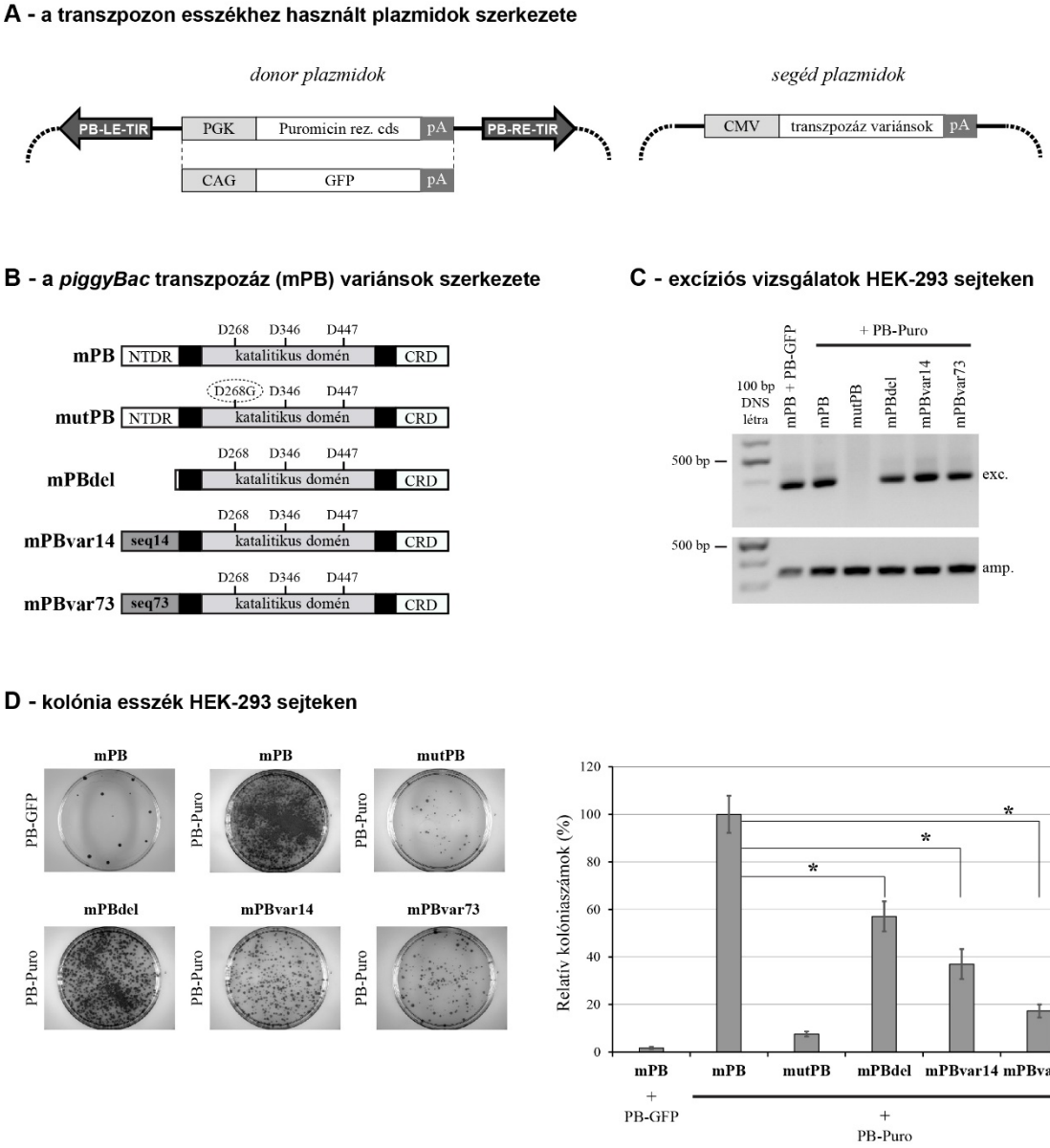
Az előző fejezetben már említésre került, hogy a *piggyBac* transzpozonos rendszert sok előnyös tulajdonsága miatt előszeretettel használják génbeviteli eljárások, illetve sokszor már génterápiás fejlesztések során is (Yusa és mtsai, 2011). A hatékonyabb és biztonságosabb alkalmazás érdekében fontos, hogy minél részletesebb ismeretekkel rendelkezzünk a transzpozíció molekuláris mechanizmusának, illetve a transzpozáz fehérje szerkezet-funkció összefüggéseinek a részleteiről. A rovar *piggyBac* transzpozáz szerkezetének vizsgálata kapcsán kiderült, hogy annak katalitikusan aktív régiója egy aszparaginsav-triádot (egy DDD motívumot) tartalmazó, „*RNase H-like fold*" szerkezetű domén, amely a retrovirális integrázok és az IS4 inszerciós szekvenciacsalád tagjaihoz hasonlóan a transzpozíció során a DNS vágási és illesztési reakciókért egyaránt felelős (Keith és mtsai, 2008b; Mitra és mtsai, 2008; Hickman és mtsai, 2010; Nesmelova és Hackett, 2010). A katalitikus domént két oldalról olyan peptidszekvenciák határolják, amelyek a fehérje térszerkezetében valójában egy egységes, úgynevezett „dimerizációs

és DNS-kötő domén”-t alkotnak, és a transzpozáz dimerek kialakításán túl részt vesznek mind a TIR szubsztrátokhoz, mind az integrációs célpont-szekvenciákhoz való kötődésben (Chen és mtsai, 2020). A fehérje karboxi-terminálisán található egy ciszteinekben gazdag régió (***c**ysteine-**r**ich **d**omain*, CRD), amely az itt található nukleáris lokalizációs szignál szekvencián keresztül egyrészt felelős a transzpozáz fehérje sejtmagba juttatásáért (Keith és mtsai, 2008a), másrészt kimutatták róla, hogy a transzpozon DNS kötéséhez és a vágásához is szükséges (Morellet és mtsai, 2018). Ez utóbbi funkciót néhány kutatócsoport cáfolja, mert bizonyos esetekben azonosítottak olyan *piggyBac* transzpozázokat, ahol a CRD régió hiányzik (Helou és mtsai, 2021a), illetve a *Paramecium tetraurelia* egysejtű domesztikált *„piggyMac”* transzpozáza esetében kimutatták, hogy a CRD nem kötődik a DNS-hez (Guerineau és mtsai, 2021).

Az ismertetett doméneken kívül van azonban a *piggyBac* transzpozáznak az N-terminálisán egy olyan hosszabb, rendezetlennek prediktált régió (***N**-**t**erminal **d**isordered **r**egion*, NTDR), amelynek a szerepe sokáig tisztázatlan volt. A transzpozáz szerkezetét részletesen vizsgáló kutatócsoport arra jutott, hogy ez a megközelítőleg 100 aminosav hosszúságú, nagyfokú rendezetlenséget mutató régió túlnyomórészt savas karakterű aminosavakat tartalmaz, és ennek kapcsán valószínűtlennek tartották, hogy ennek a szegmensnek a DNS kötésben szerepe lenne (Chen és mtsai, 2020). Mivel más fehérjék rendezetlen régiói nagyon gyakran inter- vagy intramolekuláris kölcsönhatásokban vesznek részt, amivel sok esetben biztosítják az allosztérikus reguláció molekuláris hátterét (Tompa és mtsai, 2015), ezért elképzelhető, hogy az NTDR a *piggyBac* transzpozáz esetében is hasonló szereppel bír. Kutatócsoportunkban évek óta foglalkozunk aktív és domesztikált *piggyBac* transzpozázok funkcionális analízisével, ezért úgy döntöttünk, hogy részletes vizsgálatokkal igyekszünk feltárni az NTDR szerepét a *piggyBac* transzpozíció során.

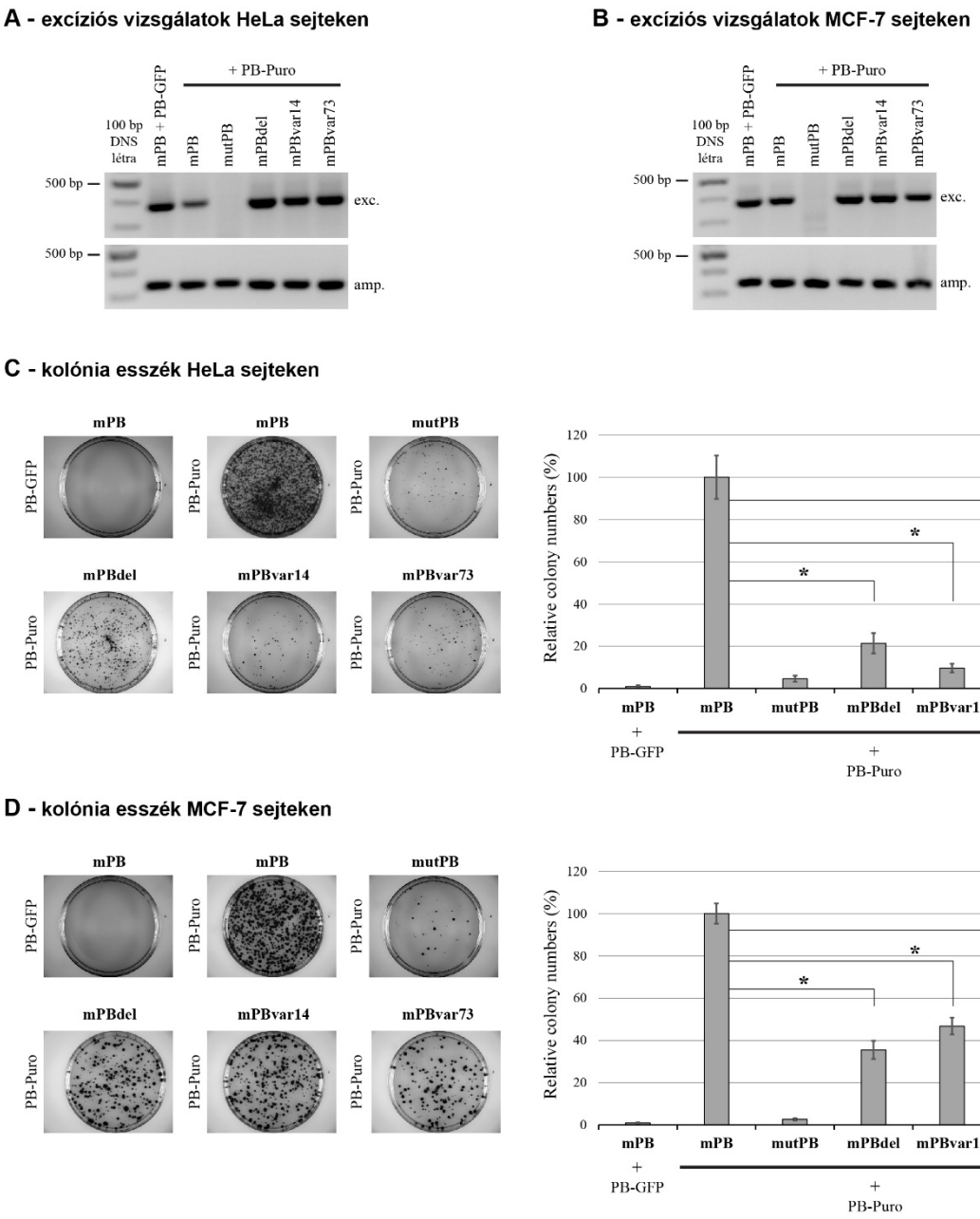
Vizsgálatainkhoz a korábban már használt rovar *piggyBac* transzpozáz emlős rendszerekre kodon-optimalizált variánsából indultunk ki (mPB), negatív kontrollként pedig a transzpozáz általunk korábban előállított katalitikusan inaktív verzióját (mutPB) használtuk (**21. ábra**). Előzetes bioinformatikai analízisünk azt mutatta, hogy több aktív és domesztikált *piggyBac*-rokon fehérje rendelkezik NTDR szegmenssel, amelyek átlagosan 100 aminosav hosszúságúak, ezért a kísérletekhez létrehoztuk az mPB első 100 aminosav hiányos deléciós változatát (mPBdel). Már a kísérletek elején lényeges kérdés volt, hogy a funkció szempontjából vajon csak a régió rendezetlen volta, vagy az NTDR szekvenciája is egyaránt fontos szereppel bír, ezért létrehoztunk több kevert

(„*scrambled* ”) variánst: ezek méretükben megegyeztek az mPB fehérjével, de az NTDR-t alkotó összes aminosavat összekeverve tartalmazták úgy, hogy a régió rendezetlen karakterisztikája továbbra is megmaradt (mPBvar14 és mPBvar73 variánsok, **21B ábra**). A létrehozott transzpozáz variánsok fehérje expressziós szintje szignifikánsan nem különbözött egymástól, így működésüket a már korábban is alkalmazott kétkomponensű, donor és segéd plazmidok transzfektálásán alapuló tesztrendszerben vizsgáltuk. Kimutattuk, hogy mindegyik mutáns esetében a transzpozíció excíziós lépése megtörténik, bár az alkalmazott diagnosztikus (nem kvantitatív) PCR a reakció hatékonyságáról nem ad információt (**21C ábra**). Az antibiotikum szelekción alapuló kolónia esszék viszont azt mutatták, hogy a transzpozíció az NTDR hiányos mPBdel transzpozáz esetében szignifikánsan kisebb hatékonysággal, a HEK-293 sejtvonal esetében közel 40%-kal csökkent mértékben megy végbe (**21D ábra**). Fontos eredmény volt továbbá, hogy a kevert mutánsok (mPBvar14 és mPBvar73) a fenotípust nem komplementálták, sőt inkább tovább csökkentették a transzpozíció hatékonyságát (**21D ábra**). A vizsgálatokat két további sejtvonalon elvégezve azt tapasztaltuk, hogy az NTDR deléció mindenhol a transzpozíciós hatékonyság csökkenéséhez vezet, de ennek a mértéke erősen függ a sejttípustól. A HeLa sejtek esetében az mPBdel transzpozázt alkalmazva nagyon markáns, az eredetihez képest ~80%-os csökkenést tapasztaltunk, míg MCF-7 sejteken a transzpozíció ~65%-kal esett vissza (**22. ábra**). Érdekes volt továbbá, hogy míg a HeLa sejtvonalon a kevert mutánsok még az mPBdel erősen lecsökkent transzpozíciós hatékonyságánál is alacsonyabb értéket mutattak, addig az MCF-7 sejtvonalon az egyik (az mPBvar14) kevert mutáns esetében az mPBdel szintjéhez közeli, vagy annál kicsit magasabb transzpozíciós rátát detektáltunk, bár a vad típusú mPB transzpozáz értékének ez még mindig kevesebb, mint 50%-át tette ki (**22D ábra**). Az eredmények tehát arra utaltak, hogy a *piggyBac* transzpozáz NTDR szegmense pozitív módon befolyásolja a transzpozíciót, és a rendezetlenség mellett ebben fontos szerepe lehet a régió szekvenciájának is, hiszen hasonló méretű és aminosav összetételű rendezetlen szakaszok ezt a moduláló funkciót nem tudták helyreállítani. Eredményeink alapján az is feltételezhető, hogy a transzpozíció ilyen fajta „finomhangolása” az NTDR más fehérjékkel történő interakcióján keresztül valósulhat meg, és ezen faktorok sejtspecifikus expressziós szintje vagy elérhetősége magyarázhatja az általunk tapasztalt sejttípus függő mintázatot.



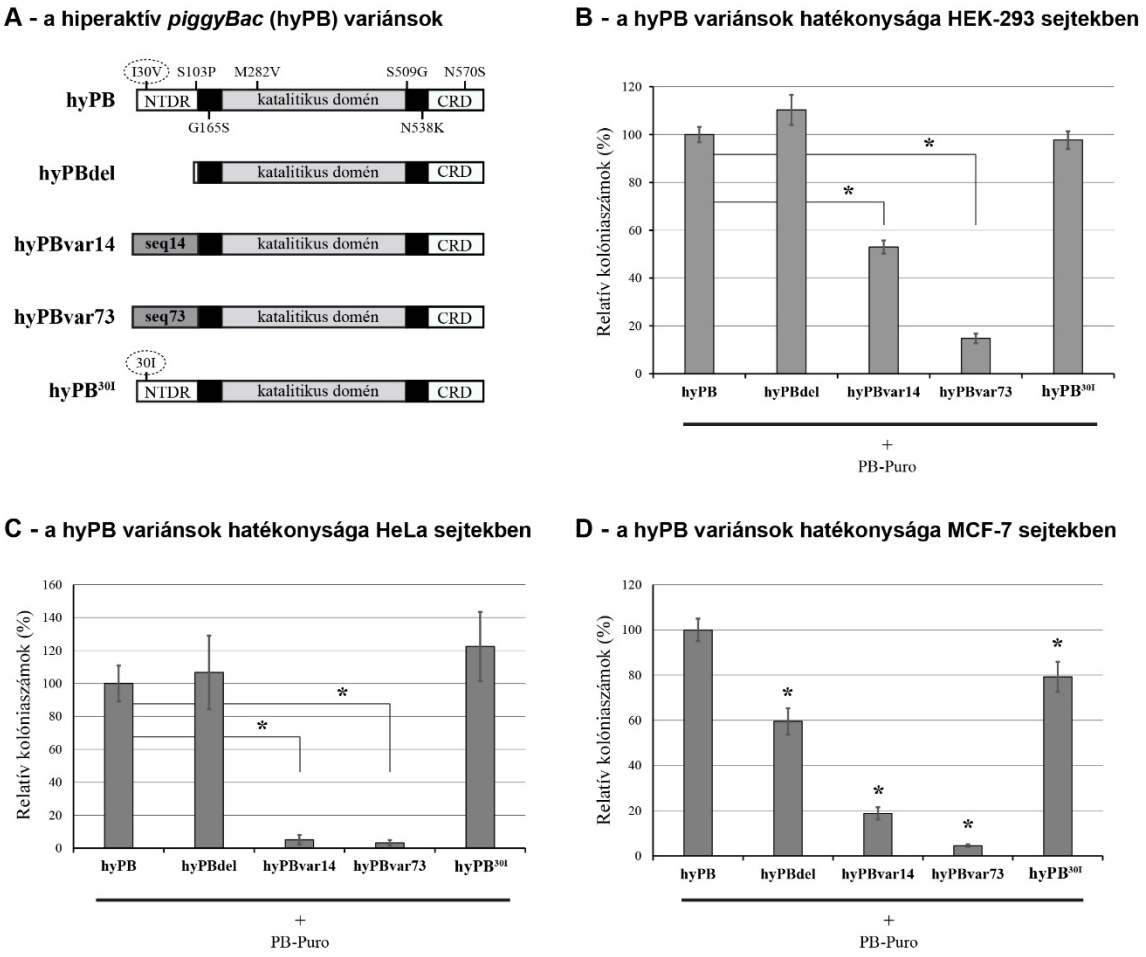
21. ábra: piggyBac transzpozáz mutánsok tesztelése HEK-293 sejtekben. (**A**) A kétkomponensű tesztrendszerben használt plazmidok szerkezete. CAG: citomegalovírus-aktin-globin mesterséges hibrid promóter; rez.: rezisztencia. (**B**) A mutáns transzpozáz fehérjék sematikus szerkezete. A katalitikus domén DDD-motívumát alkotó aminosavakat a rajzok felett külön jeleztük; a fekete téglalapok a „dimerizációs és DNS-kötő domén" két elválasztott részét jelölik. NTDR: N-terminális rendezetlen régió; CRD: ciszteinben gazdag domén; seq14/seq73: az N-terminálison kevert szekvenciát tartalmazó variánsok (magyarázat a szövegben). (**C**) A transzpozon excíziót kimutató reprezentatív PCR eredmény. exc.: az excízióra specifikus PCR termék (381 bp); amp.: kontroll PCR termék (340 bp). (**D**) A kolónia esszék reprezentatív eredményei és kvantitálásuk. Minden érték az „mPB+PB-Puro" reakció esetén kapott értékre (100%) normalizált átlagot jelent, a hibazászlók szórást jelölnek. *: $p<0,05$.

Az ábra a (Wachtl és mtsai, 2022) közleményünk alapján készült.



22. ábra: piggyBac transzpozáz mutánsok tesztelése további sejtvonalakon.
A transzpozon excíziót kimutató diagnosztikus PCR eredmények, illetve reprezentatív kolónia esszék és kvantitálásuk HeLa (**A** és **C**), illetve MCF-7 (**B** és **D**) sejteken. A **21. ábrához** hasonlóan a grafikonokon minden érték az „mPB+PB-Puro" reakció esetén kapott értékre (100%) normalizált átlagot jelent, a hibazászlók szórást jelölnek. *: $p<0,05$.
Az ábra a (Wachtl és mtsai, 2022) közleményünk alapján készült.

Az mPB transzpozáz esetében kapott eredményeket megvizsgáltuk egy „hiperaktív” *piggyBac* transzpozáz esetében is (hyPB, (Yusa és mtsai, 2011)). Ez a variáns a vad típusú mPB transzpozáztól 7 aminosavban tér el, amelyből az egyik mutáció az NTDR szegmensre esik (**23A ábra**). Érdekes módon a hyPBdel mutánst vizsgálva csak az MCF-7 sejtvonalon lehetett a transzpozíció komoly csökkenését kimutatni, a HEK-293 és a HeLa sejtvonalakon a hyPB és a hyPBdel variánsok között hatékonyság szempontjából szignifikáns különbséget nem detektáltunk (**23. ábra**). A kevert mutánsok (a hyPBvar14 és a hyPBvar73) ugyanakkor itt is jelentősen alacsonyabb transzpozíciós rátát eredményeztek, bár a HEK-293 és az MCF-7 sejtvonalakon a hyPBvar14 jobban működött, mint a hyPBvar73 mutáns (**23B és 23D ábrák**). A hyPB transzpozáz NTDR szegmensébe eső mutációjának (I30V) a vad típusra történő visszaállításával megmutattuk, hogy ez a mutáció a HEK-293 és a HeLa sejtvonalakon szignifikánsan nem befolyásolta a transzpozíciós rátát. Az MCF-7 sejtvonalon azonban ez a mutáció már önmagában is szignifikánsan hozzájárul a hyPB aktivitásához, pont abban a sejtvonalban, ahol az NTDR deléció a hyPB aktivitását is jelentősen csökkenti (**23D ábra**). Eredményeinkből arra következtettünk, hogy a „hiperaktív mutációk” bizonyos sejttípusokban az NTDR hiányában képesek kompenzálni a kiesett pozitív moduláló hatást, az NTDR szekvencia elrontása (például a kevert mutánsok esetében) azonban már domináns negatív módon befolyásolja a transzpozíciót. Mindezen eredmények ugyanakkor alátámasztották azt, hogy az NTDR-nek fontos szerepe van a *piggyBac* transzpozáz regulációjában, de jelenléte önmagában nem szükséges az alapszintű transzpozícióhoz (Wachtl és mtsai, 2022).

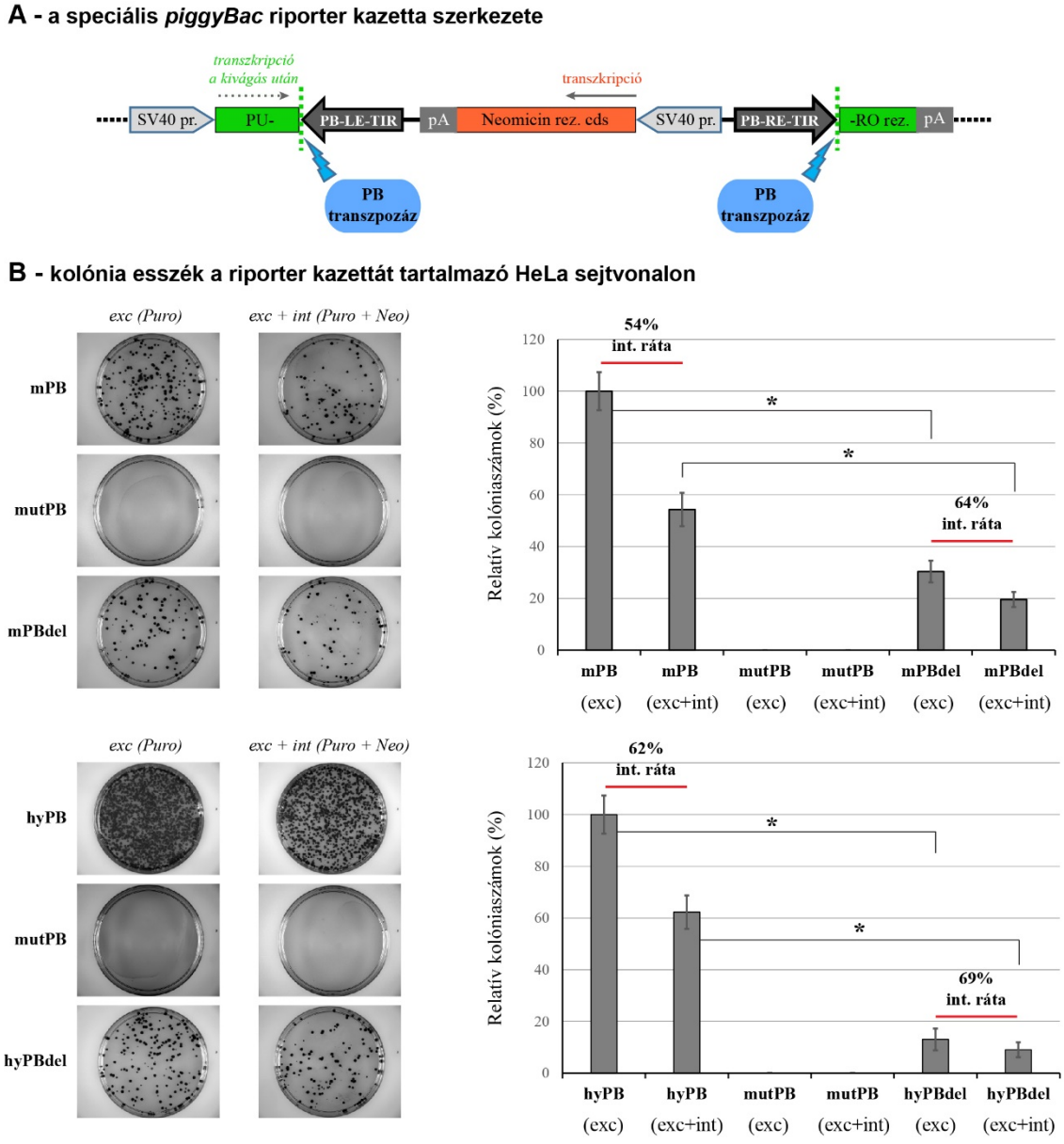


23. ábra: A hiperaktív piggyBac transzpozáz variánsok tesztelése háromféle sejtvonalon. (**A**) A vizsgált hyPB variáns fehérjék szerkezetének sematikus ábrázolása, a domének rövidítése megegyezik a **21B ábrán** bemutatottakkal. A hyPB rajza felett és alatt jelöltük azt a 7 mutációt, amelyben a fehérje eltér az mPB transzpozáztól; a hyPB^{30I} esetében pedig azt az egyet, amelyben az eltér a hyPB fehérjétől. (**B-D**) A hyPB variánsok transzpozíciós hatékonyságát bemutató reprezentatív kolónia esszék HEK-293, HeLa és MCF-7 sejtvonalakon. A grafikonokon ebben az esetben az értékek a „hyPB+PB-Puro” reakció esetén kapott értékre (100%) normalizált átlagot jelentik, a hibazászlók szórást jelölnek. *: $p<0,05$.

Az ábra a (Wachtl és mtsai, 2022) közleményünk alapján készült.

A továbbiakban arra voltunk kíváncsiak, hogy az NTDR a transzpozíció melyik lépését, az excíziót vagy az integrációt (esetleg mindkettőt) befolyásolja. A kérdés megválaszolására egy HeLa-ból származó speciális riporter sejtvonalat használtunk, amelyet Ivics Zoltán bocsátott a rendelkezésünkre. Ebben a sejtvonalban egy kópiában található meg egy génexpressziós kazetta, ahol a puromicin rezisztencia gén kódoló szekvenciáját egy olyan *piggyBac* transzpozon szakítja meg, amely önmagában tartalmaz egy neomicin rezisztencia gént kifejező transzkripciós egységet (**24A ábra**). Ha ebben a sejtvonalban egy funkcionális *piggyBac* transzpozázt expresszálunk, akkor az képes tökéletes precizitással kivágni a transzpozont, amivel a sejteket puromicin rezisztenssé teszi. A sejteket puromicinnel kezelve tehát kiszelektálhatók azok a sejtek, ahol *excízió* történt, puromicin és neomicin dupla szelekcióval pedig azok, amelyekben az *excízió* mellett *integráció* is történt. A megfelelő kezelések segítségével így kvantitálható a transzpozíció elkülönült lépéseinek, az excíziónak és az integrációnak az aránya.

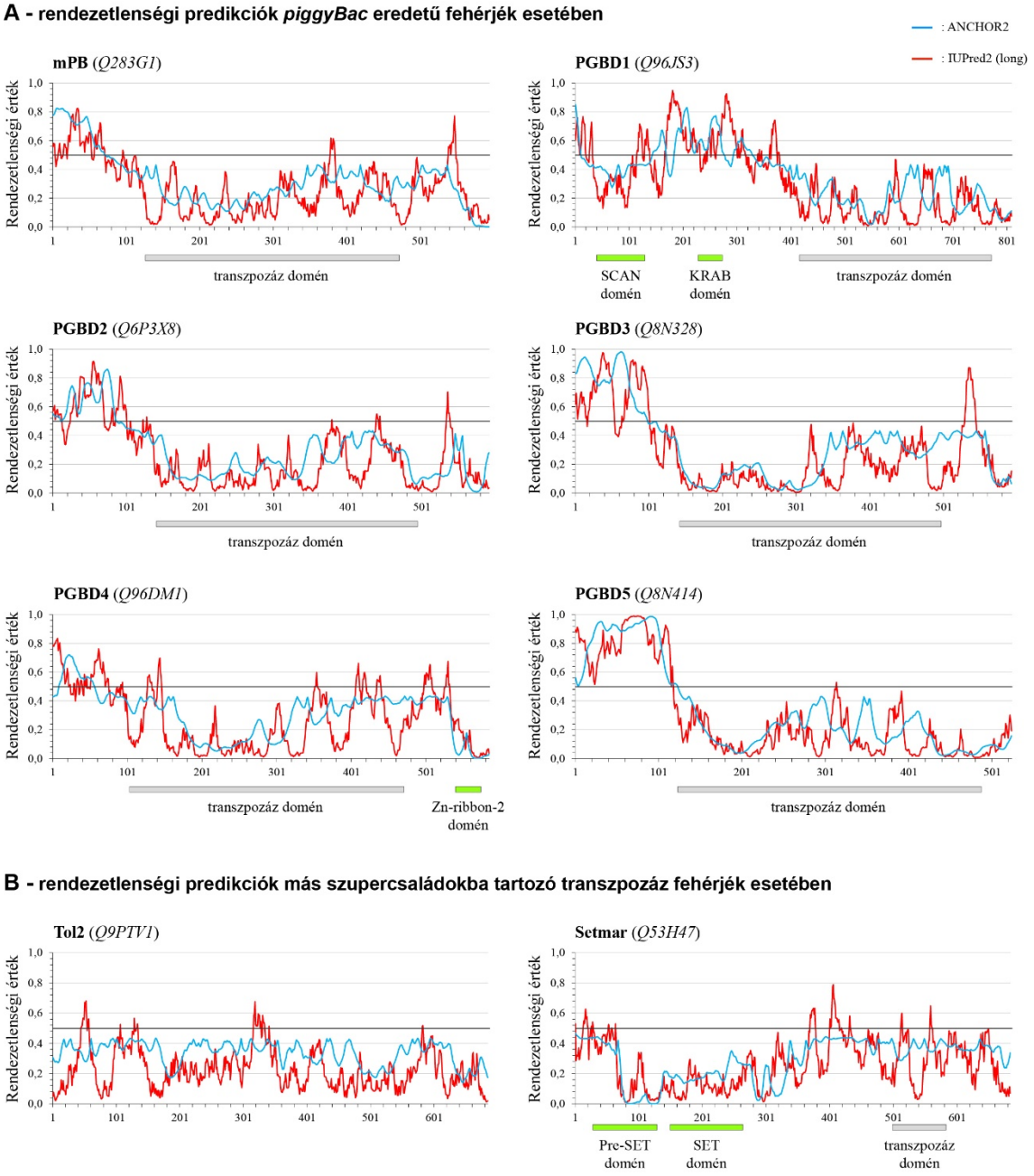
A kétkomponensű transzfekciós rendszerünket a riporter sejtvonalon alkalmazva összehasonlítottuk egymással a normál és a hiperaktív *piggyBac* transzpozázok aktivitását azok NTDR deléciós mutánsaival. Azt tapasztaltuk, hogy az mPBdel és a hyPBdel mutánsok puromicin szelekcióval mért excíziós aktivitása drámaian lecsökkent, az eredeti vad típusú fehérjékhez képest 70 vagy 90 százalékkal (**24B ábra**). Az integrációs aktivitást mérő puromicin és neomicin dupla szelekció kapcsán mindig kevesebb kolóniaszámot kaptunk, mint csak puromicinnel szelektálva, ugyanis a kivágott transzpozonok csak egy adott hányada képes funkcionális integrációra (hiszen egy részük vagy nem integrálódik, mert például a sejten belül degradálódik, vagy az integráció genotoxikus lesz, mert például egy háztartási génben okoz mutációt). Ezek alapján definiálható egy integrációs ráta, amely megmutatja a funkcionálisan integrálódó kivágott transzpozonok arányát. Vizsgálataink azt mutatták, hogy az integrációs ráta az NTDR deléciós mutánsok esetén nem csökken, sőt enyhe mértékben ugyan, de emelkedést mutat (**24B ábra**). Ezek az eredmények arra utalnak, hogy az NTDR-nek a transzpozíció excíziós lépésében van hangsúlyos szerepe, az integrációs lépést nagymértékben már nem befolyásolja, ugyanakkor a pontos molekuláris mechanizmus felderítéséhez még további kísérletekre van szükség.



24. ábra: Az mPB és a hyPB variánsok vizsgálata a speciális HeLa riporter sejtvonalon. (**A**) A sejtvonal genomjában egy kópiában található riporter kazetta sematikus szerkezete. Egy aktív *piggyBac* transzpozáz működése során a transzpozonos kazetta kivágása után helyreáll a puromicin rezisztenciát biztosító transzkripciós egység (zölddel jelölve). SV40 pr.: simian vírus 40 korai promóter. (**B**) A baloldalon látható reprezentatív kolónia esszék a jelölt transzpozáz variánsokkal történt kísérletek eredményeit mutatják. Puromicin kezeléssel az excíziós eseményekre („exc"), míg a puromicin és a neomicin együttes kezeléssel pedig az excíziót követő integrációs eseményekre („exc + int") szelektálhatunk. A jobboldali grafikonokon az értékek a pozitív kontroll reakciókhoz (mPB vagy hyPB) viszonyított relatív hatékonyságok átlagát mutatják; a hibazászlók szórást jelölnek. *: $p<0,05$.

Az ábra a (Wachtl és mtsai, 2022) közleményünk alapján készült.

A *piggyBac* transzpozáz NTDR vizsgálatához kapcsolódva kiterjedt bioinformatikai analízist végeztünk a *piggyBac* szupercsaládba tartozó fehérjéken. Az elérhető szekvenciákat az IUPred2 program (Meszaros és mtsai, 2018; Erdos és Dosztanyi, 2020) segítségével elemezve kiderült, hogy a *piggyBac*-eredetű fehérjék 95%-a tartalmaz a transzpozáz doméntől N-terminális irányban egy 10 aminosavnál hosszabb olyan régiót, amely rendezetlen karakterisztikát mutat. Ez igaz volt aktív és domesztikált fehérjékre egyaránt, olyanokra is, amelyek az evolúció során más fehérjedoménekkel fuzionálva domesztikálódtak, de a transzpozáz domén előtt megőriztek egy hosszú rendezetlen szegmenst. Ebbe a csoportba tartozik például a humán PGBD1 fehérje, amely az N-terminális régiójában található KRAB- és SCAN-domének és a C terminálisán található transzpozáz között megőrizte az NTDR-t, vagy ide tartozik a humán PGBD4 fehérje, amely a transzpozáz C-terminálisán tartalmaz egy „*Zn-ribbon*” domént, ugyanakkor szintén megőrizte a transzpozáz előtti NTDR-t. Ehhez kapcsolódva nagyon érdekes az a megfigyelés, hogy az összes humán PGBD fehérje rendelkezik a transzpozáz domén előtti ~100 aminosav hosszúságú rendezetlen régióval (**25A ábra,** piros vonalak). A rendezetlenségi predikción kívül a fehérjék vizsgálatakor alkalmaztuk az ANCHOR2 algoritmust is, amely a rendezetlen szekvenciákon belül a potenciális interakciós motívumok azonosítására szolgál (Meszaros és mtsai, 2009). A humán PGBD fehérjék esetében az NTDR-en belül ez az algoritmus nagyon magas valószínűségi értékeket mutat (**25A ábra**, kék vonalak), ugyanakkor a kísérleteink során használt kevert szekvenciájú mutánsok esetében ezek az értékek jóval alacsonyabbak. Ez magyarázattal szolgálhat azokra a korábbi eredményekre, ahol a *piggyBac* transzpozázok vizsgálatakor a kevert mutánsok nem tudták komplementálni az NTDR szekvencia hiányában detektált mutáns fenotípust. Mindemellett ezek a bioinformatikai vizsgálatok felvetik azt a lehetőséget, hogy a *piggyBac* fehérjék az evolúciós domesztikáció során még a transzpozíciós aktivitás elvesztése után is megőrzik az NTDR által biztosított molekuláris kapcsolatrendszert. Ugyanakkor ez a szerkezeti konstelláció egyértelműen a *piggyBac* szupercsalád sajátossága, hiszen más transzpozáz szupercsaládok képviselőit vizsgálva a transzpozáz domén előtti rendezetlen régió jelenlétét nem tudtuk kimutatni (**25B ábra**).



25. ábra: Transzpozáz eredetű fehérjék rendezetlenségi predikciói az IUPred programmal. (**A**) A rovar eredetű *piggyBac* transzpozáz (mPB) és a humán PGBD fehérjék rendezetlenségi predikciói. A grafikonokon a fehérjék aminosav sorrendjének függvényében az IUPred2 (piros vonal) vagy az ANCHOR2 (kék vonal) algoritmusok alapján prediktált rendezetlenségi értékeket ábrázoltuk. Egy peptid régió akkor számít rendezetlennek, ha a prediktált értékek 0,5 felettiek (a küszöbértéket egy vastagabb vízszintes vonallal jelöltük). (**B**) Rendezetlenségi predikciók más transzpozon szupercsaládok képviselőin: a *Tol2* a hAT szupercsalád, míg a *Setmar* a Tc1/Mariner szupercsalád tagja. A fehérjék UniProt azonosítói a nevük mögött zárójelben szerepelnek, az ismert doméneket pedig a grafikonok x-tengelye alatt jelöltük.

Az ábra a (Wachtl és mtsai, 2022) közleményünk alapján készült.

Összefoglalásképpen elmondható, hogy a kutatásaink során igazoltuk a *piggyBac* fehérje NTDR szegmensének pozitív reguláló szerepét a transzpozícióban, illetve bizonyítékot szolgáltattunk arra vonatkozólag, hogy ennek a régiónak a domesztikált transzpozáz fehérjék esetében továbbra is kitüntetett szerepe van (Wachtl és mtsai, 2022). Időközben egy másik kutatócsoport olyan eredményeket közölt, amelyek részben ellentmondani látszanak a mi eredményeinknek (Luo és mtsai, 2022). A *piggyBac* fehérje N-terminális régiójában a kutatók azonosítottak több kazein kináz II foszforilációs motívumot, amelyek állításuk szerint negatív módon hatnak a transzpozícióra, így eltávolításuk növeli a transzpozíció hatékonyságát. Kísérleteik jó részét ugyanakkor egy olyan *piggyBac* transzpozázzal végezték, amelynél 74 aminosav hiányzott az N-terminális régiójából (szemben a mi 100 aminosavas deléciós mutánsukkal), illetve az eredményeket egy olyan, nem természetes transzpozon szekvencián végezték, amely mindkét végén szimmetrikusan csak az egyik, az LE-TIR szekvenciát tartalmazta (szemben a mi természetes, LE- *és* RE-TIR-ket tartalmazó transzpozon szubsztrátunkkal). Néhány kísérletük során alkalmaztak még egy olyan transzpozázt, amelyről az N-terminálison 104 aminosav hiányzott, és amely variáns állításuk szerint integráció deficiens. A probléma az, hogy az eredményeket csak a szimmetrikus, LE-TIR-t tartalmazó szubsztráton mutatják meg, a vad típusú transzpozon szekvenciákon nem. Mindezek alapján igyekszünk ezeket az eredményeket szisztematikus vizsgálatokkal a mi laboratóriumunkban is megismételni, illetve célzott kísérletekkel feltárni, hogy mi lehet a különböző eredmények hátterében. Ami biztos, hogy a különböző laboratóriumok egymásra épülő további vizsgálatai mindenképpen hozzájárulnak ahhoz, hogy pontosabb képet kapjunk a *piggyBac* transzpozáz működésének molekuláris részleteiről, és ezáltal elősegítsük ennek a transzpozon alapú genetikai eszköztárnak a jövőbeli biztonságosabb használatát.

2.2. A DNS transzpozonok mint génbeviteli eszközök

A molekuláris genetika fejlődése szükségszerűen igényelte olyan technológiák kifejlesztését, amelyek segítségével gyorsan és hatékonyan lehetett különféle modellorganizmusok genetikai állományát manipulálni. A vírusokon alapuló módszerek a kezdetektől egy széles körben alkalmazott eszköztárat képviselnek, hiszen a több millió éves evolúció által kialakított, nagyon hatékony génbeviteli eljárásokat jelentenek, legyen szó akár tranziens, akár stabil genetikai módosításról, vagy külsőleg bevitt gének

expressziójáról (Li és mtsai, 2023). A virális rendszerek ugyanakkor számos olyan tulajdonsággal rendelkeznek, amelyek a stabil transzgenikus sejtek előállításánál kifejezetten hátrányt jelentenek. Ezek közé tartozik többek között a gyakran nem kívánatos integrációs profiljuk, amely kapcsán a sejtekben aktívan átíródó génekbe történő integrációs preferenciájukon keresztül növelik az inszerciós mutagenezis gyakoriságát, komolyan megnehezítve ezzel a génterápiás eljárásokban való alkalmazásukat (Schroder és mtsai, 2002). Hátrányt jelent az is, hogy a virális vektorok sokszor patogén vírusokból származnak, így a velük való munka a többszörös genetikai módosítások ellenére is egy komolyabb biztonsági kockázatot jelent (VandenDriessche és mtsai, 2003). A transzpozonok, köztük is kifejezetten a DNS transzpozonok ezzel szemben egy lehetséges és hatékony alternatívát jelentenek, igaz, a sejtekbe való bejuttatásuk sokszor technikai kihívást jelent (Ivics és Izsvak, 2006). Az alacsonyabb rendű modellorganizmusok esetében már a felfedezésüket követően nagyon hamar alkalmaztak DNS transzpozonokat, így a *C. elegans* esetében a Tc1/Mariner szupercsalád több képviselőjét (Bessereau, 2006), vagy a *Drosophila* genetikában a méltán híres P-elemet (Castro és Carareto, 2004; Ghanim és mtsai, 2020). A gerinceseken, köztük az emberi sejteken történő alkalmazásukat azonban sokáig hátráltatta az a tény, hogy nem volt ismert az ezekben a genomokban is aktívan működő DNS transzpozon. Egy magyar házaspár, Izsvák Zsuzsanna és Ivics Zoltán korszakalkotó munkájával azonban ez megváltozott: a kutatópáros halak genomjainak alapos összehasonlító elemzésével, majd az azt követő célzott és szisztematikus fejlesztéssel „felélesztett” egy inaktív transzpozont, amely működőképesnek bizonyult nemcsak halak, hanem más gerincesek sejtjeiben, köztük emberi eredetű sejtekben is (Ivics és mtsai, 1997). Az „életre csókolt” transzpozon a „Csipkerózsika” (*Sleeping Beauty*, SB) nevet kapta, és egy csapásra forradalmasította nemcsak a génbeviteli eljárásokat, hanem az erre épülő génterápiás fejlesztési irányokat is (Ivics és Izsvak, 2006). Ma már több olyan módszer áll a klinikai kipróbálás különböző fázisaiban, amely a *Sleeping Beauty* transzpozonos rendszer használatán alapul (Hudecek és mtsai, 2017). A technológia lényegét a korábbi fejezetekben már ismertetett eljárás jelenti: egy kétkomponensű rendszerben, legtöbbször két független plazmidvektor segítségével, a sejtekbe jutatott és ott kifejeződő transzpozáz a transzpozon TIR szekvenciák között elhelyezkedő expressziós kazettát stabilan beépíti a célsejt genomjába (lásd korábban, **18. ábra**). A *Sleeping Beauty* óriási előnye a többi hasonló rendszerrel szemben az, hogy a transzgén integrációs profilja nagyon közel áll a random integrációs mintázathoz, így ebben az esetben a legkisebb az inszerciós

mutagenezis, így a genotoxicitás kockázata (Grabundzija és mtsai, 2010; Huang és mtsai, 2010). Többek között ez az előnyös tulajdonság tette lehetővé, hogy az elmúlt években a *Sleeping Beauty* transzpozon alapú rendszer egy megbízható és elterjedt génbeviteli eljárássá váljon; mi több, a folyamatos fejlesztések során a kutatók egyre hatékonyabb, úgynevezett „hiperaktív" variánsokat állítottak elő, amelyeket így előszeretettel alkalmaznak génterápiás eljárások kidolgozása során is (Mates és mtsai, 2009; Narayanavari és mtsai, 2017).

Időközben a humán sejteken alkalmazható egyéb DNS transzpozon alapú rendszerek is ugrásszerű fejlődésen mentek keresztül. A génbeviteli eszköztár nemcsak olyan újabb „feltámasztott" transzpozonokkal bővült, mint a *Frog Prince* (Miskey és mtsai, 2003), a *Hsmar1* (Miskey és mtsai, 2007), vagy a *Harbinger* (Sinzelle és mtsai, 2008), de a genomikai kutatások elterjedésével más fajokban is felfedeztek olyan működő DNS transzpozonokat, amelyek humán sejtekben szintén aktívnak bizonyultak. Ezen utóbbiak közé tartozik a korábbi fejezetekben már részletesen tárgyalt rovar eredetű *piggyBac* transzpozon (Ding és mtsai, 2005; Yusa és mtsai, 2011), az ugyanebbe a szupercsaládba tartozó denevér eredetű *piggyBat* (Mitra és mtsai, 2013), vagy a szintén rovar eredetű *TcBuster* (Woodard és mtsai, 2012), de feltétlen megemlítendő a hal eredetű *tol2* transzpozon, amelyet ma már szintén széles körben alkalmaznak különféle genetikai modellrendszerek esetében (Koga és Hori, 2001; Balciunas és mtsai, 2006). A transzpozon alapú repertoár ilyen arányú örvendetes bővülése lehetővé teszi, hogy a kutatók a kísérleteikhez a számukra legoptimálisabb módszereket válasszák ki, ugyanakkor ennek kapcsán óhatatlanul felmerül a különféle rendszerek alapos karakterizálásának és szisztematikus fejlesztésének az igénye is. A következő fejezetekben ehhez kapcsolódóan olyan példákat szeretnék bemutatni, amelyekkel az elmúlt bő másfél évtizedben a kutatócsoportommal aktívan hozzájárultunk a különféle transzpozonos rendszerek fejlesztéséhez, illetve erre építve új transzgenikus modellrendszereket hoztunk létre.

2.2.1. A transzpozon alapú rendszerek tulajdonságai – a kópiaszám pontos meghatározása

A transzgenezisre használt vektorrendszerek esetében fontos kritérium, hogy amennyiben a transzgén tartósan beépülésre kerül a gazdasejt genomjába, úgy ez lehetőleg alacsony kópiaszámban, és a genom működése szempontjából leginkább

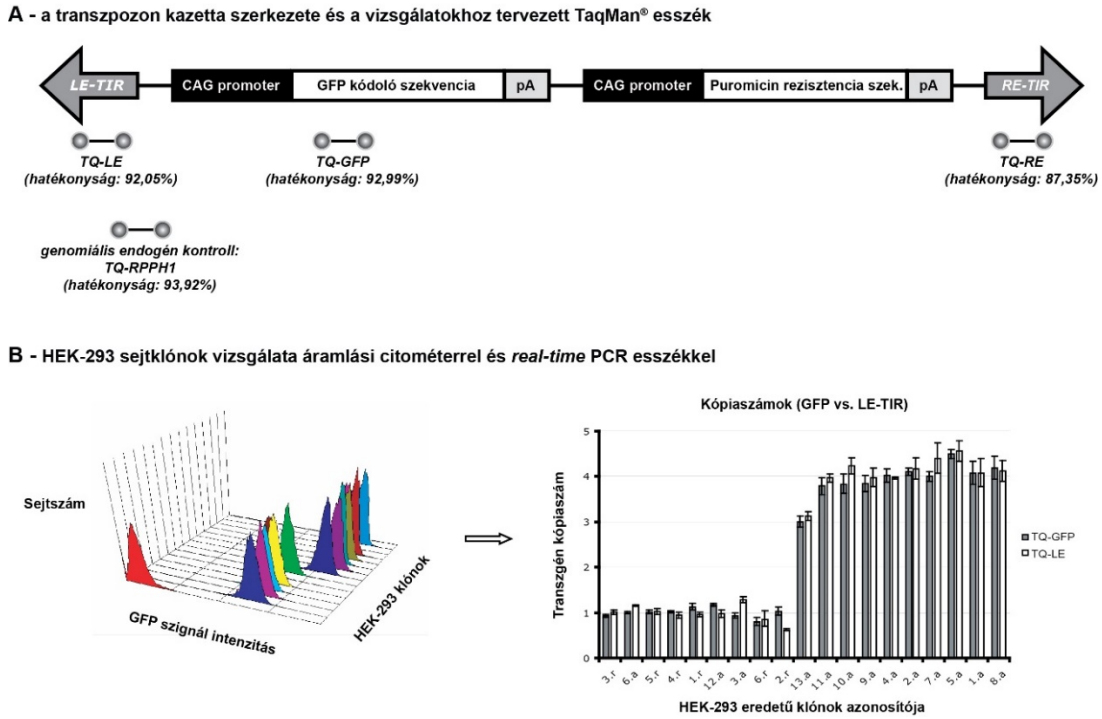
„semleges” régiókba történjen. Az alacsony kópiaszám kevésbé indukálja a sejtszintű védekezési mechanizmusokat, amelyek a transzgén csendesítését eredményezik, illetve egy nem kiugróan magas, tolerálható szintű expressziós szint hozzájárul a hosszú távon stabil transzgenikus sejtklónok kialakulásához (Bian és Belmont, 2010; Sivalingam és mtsai, 2010). A transzgenikus eljárások kapcsán számos módszerrel monitorozható a transzgének kópiaszáma: a hagyományos *Southern blot* technika mellett többféle PCR alapú eljárást is kidolgoztak, amelyek akár a vírusok, akár a transzpozonok által bejuttatott transzgének kópiaszámát kellő hatékonysággal képesek megbecsülni (például a ***L****inear-****a****mplification* ***m****ediated* PCR (LAM-PCR) vagy a *„transposon display*” technológiák) (Devon és mtsai, 1995; Wicks és mtsai, 2000; Woods és mtsai, 2003). Ezeket a módszereket hagyományosan radioaktívan jelzett nukleotidok segítségével végezték, amelyek ugyan így körülményesebbé és hosszabbá tették a méréseket, az érzékenységüknél fogva azonban még a fluoreszcens jelölési módszerek elterjedése után is sokáig előnyt élveztek. Speciális esetekben (fluoreszcens markerek, például a GFP használata esetén) az *in situ* hibridizációs technikák is megbízhatóak lehetnek a kópiaszám meghatározására, ezek alkalmazhatósága azonban nyilvánvalóan limitált (Moeller és mtsai, 2003). A felsorolt módszerek amellett hogy sokszor körülményesek és kivitelezésük speciális kémiai reagenseket és felszerelést igényel, alkalmazásuk a legtöbbször az aktuálisan használt transzgén kimutatására optimalizált, így egy újabb transzgénre vagy egy másik modellszervezetre való áttérés gyakran hosszadalmas előkísérleteket igényel.

Az előzőekben vázolt problémák a *Sleeping Beauty* transzpozonos rendszer alkalmazása során a mi laboratóriumunkban is előkerültek, így felmerült az igény egy gyors, az alkalmazott transzgéntől független kópiaszám meghatározási módszer kidolgozására. A *real-time* PCR technológia, mint megbízható kvantitálási eljárás széleskörű elterjedése minket is arra inspirált, hogy erre a platformra fejlesszünk ki egy általánosan alkalmazható technológiát. Nehézséget jelentett viszont, hogy a transzpozonos rendszernél az egyetlen állandó elem a repetitív szakaszokból álló két TIR régió, ezekre a szekvenciákra azonban kimondottan nehéz hatékony PCR esszéket tervezni. Többszöri próbálkozás eredményeképpen, a TaqMan® technológia (Holland és mtsai, 1991) segítségével azonban sikerült a *Sleeping Beauty* transzpozon jobb és bal terminális régiójára egyaránt működő *real-time* PCR esszéket terveznünk. Mivel a módszerünket olyan sejtvonalakon terveztük kipróbálni, amelyek különböző kópiában GFP-t kódoló transzpozonokat tartalmaztak, ezért a GFP szekvenciára is terveztünk egy

megfelelő TaqMan® esszét. A génexpresszió megbízható kvantitálásához emellett azonban szükség van még egy megbízható endogén referencia génre, jelen esetben egy olyan genomiális szekvenciára, amely a haploid genomban lehetőség szerint egy kópiában fordul elő. Vizsgálatainkhoz a mások által már többször megbízhatónak talált RPPH1 gént használtuk, amely a humán RNázP enzimkomplex egyik alegységének a H1 elnevezésű RNS komponensét kódolja (Baer és mtsai, 1990). Fontos volt, hogy az erre a génre tervezett *real-time* PCR esszé hatékonysága legalább 90%-os legyen, és összemérhető legyen a tesztelendő szekvenciákra specifikus esszék hatékonyságával – előzetes kísérleteink alapján a RPPH1, a GFP és az LE-TIR esszék esetén ez nagyon szépen teljesült (**26A ábra**). Adottak voltak tehát a technikai feltételek, hogy az általánosan elterjedt ΔCt módszer segítségével meghatározzuk a *Sleeping Beauty* transzpozon kópiaszámát transzgenikus sejtvonalakban.

További problémaként merült fel azonban, hogy a pontos méréseket befolyásolja az adott sejtvonal genomiális összetétele: a sejttenyésztés során használt „közkedvelt” sejtvonalak gyakran aneuploidok, és a referencia gént tartalmazó régiók esetlegesen eltérő kópiaszáma torzíthatja a méréseket. A precíz kvantitálás érdekében ezért úgy döntöttünk, hogy lehetőség szerint igyekszünk a vizsgálandó sejttípusból ismert transzpozon kópiaszámú klónokat előállítani, amelyeket referencia mintaként használva a ΔΔCt módszer (Livak és Schmittgen, 2001) segítségével a lehető legpontosabb kópiaszám meghatározást érhetjük el. Ezt elsőként HEK-293 sejtek esetén alkalmaztuk, később azonban különféle technikákkal más sejttípusokból, például humán embrionális őssejtekből, vagy patkány embriókból is előállítottunk ismert transzpozon kópiájú klónokat (lásd később, 2.2.3. fejezet). A HEK-293 esetén a sejteket olyan *Sleeping Beauty* plazmidokkal transzfektáltuk, ahol a transzpozon a GFP-t kódoló transzkripciós egység mellett egy puromicin rezisztencia gént kifejező kazettát is tartalmazott (**26A ábra**). A sejteket egy kéthetes puromicin szelekciót követően áramlási citométer segítségével a GFP expresszió intenzitása alapján egysejt klónoztuk (**26B ábra**). Feltételeztük, hogy ha GFP transzgén nem csendesül, akkor az expresszió mértéke arányos lesz a kópiaszámmal, és sok alacsony GFP expressziót mutató klón összehasonlításával kiválaszthatjuk a valószínűleg egykópiás klónokat, majd ezeket referenciaként használva nagyobb transzpozon kópiaszámú klónokat is azonosíthatunk. A módszer bevált: a GFP és a LE-TIR szekvenciákra specifikus TaqMan® esszék segítségével sikerült többféle, ismert kópiaszámú klónt előállítanunk, amelyeket később sikerrel alkalmaztunk más kísérletek során előállított sejtklónok vizsgálatánál (Kolacsek és mtsai, 2011). A módszer arra is

alkalmas, hogy vegyes sejtpopulációk esetén megbecsüljük a sejtekben található transzgének átlagos kópiaszámát, amely sok vizsgálat esetén elegendő információ a kísérletek további tervezéséhez. Hogy bizonyítsuk az általunk kifejlesztett eljárás megbízhatóságát, a módszerünket sikeresen teszteltük olyan a *Sleeping Beauty* rendszerrel előállított sejtklónokon, amelyeken más módszerrel mi vagy mások korábban már megmérték a transzgének kópiaszámát (**1. táblázat**). Összességében elmondhatjuk, hogy kidolgoztunk egy megbízható, az alkalmazott transzgéntől független, a *Sleeping Beauty* transzpozonos rendszerre specifikus kópiaszám meghatározási módszert, amely a transzgenikus sejtek előállításának fontos lépéseként sikeresen alkalmazható adott kópiaszámú klónok azonosításához és kiválogatásához (Kolacsek és mtsai, 2011).



26. ábra: A Sleeping Beauty transzpozonra specifikus kópiaszám meghatározás kidolgozása. (**A**) A módszerhez használt transzpozon sematikus szerkezete, alul jelölve a transzpozonra, illetve a transzgénre specifikus esszék tapadási helyét. Legalul feltüntettük az RPPH1 génre specifikus kontroll esszét is, illetve minden esszé esetben a hatékonysági értéket. TQ: TaqMan®. (**B**) A sejtklónok áramlási citométerrel meghatározott GFP expressziós szintjei jól korrelálnak a kópiaszámmal (bár természetesen a genomi pozíciótól is függnek; pirossal a kontroll sejtvonal autofluoreszcenciája látható). A kétféle esszével végzett mérések az esetek túlnyomó többségében jó korrelációt mutatnak, és az eredmények alapján kiválasztottunk ismert kópiaszámú klónokat, amelyeket a későbbiekben kontrollként használtunk. A grafikonon a mérések átlagát tüntettük fel, a hibazászlók pedig szórást jelölnek.

Az ábra a (Kolacsek és mtsai, 2011) közleményünk alapján készült.

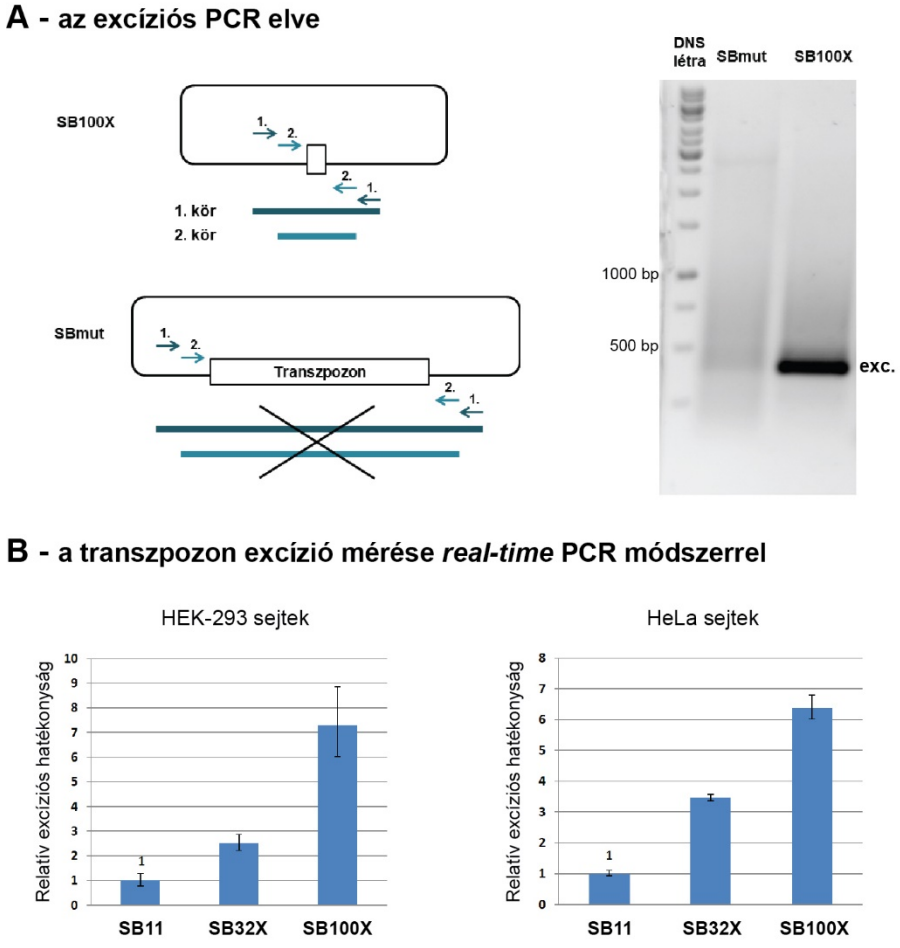
A sejtklón azonosítója	A hagyományos módszer neve	Kópiaszámok	
		hagyományos módszer	*real-time* PCR
2/1	*Transposon display* és *Southern Blot*	8-10	8
2/2	*Transposon display* és *Southern Blot*	3	4
2/3	*Transposon display* és *Southern Blot*	10-12	10
2/9	*Transposon display* és *Southern Blot*	1	1
1	*Transposon display* és *Southern Blot*	12-13	13
4	Dot blot	52	50
5	*Transposon display* és *Southern Blot*	15	15
6	*Transposon display* és *Southern Blot*	12	11
7	*Transposon display* és *Southern Blot*	1	1
8	*Transposon display* és *Southern Blot*	2	2
9	*Transposon display* és *Southern Blot*	1	1
A3	*Splinkerette* és inverz PCR	2	2
A4	*Splinkerette* és inverz PCR	4	4
A5	*Splinkerette* és inverz PCR	4	4.5*
A6	*Splinkerette* és inverz PCR	2	2
B1	*Splinkerette* és inverz PCR	1	2
B3	*Splinkerette* és inverz PCR	3	3
B5	*Splinkerette* és inverz PCR	2	2

1. táblázat: Néhány hagyományos módszer összehasonlítása az általunk kifejlesztett real-time PCR eljárással. A sejtklónok a *Sleeping Beauty* transzpozonos rendszer segítségével lettek előállítva és különböző transzgéneket tartalmaztak. A vizsgálatok minden esetben ugyanabból az adott klónból izolált genomiális DNS mintán történtek. Az általunk kifejlesztett módszerrel kapott eredmények jól korreláltak a többi módszer eredményeivel. Az eltérések esetén további vizsgálatok igazolták, hogy a *real-time* PCR alapú mérés a pontos: a *-gal jelzett esetben például kiderült, hogy az 'A5' jelű minta nem klón, hanem egy kevert sejtpopuláció.
A táblázat a (Kolacsek és mtsai, 2011) közleményünk alapján készült.

2.2.2. A transzpozon alapú rendszerek tulajdonságai – a génbeviteli hatékonyság finomhangolása

A transzpozon alapú génbeviteli rendszerek esetében is nagyon fontos kérdés a módszer hatékonysága, pontosabban annak a kérdése, hogy mennyire rugalmasan szabályozhatók a génbevitel aspektusai különböző sejtes rendszerek esetén, vagy egy-egy adott modellsejten különféle célok elérése érdekében. Egy minél kiterjedtebb mutagenezis esetén például előnyösek azok a rendszerek, amelyek a transzgént magas

kópiaszámban építik be, lehetőleg minél gyakrabban integrálódva a genom kódoló vagy szabályozó régióiba. Ezzel szemben a stabil sejtvonalakat célzó transzgenezis esetében kívánatos az alacsony kópiaszám elérése, az integrációs profilt tekintve pedig a „semleges" régiók előnyben részesítése. A hiperaktív transzpozázok célzott fejlesztése mindenképpen növeli a technológiai repertoárt, sokszor azonban nem egyértelmű, hogy a létrehozott variánsok a transzpozíciós reakció melyik lépését befolyásolják, és alkalmazásukat tekintve milyen területen jelentenek előnyt (Zayed és mtsai, 2004; Baus és mtsai, 2005; Pledger és Coates, 2005). A *Sleeping Beauty* rendszer esetében például az eddigi leghatékonyabb transzpozáz (az ún. „SB100X" variáns) bizonyos körülmények között az eredetileg „feltámasztott" fehérjénél százszor nagyobb aktivitást mutat, más sejtes esszékben azonban ez az előny kevésbé kifejezett (Mates és mtsai, 2009). Felmerül a kérdés: milyen vizsgálatokkal lehet a lehető legegyszerűbben, de a lehetőségekhez képest a legmegbízhatóbban jellemezni egy transzpozáz hatékonyságát? Kísérletes rendszereink optimalizálása során ehhez kapcsolódóan azt vizsgáltuk, hogy a transzpozíció kezdeti lépésének, a transzpozon kivágásának (az excíziónak) a kvantitatív jellemzése vajon mennyire ad megbízható támpontot egy transzpozáz hatékonyságának a megítélésében. A transzpozázok működését gyakran tesztelik az excízió vizsgálatával, legtöbbször egy diagnosztikus „*nested* " PCR segítségével (**27A ábra**, illetve lásd még korábban a 2.1.1. és a 2.1.2. fejezetek ábráit), ez a hagyományos PCR módszer azonban önmagában nem kvantitatív. Korábbi tapasztalataink alapján ugyanakkor úgy gondoltuk, hogy a donor plazmidon található, a transzpozont határoló szekvenciákra specifikus primerek felhasználásával kifejleszthető egy olyan *real-time* PCR esszé, amely a kivágás hatékonyságát megbízhatóan méri. A transzpozáz reakció sajátossága, hogy a kivágás nem 100%-os precizitással történik, így mivel a „hátrahagyott" szekvencia nem mindig pontosan ugyanaz, a TaqMan® technológia itt nem alkalmazható megbízhatóan (ugyan csak néhány nukleotidos különbségekről van szó, a fluoreszcensen jelölt próba azonban így nem minden esetben tapad ki az összezárult plazmidok heterogén szekvenciáira, **27A ábra**). A *real-time* PCR módszer ugyanakkor más módon, például a *SYBR Green* festék alkalmazásával is elvégezhető, amely esetben elegendő a célszekvenciát határoló primereket alkalmazni: az apróbb szekvenciális különbségek itt nem okoznak gondot, és a detektált szekvenciapopuláció mennyisége megbízhatóan reprezentálja a kivágás hatékonyságát.



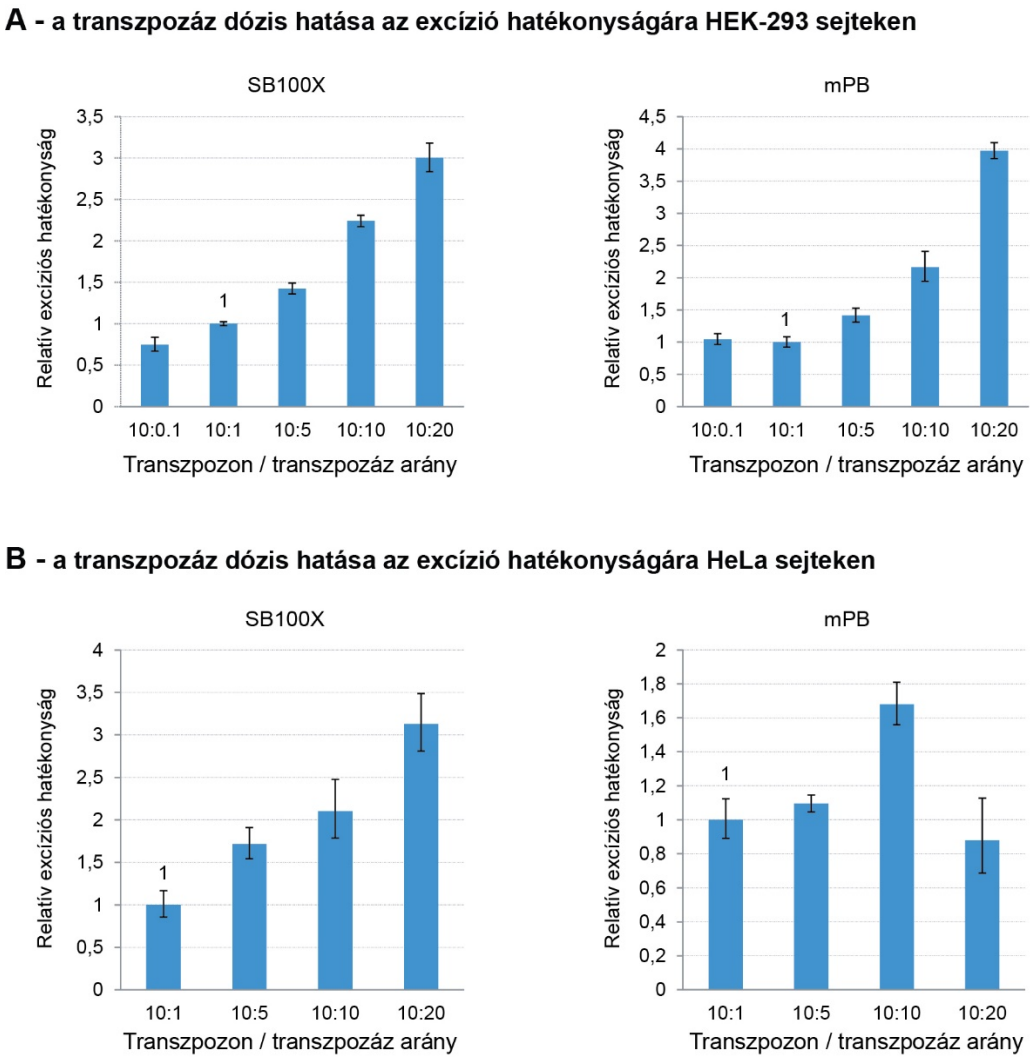
27. ábra: A transzpozon excízió kvantitatív mérésének kidolgozása. (**A**) A baloldali rajzok az excíziós PCR elvét mutatják be, a donor plazmidok sematikus ábrázolásával. Ha a sejtben egy aktív transzpozáz van jelen (itt az SB100X variáns, felső rajz), akkor a transzpozon kivágódik a plazmidról, az excízió helyét pedig a sejt DNS *repair* apparátusa javítja, az illesztésnél esetleg apróbb hibákat generálva (az ún. transzpozíciós *footprint*, kis téglalappal jelölve). Mutáns transzpozáz esetén (SBmut, alsó rajz) viszont nem történik kivágás, így a határoló szekvenciákra specifikus primerek az adott körülmények között nem eredményeznek PCR terméket sem az első, sem a második körben. Az ábra jobboldali panelén a *nested* PCR 2. körének termékei láthatók egy agaróz gélen való futtatást követően („exc.", 381 bp). (**B**) Háromféle *Sleeping Beauty* transzpozáz variáns excíziós aktivitásának összehasonlítása két sejtvonalon. A *nested* PCR 2. körét *SYBR Green* alapú *real-time* PCR esszével vizsgáltuk, referencia mintának pedig az SB11 variánssal végzett vizsgálatot választottuk (magyarázat a szövegben). A grafikonon egy reprezentatív mérés szerepel, az értékek három technikai párhuzamos átlagát jelentik, a hibazászlók a 95%-os konfidencia intervallumot mutatják.

Az ábra a (Kolacsek és mtsai, 2014b) közleményünk alapján készült.

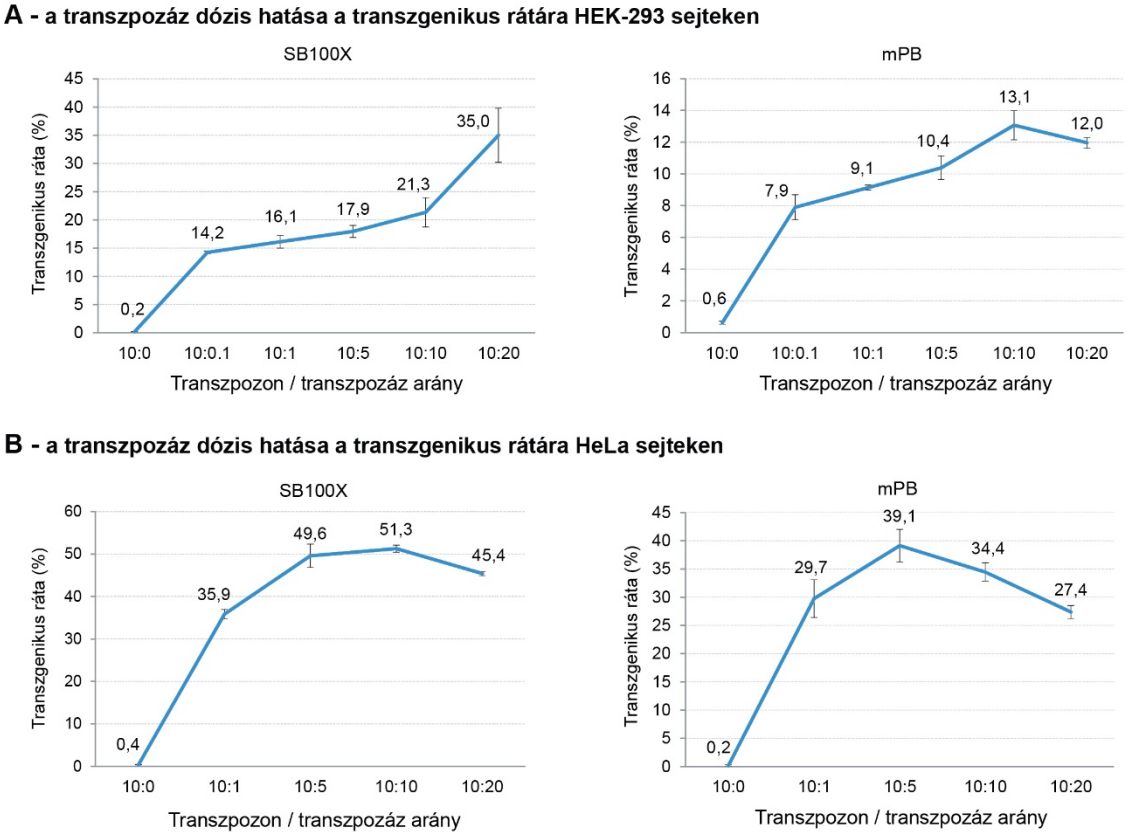
A reakció körülményeit természetesen több szempontból is optimalizálni kellett: az 1. kör hagyományos PCR-t még a reakció logaritmikus fázisában leállítottuk, és a megfelelően hígított terméket a 2. körben most már a *real-time* PCR segítségével kvantitáltuk. Kontrollként a plazmidgerincen található ampicillin szekvenciát használtuk, amely az eredeti és a kivágáson átesett plazmidok mennyiségét együttesen méri, így az excíziós termék mennyiségét a transzfekció során bejuttatott plazmidpopulációra normáltuk. A módszer működött és a segítségével megbízhatóan össze tudtunk hasonlítani különféle *Sleeping Beauty* variánsokat, így az eredetihez legközelebb álló SB11 variánst az SB32X és az SB100X hiperaktív transzpozázokkal (**27B ábra**). A transzpozíciót HEK-293 és HeLa sejtvonalakban vizsgálva jól látszik, hogy a hiperaktív transzpozázok esetében az excízió hatékonysága függ a sejttípustól, illetve ebben a tekintetben például az SB100X is „csak" maximum 7x hatékonyabb az SB11 variánsnál (természetesen az excíziót követő integráció hatékonyságát is figyelembe véve az SB100X azért ennél jóval magasabb arányú transzpozíciós rátát eredményez). Az általunk kidolgozott módszer tehát alkalmas különböző transzpozázok kivágási hatékonyságának jellemzésére (Kolacsek és mtsai, 2014b), de amint azt látni fogjuk, kiválóan működik egy adott transzpozáz variáns különböző körülmények között végbemenő excíziós aktivitásának monitorozására is (lásd a **28. ábrát**).

A transzpozon alapú génbevitel egyik sarkalatos kérdése a bevitt transzpozáz optimális mennyiségének a beállítása. Régről ismert ugyanis, hogy a magas transzpozáz dózis túltermeléses gátláshoz, az ún. „*overproduction inhibition*" jelenségéhez vezet: az enzimkoncentráció növelése a reakciót először telítésbe viszi, egy ponton túl azonban egyre nagyobb mértékben gátolja (Lohe és Hartl, 1996; Hartl és mtsai, 1997; Lampe és mtsai, 1998; Wilson és mtsai, 2005). A kétkomponensű plazmidos rendszer (lásd korábban, **18A ábra**) használatának előnye, hogy a transzpozon és transzpozáz abszolút és egymáshoz viszonyított mennyisége több kombinációban is kipróbálható, és kiválasztható az adott rendszerben legoptimálisabb keverési arány. A *Sleeping Beauty* és a *piggyBac* transzpozonos rendszerek ebből a szempontból történő vizsgálata érdekes eredményeket hozott. Az excíziós hatékonyságot vizsgálva a korábbi eredményekkel ellentétben az SB100X transzpozáz esetében nem tapasztaltunk túltermeléses gátlást, az mPB transzpozázt vizsgálva HEK-293 sejtekben nem, HeLa sejtekben azonban ki tudtuk mutatni az „*overproduction inhibition*" jelenségét (**28. ábra**). Az antibiotikum rezisztencián alapuló, kolónia esszék segítségével mért transzpozíciós hatékonyságot (tulajdonképpen a transzgenikus rátát) vizsgálva azonban más eredményeket kaptunk. Az

SB100X transzpozáz esetében a HeLa sejtekben már jelentkezett a túltermeléses gátlás, az mPB transzpozázt alkalmazva pedig a HeLa sejtek mellett már a HEK-293 sejtvonalon is kimutatható volt a jelenség (**29A és 29B ábrák**).



28. ábra: A transzpozon kivágás hatékonysága függhet a transzpozáz mennyiségétől. Az excíziós hatékonyság vizsgálata *Sleeping Beauty* (SB100x) és *piggyBac* (mPB) transzpozáz variánsok esetén, *real-time* PCR módszerrel, HEK-293 (**A**), illetve HeLa (**B**) sejtvonalon. A túltermeléses gátlás jelenségét csak az mPB esetén és csak a HeLa sejtvonalon detektáltuk. A méréseknél a 10:1 transzpozon:transzpozáz arányt tekintettük referenciának (1-gyel jelölve), az értékek min. 3 független mérés átlagát mutatják, a hibazászlók a szórást jelölik.
Az ábra a (Kolacsek és mtsai, 2014a) közleményünk alapján készült.

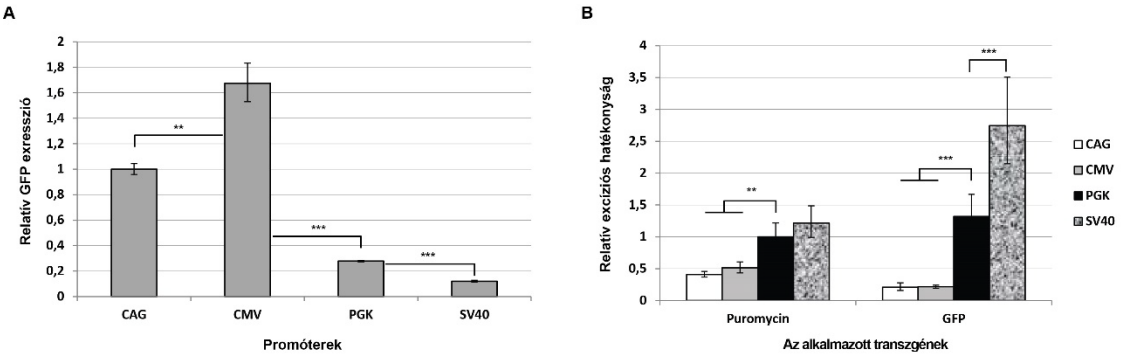


29. ábra: A transzpozíció hatékonyságát a transzpozáz mennyisége jelentősen befolyásolhatja.

A transzpozáz dózis hatása a transzgenikus rátára HEK-293 (**A**) és HeLa (**B**) sejteken. A túltermeléses gátlást HEK-293 sejtvonalon csak a *piggyBac* transzpozáz esetében tudtuk kimutatni, míg a HeLa sejtvonalon már a *Sleeping Beauty* transzpozáz esetében is jelentkezik. Az alkalmazott kísérleti körülmények között az SB100x variáns mindig magasabb transzgenikus rátát eredményezett, mint az mPB transzpozáz. (**C**) A transzpozon kópiaszám a donor plazmid csökkentése esetén drasztikusan leesik, amely bizonyos applikációk (például stabil vonalak előállítása) esetén kifejezetten előnyös lehet. Az adatok minden esetben legalább három mérés átlagát jelentik, a hibazászlók szórást jelölnek.

Az ábra a (Kolacsek és mtsai, 2014a) közleményünk alapján készült.

A túltermeléses gátlás természetesen komolyan befolyásolhatja az integrálódó transzgén kópiaszámát is: magas transzpozáz dózist alkalmazva minden általunk vizsgált transzpozáz variáns esetén a transzpozíciót követően a sejtpopuláción mérhető átlagos kópiaszám jelentős mértékben csökkent (**29C ábra**). Részletes vizsgálataink arra utaltak, hogy minden modellrendszer esetében, még a hiperaktív transzpozáz variánsok alkalmazása esetén is rendkívül fontos a reakciókörülmények körültekintő optimalizálása, hiszen a sikeres transzgenezisnek minden esetben ez a kulcsa (Kolacsek és mtsai, 2014a).



30. ábra: A promóter aktivitás hatása a transzpozon excízióra. (**A**) A promóterek által kiváltott mRNS expressziós szintek vizsgálata *real-time* PCR módszerrel, GFP transzgén esetén, HEK-293 sejteken. A vizsgálatok során a CAG promóterhez tartozó értéket tekintettük referenciának. (**B**) A különféle promóterekkel meghajtott transzgéneket tartalmazó transzpozonok excíziós hatékonyságának összehasonlítása *real-time* PCR módszerrel. Látható, hogy a gyengébb promóterek esetében magasabb excíziós hatékonyságot detektáltunk. Ezekben a mérésekben a PGK promóter esetében kapott értéket tekintettük referenciának. A grafikonokon bemutatott értékek három párhuzamos mérés átlagát jelentik, a hibazászlók a 95%-os konfidencia intervallumot mutatják. **: p<0,01; ***: p<0,001.
Az ábra a (Kolacsek és Orban, 2018) közleményünk alapján készült.

Egy másik érdekes jelenségre akkor figyeltünk fel, amikor tesztelni kezdtük a transzpozonnal bevitt transzkripciós kazettában jelenlévő promóter típusát. A stabil sejtvonalak kialakításában sokszor előnyös, ha gyengébb promótert használunk, mert a transzgén alacsonyabb transzkripciós aktivitása a sejtek számára gyakran jobban tolerálható. Vizsgálataink során ugyanakkor azt tapasztaltuk, hogy az erősebb transzkripció a transzpozíció hatékonyságát is befolyásolja: az SB100X rendszer használata esetén az aktívabb promóterek egyre nagyobb mértékben gátolták az excízió

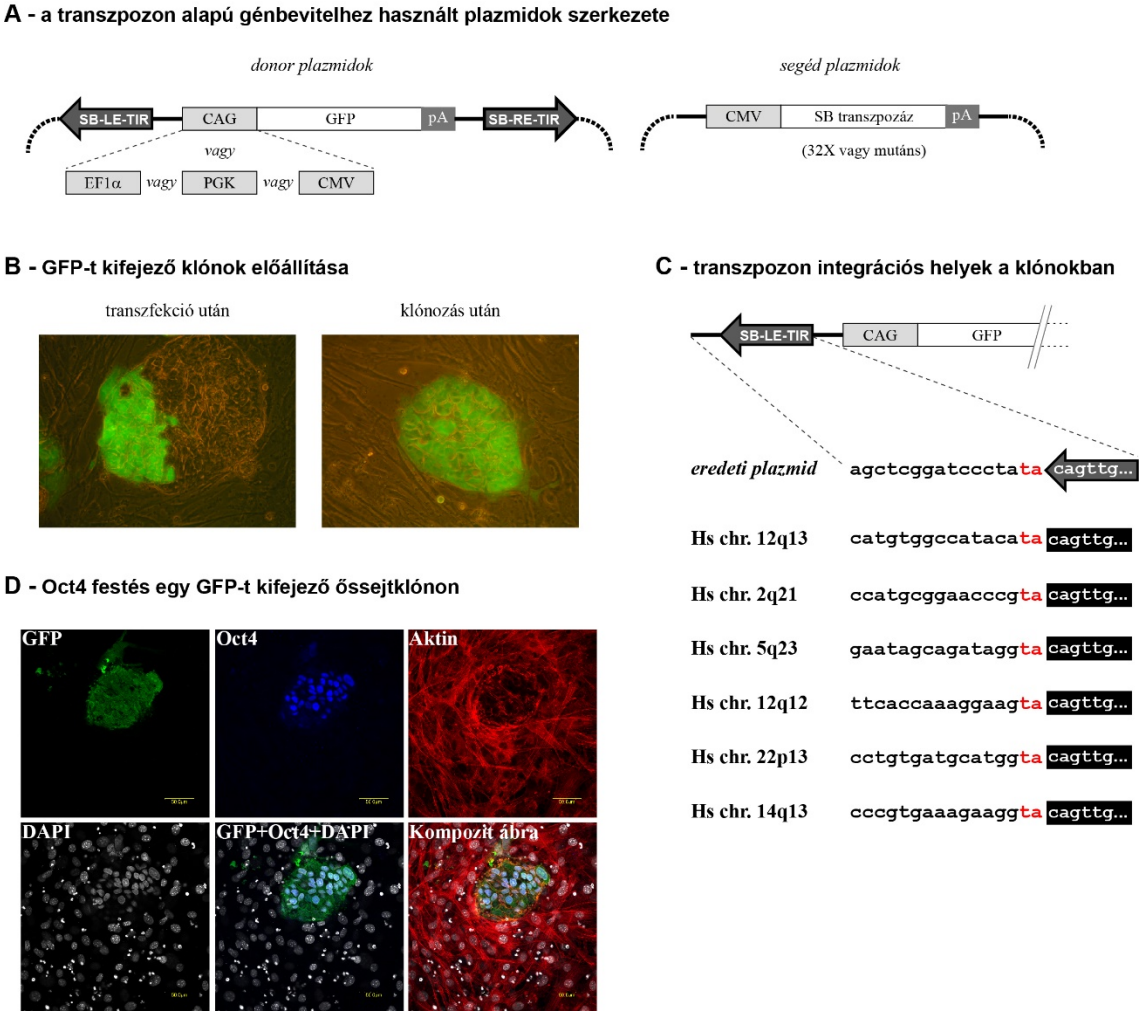
hatékonyságát, amely különböző transzgének alkalmazása során különböző mértékben, de mindig detektálható volt (**30. ábra**). A jelenséget eddig csak a *Sleeping Beauty* rendszer esetében vizsgáltuk, de ennek karakterizálása más transzpozonos rendszerek alkalmazása során is indokolt lehet (Kolacsek és Orban, 2018). Összességében elmondható, hogy az ebben a fejezetben bemutatott kísérleteink rávilágítottak arra: a transzpozon alapú génbevitel nagyon sok szempontból optimalizálható, ami nemcsak a génbevitel hatékonyságát, hanem például a létrehozott transzgenikus sejtek hosszú távú stabilitását is befolyásolhatja, és amely így különösen fontos a génterápiás célú alkalmazások fejlesztése során.

2.2.3. Transzgenikus sejt- és állatmodellek transzpozonokkal: a szívizomtól a patkányveséig

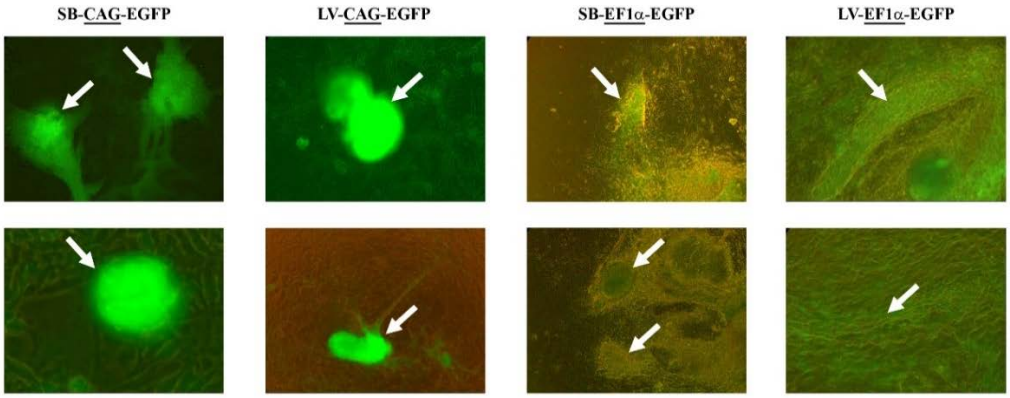
Az elmúlt években a *Sleeping Beauty* transzpozonos rendszer segítségével számos transzgenikus sejtmodellt hoztunk létre, amely közül kétségkívül az embrionális őssejtek módosítása jelentette a legizgalmasabb feladatot, de egyben a legnagyobb kihívást is. Amikor 2007 környékén ezekbe a kísérletekbe belefogtunk, még nem volt ismert, hogy a kiterjedt genomiális védelmi mechanizmusokkal rendelkező embrionális őssejtek vajon képesek-e gátolni egy ismeretlen DNS transzpozon, jelen esetben a *Sleeping Beauty* működését (Krishnan és mtsai, 2006; Liew és mtsai, 2007). Ezek alapján elsőként azt tűztük ki célul, hogy az intézetünkben ekkor már megbízhatóan tenyésztett humán embrionális őssejtvonalak (például a HuES9) felhasználásával létrehozzunk többféle, riporter géneket kifejező stabil őssejtklónokat. Mivel azt lehetett tudni, hogy bizonyos virális eredetű promóterek az őssejtekben csendesülnek (Chung és mtsai, 2002; Zeng és mtsai, 2003; Xia és mtsai, 2007), ezért kísérleteinkben a transzpozon kazettákban már eleve többféle promótert próbáltunk ki. Ezek között volt erős aktivitást biztosító virális promóter (CMV, lásd korábban a **18A ábrán**), viszonylag magas szintű aktivitást mutató mesterséges hibrid promóter (CAG, **21A ábra**), valamint voltak olyan humán endogén promóter szekvenciák (EF1α és PGK, **13C és 18A ábrák**), amelyek alacsonyabb szintű, de éppen ezért több sejttípusban is tolerálható mértékű transzkripciós aktivitást biztosítanak. A várakozásnak megfelelően az előkísérletekből az derült ki, hogy a CMV promóter az őssejtekben egy rövid idejű aktivitást követően hamar csendesül, így a továbbiakban a többi promótert alkalmaztuk a transzgént kifejező, transzpozon alapú vektorokban. A CMV-t a rövid idejű aktivitása ugyanakkor alkalmassá tette a transzpozáz

gén kifejezésére, ebben az esetben ugyanis el akartuk kerülni a fehérje hosszú távú expresszióját, hiszen arra csak a transzfekciót követő szűkebb „időablakban” van szükség. Az eredmények alapján úgy döntöttünk, hogy a kiválasztott promóterek segítségével megpróbálunk GFP riportergént stabilan kifejező őssejtklónokat előállítani (**31A ábra**). Az akkori lipid-alapú módszerekkel kezdetben alacsony transzfekciós hatékonyságot értünk el, ugyanakkor a fluoreszcens szignálok alapján, áramlási citométer segítségével sikerült a sejteket dúsítani, és stabil transzgenikus őssejtklónokat előállítani (**31B ábra**). A korábban általunk kidolgozott módszerrel (lásd 2.2.1 fejezet) meghatároztuk a beépült transzpozonok kópiaszámát, az integrációs helyek meghatározásával pedig igazoltuk, hogy a transzgenikus klónok kialakulása valóban transzpozíciós eseményeknek köszönhető (**31C ábra**). Igazolnunk kellett ugyanakkor, hogy a klónok továbbra is megőrizték „őssejt mivoltukat”, vagyis kifejezik az erre az állapotra jellemző markereket, illetve megőrzik a három csíralemez irányába történő differenciációs képességüket. Az őssejt-specifikus markerekre példaként megmutattuk, hogy a GFP pozitív klónok expesszálják az Oct4 transzkripciós faktort (**31D ábra**), az embriótesteken keresztül történő spontán differenciáció (Thomson és mtsai, 1998; Apati és mtsai, 2008) során pedig sikerült belőlük olyan sejttípusokat előállítani, amelyek az ektoderma, az endoderma, vagy a mezoderma irányú differenciáció során keletkeznek. Úgy tűnt tehát, hogy a *Sleeping Beauty* rendszer kifejezetten alkalmas az embrionális őssejtek genetikai módosítására, és a transzpozonok átmeneti aktivitása nincs negatív hatással az őssejtek állapotára, vagy differenciációs képességére (Orbán és mtsai, 2009; Orbán és mtsai, 2011).

A transzgenikus őssejtekkel folytatott differenciációs kísérleteink érdekes hozadéka volt ugyanakkor egy olyan új megfigyelés, miszerint az általunk használt CAG promóter variáns hatására a spontán összehúzódásra képes szívizomsejtekben nagyon megemelkedett a transzgén expressziója. A GFP esetében a fluoreszcens szignál intenzitása például olyan mértékben megnőtt, hogy az ilyen típusú sejtek a tenyésztőedényben fluoreszcencia mikroszkóp nélkül, szabad szemmel is azonosíthatóak voltak. Ez azért is volt izgalmas megfigyelés, mert ezzel felfedeztünk egy „egyet fizet – kettőt kap” („*double-feature*”) típusú promótert, amely egyszerre használható a génbevitel és a transzfekciós hatékonyság ellenőrzésére (hiszen a legtöbb sejttípusban aktív), ugyanakkor az aktivitás szintje alapján bizonyos sejtek megbízhatóan elkülöníthetőek más sejttípusoktól (itt például a szívizomsejtek más sejtektől).



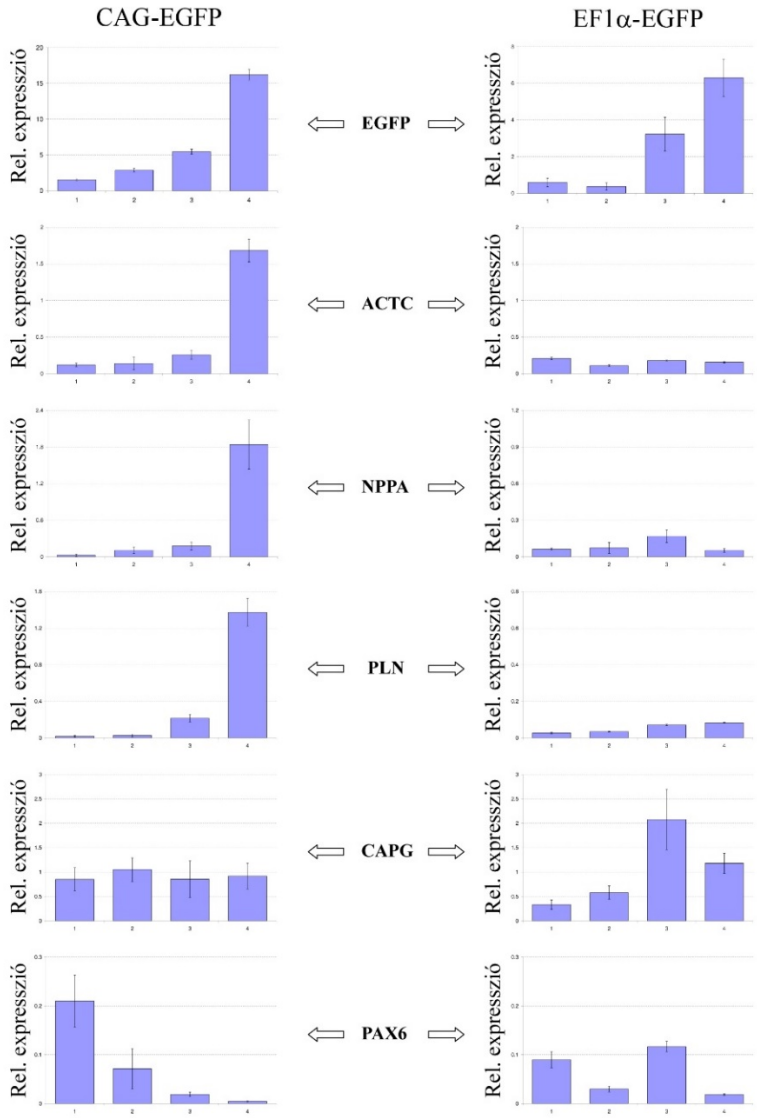
31. ábra: GFP transzgént kifejező humán embrionális őssejtklónok létrehozása a Sleeping Beauty transzpozonos rendszerrel. (**A**) A kétkomponensű rendszerben használt plazmidok sematikus szerkezete: a donor plazmidon a négyféle promóterből mindig csak az egyik található meg, míg a segéd plazmidon mindig a CMV promóter hajtja meg a normál (itt az SB32X variáns) vagy a mutáns (negatív kontroll) transzpozázt. SB: *Sleeping Beauty*. (**B**) Fénymikroszkópos felvételek a tápláló („*feeder*”) sejtrétegen növesztett őssejtcsomókról („*clump*”-okról), áteső és kék fénnyel megvilágítva. A transzfekciót követően a GFP^+ sejteket áramlási citométer segítségével leválogattuk és egyesével klónoztuk. (**C**) A GFP^+ őssejtklónokon *splinkerette* és inverz PCR segítségével meghatároztunk több transzpozon integrációs helyet (egy klónban több kópia is előfordulhat). A *Sleeping Beauty* transzpozon integrációja során egy ’TA’ dinukleotid megkettőződik („*footprint*”), ezeket a bal transzpozonvégek példáján bemutatva az ábrán pirossal jelöltük (balra a genomiális, a ’TA’-tól jobbra pedig a transzpozon szekvencia található). (**D**) Konfokális mikroszkópos felvétel, amelyen az egér tápláló sejtrétegen növesztett humán embrionális őssejtklón látható. Az immunfestés igazolta, hogy az Oct4 transzkripciós faktor a GFP^+ sejtek magjában található (DAPI jelölés); az aktin fehérje a tápláló sejtekben mutatható ki. Az ábra az (Orban és mtsai, 2009) közleményünk alapján készült.



32. ábra: A CAG promóter „double-feature" tulajdonsága független a génbeviteli eljárástól. A fluoreszcencia mikroszkóppal készült felvételeken különféle őssejtklónok spontán differenciációja során keletkezett, szívizom jellegű sejtek láthatók (áteső és kék fénnyel megvilágítva). A fehér nyilak a spontán összehúzódásra képes sejtcsoportokra mutatnak, amelyekben a CAG promóter esetében az EGFP fluoreszcenciája a környező sejtekhez képest jelentősen magasabb. Az EF1α promóter esetében ez a markáns intenzitás különbség nem látható, és a jelenség független attól, hogy a kiindulási transzgenikus őssejtklón a *Sleeping Beauty* (SB) transzpozonos rendszerrel, vagy egy lentivírus alapú (LV) génbeviteli eljárással lett létrehozva. Az ábra az (Orban és mtsai, 2009) közleményünk alapján készült.

Mindenképpen szerettük volna a jelenség molekuláris hátterét feltárni, és első körben azt bizonyítottuk, hogy a szívizomsejtekben a CAG promóter hatására megemelkedő expresszió lentivírusokkal létrehozott őssejtklónokban is detektálható (**32. ábra**). Ezt követően azt kellett igazolni, hogy a transzgén (jelen esetben a GFP) expressziója transzkripciós vagy transzlációs szinten kifejezetten a szívizomsejtekben emelkedik meg. Ennek bizonyításhoz a különféle promóterek által GFP-t kifejező őssejtklónokat spontán differenciáltattuk, majd a keletkezett keverék sejtpopulációt áramlási citométer segítségével, a GFP fluoreszcencia intenzitása alapján 4 frakcióra különítettük el. A frakciókban lévő sejtekből teljes RNS-t izoláltunk, és *real-time* PCR segítségével meghatároztuk a GFP, valamint szív-, bőrsejt- és neuron-specifikus markergének mRNS szintű expresszióját. A vizsgálatokban egyértelműen igazoltuk, hogy a különböző frakciókban a fluoreszcencia intenzitása pozitív módon korrelál a GFP mRNS szintjével, így a frakciók közötti különbség a transzkripciós aktivitásból, nem pedig például a fokozott transzlációból adódott. Azt is sikerült igazolnunk, hogy a CAG promótert tartalmazó klónok utódaiban a szívizom-specifikus markergének (például az ACTC, az NPPA – *natriuretic peptide precursor A*, vagy a PLN – *phospholamban*) mRNS szintje a legmagasabb GFP intenzitást mutató frakcióban szembetűnően magas volt, míg a többi

frakcióban nagyságrendekkel alacsonyabb. Ezzel szemben a bőrspecifikus CAPG (*actin filament capping protein, gelsolin like*) mRNS szintje egyenletesen oszlott el a frakciók között, a neuronspecifikus PAX6 (*paired box gene 6*) mRNS pedig inverz korrelációt mutatott a GFP intenzitással (**33. ábra**).



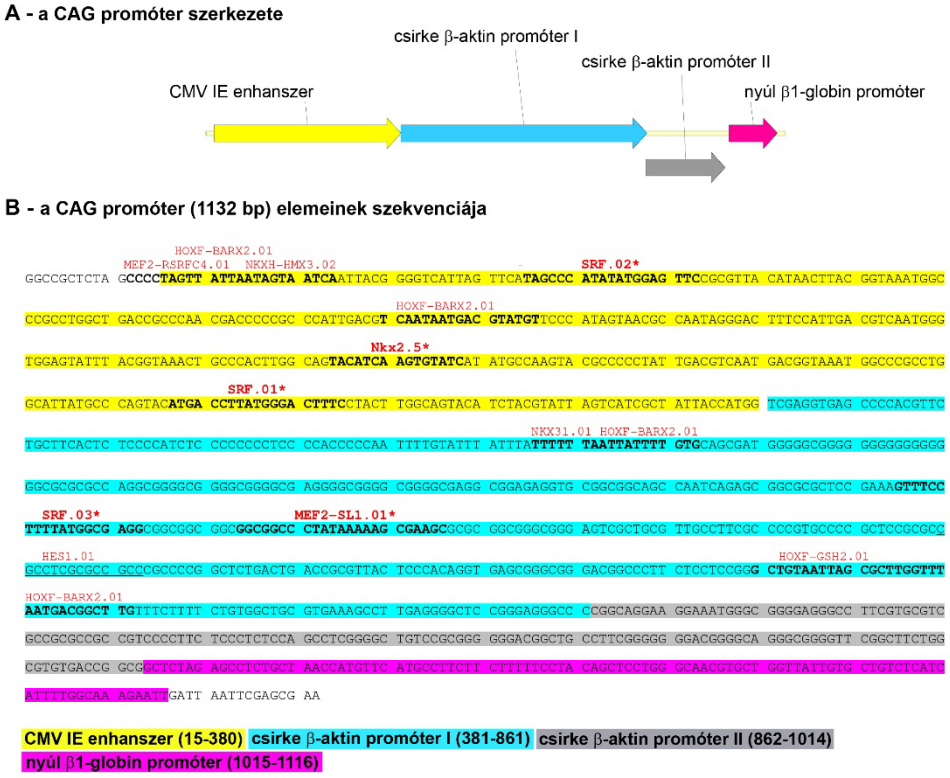
33. ábra: A különféle őssejtklónokból differenciáltatott sejtpopulációk expressziós profilja. Az EGFP fluoreszcencia intenzitása alapján szétválogatott négy sejtcsoporton (1-4) *real-time* PCR segítségével megmértük az EGFP transzgén, több szívizom-specifikus gén (ACTC, NPPA, PLN), valamint egy bőrspecifikus (CAPG) és egy neuronspecifikus (PAX6) gén mRNS szintű expresszióját. A baloldali panelsorozat a CAG-EGFP, míg a jobboldali panelsorozat az EF1α-EGFP transzgént tartalmazó sejteken végzett méréseket mutatja be. A grafikonokon a P0 riboszomális kontrollgénre vonatkoztatott értékek átlagát tüntettük fel, a hibazászlók szórást jelölnek.

Az ábra az (Orban és mtsai, 2009) közleményünk alapján készült.

A CAG promóter specificitására a végső bizonyítékot az endogén eredetű EF1α promótert tartalmazó őssejtklónok hasonló vizsgálata szolgáltatta: ebben az esetben a szívizom-specifikus gének mRNS szintjei egyenletes eloszlást mutattak a különböző intenzitású GFP frakciókban, és a bőrsejt- vagy neuronspecifikus gének eloszlása úgyszintén nem korrelált a GFP mRNS szinttel (**33. ábra**). Eredményeink egyértelműen igazolták a CAG promóter kettős tulajdonságát: egyfajta konstitutív promóterként minden sejttípusban működik, de kifejezetten erős aktivitást mutat szívizomsejtekben, így azok ezen promóter aktivitása alapján megbízhatóan elkülöníthetők más sejtektől. A CAG promóter ezen karakterisztikáját tovább erősítette, hogy a jelenséget humán eredetű sejtek mellett a későbbiekben egér embrionális őssejtekből differenciált szívizomsejteken is sikerült igazolni (Orban és mtsai, 2009).

De milyen szekvenciális okok állhatnak a CAG promóter ezen nem várt tulajdonsága mögött? A kérdés megválaszolásához a szakirodalomban utánanéztünk annak, hogy honnan is származik a sok laboratórium által előszeretettel használt CAG promóter, és hogy valaha tapasztaltak-e ilyen erős transzkripciós aktivitást kifejezetten szívizom sejtekben – és az eredmények minket is megleptek. Az eredetileg létrehozott hibrid promóterrel a cél valóban az volt, hogy létrehozzanak egy általánosan használható, a legtöbb sejtben konstitutívan aktív promótert. Ehhez felhasználták a **C**MV vírus *immediate early* enhanszer szekvenciáját, amelyet fuzionáltak a csirke β-**a**ktin promóter egy hosszabb, illetve a nyúl β1-**g**lobin promóter egy rövidebb szekvenciájával. Az így létrehozott mesterséges hibrid promóter valóban rendelkezett a kívánt konstitutív aktivitással, a rövidített neve pedig a felépítő „modulok” kezdőbetűiből állt össze: C-A-G (Niwa és mtsai, 1991). A probléma ott kezdődött, hogy a molekuláris klónozás elterjedésével a különböző laboratóriumok a saját konstrukcióik előállításakor a saját klónozási stratégiáiknak (és kényszereiknek) megfelelően meglehetősen „rugalmasan” alakítgatták a hibrid promóter különféle régióit, így mára már nagyon sokféle „CAG promóter” létezik, amelyek sokszor szekvenciálisan is nagyon különböznek egymástól: némelyikből akár egy teljes modul (például a nyúl β1-globin promóter) is hiányzik (Sawicki és mtsai, 1998; Xu és mtsai, 2001). Ennek megfelelően még ha találtunk is volna az erős szívizom-specificitásra utaló adatot, akkor az abban a kísérletben alkalmazott CAG promóter variánst kellett volna összehasonlítani az általunk használt promóterrel – viszont a közleményekből sokszor már nem derül ki a szekvenciális összetétel, hiszen a szerzők csak „CAG”-ként hivatkoznak a vektorukban található promóterre. Az őssejtes

vizsgálatokban általunk használt CAG promóter variáns tartalmazza az összes eredeti C-A-G modult, bár az aktin régió rövidebb, mert az eredeti szekvencia középső régiója hiányzik (**34A ábra**). A variánst korábban egy izraeli kutatócsoporttól kaptuk (Evelyn Zeira, *Hadassah Medical Organization*, Izrael), azonban a pontos eredetét sajnos nem sikerült kideríteni. Arra gondoltunk, hogy ebben a variánsban esetleg a modulszekvenciák egy szerencsés konstellációja okozhatja a tapasztalt fenotípust, ezért bioinformatikai módszerekkel próbáltunk olyan szekvencia motívumokat keresni, amelyek szívizom-specifikus transzkripciós faktorok kötőhelyei lehetnek. A szigorú feltételek mellett elvégzett predikciók kapcsán több ilyen motívumot is sikerült azonosítanunk, amelyek a vírus enhanszer régióban, illetve az aktin modulban találhatók (**34B ábra**). Az egyik legígéretesebb találat a vírus enhanszer régióban azonosított Nkx2.5 kötőhely, mert ennek a transzkripciós faktornak kiemelkedő szerepe van a szívizom irányú differenciációban, a korai neonatális fejlődéstől kezdve az érett kardiomiociták kialakulásáig (Chen és Schwartz, 1996; Sepulveda és mtsai, 1998; Linhares és mtsai, 2004; Christoforou és mtsai, 2008). A predikciók alapján fontos szerepe lehet még a három azonosított szérum reszponzív faktor (SRF.01-03) és a MEF2 (MEF2-SL1.01) kötőhelyeknek, mert az ide kapcsolódó fehérjék egymással, illetve más szívizom-specifikus faktorokkal (például a miokardinnal) asszociálódva szintén kitüntetett szereppel bírnak a kardiomiocita differenciáció során (Wang és mtsai, 2001). A bioinformatikai elemzést követő kísérletes vizsgálatokkal azonban egyelőre még nem sikerült minden kétséget kizárólag bizonyítanunk a fenotípus mögött rejlő szekvenciális hátteret. Maga a jelenség ugyanakkor nem csak tudományos, hanem orvosbiológiai szempontból is fontosnak bizonyult, amit az is bizonyít, hogy a szívizomsejtek ezen módszerrel történő izolálására beadott szabadalmi igényünket Európában elfogadták (az európai regionális fázis alapszáma: *EP 08839475.4*). A jövő kutatásait illetően a „*double feature*” tulajdonság mögötti szekvenciális háttér megértése pedig azért is fontos és izgalmas kérdés, hiszen annak ismeretében elvileg más sejttípusra specifikus „kettős arcú” promótereket is elő lehetne állítani, amelyeknek a jövőben a sejt- és génterápiák tekintetében komoly jelentőségük lehet (Orban és mtsai, 2009).



34. ábra: A CAG promóter szerkezete és szekvenciális összetevői. (**A**) Az általunk használt CAG promóter moduljainak méretarányos ábrázolása. CMV IE: citomegalovírus *immediate early* (enhanszer). (**B**) A CAG promóter modulok szekvenciája, jelölve a prediktált transzkripciós faktor kötőhelyek pozícióit. A legerősebb predikciókat *-gal jelöltük, ezek közül a 234-248-as pozícióban található Nkx2.5 kötőhely tűnik szívizom-specificitás szempontjából a legígéretesebbnek (magyarázat a szövegben).
Az ábra az (Orban és mtsai, 2009) közleményünk alapján készült.

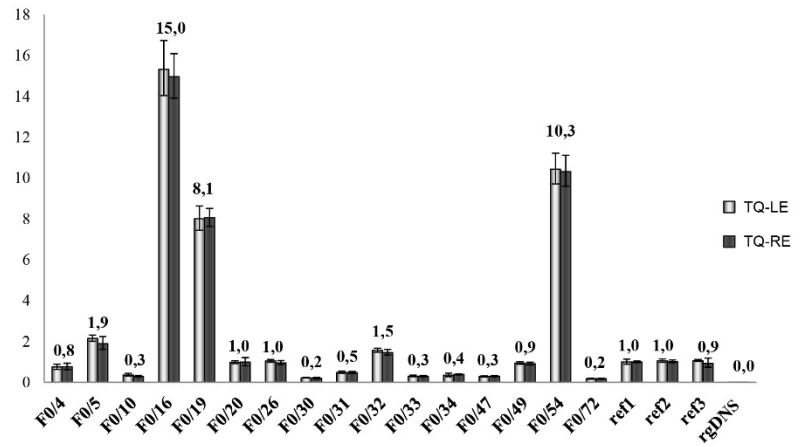
Az elmúlt évtizedben a transzpozonos rendszerek segítségével nem csak transzgenikus sejtvonalakat, hanem különféle transzgenikus állatmodelleket is sikerült előállítanunk. Ezek közül mindenképp kiemelném azt a farmakológiai kutatási célokra létrehozott patkányvonalat, amely homozigóta formában minden sejtjében hordozza a GCaMP2 elnevezésű fluoreszcens kalcium szenzorfehérje génjét (Nakai és mtsai, 2001; Tallini és mtsai, 2006). A széleskörű tudományos együttműködés keretében létrehozott patkányvonal előállításánál a transzpozonos rendszer alkotóelemeit közvetlenül a megtermékenyítést követően a zigótába injektáltuk. Célunk az volt, hogy a létrehozandó vonal egyedeiben a transzgén minden sejttípusban kifejeződjön, de az expresszió a szívizomsejtekben kifejezetten magas legyen – ennek érdekében a *Sleeping Beauty* transzpozon alapú donor vektorban a GCaMP2 fehérje génjét a már említett CAG

promóter variánssal hajtottuk meg. Az injektálásnál a transzpozázt a zigótába nem plazmid, hanem transzlációképes mRNS formájában juttattuk be, amivel a lehető legkisebb mértékűre szerettük volna csökkenteni az egyedek mozaikosságát. A rágcsálók embrionális fejlődésére ugyanis általánosan jellemző, hogy a zigóta állapotot követő első néhány osztódás alatt a genom transzkripcionálisan inaktív, és a sejtek működéséhez szükséges információt az anyai eredetű fehérjék és RNS-ek szolgáltatják. Ha a transzpozázt DNS (vagyis plazmid) formában juttatjuk a zigótába, arról működőképes fehérjetermék csak az első néhány osztódást követően a különböző utódsejtekben lesz, így a transzpozíciós események a sejtek különféle leszármazási vonalaiban függetlenül, eltérő kópiaszámban és integrációs pozíciókban történnek – vagyis az egyed a transzgén szempontjából mozaikos lesz. Ha azonban már az osztódások előtt jelen van a transzpozázt kódoló mRNS, akkor komoly eséllyel már a zigótában is történik transzpozíció, így lesznek olyan integrálódott transzpozonok, amelyek az egyed minden sejtjében ugyanabban a pozícióban lesznek jelen.

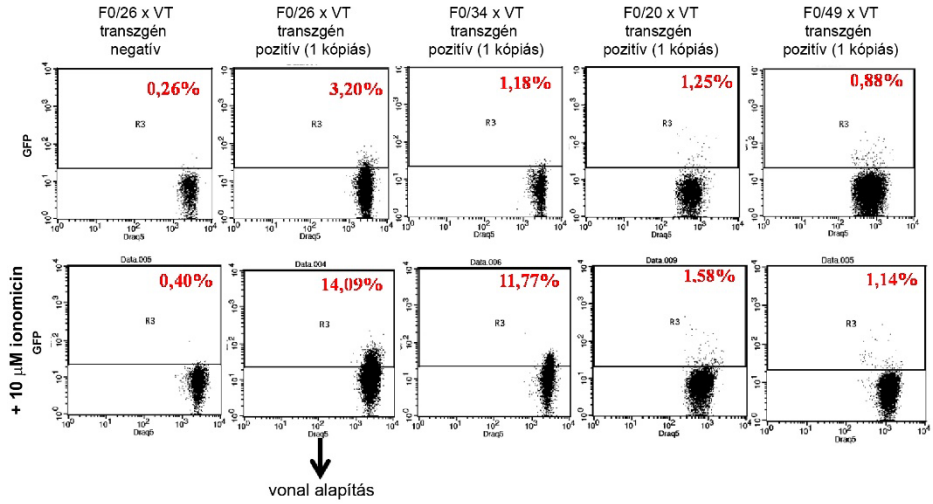
Az injektálást követően a zigóták álterhes nőstényekbe lettek beültetve, amelynek eredményeképpen megszülettek az ún. F0 (alapító) generáció egyedei. A génbeviteli módszerünk hatékonynak bizonyult: a farokszövetből származó genomiális DNS minták vizsgálata alapján 75 újszülöttből 16 hordozót azonosítottunk (~21%), ami a hasonló módszereknél tapasztalt transzgenikus rátákkal összehasonlítva nagyon jónak mondható. Azért, hogy a stabil patkányvonal előállítását minél hatékonyabbá és gyorsabbá tegyük, a pozitív mintákban a korábban kidolgozott *real-time* PCR módszer (2.2.1 fejezet) segítségével meghatároztuk a transzpozonok kópiaszámát is (**35A ábra**). Az egyedek közül azokat válogattuk ki további tenyésztésre, amelyek 1 kópiában tartalmazták a bevitt transzgént, hiszen ezektől várható a több generáción keresztül történő stabil örökítés, és ezekből lehet a későbbi keresztezések során homozigóta vonalakat előállítani. A tenyésztés során azonban komoly problémát jelent, hogy nem tudjuk, vajon a kérdéses vonalban a transzgén ténylegesen kifejeződik-e, vagy esetleg csendesített állapotban öröklődik, ezért a további keresztezések előtt funkcionális teszttel igyekeztünk meghatározni az aktív transzgént hordozó egyedeket. Mivel a konstrukcióban alkalmazott CAG promóter a „*double feature*” jellegből adódóan egy konstitutív jellegű promóter is, ezért feltételeztük, hogy különféle vérsejtekben is kifejeződik, így izolált leukocitákban ellenőrizni tudtuk a működését. A GCaMP2 egy GFP eredetű kalcium szenzorfehérje, amelynek a fluoreszcencia aktivitása a kalcium szint emelkedésével nő (Tallini és mtsai, 2006). A vizsgálandó egyedek vérmintáiból leukocitákat izoláltunk, majd áramlási

citométer segítségével meghatároztuk a fluoreszcens jelet kezeletlen, illetve ionomicinnel kezelt sejteken. A kezelés hatására a sejtek endogén kalcium szintje megnő, amivel együtt a GCaMP2 fehérje fluoreszcenciája is szignifikánsan megemelkedik (**35B ábra**). A mérések segítségével ki tudtuk szűrni azokat az 1 kópiás egyedeket, amelyekben a transzgén aktívan működött: ezzel egyrészt kezelhető mennyiségűre tudtuk csökkenteni a további tenyésztésre alkalmas állatok számát, illetve jelentősen fel tudtuk gyorsítani a stabil transzgenikus vonal előállításának az idejét. A kiválogatott F0 generációs utódok vad típusú egyedekkel való keresztezését („tisztázását") követően az azonos eredetű transzgént tartalmazó alomtársakat egymással párosítva sikerült előállítanunk olyan homozigóta egyedeket, amelyek a haploid genomban 1 kópiában tartalmazták a CAG-GCaMP2 transzgént. Ezek közül kiválasztottuk azt a vonalat, amely a különböző szövetek expressziós mintázatát tekintve is a legalkalmasabbnak tűnt, és amelyben a transzpozon szekvenciák segítségével pontosan meghatároztuk a transzgén integrációs helyét a genomban (**35C ábra**). Sikerült tehát létrehoznunk egy olyan patkányvonalat, amely több generáción át stabilan örökítette a kalcium szenzorfehérjét kifejező transzgént, és amely további funkcionális vizsgálatokra is alkalmas volt (Szebenyi és mtsai, 2015b). A vonal egyedein ezt követően többféle szöveti működéshez kapcsolódó élettani és farmakológiai vizsgálatsorozatot végeztünk: teszteltük például különféle hatóanyagok kardiovaszkuláris mellékhatásait (Szebenyi és mtsai, 2015b), elemeztük a szívműködés hemodinamikai jellemzőit (Olah és mtsai, 2019), illetve a GCaMP2 kalciumszenzor segítségével vizsgáltuk a májsejtek toxikológiai működését (Jemnitz és mtsai, 2017), vagy az idegrendszerben az asztrociták szikronizációját (Szabo és mtsai, 2017).

A - átlagos transzpozon kópiaszámok az alapító (F0) generáció hordozó egyedeiben



B - GCaMP2 aktivitás vizsgálata az első utódnemzedék leukocitáin, áramlási citométerrel



C - transzpozon integrációs hely a létrehozott stabil vonalban

GGGACTAGGTTGGGCTAAGAGTGAAGACTCTTTAGC
TGTCGTTCTATGGCAATCCTGACAGGATTCCACTCC
CCTTGTAAAGCAGGTA***CAGTTGAAGTCGGAAGTTTA***
CATACACCTTAGCCAATCACTAGTGAATTCGCGGCC
GCCT

normál: intergenikus régió a patkány 9. kromoszómáján (ref. szek.: NC_005108.3; nukleotid pozíció: 78819834)

kiemelt: a *Sleeping Beauty* transzpozon LE-TIR szekvenciája

35. ábra: A GCAMP2-t hordozó patkányvonal létrehozása. (**A**) A kópiaszám meghatározása farokszövetből történt, a *Sleeping Beauty* transzpozon végszekvenciáira specifikus *real-time* PCR próbákkal (lásd **26. ábra**); az <1 kópiaszámú egyedek genetikailag mozaikosak. ref1-3: ismert kópiaszámú referenciaminták; rgDNS: patkány genomi DNS (negatív kontroll). A grafikonon átlagértékek ± konfidencia intervallumok szerepelnek. (**B**) A jelölt F0 egyedek vad típusú (VT) állatokkal történő keresztezéséből származó egyedek vérsejtjeinek funkcionális vizsgálata, a piros számok a GFP^+ sejtek arányát jelölik. Draq5 (x-tengely): sejtmagot jelölő festék. (**C**) A létrehozott, a GCaMP2 transzgénre homozigóta patkányvonalban a *splinkerette* PCR segítségével meghatározott integrációs hely szekvenciakörnyezete.

Az ábra a (Szebenyi és mtsai, 2015b) közleményünk alapján készült.

Az eddigi eredmények mellett izgalmas adalékként ismét fény derült az alkalmazott CAG promóter egy új tulajdonságára: patkányban a szívizom sejtek mellett igen erőteljes aktivitást mutatott a vese proximális tubulussejtjeiben. Ez a megfigyelés egy kollaborációs munka kereteiben lehetőséget nyújtott arra, hogy modern képalkotó módszerekkel, eddig nem hozzáférhető módon tanulmányozzuk a vese működését, a proximális tubulusok aktivitását. Konfokális és két-foton mikroszkópia segítségével sikerült bizonyítani, hogy a veseszövetben egy adott időben a proximális tubulusok csak egy része végez aktív szűrést, és ezeket a sejtjeik magasabb citoplazmatikus kalcium szintje miatt a GCaMP2 szenzor rendszerrel vizualizálni lehet (Szebenyi és mtsai, 2015a). A transzgenikus modell által biztosított módszer erejét jól mutatja, hogy a proximális tubulusok működésének ilyen részletes és látványos bemutatására korábban nem volt lehetőség, és talán nem véletlen, az általunk bemutatott eljárást a veseműködést ismertető komoly áttekintő munkákban is idézik (Mullins és mtsai, 2016; Martins és mtsai, 2021).

Nagyon messzire jutottunk a transzpozon alapú génbeviteli eljárások optimalizálásától a sejt- és állatmodellek létrehozásán keresztül az élettani működések részletes vizsgálatáig. Mindezek alapján azt gondolom, hogy az ebben a fejezetben ismertetett kutatási eredmények jól mutatják a molekuláris biológiai kutatások erejét és a bennük rejlő potenciált: az eredetileg Barbara McClintock által azonosított transzpozonok molekuláris hátterének a tudományos kíváncsiság vezérelte feltárása, majd különféle sejtes rendszerekben történő alkalmazása elvezetett minket olyan élettani és farmakológiai jelenségek megértéséhez, amelyekről kijelenthetjük, hogy jó eséllyel hozzájárulhatnak az orvosbiológiai vizsgálatok és kezelések jövőbeli fejlesztéséhez.

2.3. Kitekintés – transzpozonok: a velünk élő történelem

A transzpozonok felfedezése egy rendkívül fontos mérföldkő volt a genetika, illetve a molekuláris biológia történetében. Túlzás nélkül állíthatjuk, hogy a Barbara McClintock által feltérképezett mobilis genetikai elemek egy komoly szemléletbeli paradigma váltást hoztak magukkal, amelyet a kortárs kutatók nagyon nehezen fogadtak el (lehet, hogy nem is értettek igazán) – talán hasonlóan a fizika tudományában a XX. század elején megjelent rendkívül újszerű és az addigi világképet átformáló elméletekhez. A további kutatások pedig újabb izgalmas eredményeket hoztak: a kezdetben molekuláris parazitáknak titulált transzpozonokról nem csak az derült ki, hogy az élővilág minden nagyobb doménjében jelen vannak, de az „önző" invazív tulajdonságuk mellett sokszor pozitív hatásaik is vannak a genomok evolúciójára. Arra is rájöttünk, hogy a mobilis elemek jelentős mértékben hozzájárulnak nemcsak a genomok klasszikus értelemben vett nem-kódoló régióinak, hanem bizony a fehérjekódoló régióknak az alakulásához, fejlődéséhez is. Az egyre nagyobb számban leírt transzpozon domesztikációk kapcsán sikerült bepillantanunk a genomokban bekövetkező változások rejtelmeibe, köztük az emberi örökítőanyag evolúciós múltjába is. Nem szabad ugyanakkor elfeledkezni a transzpozonok mutagén hatásairól sem, amely tetten érhető például a rákos daganatok kialakulásában (Hancks és Kazazian, 2016), vagy akár az öregedés folyamatainak a hátterében is (De Cecco és mtsai, 2019; Simon és mtsai, 2019). Érdemes viszont azt is kiemelni, hogy a „korlátok közé szorított" mutagén transzpozonok nagyon hasznos szolgálatot tettek számos sejtbiológiai folyamat molekuláris részleteinek a feltárásában, a „megszelídített" transzpozonokat pedig – például bizonyos génterápiás fejlesztések kapcsán – sikerült a modern gyógyítás szolgálatába állítani.

Az elmondottak mellett persze fontos hangsúlyozni, hogy nagyon sok mindent nem tudunk még a transzpozonok világáról, és sok izgalmas felfedezés áll még előttünk e téren is. Az értekezésem második felében bemutatott eredmények kapcsán azonban talán érzékelhető, hogy mint annyi más tudományos probléma és kihívás, így a transzpozonok világának felderítése is túlmutat a puszta tudományos kíváncsiságon és intellektuális „játékokon" – én pedig személy szerint örülök, hogy ha csak egy kicsit is, de részese lehettem ezeknek a molekuláris biológiai felfedezéseknek.

AZ ÚJ TUDOMÁNYOS EREDMÉNYEK ÖSSZEFOGLALÁSA

Az elmúlt évekre visszatekintve úgy érzem, hogy mind posztdoktor kutatóként, mind később laboratórium vezetőként sikerült kamatoztatnom a PhD éveim alatt megszerzett molekuláris biológiai szemléletmódot és tudást. Szerencsésnek mondhatom magam, hiszen az elmúlt évtizedekben részese lehettem az általam is művelt tudományterület robbanásszerű fejlődésének, és olyan izgalmas tudományos témákkal foglalkozhattam, amelyek a kíváncsiság vezérelte kutatás mellett lehetőséget biztosítottak a megismert jelenségek és a kifejlesztett módszerek orvosbiológiai hasznosítására is. A doktori értekezésemben bemutatott eredményeket két főbb kutatási téma, az RNS interferencia, illetve a DNS transzpozonok működése köré lehet csoportosítani, és a kutatásaim eredményeit a következő pontokban lehet összefoglalni:

1. Hozzájárultunk az RNS interferencia folyamatainak pontosabb megértéséhez, a molekuláris mechanizmusok részleteinek feltárásához, ezen belül:

❶ Leírtuk az siRNS útvonal által kijelölt mRNS molekulák degradációs mechanizmusát. Igazoltuk, hogy a RISC által kiváltott endonukleolitikus hasítást követően a keletkező termékek a sejt endogén lebontó hálózatán keresztül eliminálódnak: az 5’ intermediert a vágási ponttól „visszafele” (3’→5’ irányban) a Ski-komplex segítségével az exoszóma-komplex bontja le, míg a 3’ intermedier degradációját a vágási ponttól 5’→3’ irányban az Xrn1 fehérje végzi el. Ez a mechanizmus később általánosnak bizonyult minden genetikai modellrendszer esetében.

❷ Részletesen vizsgáltuk egy alternatív útvonalon érő miRNS csoport, a mirtronok képződésének molekuláris részleteit. Ezen miRNS-ek prekurzorai más gének rövid intronjaként találhatók meg a genomban, és az érésük kezdeti lépését nem a Mikroprocesszor-komplex, hanem a splicing apparátus végzi („Drosha-független” miRNS-ek). Más kutatócsoportokkal egy időben igazoltuk, hogy az útvonal létezik emlős rendszerekben is, ugyanakkor felhívtuk a figyelmet a mirtronok bioinformatikai predikciójának elégtelen voltára, és a kísérletes validálás szükségességére.

❸ Igazoltuk, hogy a miRNS gének terméke a kezdeti elképzelésekkel ellentétben nem egy diszkrét RNS molekula, hanem egy komplex kisRNS populáció. Az érési folyamatokban nagyon fontos szerepe van az 5’ vagy a 3’ kar kiválasztásának, de több miRNS gén esetén sokszor mindkét kar egyidejű érése megfigyelhető. A diverzitás fontos szabályozó lépése az izomiR-ek képződése, és igazoltuk, hogy ezek kimutatása vagy független kvantitálása komoly technikai kihívást jelent.

❹ Megmutattuk, hogy a klaszterekben elhelyezkedő miRNS-ek érését speciális tényezők befolyásolják. Az egyedi miRNS-ekre ható faktorok mellett a Mikroprocesszor-komplex komponenseinek, a Drosha és a DGCR8 fehérjéknek a lokális koncentrációja limitáló tényező lehet, és a klaszterről képződő miRNS-ek szekvenciális sorrendjének megfelelő kisRNS mennyiségi grádiens kialakulását, egyfajta pozíció hatást eredményezhet.

2. Részletesen vizsgáltuk a DNS transzpozonok működését és evolúcióját, ezen belül:

❶ Szisztematikus vizsgálatokkal bizonyítottuk, hogy a humán genomban található *piggyBac*-eredetű fehérjék (PGBD1-5) nem képesek transzpozícióra, és több funkcionális vizsgálattal is igazoltuk, hogy ezek az elemek endogén funkciókra domesztikálódtak. A rovar *piggyBac* fehérje N-terminális rendezetlen régiójáról megmutattuk, hogy részt vesz a transzpozíció finomhangolásában, és minden valószínűség szerint a transzpozíció excíziós lépésére van jelentős hatással. Mivel az N-terminális rendezetlen régió a *piggyBac* szupercsalád képviselőinek túlnyomó többségében megtalálható, ezért feltételezzük, hogy az általa biztosított molekuláris interakciók a domesztikált fehérjék szabályozásában is jelentős szereppel bírnak.

❷ A rovar *piggyBac* és a feltámasztott *Sleeping Beauty* rendszerek részletes vizsgálatával behatóan elemeztük a transzpozon alapú génbeviteli eljárások tulajdonságait, és többféle sejtes rendszerben meghatároztuk a hatékony transzgenezis optimális paramétereit. Kidolgoztunk olyan *real-time* PCR alapú módszereket, amelyekkel egyrészt a beépült transzpozonok kópiaszáma, illetve a transzpozíció excíziós lépésének hatékonysága nagyon pontosan kvantitálható.

❸ Az általunk optimalizált *Sleeping Beauty* transzpozonos rendszerrel sokféle sejt- és állatmodellt hoztunk létre. Ezek közül komoly kihívást jelentett a humán embrionális őssejtvonalak módosításának kidolgozása, amelynek izgalmas „melléktermékeként" azonosítottunk egy szokatlan tulajdonságú, ún. „*double-feature*" promótert. Ez a CAG promóter variáns konstitutív promóterként a génbevitel indikátoraként is használható, ugyanakkor a szívizomsejtekben tapasztalható különösen erős aktivitását felhasználva ezen sejttípusok megbízhatóan elkülöníthetők más sejtektől. Az általunk létrehozott transzgenikus állatmodellek közül kiemelendő a GCaMP2 kalcium szenzorfehérjét kifejező patkányvonal, amely az alkalmazott promóter tulajdonságai alapján kifejezetten alkalmas szívizom-, illetve vesesejtek élettani és farmakológiai vizsgálatára.

A doktori értekezésemben bemutatott eredmények közös metszetét a genom „nem-kódolónak" titulált elemei jelentik – ez a definíció azonban úgy érzem, hogy lassan kezd idejétmúlttá válni. Ahogy egyre többet tárunk fel genomunk titkaiból, úgy kezdik lassan a „nem-kódoló" gének számosságban is felülmúlni a fehérjekódoló géneket (Mattick és mtsai, 2023). Ugyanakkor egyre világosabbá válik, hogy ezekre az általunk generált mesterséges kategóriákra nem független entitásokként kell tekintenünk, hiszen ezek összehangolt működése tette lehetővé a minket körülvevő élőlények végtelen változatosságát. A kódoló és nem-kódoló szekvenciák pazar összjátékának az egyik legszebb példái éppen a transzpozonok, ahol a transzpozáz fehérjék a repetitív DNS szakaszokkal együttesen képesek egyfajta parazita viselkedésre, másrészt viszont pozitív módon is képesek hozzájárulni a genomok fejlődéséhez. Külön öröm számomra, hogy az RNS alapú szabályozási mechanizmusok feltárása mellett a mobilis genetikai elemek molekuláris vizsgálatával együtt bepillanthattam a genomok, azok között is az emberi genom evolúciójának eddig ismeretlen rétegeibe, és a tudományos pályám még előttem álló részében is arra fogok törekedni, hogy ezt az izgalmas utazást tovább folytassam.

AZ ALKALMAZOTT METODIKÁK ÁTTEKINTÉSE

A doktori értekezésemben bemutatott kutatásokban sokféle molekuláris biológiai és sejtbiológiai módszert alkalmaztunk, amelyek részletes bemutatása megtalálható a disszertáció alapját képező tudományos közleményekben. Ahol a kísérletek pontos megértése ezt megköveteli, ott az értekezés megfelelő részében részletesebben is kitértem az adott metodikákra, a jelen fejezetben ugyanakkor csak röviden, tematikailag csoportosítva sorolom fel a fontosabb anyagokat és módszereket.

Baktériumtörzsek és plazmidvektorok

Baktérium gazdasejtként az *Escherichia coli* K12 törzsének az XL1-Blue és a DH5α változatait használtuk, és a szokásos mikrobiológiai körülmények között tenyésztettük, illetve tettük plazmid felvételre kompetenssé. A kísérleteink során alkalmazott plazmidvektorokat a hagyományos molekuláris klónozással (restrikciós endonukleázok, ligázok, stb. segítségével), vagy a modern klónozási módszerekkel (pl. *Gibson Assembly* módszer) állítottuk elő. A konstrukciók végső ellenőrzése *Sanger*-szekvenálással történt.

Eukarióta sejttenyészetek fenntartása és manipulálása

A kísérletekben a tudományos kutatásokban általánosan elfogadott eukarióta sejtvonalakat használtunk (HEK-293, HeLa, MCF-7, stb.), amelyeket a nemzetközi standardoknak megfelelő módszerekkel és körülmények között tenyésztettünk. A humán pluripotens őssejtvonalak fenntartását és differenciáltatását Apáti Ágota laboratóriumában végezték. A sejtvonalakon történő kezelések (pl. különféle stresszkezelések) és nukleinsav bejuttatási módszerek (pl. transzfekció) részletei, és az adott kísérleti rendszerekre történő legoptimálisabb paraméterek a disszertációhoz kapcsolódó közleményekben részletesen megtalálhatók.

Transzgenikus patkányvonalak előállítása

Az általunk kidolgozott módszertanról szól a (Szebenyi és mtsai, 2015b) közleményünk, amelyben a technikai paraméterek kellő részletességgel megtalálhatók.

Génexpressziós vizsgálatok

A különféle RNS és fehérje molekulák kimutatására mindig az adott kísérleti rendszerre leginkább optimális módszereket alkalmaztuk. Az RNS molekulák kvantitálására *Northern* analízist és *real-time* PCR módszereket, a fehérjék kimutatására pedig *Western* analízist és immunprecipitációs módszereket használtunk, a nemzetközileg elfogadott módszertani ajánlások figyelembevételével. A kisRNS molekulák funkcionális vizsgálatára luciferáz-esszéket alkalmaztunk.

A DNS transzpozonok működésének vizsgálata

A kísérleteinkben használt DNS transzpozonok működését többféle molekuláris módszerrel vizsgáltuk: az excíziós és kolónia esszék részletei a disszertációban is leírásra kerültek (lásd a 2.1. részben található fejezeteket). A transzpozonok kópiaszámának pontos meghatározására kidolgozott módszerünk az erről szóló közleményünkben található meg részletesen (Kolacsek és mtsai, 2011).

Bioinformatikai analízisek és statisztikai elemzések

A releváns kísérleteinkben használt bioinformatikai predikciókat és elemzéseket az interneten szabadon hozzáférhető platformokon és szoftverekkel végeztük, a részletek az egyes közleményekben megtalálhatók. A vizsgálatok és mérések során generált adatok statisztikai analízisei a biológiai mintákra adaptált nemzetközi ajánlások alapján készültek, a részletek a közleményeken kívül az adatokat bemutató ábrák magyarázatainál is megtalálhatók.

A DISSZERTÁCIÓ ALAPJÁT KÉPEZŐ KÖZLEMÉNYEK

A doktori értekezésem alapját olyan, a PhD fokozat megszerzését (2003) követő években megjelent közlemények jelentik, amelyekben meghatározó (első vagy utolsó/levelező (*)) szerző vagyok. Ezek a közlemények időrendben a következők:

1. **Orban, T. I.**, and E. Izaurralde: Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. *RNA* 2005, 11(4):459-469.
   doi: https://doi.org/10.1261/rna.7231505.
   impakt faktor: 6,140 (D1)
   független idézetek száma: 265

2. **Orban, T. I.**, A. Apati, A. Nemeth, N. Varga, V. Krizsik, A. Schamberger, K. Szebenyi, Z. Erdei, G. Varady, E. Karaszi, L. Homolya, K. Nemet, E. Gocza, C. Miskey, L. Mates, Z. Ivics, Z. Izsvak, and B. Sarkadi: Applying a ”double-feature” promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery. *Stem Cells* 2009, 27(5):1077-1087.
   doi: https://doi.org/10.1002/stem.45.
   impakt faktor: 7,750 (D1)
   független idézetek száma: 25

3. Kolacsek, O., V. Krizsik, A. Schamberger, Z. Erdei, A. Apati, G. Varady, L. Mates, Z. Izsvak, Z. Ivics, B. Sarkadi, and **T. I. Orban***: Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. *Mob DNA* 2011, 2(1):5. doi: https://doi.org/10.1186/1759-8753-2-5.
   impakt faktor: 2,430 (Q2)
   független idézetek száma: 19

4. Schamberger, A., B. Sarkadi, and **T. I. Orban***: Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. *RNA Biol* 2012, 9(9):1177-1185. (***cover page story***)
   doi: https://doi.org/10.4161/rna.21359.
   impakt faktor: 4,841 (Q1)
   független idézetek száma: 15

5. Schamberger, A., and **T. I. Orban***: 3' IsomiR species and DNA contamination influence reliable quantification of microRNAs by stem-loop quantitative PCR. *PLoS One* 2014, 9(8):e106315.
doi: https://doi.org/10.1371/journal.pone.0106315.

impakt faktor: 3,234 (D1)
független idézetek száma: 46

6. Schamberger, A., and **T. I. Orban***: Experimental validation of predicted mammalian microRNAs of mirtron origin. *Methods Mol Biol* 2014, 1182:245-263. doi: https://doi.org/10.1007/978-1-4939-1062-5_22.

impakt faktor: - (*könyvfejezet*)
független idézetek száma: 1

7. Kolacsek, O., Z. Erdei, A. Apati, S. Sandor, Z. Izsvak, Z. Ivics, B. Sarkadi, and **T. I. Orban***: Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. *Hum Gene Ther Methods* 2014, 25(4):241-252.
doi: https://doi.org/10.1089/hgtb.2013.149.

impakt faktor: 2,436 (Q1)
független idézetek száma: 12

8. Kolacsek, O., Z. Izsvák, Z. Ivics, B. Sarkadi, and **T. I. Orbán***: Quantitative analysis of DNA transposon-mediated gene delivery: the Sleeping Beauty system as an example. in: *Genomics III - Methods, Techniques and Applications*; iConcept Press Ltd. 2014, pages: 97-123.

impakt faktor: - (*könyvfejezet*)
független idézetek száma: -

9. Szebenyi, K., A. Furedi, O. Kolacsek, E. Pergel, Z. Bosze, B. Bender, P. Vajdovich, J. Tovari, L. Homolya, G. Szakacs, L. Heja, A. Enyedi, B. Sarkadi, A. Apati, and **T. I. Orban***: Generation of a Homozygous Transgenic Rat Strain Stably Expressing a Calcium Sensor Protein for Direct Examination of Calcium Signaling. *Sci Rep* 2015, 5:12645. doi: https://doi.org/10.1038/srep12645.

impakt faktor: 5,228 (D1)
független idézetek száma: 4

10. Kolacsek, O., E. Pergel, N. Varga, A. Apati, and **T. I. Orban***: Ct shift: A novel and accurate real-time PCR quantification model for direct comparison of different nucleic acid sequences and its application for transposon quantifications. *Gene* 2017, 598:43-49. doi: https://doi.org/10.1016/j.gene.2016.10.035.

    impakt faktor: 2,498 (Q1)

    független idézetek száma: 6

11. Kolacsek, O., and **T. I. Orban***: Transcription activity of transposon sequence limits Sleeping Beauty transposition. *Gene* 2018, 676:184-188.
    doi: https://doi.org/10.1016/j.gene.2018.07.045.

    impakt faktor: 2,638 (Q1)

    független idézetek száma: 1

12. Fothi, A., O. Biro, Z. Erdei, A. Apati, and **T. I. Orban***: Tissue-specific and transcription-dependent mechanisms regulate primary microRNA processing efficiency of the human chromosome 19 MicroRNA cluster. *RNA Biol* 2021, 18(8):1170-1180.
    doi: https://doi.org/10.1080/15476286.2020.1836457.

    impakt faktor: 4,766 (Q1)

    független idézetek száma: 6

13. Kolacsek, O., G. Wachtl, A. Fothi, A. Schamberger, S. Sandor, E. Pergel, N. Varga, T. Rasko, Z. Izsvak, A. Apati, and **T. I. Orban***: Functional indications for transposase domestications - Characterization of the human piggyBac transposase derived (PGBD) activities. *Gene* 2022, 834:146609.
    doi: https://doi.org/10.1016/j.gene.2022.146609.

    impakt faktor: 3,500 (Q2)

    független idézetek száma: 3

14. Wachtl, G., E. Schad, K. Huszar, A. Palazzo, Z. Ivics, A. Tantos, and **T. I. Orban***: Functional Characterization of the N-Terminal Disordered Region of the piggyBac Transposase. *Int J Mol Sci* 2022, 23(18).
doi: https://doi.org/10.3390/ijms231810317.

impakt faktor: 5,600 (D1)
független idézetek száma: 1

15. **Orban, T. I.***: One locus, several functional RNAs-emerging roles of the mechanisms responsible for the sequence variability of microRNAs. *Biol Futur* 2023, 74(1-2):17-28. (***invited review – Special Issue: Novel Contributions to the RNA Story: Karikó's Legacy***)
doi: https://doi.org/10.1007/s42977-023-00154-7.

impakt faktor ('22): 2,100 (Q2)
független idézetek száma: 1

A DISSZERTÁCIÓHOZ KAPCSOLÓDÓ EGYÉB KÖZLEMÉNYEK

1. **Orbán, T.**, A. Apáti, and B. Sarkadi: Új remény a gén- és őssejtterápiában: a transzpozon alapú génbeviteli eljárások. *Biokémia* 2009, 33(1):49-58.
2. **Orbán, T. I.**, A. Apáti, Z. Izsvák, Z. Ivics, and B. Sarkadi: Use of Transposon-Transposase Systems for Stable Genetic Modification of Embryonic Stem Cells. in: *Methodological Advances in the Culture, Manipulation and Utilization of Embryonic Stem Cells for Basic and Practical Applications*; InTech Ed. 2011, pages: 259-274.
3. Szebenyi, K., A. Furedi, O. Kolacsek, R. Csohany, A. Prokai, K. Kis-Petik, A. Szabo, Z. Bosze, B. Bender, J. Tovari, A. Enyedi, **T. I. Orban**, A. Apati, and B. Sarkadi: Visualization of Calcium Dynamics in Kidney Proximal Tubules. *J Am Soc Nephrol* 2015, 26(11):2731-2740.
   doi: https://doi.org/10.1681/ASN.2014070705.
4. Sandor, S., T. Jordanidisz, A. Schamberger, G. Varady, Z. Erdei, A. Apati, B. Sarkadi, and **T. I. Orban***: Functional characterization of the ABCG2 5' non-coding exon variants: Stem cell specificity, translation efficiency and the influence of drug selection. *Biochim Biophys Acta* 2016, 1859(7):943-951.
   doi: https://doi.org/10.1016/j.bbagrm.2016.05.007.
5. Szabo, Z., L. Heja, G. Szalay, O. Kekesi, A. Furedi, K. Szebenyi, A. Dobolyi, **T. I. Orban**, O. Kolacsek, T. Tompa, Z. Miskolczy, L. Biczok, B. Rozsa, B. Sarkadi, and J. Kardos: Extensive astrocyte synchronization advances neuronal coupling in slow wave activity in vivo. *Sci Rep* 2017, 7(1):6018.
   doi: https://doi.org/10.1038/s41598-017-06073-7.
6. Jemnitz, K., A. Batai-Konczos, M. Szabo, E. Ioja, O. Kolacsek, **T. I. Orban**, G. Torok, L. Homolya, E. Kovacs, I. Jablonkai, and Z. Veres: A transgenic rat hepatocyte - Kupffer cell co-culture model for evaluation of direct and macrophage-related effect of poly(amidoamine) dendrimers. *Toxicol In Vitro* 2017, 38:159-169.
   doi: https://doi.org/10.1016/j.tiv.2016.09.016.

7. Szadeczky-Kardoss, I., T. Csorba, A. Auber, A. Schamberger, T. Nyiko, J. Taller, **T. I. Orban**, J. Burgyan, and D. Silhavy: The nonstop decay and the RNA silencing systems operate cooperatively in plants. *Nucleic Acids Res* 2018, 46(9):4632-4648. doi: https://doi.org/10.1093/nar/gky279.

8. Biro, O., A. Fothi, B. Alasztics, B. Nagy, **T. I. Orban***, and J. Rigo, Jr.: Circulating exosomal and Argonaute-bound microRNAs in preeclampsia. *Gene* 2019, 692:138-144. doi: https://doi.org/10.1016/j.gene.2019.01.012.

9. Ree, D., A. Borsy, A. Fothi, **T. I. Orban**, G. Varady, Z. Erdei, B. Sarkadi, J. Rethelyi, N. Varga, and A. Apati: Establishing a human embryonic stem cell clone with a heterozygous mutation in the DGCR8 gene. *Stem Cell Res* 2020, 50:102134. doi: https://doi.org/10.1016/j.scr.2020.102134.

10. Schamberger, A., G. Varady, A. Fothi, and **T. I. Orban***: Posttranscriptional Regulation of the Human ABCG2 Multidrug Transporter Protein by Artificial Mirtrons. *Genes (Basel)* 2021, 12(7):1068. doi: https://doi.org/10.3390/genes12071068.

11. Ree, D., A. Fothi, N. Varga, O. Kolacsek, **T. I. Orban***, and A. Apati: Partial Disturbance of Microprocessor Function in Human Stem Cells Carrying a Heterozygous Mutation in the DGCR8 Gene. *Genes (Basel)* 2022, 13(11):1925. (***cover page story***) doi: https://doi.org/10.3390/genes13111925.

12. Rasko, T., A. Pande, K. Radscheit, A. Zink, M. Singh, C. Sommer, G. Wachtl, O. Kolacsek, G. Inak, A. Szvetnik, S. Petrakis, M. Bunse, V. Bansal, M. Selbach, **T. I. Orban**, A. Prigione, L. D. Hurst, and Z. Izsvak: A Novel Gene Controls a New Structure: PiggyBac Transposable Element-Derived 1, Unique to Mammals, Controls Mammal-Specific Neuronal Paraspeckles. *Mol Biol Evol* 2022, 39(10):msac175. doi: https://doi.org/10.1093/molbev/msac175.

13. Gal, L., A. Fothi, G. Orosz, S. Nagy, N. G. Than, and **T. I. Orban***: Exosomal small RNA profiling in first-trimester maternal blood explores early molecular pathways of preterm preeclampsia. *Front Immunol* 2024, 15:1321191. doi: https://doi.org/10.3389/fimmu.2024.1321191.

A DISSZERTÁCIÓHOZ KAPCSOLÓDÓ SZABADALMI BEADVÁNYOK

1. **Tamás Orbán**, Ágota Apáti, Balázs Sarkadi, Zsuzsanna Izsvák, Katalin Német (*inventors*): Genetically modified stem cells and methods for identifying tissues differentiated therefrom. 2008. October 15th.
   nemzetközi bejelentés száma: PCT/IB2008/054238
   európai regionális fázis alapszáma: EP 08839475.4 (megadott – 2014. november 5.)
   USA nemzeti fázis alapszáma: US 12/780315
2. Ágota Apáti, Kornélia Szebényi, Orsolya Kolacsek, András Füredi, **Tamás Orbán**, Balázs Sarkadi (*inventors*): A transgenic laboratory rat strain. 2014. October 21st.
   hazai bejelentés száma: P1400500
3. Kornélia Szebényi, **Tamás Orbán**, Balázs Sarkadi, Ágota Apáti (*inventors*): Generation of Cardiac cells from Cardiac Progenitors in Culture. 2015. May 15th.
   hazai bejelentés száma: P1500235

KÖSZÖNETNYILVÁNÍTÁS

Ennek a disszertációnak a megírása szakmai és érzelmi szempontból is felkavaró volt, hiszen szembesülnöm kellett a PhD fokozatom megszerzése (2003) óta eltelt két évtized emlékeivel, tapasztalataival. A közlemények és a tudományos eredmények újbóli áttekintése egyfelől óriási élményt jelentett, hiszen újra felidézhettem a kísérletezés okozta intenzív érzelmi „hullámvasutazást”, és újra átélhettem egy-egy, számomra oly kedves tudományos felfedezés örömét is. Szerencsés vagyok, hiszen sok izgalmas projektben vehettem részt, ugyanakkor komoly kihívást jelentett kiválogatni azokat az eredményeket, amelyek egy nehéz szívvel történő szelekció után végül egy koherens tudományos történetté, egy befogadható, de egyben talán olvasmányos dolgozattá formálódtak. A modern kori kutatások, a biológia területén különösképpen, ma már elképzelhetetlenek jó csapatmunka, és az ehhez tartozó kiváló mentorok nélkül. Ennek megfelelően én is számtalan embernek tartozom köszönettel azért, hogy eljuthattam idáig.

A posztdoktori kutatásaim meghatározó mentora volt Elisa Izaurralde, akinek a németországi laboratóriumában, a heidelbergi EMBL-ben eltöltött évek alatt, a sok nehézséggel is fűszerezett rengeteg tapasztalat kapcsán eldöntöttem, hogy a tudományos pályán szeretnék maradni. Sajnos ő pár éve már nincs közöttünk, így ennek az értekezésnek a megszületését sem érhette meg.

Tudományos karrierem és szakmai munkám során legfontosabb mentoromnak Sarkadi Balázst tartom, aki a posztdoktori kutatásaim második, magyarországi fázisától kezdődően a mai napig egyengette az utamat, és akivel a tudományos problémákat és az élet nagy kérdéseit egyaránt meg lehet vitatni. És bár mindvégig támogatott, és pontosan ismeri a kutatási projektjeinket, nekem az az érzésem, hogy a mai napig nem hisz (teljesen) az RNS-ekben…

Rengeteg köszönettel tartozom volt és jelenlegi hallgatóimnak, akik komolyan hozzájárultak tudományos sikereimhez, és saját laborvezetői képességeim csiszolásához, javításához. Hálás vagyok ezért Schamberger Anitának, Kolacsek Orsolyának, Sándor Sárának, Fóthi Ábelnek, Wachtl Gerdának és Gál Lucának (időrendben), hogy PhD hallgatóként meghatározó formálói voltak az általam vezetett labornak. Köszönöm továbbá Jordanidisz Theodórának, Pergel Enikőnek, Bolyácz Mártonnak, Tibori Kingának, Kósa Juditnak, Kerekes Péternek, Balázs Sámuelnek és Lávay Petrának, hogy MSc vagy BSc hallgatóként szintén hozzájárultak az általam vezetett szakmai munkához. Egy laboratórium működése elképzelhetetlen jó asszisztens nélkül – különösen igaz ez

Némethy Kornéliára, aki fáradhatatlan lelkesedéssel és megfelelő szigorral képes kordában tartani a kísérletek lázában gyakorta elszabaduló kutatói zűrzavart. Köszönöm neki, hogy lassan egy évtizede segít megteremteni azt a stabil laborhátteret, amely nélkül a kutatásaink elképzelhetetlenek lennének. Köszönöm korábbi asszisztenseimnek, Sebestyén Zsuzsannának és Nagy Zsuzsannának, hogy odaadóan segítették a munkánkat, amikor a kezdetekben fel kellett építeni és ki kellett alakítani az akkor formálódó molekuláris biológiai laboratóriumot.

Óriási hálával tartozom „szakmai szobatársaimnak", Apáti Ágotának és Szakács Gergelynek, hogy tudományos éleslátásukkal, kezdeményező- és segítőkészségükkel hozzájárultak közös szakmai sikereinkhez, humorukkal pedig a vidám mindennapok hangulatához. Az elmúlt évek meghatározó szakmai pezsgését köszönhetem a Sarkadi Balázs és Váradi András által létrehozott „Biomembrán" közösségnek, ahol a két alapító mellett rengeteg szakmai és baráti kapcsolat is szövődött: közülük kiemelném Homolya Lászlót, Német Katalint, Hegedűs Tamást és Welker Ervint, akikkel rengeteg vidám szakmai időt tölthettem együtt. Köszönöm a szakmai együttműködéseket Apáti Ágota laborjából Szebényi Kornéliának, Erdei Zsuzsának és Varga Nórának, Szakács Gergely laborjából Füredi Andrásnak, illetve Welker Ervin laborjából Huszár Krisztinának. A munkánkhoz tartozó adminisztratív feladatok és a mindennapi szervezési munkák átvállalása óriási segítséget jelent, amiért őszintén hálás vagyok elsősorban Rieth Katalinnak és Lesti Juditnak, illetve Mohos Krisztinának. Köszönöm Kardos Juliannának, Buday Lászlónak és Hunyady Lászlónak, hogy intézet igazgatóként láttak bennem fantáziát, és csoportvezetőként mindig támogatták a munkámat.

Rengeteg köszönet illeti szakmai együttműködő partnereimet, akik nélkül szintén elképzelhetetlen lett volna ezen izgalmas tudományos eredmények elérése: Izsvák Zsuzsanna és Ivics Zoltán, valamint Mátés Lajos, Miskey Csaba és Krízsik Virág a transzpozonos projektekben voltak segítségemre; Silhavy Dániellel, Várallyay Évával és Havelda Zoltánnal a kisRNS-ekhez kapcsolódó projektekben, míg Bender Balázzsal, Hoffmann Orsolyával, Gócza Elennel és Héja Lászlóval a transzgenikus patkánymodell kapcsán dolgoztunk együtt. Hálás vagyok Biró Orsolyának, Rigó Jánosnak és Than Gábor Nándornak a terhességi kórképekhez kapcsolódó RNS interferencia kutatásokhoz nyújtott segítségükért; Vellai Tibornak és Vellai-Takács Krisztinának a közös kutatásokért és a szakmai beszélgetésekért; Tantos Ágnesnek és Schád Évának pedig, hogy megnyitották előttünk a rendezetlen fehérjék izgalmas világát. Köszönöm a céges partnereinknek a sok támogatást: Hirka Gábornak a Toxi-Coop Zrt-től, Prém Péternének

az Akadimpex Zrt-től, és Kacskovics Imrének az ImmunoGenes Zrt-től. Ezúton mondok köszönetet minden szervezetnek, amelyek pénzügyileg támogatták a kutatásaimat (Magyar Tudományos Akadémia, OTKA, NKFIH, RNA Society, Eötvös Loránd Kutatási Hálózat).

A szakmai munka hátterében mindig megbújik egy támogató és önzetlen család: nagy szeretettel és hálával tartozom édesanyámnak, feleségemnek és három gyermekemnek, valamint testvéremnek, akik mindig elfogadták, sőt támogatták a kutatásaimmal járó nélkülözéseket. Azoknak a családtagoknak is köszönöm, akik a mű elkészültét sajnos már nem érhették meg.

És végezetül persze külön köszönet illeti a „nem-kódoló” RNS-eket és a transzpozonokat, amelyek nélkül sem ez a mű, sem ennek írója, de még az olvasói sem jöhettek volna létre…

IRODALOMJEGYZÉK

Almeida, M. V., G. Vernaz, A. L. K. Putman, and E. A. Miska: **Taming transposable elements in vertebrates: from epigenetic silencing to domestication**. *Trends Genet* 2022, 38(6):529-553.

Altuvia, Y., P. Landgraf, G. Lithwick, N. Elefant, S. Pfeffer, A. Aravin, M. J. Brownstein, T. Tuschl, and H. Margalit: **Clustering and conservation patterns of human microRNAs**. *Nucleic Acids Res* 2005, 33(8):2697-2706.

Alzohairy, A. M., G. Gyulai, R. K. Jansen, and A. Bahieldin: **Transposable elements domesticated and neofunctionalized by eukaryotic genomes**. *Plasmid* 2013, 69(1):1-15.

Apati, A., T. I. Orban, N. Varga, A. Nemeth, A. Schamberger, V. Krizsik, B. Erdelyi-Belle, L. Homolya, G. Varady, R. Padanyi, E. Karaszi, E. W. Kemna, K. Nemet, and B. Sarkadi: **High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells**. *Biochim Biophys Acta* 2008, 1778(12):2700-2709.

Arora, S., R. Rana, A. Chhabra, A. Jaiswal, and V. Rani: **miRNA-transcription factor interactions: a combinatorial regulation of gene expression**. *Mol Genet Genomics* 2013, 288(3-4):77-87.

Babiarz, J. E., J. G. Ruby, Y. Wang, D. P. Bartel, and R. Blelloch: **Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs**. *Genes Dev* 2008, 22(20):2773-2785.

Baer, M., T. W. Nilsen, C. Costigan, and S. Altman: **Structure and transcription of a human gene for H1 RNA, the RNA component of human RNase P**. *Nucleic Acids Res* 1990, 18(1):97-103.

Baertsch, R., M. Diekhans, W. J. Kent, D. Haussler, and J. Brosius: **Retrocopy contributions to the evolution of the human genome**. *BMC Genomics* 2008, 9:466.

Bailey, A. D., L. T. Gray, T. Pavelitz, J. C. Newman, K. Horibata, K. Tanaka, and A. M. Weiner: **The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells**. *DNA Repair (Amst)* 2012, 11(5):488-501.

Balciunas, D., K. J. Wangensteen, A. Wilber, J. Bell, A. Geurts, S. Sivasubbu, X. Wang, P. B. Hackett, D. A. Largaespada, R. S. McIvor, and S. C. Ekker: **Harnessing a high cargo-capacity transposon for genetic applications in vertebrates**. *PLoS Genet* 2006, 2(11):e169.

Baldari, C. T., and F. Amaldi: **DNA reassociation kinetics in relation to genome size in four amphibian species**. *Chromosoma* 1976, 59(1):13-22.

Bar, M., S. K. Wyman, B. R. Fritz, J. Qi, K. S. Garg, R. K. Parkin, E. M. Kroh, A. Bendoraite, P. S. Mitchell, A. M. Nelson, W. L. Ruzzo, C. Ware, J. P. Radich, R. Gentleman, H. Ruohola-Baker, and M. Tewari: **MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries**. *Stem Cells* 2008, 26(10):2496-2505.

Baralle, F. E., and J. Giudice: **Alternative splicing as a regulator of development and tissue identity**. *Nat Rev Mol Cell Biol* 2017, 18(7):437-451.

Barton, N. H.: **The "New Synthesis"**. *Proc Natl Acad Sci U S A* 2022, 119(30):e2122147119.

Baus, J., L. Liu, A. D. Heggestad, S. Sanz, and B. S. Fletcher: **Hyperactive transposase mutants of the Sleeping Beauty transposon**. *Mol Ther* 2005, 12(6):1148-1156.

Beadle, G. W., and E. L. Tatum: **Genetic Control of Biochemical Reactions in Neurospora**. *Proc Natl Acad Sci U S A* 1941, 27(11):499-506.

Beckermann, T. M., W. Luo, C. M. Wilson, R. A. Veach, and M. H. Wilson: **Cognate restriction of transposition by piggyBac-like proteins**. *Nucleic Acids Res* 2021, 49(14):8135-8144.

Berezikov, E., W. J. Chung, J. Willis, E. Cuppen, and E. C. Lai: **Mammalian mirtron genes**. *Mol Cell* 2007, 28(2):328-336.

Berget, S. M., C. Moore, and P. A. Sharp: **Spliced segments at the 5' terminus of adenovirus 2 late mRNA**. *Proc Natl Acad Sci U S A* 1977, 74(8):3171-3175.

Bessereau, J. L.: **Transposons in C. elegans**. *WormBook* 2006:1-13.

Bian, Q., and A. S. Belmont: **BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression**. *Nucleic Acids Res* 2010, 38(11):e127.

Biemont, C., and C. Vieira: **Genetics: junk DNA as an evolutionary force**. *Nature* 2006, 443(7111):521-524.

Biro, O., A. Fothi, B. Alasztics, B. Nagy, T. I. Orban, and J. Rigo, Jr.: **Circulating exosomal and Argonaute-bound microRNAs in preeclampsia**. *Gene* 2019, 692:138-144.

Bofill-De Ros, X., A. Yang, and S. Gu: **IsomiRs: Expanding the miRNA repression toolbox beyond the seed**. *Biochim Biophys Acta Gene Regul Mech* 2020, 1863(4):194373.

Bortolin-Cavaille, M. L., M. Dance, M. Weber, and J. Cavaille: **C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts**. *Nucleic Acids Res* 2009, 37(10):3464-3473.

Bouallegue, M., J. D. Rouault, A. Hua-Van, M. Makni, and P. Capy: **Molecular Evolution of piggyBac Superfamily: From Selfishness to Domestication**. *Genome Biol Evol* 2017, 9(2):323-339.

Burns, K. H.: **Our Conflict with Transposable Elements and Its Implications for Human Disease**. *Annu Rev Pathol* 2020, 15:51-70.

Burroughs, A. M., Y. Ando, M. J. de Hoon, Y. Tomaru, H. Suzuki, Y. Hayashizaki, and C. O. Daub: **Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin**. *RNA Biol* 2011, 8(1):158-177.

Cadinanos, J., and A. Bradley: **Generation of an inducible and optimized piggyBac transposon system**. *Nucleic Acids Res* 2007, 35(12):e87.

Cary, L. C., M. Goebel, B. G. Corsaro, H. G. Wang, E. Rosen, and M. J. Fraser: **Transposon mutagenesis of baculoviruses: analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses**. *Virology* 1989, 172(1):156-169.

Castro, J. P., and C. M. Carareto: **Drosophila melanogaster P transposable elements: mechanisms of transposition and regulation**. *Genetica* 2004, 121(2):107-118.

Cheetham, S. W., G. J. Faulkner, and M. E. Dinger: **Overcoming challenges and dogmas to understand the functions of pseudogenes**. *Nat Rev Genet* 2020, 21(3):191-201.

Chen, C. Y., and R. J. Schwartz: **Recruitment of the tinman homolog Nkx-2.5 by serum response factor activates cardiac alpha-actin gene transcription**. *Mol Cell Biol* 1996, 16(11):6372-6384.

Chen, K., and N. Rajewsky: **The evolution of gene regulation by transcription factors and microRNAs**. *Nat Rev Genet* 2007, 8(2):93-103.

Chen, Q., W. Luo, R. A. Veach, A. B. Hickman, M. H. Wilson, and F. Dyda: **Structural basis of seamless excision and specific targeting by piggyBac transposase**. *Nat Commun* 2020, 11(1):3446.

Chow, L. T., R. E. Gelinas, T. R. Broker, and R. J. Roberts: **An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA**. *Cell* 1977, 12(1):1-8.

Christoforou, N., R. A. Miller, C. M. Hill, C. C. Jie, A. S. McCallion, and J. D. Gearhart: **Mouse ES cell-derived cardiac precursor cells are multipotent and facilitate identification of novel cardiac genes**. *J Clin Invest* 2008, 118(3):894-903.

Chung, S., T. Andersson, K. C. Sonntag, L. Bjorklund, O. Isacson, and K. S. Kim: **Analysis of different promoter systems for efficient transgene expression in mouse embryonic stem cell lines**. *Stem Cells* 2002, 20(2):139-145.

Chuong, E. B., N. C. Elde, and C. Feschotte: **Regulatory activities of transposable elements: from conflicts to benefits**. *Nat Rev Genet* 2017, 18(2):71-86.

Claeys Bouuaert, C., K. Lipkow, S. S. Andrews, D. Liu, and R. Chalmers: **The autoregulation of a eukaryotic DNA transposon**. *Elife* 2013, 2:e00668.

Consortium, E. P.: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, 489(7414):57-74.

Cordaux, R., S. Udit, M. A. Batzer, and C. Feschotte: **Birth of a chimeric primate gene by capture of the transposase gene from a mobile element**. *Proc Natl Acad Sci U S A* 2006, 103(21):8101-8106.

Cosby, R. L., N. C. Chang, and C. Feschotte: **Host-transposon interactions: conflict, cooperation, and cooption**. *Genes Dev* 2019, 33(17-18):1098-1116.

Cosby, R. L., J. Judd, R. Zhang, A. Zhong, N. Garry, E. J. Pritham, and C. Feschotte: **Recurrent evolution of vertebrate transcription factors by transposase capture**. *Science* 2021, 371(6531).

Coufal, N. G., J. L. Garcia-Perez, G. E. Peng, G. W. Yeo, Y. Mu, M. T. Lovci, M. Morell, K. S. O'Shea, J. V. Moran, and F. H. Gage: **L1 retrotransposition in human neural progenitor cells**. *Nature* 2009, 460(7259):1127-1131.

Crain, W. R., E. H. Davidson, and R. J. Britten: **Contrasting patterns of DNA sequence arrangement in Apis mellifera (honeybee) and Musca domestica (housefly)**. *Chromosoma* 1976, 59(1):1-12.

Darnell, J. E., Jr.: **Reflections on the history of pre-mRNA processing and highlights of current knowledge: a unified picture**. *RNA* 2013, 19(4):443-460.

Dawkins, R. 2021. Az önző gén. Kossuth Kiadó, Budapest.

De Cecco, M., T. Ito, A. P. Petrashen, A. E. Elias, N. J. Skvir, S. W. Criscione, A. Caligiana, G. Brocculi, E. M. Adney, J. D. Boeke, O. Le, C. Beausejour, J. Ambati, K. Ambati, M. Simon, A. Seluanov, V. Gorbunova, P. E. Slagboom, S. L. Helfand, N. Neretti, *et al.*: **L1 drives IFN in senescent cells and promotes age-associated inflammation**. *Nature* 2019, 566(7742):73-78.

Devon, R. S., D. J. Porteous, and A. J. Brookes: **Splinkerettes--improved vectorettes for greater efficiency in PCR walking**. *Nucleic Acids Res* 1995, 23(9):1644-1645.

Ding, S., X. Wu, G. Li, M. Han, Y. Zhuang, and T. Xu: **Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice**. *Cell* 2005, 122(3):473-483.

Djebali, S., C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, *et al.*: **Landscape of transcription in human cells**. *Nature* 2012, 489(7414):101-108.

Dykxhoorn, D. M., C. D. Novina, and P. A. Sharp: **Killing the messenger: short RNAs that silence gene expression**. *Nat Rev Mol Cell Biol* 2003, 4(6):457-467.

Ebhardt, H. A., H. H. Tsang, D. C. Dai, Y. Liu, B. Bostan, and R. P. Fahlman: **Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications**. *Nucleic Acids Res* 2009, 37(8):2461-2470.

Elbashir, S. M., W. Lendeckel, and T. Tuschl: **RNA interference is mediated by 21- and 22-nucleotide RNAs**. *Genes Dev* 2001, 15(2):188-200.

Erdos, G., and Z. Dosztanyi: **Analyzing Protein Disorder with IUPred2A**. *Curr Protoc Bioinformatics* 2020, 70(1):e99.

Feschotte, C., and E. J. Pritham: **DNA transposons and the evolution of eukaryotic genomes**. *Annu Rev Genet* 2007, 41:331-368.

Fire, A., S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello: **Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans**. *Nature* 1998, 391(6669):806-811.

Firth, N., and R. A. Skurray: **Mobile elements in the evolution and spread of multiple-drug resistance in staphylococci**. *Drug Resist Updat* 1998, 1(1):49-58.

Flavell, R. B., M. D. Bennett, J. B. Smith, and D. B. Smith: **Genome size and the proportion of repeated nucleotide sequence DNA in plants**. *Biochem Genet* 1974, 12(4):257-269.

Flor, I., and J. Bullerdiek: **The dark side of a success story: microRNAs of the C19MC cluster in human tumours**. *J Pathol* 2012, 227(3):270-274.

Fothi, A., O. Biro, Z. Erdei, A. Apati, and T. I. Orban: **Tissue-specific and transcription-dependent mechanisms regulate primary microRNA processing efficiency of the human chromosome 19 MicroRNA cluster**. *RNA Biol* 2021, 18(8):1170-1180.

Franco, S., R. Pluvinet, J. F. Sanchez-Herrero, L. Sumoy, and M. A. Martinez: **Rapid and accurate quantification of isomiRs by RT-qPCR**. *Sci Rep* 2022, 12(1):17220.

Fraser, M. J., G. E. Smith, and M. D. Summers: **Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses**. *J Virol* 1983, 47(2):287-300.

Gates, A. J., D. M. Gysi, M. Kellis, and A. L. Barabasi: **A wealth of discovery built on the Human Genome Project - by the numbers**. *Nature* 2021, 590(7845):212-215.

Gebert, L. F. R., and I. J. MacRae: **Regulation of microRNA function in animals**. *Nat Rev Mol Cell Biol* 2019, 20(1):21-37.

Ghanim, G. E., D. C. Rio, and F. K. Teixeira: **Mechanism and regulation of P element transposition**. *Open Biol* 2020, 10(12):200244.

Ghildiyal, M., and P. D. Zamore: **Small silencing RNAs: an expanding universe**. *Nat Rev Genet* 2009, 10(2):94-108.

Grabundzija, I., M. Irgang, L. Mates, E. Belay, J. Matrai, A. Gogol-Doring, K. Kawakami, W. Chen, P. Ruiz, M. K. Chuah, T. VandenDriessche, Z. Izsvak, and Z. Ivics: **Comparative analysis of transposable element vector systems in human cells**. *Mol Ther* 2010, 18(6):1200-1209.

Graur, D., Y. Zheng, N. Price, R. B. Azevedo, R. A. Zufall, and E. Elhaik: **On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE**. *Genome Biol Evol* 2013, 5(3):578-590.

Gray, L. T., K. K. Fong, T. Pavelitz, and A. M. Weiner: **Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans**. *PLoS Genet* 2012, 8(9):e1002972.

Grimson, A., M. Srivastava, B. Fahey, B. J. Woodcroft, H. R. Chiang, N. King, B. M. Degnan, D. S. Rokhsar, and D. P. Bartel: **Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals**. *Nature* 2008, 455(7217):1193-1197.

Gu, S., L. Jin, Y. Zhang, Y. Huang, F. Zhang, P. N. Valdmanis, and M. A. Kay: **The loop position of shRNAs and pre-miRNAs is critical for the accuracy of dicer processing in vivo**. *Cell* 2012, 151(4):900-911.

Guerineau, M., L. Bessa, S. Moriau, E. Lescop, F. Bontems, N. Mathy, E. Guittet, J. Bischerour, M. Betermier, and N. Morellet: **The unusual structure of the PiggyMac cysteine-rich domain reveals zinc finger diversity in PiggyBac-related transposases**. *Mob DNA* 2021, 12(1):12.

Hancks, D. C., and H. H. Kazazian, Jr.: **Roles for retrotransposon insertions in human disease**. *Mob DNA* 2016, 7:9.

Harmer, C. J., C. H. Pong, and R. M. Hall: **Structures bounded by directly-oriented members of the IS26 family are pseudo-compound transposons**. *Plasmid* 2020, 111:102530.

Hartl, D. L., E. R. Lozovskaya, D. I. Nurminsky, and A. R. Lohe: **What restricts the activity of mariner-like transposable elements**. *Trends Genet* 1997, 13(5):197-201.

Helou, L., L. Beauclair, H. Dardente, P. Arensburger, N. Buisine, Y. Jaszczyszyn, F. Guillou, T. Lecomte, A. Kentsis, and Y. Bigot: **The C-terminal Domain of piggyBac Transposase Is Not Required for DNA Transposition**. *J Mol Biol* 2021a, 433(7):166805.

Helou, L., L. Beauclair, H. Dardente, B. Piegu, L. Tsakou-Ngouafo, T. Lecomte, A. Kentsis, P. Pontarotti, and Y. Bigot: **The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements Tcr-pble and Ifp2**. *J Mol Biol* 2021b, 433(7):166839.

Henssen, A. G., E. Henaff, E. Jiang, A. R. Eisenberg, J. R. Carson, C. M. Villasante, M. Ray, E. Still, M. Burns, J. Gandara, C. Feschotte, C. E. Mason, and A. Kentsis: **Genomic DNA transposition induced by human PGBD5**. *Elife* 2015, 4.

Henssen, A. G., E. Jiang, J. Zhuang, L. Pinello, N. D. Socci, R. Koche, M. Gonen, C. M. Villasante, S. A. Armstrong, D. E. Bauer, Z. Weng, and A. Kentsis: **Forward genetic screen of human transposase genomic rearrangements**. *BMC Genomics* 2016, 17:548.

Henssen, A. G., R. Koche, J. Zhuang, E. Jiang, C. Reed, A. Eisenberg, E. Still, I. C. MacArthur, E. Rodriguez-Fos, S. Gonzalez, M. Puiggros, A. N. Blackford, C. E. Mason, E. de Stanchina, M.

Gonen, A. K. Emde, M. Shah, K. Arora, C. Reeves, N. D. Socci, *et al.*: **PGBD5 promotes site-specific oncogenic mutations in human tumors**. *Nat Genet* 2017a, 49(7):1005-1014.

Henssen, A. G., C. Reed, E. Jiang, H. D. Garcia, J. von Stebut, I. C. MacArthur, P. Hundsdoerfer, J. H. Kim, E. de Stanchina, Y. Kuwahara, H. Hosoi, N. J. Ganem, F. Dela Cruz, A. L. Kung, J. H. Schulte, J. H. Petrini, and A. Kentsis: **Therapeutic targeting of PGBD5-induced DNA repair dependency in pediatric solid tumors**. *Sci Transl Med* 2017b, 9(414).

Heo, I., M. Ha, J. Lim, M. J. Yoon, J. E. Park, S. C. Kwon, H. Chang, and V. N. Kim: **Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs**. *Cell* 2012, 151(3):521-532.

Heo, I., C. Joo, Y. K. Kim, M. Ha, M. J. Yoon, J. Cho, K. H. Yeom, J. Han, and V. N. Kim: **TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation**. *Cell* 2009, 138(4):696-708.

Hickman, A. B., M. Chandler, and F. Dyda: **Integrating prokaryotes and eukaryotes: DNA transposases in light of structure**. *Crit Rev Biochem Mol Biol* 2010, 45(1):50-69.

Holland, P. M., R. D. Abramson, R. Watson, and D. H. Gelfand: **Detection of specific polymerase chain reaction product by utilizing the 5'----3' exonuclease activity of Thermus aquaticus DNA polymerase**. *Proc Natl Acad Sci U S A* 1991, 88(16):7276-7280.

Honda, S., and Y. Kirino: **Dumbbell-PCR: a method to quantify specific small RNA variants with a single nucleotide resolution at terminal sequences**. *Nucleic Acids Res* 2015, 43(12):e77.

Huang, X., H. Guo, S. Tammana, Y. C. Jung, E. Mellgren, P. Bassi, Q. Cao, Z. J. Tu, Y. C. Kim, S. C. Ekker, X. Wu, S. M. Wang, and X. Zhou: **Gene transfer efficiency and genome-wide integration profiling of Sleeping Beauty, Tol2, and piggyBac transposons in human primary T cells**. *Mol Ther* 2010, 18(10):1803-1813.

Hudecek, M., Z. Izsvak, S. Johnen, M. Renner, G. Thumann, and Z. Ivics: **Going non-viral: the Sleeping Beauty transposon system breaks on through to the clinical side**. *Crit Rev Biochem Mol Biol* 2017, 52(4):355-380.

Huntzinger, E., and E. Izaurralde: **Gene silencing by microRNAs: contributions of translational repression and mRNA decay**. *Nat Rev Genet* 2011, 12(2):99-110.

Hurst, G. D., and J. H. Werren: **The role of selfish genetic elements in eukaryotic evolution**. *Nat Rev Genet* 2001, 2(8):597-606.

Hurst, L. D.: **The Ka/Ks ratio: diagnosing the form of sequence evolution**. *Trends Genet* 2002, 18(9):486.

Hutvagner, G.: **Small RNA asymmetry in RNAi: function in RISC assembly and gene regulation**. *FEBS Lett* 2005, 579(26):5850-5857.

Ivics, Z.: **Endogenous Transposase Source in Human Cells Mobilizes piggyBac Transposons**. *Mol Ther* 2016, 24(5):851-854.

Ivics, Z., P. B. Hackett, R. H. Plasterk, and Z. Izsvak: **Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells**. *Cell* 1997, 91(4):501-510.

Ivics, Z., and Z. Izsvak: **Transposons for gene therapy!** *Curr Gene Ther* 2006, 6(5):593-607.

Jangam, D., C. Feschotte, and E. Betran: **Transposable Element Domestication As an Adaptation to Evolutionary Conflicts**. *Trends Genet* 2017, 33(11):817-831.

Jemnitz, K., A. Batai-Konczos, M. Szabo, E. Ioja, O. Kolacsek, T. I. Orban, G. Torok, L. Homolya, E. Kovacs, I. Jablonkai, and Z. Veres: **A transgenic rat hepatocyte - Kupffer cell co-culture model for evaluation of direct and macrophage-related effect of poly(amidoamine) dendrimers**. *Toxicol In Vitro* 2017, 38:159-169.

Johnson, J. M., S. Edwards, D. Shoemaker, and E. E. Schadt: **Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments**. *Trends Genet* 2005, 21(2):93-102.

Jones, M. R., M. T. Blahna, E. Kozlowski, K. Y. Matsuura, J. D. Ferrari, S. A. Morris, J. T. Powers, G. Q. Daley, L. J. Quinton, and J. P. Mizgerd: **Zcchc11 uridylates mature miRNAs to enhance neonatal IGF-1 expression, growth, and survival**. *PLoS Genet* 2012, 8(11):e1003105.

Kapitonov, V. V., and J. Jurka: **RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons**. *PLoS Biol* 2005, 3(6):e181.

Kazazian, H. H., Jr.: **Mobile elements: drivers of genome evolution**. *Science* 2004, 303(5664):1626-1632.

Keith, J. H., T. S. Fraser, and M. J. Fraser, Jr.: **Analysis of the piggyBac transposase reveals a functional nuclear targeting signal in the 94 c-terminal residues**. *BMC Mol Biol* 2008a, 9:72.

Keith, J. H., C. A. Schaeper, T. S. Fraser, and M. J. Fraser, Jr.: **Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the piggyBac transposase**. *BMC Mol Biol* 2008b, 9:73.

Keynes, M., and T. M. Cox: **William Bateson, the rediscoverer of Mendel**. *J R Soc Med* 2008, 101(3):104.

Khvorova, A., A. Reynolds, and S. D. Jayasena: **Functional siRNAs and miRNAs exhibit strand bias**. *Cell* 2003, 115(2):209-216.

Kim, H., J. Kim, S. Yu, Y. Y. Lee, J. Park, R. J. Choi, S. J. Yoon, S. G. Kang, and V. N. Kim: **A Mechanism for microRNA Arm Switching Regulated by Uridylation**. *Mol Cell* 2020, 78(6):1224-1236 e1225.

Kim, Y. K., and V. N. Kim: **Processing of intronic microRNAs**. *Embo J* 2007, 26(3):775-783.

Kim, Y. K., J. Yu, T. S. Han, S. Y. Park, B. Namkoong, D. H. Kim, K. Hur, M. W. Yoo, H. J. Lee, H. K. Yang, and V. N. Kim: **Functional links between clustered microRNAs: suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer**. *Nucleic Acids Res* 2009, 37(5):1672-1681.

Kimura, K., A. Wakamatsu, Y. Suzuki, T. Ota, T. Nishikawa, R. Yamashita, J. Yamamoto, M. Sekine, K. Tsuritani, H. Wakaguri, S. Ishii, T. Sugiyama, K. Saito, Y. Isono, R. Irie, N. Kushida, T. Yoneyama, R. Otsuka, K. Kanda, T. Yokoi, *et al.*: **Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes**. *Genome Res* 2006, 16(1):55-65.

Kleckner, N.: **Transposable elements in prokaryotes**. *Annu Rev Genet* 1981, 15:341-404.

Kock, K. H., K. W. Kong, S. Hoon, and Y. Seow: **Functional VEGFA knockdown with artificial 3'-tailed mirtrons defined by 5' splice site and branch point**. *Nucleic Acids Res* 2015, 43(13):6568-6578.

Koga, A., and H. Hori: **The Tol2 transposable element of the medaka fish: an active DNA-based element naturally occurring in a vertebrate genome**. *Genes Genet Syst* 2001, 76(1):1-8.

Kolacsek, O., Z. Erdei, A. Apati, S. Sandor, Z. Izsvak, Z. Ivics, B. Sarkadi, and T. I. Orban: **Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons**. *Hum Gene Ther Methods* 2014a, 25(4):241-252.

Kolacsek, O., Z. Izsvák, Z. Ivics, B. Sarkadi, and T. I. Orbán: **Quantitative analysis of DNA transposon-mediated gene delivery: the Sleeping Beauty system as an example**. in: *Genomics III - Methods, Techniques and Applications; iConcept Press Ltd.* 2014b, pages: 97-123.

Kolacsek, O., V. Krizsik, A. Schamberger, Z. Erdei, A. Apati, G. Varady, L. Mates, Z. Izsvak, Z. Ivics, B. Sarkadi, and T. I. Orban: **Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers**. *Mob DNA* 2011, 2(1):5.

Kolacsek, O., and T. I. Orban: **Transcription activity of transposon sequence limits Sleeping Beauty transposition**. *Gene* 2018, 676:184-188.

Kolacsek, O., E. Pergel, N. Varga, A. Apati, and T. I. Orban: **Ct shift: A novel and accurate real-time PCR quantification model for direct comparison of different nucleic acid sequences and its application for transposon quantifications**. *Gene* 2017, 598:43-49.

Kolacsek, O., G. Wachtl, A. Fothi, A. Schamberger, S. Sandor, E. Pergel, N. Varga, T. Rasko, Z. Izsvak, A. Apati, and T. I. Orban: **Functional indications for transposase domestications - Characterization of the human piggyBac transposase derived (PGBD) activities**. *Gene* 2022, 834:146609.

Kong, X., M. Yang, B. H. Le, W. He, and Y. Hou: **The master role of siRNAs in plant immunity**. *Mol Plant Pathol* 2022, 23(10):1565-1574.

Krishnan, M., J. M. Park, F. Cao, D. Wang, R. Paulmurugan, J. R. Tseng, M. L. Gonzalgo, S. S. Gambhir, and J. C. Wu: **Effects of epigenetic modulation on reporter gene expression: implications for stem cell imaging**. *FASEB J* 2006, 20(1):106-108.

Kroutter, E. N., V. P. Belancio, B. J. Wagstaff, and A. M. Roy-Engel: **The RNA polymerase dictates ORF1 requirement and timing of LINE and SINE retrotransposition**. *PLoS Genet* 2009, 5(4):e1000458.

Kuchenbauer, F., R. D. Morin, B. Argiropoulos, O. I. Petriv, M. Griffith, M. Heuser, E. Yung, J. Piper, A. Delaney, A. L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. L. Hansen, M. A. Marra, and R. K. Humphries: **In-depth characterization of the microRNA transcriptome in a leukemia progression model**. *Genome Res* 2008, 18(11):1787-1797.

Lampe, D. J., T. E. Grant, and H. M. Robertson: **Factors affecting transposition of the Himar1 mariner transposon in vitro**. *Genetics* 1998, 149(1):179-187.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, 409(6822):860-921.

Lee, L. W., S. Zhang, A. Etheridge, L. Ma, D. Martin, D. Galas, and K. Wang: **Complexity of the microRNA repertoire revealed by next-generation sequencing**. *RNA* 2010, 16(11):2170-2180.

Lee, R. C., R. L. Feinbaum, and V. Ambros: **The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14**. *Cell* 1993, 75(5):843-854.

Lehnert, S., P. Van Loo, P. J. Thilakarathne, P. Marynen, G. Verbeke, and F. C. Schuit: **Evidence for co-evolution between human microRNAs and Alu-repeats**. *PLoS One* 2009, 4(2):e4456.

Li, N., X. You, T. Chen, S. D. Mackowiak, M. R. Friedlander, M. Weigt, H. Du, A. Gogol-Doring, Z. Chang, C. Dieterich, Y. Hu, and W. Chen: **Global profiling of miRNAs and the hairpin precursors: insights into miRNA processing and novel miRNA discovery**. *Nucleic Acids Res* 2013, 41(6):3619-3634.

Li, X., Y. Le, Z. Zhang, X. Nian, B. Liu, and X. Yang: **Viral Vector-Based Gene Therapy**. *Int J Mol Sci* 2023, 24(9).

Liew, C. G., J. S. Draper, J. Walsh, H. Moore, and P. W. Andrews: **Transient and stable transgene expression in human embryonic stem cells**. *Stem Cells* 2007, 25(6):1521-1528.

Linhares, V. L., N. A. Almeida, D. C. Menezes, D. A. Elliott, D. Lai, E. C. Beyer, A. C. Campos de Carvalho, and M. W. Costa: **Transcriptional regulation of the murine Connexin40 promoter by cardiac factors Nkx2-5, GATA4 and Tbx5**. *Cardiovasc Res* 2004, 64(3):402-411.

Liu, J., M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J. J. Song, S. M. Hammond, L. Joshua-Tor, and G. J. Hannon: **Argonaute2 is the catalytic engine of mammalian RNAi**. *Science* 2004, 305(5689):1437-1441.

Livak, K. J., and T. D. Schmittgen: **Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method**. *Methods* 2001, 25(4):402-408.

Lohe, A. R., and D. L. Hartl: **Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation**. *Mol Biol Evol* 1996, 13(4):549-555.

Loher, P., E. R. Londin, and I. Rigoutsos: **IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies**. *Oncotarget* 2014, 5(18):8790-8802.

Long, M., N. W. VanKuren, S. Chen, and M. D. Vibranovski: **New gene evolution: little did we know**. *Annu Rev Genet* 2013, 47:307-333.

Lu, J., Y. Zhu, S. Williams, M. Watts, M. A. Tonta, H. A. Coleman, H. C. Parkington, and C. Claudianos: **Autism-associated miR-873 regulates ARID1B, SHANK3 and NRXN2 involved in neurodevelopment**. *Transl Psychiatry* 2020, 10(1):418.

Luo, W., A. B. Hickman, P. Genzor, R. Ghirlando, C. M. Furman, A. Menshikh, A. Haase, F. Dyda, and M. H. Wilson: **Transposase N-terminal phosphorylation and asymmetric transposon ends inhibit piggyBac transposition in mammalian cells**. *Nucleic Acids Res* 2022, 50(22):13128-13142.

Majumdar, S., A. Singh, and D. C. Rio: **The human THAP9 gene encodes an active P-element DNA transposase**. *Science* 2013, 339(6118):446-448.

Malnou, E. C., D. Umlauf, M. Mouysset, and J. Cavaille: **Imprinted MicroRNA Gene Clusters in the Evolution, Development, and Functions of Mammalian Placenta**. *Front Genet* 2019, 9:706.

Mandal, P. K., and H. H. Kazazian, Jr.: **SnapShot: Vertebrate transposons**. *Cell* 2008, 135(1):192-192 e191.

Manuri, P. V., M. H. Wilson, S. N. Maiti, T. Mi, H. Singh, S. Olivares, M. J. Dawson, H. Huls, D. A. Lee, P. H. Rao, J. M. Kaminski, Y. Nakazawa, S. Gottschalk, P. Kebriaei, E. J. Shpall, R. E. Champlin, and L. J. Cooper: **piggyBac transposon/transposase system to generate CD19-specific T cells for the treatment of B-lineage malignancies**. *Hum Gene Ther* 2010, 21(4):427-437.

Marasco, L. E., and A. R. Kornblihtt: **The physiology of alternative splicing**. *Nat Rev Mol Cell Biol* 2022.

Martinez, J., A. Patkaniowska, H. Urlaub, R. Luhrmann, and T. Tuschl: **Single-stranded antisense siRNAs guide target RNA cleavage in RNAi**. *Cell* 2002, 110(5):563-574.

Martinez, J., and T. Tuschl: **RISC is a 5' phosphomonoester-producing RNA endonuclease**. *Genes Dev* 2004, 18(9):975-980.

Martins, J. R., D. Haenni, M. Bugarski, M. Polesel, C. Schuh, and A. M. Hall: **Intravital kidney microscopy: entering a new era**. *Kidney Int* 2021, 100(3):527-535.

Mates, L., M. K. Chuah, E. Belay, B. Jerchow, N. Manoj, A. Acosta-Sanchez, D. P. Grzela, A. Schmitt, K. Becker, J. Matrai, L. Ma, E. Samara-Kuko, C. Gysemans, D. Pryputniewicz, C. Miskey, B. Fletcher, T. Vandendriessche, Z. Ivics, and Z. Izsvak: **Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates**. *Nat Genet* 2009, 41(6):753-761.

Mattick, J. S., P. P. Amaral, P. Carninci, S. Carpenter, H. Y. Chang, L. L. Chen, R. Chen, C. Dean, M. E. Dinger, K. A. Fitzgerald, T. R. Gingeras, M. Guttman, T. Hirose, M. Huarte, R. Johnson, C. Kanduri, P. Kapranov, J. B. Lawrence, J. T. Lee, J. T. Mendell, *et al.*: **Long non-coding RNAs: definitions, functions, challenges and recommendations**. *Nat Rev Mol Cell Biol* 2023, 24(6):430-447.

McClintock, B.: **The origin and behavior of mutable loci in maize**. *Proc Natl Acad Sci U S A* 1950, 36(6):344-355.

McClintock, B.: **Induction of Instability at Selected Loci in Maize**. *Genetics* 1953, 38(6):579-599.

Meister, G., M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl: **Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs**. *Mol Cell* 2004, 15(2):185-197.

Meszaros, B., G. Erdos, and Z. Dosztanyi: **IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding**. *Nucleic Acids Res* 2018, 46(W1):W329-W337.

Meszaros, B., I. Simon, and Z. Dosztanyi: **Prediction of protein binding regions in disordered proteins**. *PLoS Comput Biol* 2009, 5(5):e1000376.

Michlewski, G., and J. F. Caceres: **Post-transcriptional control of miRNA biogenesis**. *RNA* 2019, 25(1):1-16.

Miskei, M., A. Horvath, L. Viola, L. Varga, E. Nagy, O. Fero, Z. Karanyi, J. Roszik, C. Miskey, Z. Ivics, and L. Szekvolgyi: **Genome-wide mapping of binding sites of the transposase-derived SETMAR protein in the human genome**. *Comput Struct Biotechnol J* 2021, 19:4032-4041.

Miskey, C., Z. Izsvak, R. H. Plasterk, and Z. Ivics: **The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells**. *Nucleic Acids Res* 2003, 31(23):6873-6881.

Miskey, C., B. Papp, L. Mates, L. Sinzelle, H. Keller, Z. Izsvak, and Z. Ivics: **The ancient mariner sails again: transposition of the human Hsmar1 element by a reconstructed transposase and activities of the SETMAR protein on transposon ends**. *Mol Cell Biol* 2007, 27(12):4589-4600.

Mitra, R., J. Fain-Thornton, and N. L. Craig: **piggyBac can bypass DNA synthesis during cut and paste transposition**. *Embo J* 2008, 27(7):1097-1109.

Mitra, R., X. Li, A. Kapusta, D. Mayhew, R. D. Mitra, C. Feschotte, and N. L. Craig: **Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon**. *Proc Natl Acad Sci U S A* 2013, 110(1):234-239.

Miyoshi, K., T. Miyoshi, and H. Siomi: **Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production**. *Mol Genet Genomics* 2010, 284(2):95-103.

Moeller, F., F. C. Nielsen, and L. B. Nielsen: **New tools for quantifying and visualizing adoptively transferred cells in recipient mice**. *J Immunol Methods* 2003, 282(1-2):73-82.

Morales-Prieto, D. M., S. Ospina-Prieto, W. Chaiwangyen, M. Schoenleben, and U. R. Markert: **Pregnancy-associated miRNA-clusters**. *J Reprod Immunol* 2013, 97(1):51-61.

Morales Poole, J. R., S. F. Huang, A. Xu, J. Bayet, and P. Pontarotti: **The RAG transposon is active through the deuterostome evolution and domesticated in jawed vertebrates**. *Immunogenetics* 2017, 69(6):391-400.

Moran, Y., M. Agron, D. Praher, and U. Technau: **The evolutionary origin of plant and animal microRNAs**. *Nat Ecol Evol* 2017, 1(3):27.

Morellet, N., X. Li, S. A. Wieninger, J. L. Taylor, J. Bischerour, S. Moriau, E. Lescop, B. Bardiaux, N. Mathy, N. Assrir, M. Betermier, M. Nilges, A. B. Hickman, F. Dyda, N. L. Craig, and E. Guittet: **Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase**. *Nucleic Acids Res* 2018, 46(5):2660-2677.

Morin, R. D., M. D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A. L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. J. Eaves, and M. A. Marra: **Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells**. *Genome Res* 2008, 18(4):610-621.

Mullins, L. J., B. R. Conway, R. I. Menzies, L. Denby, and J. J. Mullins: **Renal disease pathophysiology and treatment: contributions from the rat**. *Dis Model Mech* 2016, 9(12):1419-1433.

Muotri, A. R., V. T. Chu, M. C. Marchetto, W. Deng, J. V. Moran, and F. H. Gage: **Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition**. *Nature* 2005, 435(7044):903-910.

Nakai, J., M. Ohkura, and K. Imoto: **A high signal-to-noise Ca(2+) probe composed of a single green fluorescent protein**. *Nat Biotechnol* 2001, 19(2):137-141.

Napoli, C., C. Lemieux, and R. Jorgensen: **Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans**. *Plant Cell* 1990, 2(4):279-289.

Narayanavari, S. A., S. S. Chilkunda, Z. Ivics, and Z. Izsvak: **Sleeping Beauty transposition: from biology to applications**. *Crit Rev Biochem Mol Biol* 2017, 52(1):18-44.

Nei, M., and A. P. Rooney: **Concerted and birth-and-death evolution of multigene families**. *Annu Rev Genet* 2005, 39:121-152.

Neilsen, C. T., G. J. Goodall, and C. P. Bracken: **IsomiRs--the overlooked repertoire in the dynamic microRNAome**. *Trends Genet* 2012, 28(11):544-549.

Nesmelova, I. V., and P. B. Hackett: **DDE transposases: Structural similarity and diversity**. *Advanced drug delivery reviews* 2010, 62(12):1187-1195.

Newman, J. C., A. D. Bailey, H. Y. Fan, T. Pavelitz, and A. M. Weiner: **An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome**. *PLoS Genet* 2008, 4(3):e1000031.

Nguyen, T. A., M. H. Jo, Y. G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J. S. Woo: **Functional Anatomy of the Human Microprocessor**. *Cell* 2015, 161(6):1374-1387.

Nikolova, M., M. Naydenov, I. Glogovitis, A. Apostolov, M. Saare, N. Boggavarapu, A. Salumets, V. Baev, and G. Yahubyan: **Coupling miR/isomiR and mRNA Expression Signatures Unveils New Molecular Layers of Endometrial Receptivity**. *Life (Basel)* 2021, 11(12).

Nilsen, T. W., and B. R. Graveley: **Expansion of the eukaryotic proteome by alternative splicing**. *Nature* 2010, 463(7280):457-463.

Niwa, H., K. Yamamura, and J. Miyazaki: **Efficient selection for high-expression transfectants with a novel eukaryotic vector**. *Gene* 1991, 108(2):193-199.

Noguchi, S., T. Arakawa, S. Fukuda, M. Furuno, A. Hasegawa, F. Hori, S. Ishikawa-Kato, K. Kaida, A. Kaiho, M. Kanamori-Katayama, T. Kawashima, M. Kojima, A. Kubosaki, R. I. Manabe, M. Murata, S. Nagao-Sato, K. Nakazato, N. Ninomiya, H. Nishiyori-Sueki, S. Noma, *et al.*: **FANTOM5 CAGE profiles of human and mouse samples**. *Sci Data* 2017, 4:170112.

Noguer-Dance, M., S. Abu-Amero, M. Al-Khtib, A. Lefevre, P. Coullin, G. E. Moore, and J. Cavaille: **The primate-specific microRNA gene cluster (C19MC) is imprinted in the placenta**. *Hum Mol Genet* 2010, 19(18):3566-3582.

Noland, C. L., and J. A. Doudna: **Multiple sensors ensure guide strand selection in human RNAi pathways**. *RNA* 2013, 19(5):639-648.

Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, S. J. Hoyt, M. Diekhans, G. A. Logsdon, M. Alonge, S. E. Antonarakis, M. Borchers, G. G. Bouffard, S. Y. Brooks, G. V. Caldas, *et al.*: **The complete sequence of a human genome**. *Science* 2022, 376(6588):44-53.

Nykanen, A., B. Haley, and P. D. Zamore: **ATP requirements and small interfering RNA structure in the RNA interference pathway**. *Cell* 2001, 107(3):309-321.

Okamura, K., J. W. Hagen, H. Duan, D. M. Tyler, and E. C. Lai: **The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila**. *Cell* 2007, 130(1):89-100.

Okamura, K., M. D. Phillips, D. M. Tyler, H. Duan, Y. T. Chou, and E. C. Lai: **The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution**. *Nat Struct Mol Biol* 2008, 15(4):354-363.

Olah, A., M. Ruppert, T. I. Orban, A. Apati, B. Sarkadi, B. Merkely, and T. Radovits: **Hemodynamic characterization of a transgenic rat strain stably expressing the calcium sensor protein GCaMP2**. *Am J Physiol Heart Circ Physiol* 2019, 316(5):H1224-H1228.

Onishi, R., S. Yamanaka, and M. C. Siomi: **piRNA- and siRNA-mediated transcriptional repression in Drosophila, mice, and yeast: new insights and biodiversity**. *EMBO Rep* 2021, 22(10):e53062.

Orbán, T., A. Apáti, and B. Sarkadi: **Új remény a gén- és őssejtterápiában: a transzpozon alapú génbeviteli eljárások**. *Biokémia* 2009, 33(1):49-58.

Orban, T. I.: **One locus, several functional RNAs-emerging roles of the mechanisms responsible for the sequence variability of microRNAs**. *Biol Futur* 2023, 74(1-2):17-28.

Orbán, T. I., A. Apáti, Z. Izsvák, Z. Ivics, and B. Sarkadi: **Use of Transposon-Transposase Systems for Stable Genetic Modification of Embryonic Stem Cells**. in: *Methodological Advances in the Culture, Manipulation and Utilization of Embryonic Stem Cells for Basic and Practical Applications; InTech Ed.* 2011, pages: 259-274.

Orban, T. I., A. Apati, A. Nemeth, N. Varga, V. Krizsik, A. Schamberger, K. Szebenyi, Z. Erdei, G. Varady, E. Karaszi, L. Homolya, K. Nemet, E. Gocza, C. Miskey, L. Mates, Z. Ivics, Z. Izsvak, and B. Sarkadi: **Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery**. *Stem Cells* 2009, 27(5):1077-1087.

Orban, T. I., and E. Izaurralde: **Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome**. *RNA* 2005, 11(4):459-469.

Packer, A. N., Y. Xing, S. Q. Harper, L. Jones, and B. L. Davidson: **The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease**. *J Neurosci* 2008, 28(53):14341-14346.

Pardue, M. L., and P. G. DeBaryshe: **Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres**. *Annu Rev Genet* 2003, 37:485-511.

Patthy, L.: **Exon Shuffling Played a Decisive Role in the Evolution of the Genetic Toolkit for the Multicellular Body Plan of Metazoa**. *Genes (Basel)* 2021, 12(3).

Pavelitz, T., L. T. Gray, S. L. Padilla, A. D. Bailey, and A. M. Weiner: **PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans**. *Mob DNA* 2013, 4(1):23.

Pavlicek, A., J. Paces, D. Elleder, and J. Hejnar: **Processed pseudogenes of human endogenous retroviruses generated by LINEs: their integration, stability, and distribution**. *Genome Res* 2002a, 12(3):391-399.

Pavlicek, A., J. Paces, R. Zika, and J. Hejnar: **Length distribution of long interspersed nucleotide elements (LINEs) and processed pseudogenes of human endogenous retroviruses: implications for retrotransposition and pseudogene detection**. *Gene* 2002b, 300(1-2):189-194.

Pedersen, I. M., and D. G. Zisoulis: **Transposable elements and miRNA: Regulation of genomic stability and plasticity**. *Mobile genetic elements* 2016, 6(3):e1175537.

Pennisi, E.: **Why do humans have so few genes?** *Science* 2005, 309(5731):80.

Picard, G., J. C. Bregliano, A. Bucheton, J. M. Lavige, A. Pelisson, and M. G. Kidwell: **Non-mendelian female sterility and hybrid dysgenesis in Drosophila melanogaster**. *Genet Res* 1978, 32(3):275-287.

Pillai, R. S.: **MicroRNA function: multiple mechanisms for a tiny RNA?** *RNA* 2005, 11(12):1753-1761.

Piriyapongsa, J., L. Marino-Ramirez, and I. K. Jordan: **Origin and evolution of human microRNAs from transposable elements**. *Genetics* 2007, 176(2):1323-1337.

Pledger, D. W., and C. J. Coates: **Mutant Mos1 mariner transposons are hyperactive in Aedes aegypti**. *Insect Biochem Mol Biol* 2005, 35(10):1199-1207.

Prata, D. P., B. Costa-Neves, G. Cosme, and E. Vassos: **Unravelling the genetic basis of schizophrenia and bipolar disorder with GWAS: A systematic review**. *J Psychiatr Res* 2019, 114:178-207.

Pritham, E. J., C. Feschotte, and S. R. Wessler: **Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans**. *Mol Biol Evol* 2005, 22(9):1751-1763.

Qin, S., P. Jin, X. Zhou, L. Chen, and F. Ma: **The Role of Transposable Elements in the Origin and Evolution of MicroRNAs in Human**. *PLoS One* 2015, 10(6):e0131365.

Ramalingam, P., J. K. Palanichamy, A. Singh, P. Das, M. Bhagat, M. A. Kassab, S. Sinha, and P. Chattopadhyay: **Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing**. *RNA* 2014, 20(1):76-87.

Rand, T. A., K. Ginalski, N. V. Grishin, and X. Wang: **Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity**. *Proc Natl Acad Sci U S A* 2004, 101(40):14385-14389.

Ransohoff, J. D., Y. Wei, and P. A. Khavari: **The functions and unique features of long intergenic non-coding RNA**. *Nat Rev Mol Cell Biol* 2018, 19(3):143-157.

Rasko, T., A. Pande, K. Radscheit, A. Zink, M. Singh, C. Sommer, G. Wachtl, O. Kolacsek, G. Inak, A. Szvetnik, S. Petrakis, M. Bunse, V. Bansal, M. Selbach, T. I. Orban, A. Prigione, L. D. Hurst, and Z. Izsvak: **A Novel Gene Controls a New Structure: PiggyBac Transposable Element-Derived 1, Unique to Mammals, Controls Mammal-Specific Neuronal Paraspeckles**. *Mol Biol Evol* 2022, 39(10):msac175.

Ree, D., A. Fothi, N. Varga, O. Kolacsek, T. I. Orban, and A. Apati: **Partial Disturbance of Microprocessor Function in Human Stem Cells Carrying a Heterozygous Mutation in the DGCR8 Gene**. *Genes (Basel)* 2022, 13(11):1925.

Ren, J., P. Jin, E. Wang, F. M. Marincola, and D. F. Stroncek: **MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells**. *J Transl Med* 2009, 7:20.

Ro, S., C. Park, D. Young, K. M. Sanders, and W. Yan: **Tissue-dependent paired expression of miRNAs**. *Nucleic Acids Res* 2007, 35(17):5944-5953.

Rojas-Duran, M. F., and W. V. Gilbert: **Alternative transcription start site selection leads to large differences in translation activity in yeast**. *RNA* 2012, 18(12):2299-2305.

Romano, N., and G. Macino: **Quelling: transient inactivation of gene expression in Neurospora crassa by transformation with homologous sequences**. *Mol Microbiol* 1992, 6(22):3343-3353.

Rotival, M., K. J. Siddle, M. Silvert, J. Pothlichet, H. Quach, and L. Quintana-Murci: **Population variation in miRNAs and isomiRs and their impact on human immunity to infection**. *Genome Biol* 2020, 21(1):187.

Rubin, G. M., M. G. Kidwell, and P. M. Bingham: **The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations**. *Cell* 1982, 29(3):987-994.

Rubin, G. M., and A. C. Spradling: **Genetic transformation of Drosophila with transposable element vectors**. *Science* 1982, 218(4570):348-353.

Ruby, J. G., C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans**. *Cell* 2006, 127(6):1193-1207.

Ruby, J. G., C. H. Jan, and D. P. Bartel: **Intronic microRNA precursors that bypass Drosha processing**. *Nature* 2007, 448(7149):83-86.

Saha, S., L. E. Woodard, E. M. Charron, R. C. Welch, C. M. Rooney, and M. H. Wilson: **Evaluating the potential for undesired genomic effects of the piggyBac transposon system in human cells**. *Nucleic Acids Res* 2015, 43(3):1770-1782.

Salem, O., N. Erdem, J. Jung, E. Munstermann, A. Worner, H. Wilhelm, S. Wiemann, and C. Korner: **The highly expressed 5'isomiR of hsa-miR-140-3p contributes to the tumor-suppressive effects of miR-140 by reducing breast cancer proliferation and migration**. *BMC Genomics* 2016, 17:566.

Sandor, S., T. Jordanidisz, A. Schamberger, G. Varady, Z. Erdei, A. Apati, B. Sarkadi, and T. I. Orban: **Functional characterization of the ABCG2 5' non-coding exon variants: Stem cell specificity, translation efficiency and the influence of drug selection**. *Biochim Biophys Acta* 2016, 1859(7):943-951.

Sarkar, A., C. Sim, Y. S. Hong, J. R. Hogan, M. J. Fraser, H. M. Robertson, and F. H. Collins: **Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences**. *Molecular genetics and genomics* 2003, 270(2):173-180.

Sawicki, J. A., R. J. Morris, B. Monks, K. Sakai, and J. Miyazaki: **A composite CMV-IE enhancer/beta-actin promoter is ubiquitously expressed in mouse cutaneous epithelium**. *Exp Cell Res* 1998, 244(1):367-369.

Schamberger, A., and T. I. Orban: **3' IsomiR species and DNA contamination influence reliable quantification of microRNAs by stem-loop quantitative PCR**. *PLoS One* 2014a, 9(8):e106315.

Schamberger, A., and T. I. Orban: **Experimental validation of predicted mammalian microRNAs of mirtron origin**. *Methods Mol Biol* 2014b, 1182:245-263.

Schamberger, A., B. Sarkadi, and T. I. Orban: **Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner**. *RNA Biol* 2012, 9(9):1177-1185.

Scheer, H., H. Zuber, C. De Almeida, and D. Gagliardi: **Uridylation Earmarks mRNAs for Degradation... and More**. *Trends Genet* 2016, 32(10):607-619.

Scheper, M., A. Romagnolo, Z. M. Besharat, A. M. Iyer, R. Moavero, C. Hertzberg, B. Weschke, K. Riney, M. Feucht, T. Scholl, B. Petrak, A. Maulisova, R. Nabbout, A. C. Jansen, F. E. Jansen, L. Lagae, M. Urbanska, E. Ferretti, A. Tempes, M. Blazejczyk, *et al.*: **miRNAs and isomiRs: Serum-Based Biomarkers for the Development of Intellectual Disability and Autism Spectrum Disorder in Tuberous Sclerosis Complex**. *Biomedicines* 2022, 10(8).

Schmid, M., and T. H. Jensen: **Nuclear quality control of RNA polymerase II transcripts**. *Wiley interdisciplinary reviews. RNA* 2010, 1(3):474-485.

Schrider, D. R., F. C. Navarro, P. A. Galante, R. B. Parmigiani, A. A. Camargo, M. W. Hahn, and S. J. de Souza: **Gene copy-number polymorphism caused by retrotransposition in humans**. *PLoS Genet* 2013, 9(1):e1003242.

Schroder, A. R., P. Shinn, H. Chen, C. Berry, J. R. Ecker, and F. Bushman: **HIV-1 integration in the human genome favors active genes and local hotspots**. *Cell* 2002, 110(4):521-529.

Schwarz, D. S., G. Hutvagner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore: **Asymmetry in the assembly of the RNAi enzyme complex**. *Cell* 2003, 115(2):199-208.

Sepulveda, J. L., N. Belaguli, V. Nigam, C. Y. Chen, M. Nemer, and R. J. Schwartz: **GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: role for regulating early cardiac gene expression**. *Mol Cell Biol* 1998, 18(6):3405-3415.

Shapiro, J. A.: **Mutations caused by the insertion of genetic material into the galactose operon of Escherichia coli**. *J Mol Biol* 1969, 40(1):93-105.

Shenasa, H., and K. J. Hertel: **Combinatorial regulation of alternative splicing**. *Biochim Biophys Acta Gene Regul Mech* 2019, 1862(11-12):194392.

Simon, M., M. Van Meter, J. Ablaeva, Z. Ke, R. S. Gonzalez, T. Taguchi, M. De Cecco, K. I. Leonova, V. Kogan, S. L. Helfand, N. Neretti, A. Roichman, H. Y. Cohen, M. V. Meer, V. N. Gladyshev, M. P. Antoch, A. V. Gudkov, J. M. Sedivy, A. Seluanov, and V. Gorbunova: **LINE1 Derepression in Aged Wild-Type and SIRT6-Deficient Mice Drives Inflammation**. *Cell Metab* 2019, 29(4):871-885 e875.

Singer, T., M. J. McConnell, M. C. Marchetto, N. G. Coufal, and F. H. Gage: **LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes?** *Trends Neurosci* 2010, 33(8):345-354.

Sinzelle, L., Z. Izsvak, and Z. Ivics: **Molecular domestication of transposable elements: from detrimental parasites to useful host genes**. *Cellular and molecular life sciences : CMLS* 2009, 66(6):1073-1093.

Sinzelle, L., V. V. Kapitonov, D. P. Grzela, T. Jursch, J. Jurka, Z. Izsvak, and Z. Ivics: **Transposition of a reconstructed Harbinger element in human cells and functional homology with two transposon-derived cellular genes**. *Proc Natl Acad Sci U S A* 2008, 105(12):4715-4720.

Sivalingam, J., S. Krishnan, W. H. Ng, S. S. Lee, T. T. Phan, and O. L. Kon: **Biosafety assessment of site-directed transgene integration in human umbilical cord-lining cells**. *Mol Ther* 2010, 18(7):1346-1356.

Siwaszek, A., M. Ukleja, and A. Dziembowski: **Proteins involved in the degradation of cytoplasmic mRNA in the major eukaryotic model systems**. *RNA Biol* 2014, 11(9):1122-1136.

Souret, F. F., J. P. Kastenmayer, and P. J. Green: **AtXRN4 degrades mRNA in Arabidopsis and its substrates include selected miRNA targets**. *Mol Cell* 2004, 15(2):173-183.

Starega-Roslan, J., P. Galka-Marciniak, and W. J. Krzyzosiak: **Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer**. *Nucleic Acids Res* 2015a, 43(22):10939-10951.

Starega-Roslan, J., T. M. Witkos, P. Galka-Marciniak, and W. J. Krzyzosiak: **Sequence features of Drosha and Dicer cleavage sites affect the complexity of isomiRs**. *Int J Mol Sci* 2015b, 16(4):8110-8127.

Stefansson, H., R. A. Ophoff, S. Steinberg, O. A. Andreassen, S. Cichon, D. Rujescu, T. Werge, O. P. Pietilainen, O. Mors, P. B. Mortensen, E. Sigurdsson, O. Gustafsson, M. Nyegaard, A. Tuulio-Henriksson, A. Ingason, T. Hansen, J. Suvisaari, J. Lonnqvist, T. Paunio, A. D. Borglum, *et al.*: **Common variants conferring risk of schizophrenia**. *Nature* 2009, 460(7256):744-747.

Suzuki, H. I., A. Katsura, T. Yasuda, T. Ueno, H. Mano, K. Sugimoto, and K. Miyazono: **Small-RNA asymmetry is directly driven by mammalian Argonautes**. *Nat Struct Mol Biol* 2015, 22(7):512-521.

Suzuki, H. I., R. A. Young, and P. A. Sharp: **Super-Enhancer-Mediated RNA Processing Revealed by Integrative MicroRNA Network Analysis**. *Cell* 2017, 168(6):1000-1014 e1015.

Szabo, Z., L. Heja, G. Szalay, O. Kekesi, A. Furedi, K. Szebenyi, A. Dobolyi, T. I. Orban, O. Kolacsek, T. Tompa, Z. Miskolczy, L. Biczok, B. Rozsa, B. Sarkadi, and J. Kardos: **Extensive astrocyte synchronization advances neuronal coupling in slow wave activity in vivo**. *Sci Rep* 2017, 7(1):6018.

Szadeczky-Kardoss, I., T. Csorba, A. Auber, A. Schamberger, T. Nyiko, J. Taller, T. I. Orban, J. Burgyan, and D. Silhavy: **The nonstop decay and the RNA silencing systems operate cooperatively in plants**. *Nucleic Acids Res* 2018, 46(9):4632-4648.

Szebenyi, K., A. Furedi, O. Kolacsek, R. Csohany, A. Prokai, K. Kis-Petik, A. Szabo, Z. Bosze, B. Bender, J. Tovari, A. Enyedi, T. I. Orban, A. Apati, and B. Sarkadi: **Visualization of Calcium Dynamics in Kidney Proximal Tubules**. *J Am Soc Nephrol* 2015a, 26(11):2731-2740.

Szebenyi, K., A. Furedi, O. Kolacsek, E. Pergel, Z. Bosze, B. Bender, P. Vajdovich, J. Tovari, L. Homolya, G. Szakacs, L. Heja, A. Enyedi, B. Sarkadi, A. Apati, and T. I. Orban: **Generation of a Homozygous Transgenic Rat Strain Stably Expressing a Calcium Sensor Protein for Direct Examination of Calcium Signaling**. *Sci Rep* 2015b, 5:12645.

Tallini, Y. N., M. Ohkura, B. R. Choi, G. Ji, K. Imoto, R. Doran, J. Lee, P. Plan, J. Wilson, H. B. Xin, A. Sanbe, J. Gulick, J. Mathai, J. Robbins, G. Salama, J. Nakai, and M. I. Kotlikoff: **Imaging cellular signals in the heart in vivo: Cardiac expression of the high-signal Ca2+ indicator GCaMP2**. *Proc Natl Acad Sci U S A* 2006, 103(12):4753-4758.

Telonis, A. G., R. Magee, P. Loher, I. Chervoneva, E. Londin, and I. Rigoutsos: **Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32 TCGA cancer types**. *Nucleic Acids Res* 2017, 45(6):2973-2985.

Thomson, J. A., J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones: **Embryonic stem cell lines derived from human blastocysts**. *Science* 1998, 282(5391):1145-1147.

Tomasello, L., R. Distefano, G. Nigita, and C. M. Croce: **The MicroRNA Family Gets Wider: The IsomiRs Classification and Role**. *Front Cell Dev Biol* 2021, 9:668648.

Tompa, P., E. Schad, A. Tantos, and L. Kalmar: **Intrinsically disordered proteins: emerging interaction specialists**. *Curr Opin Struct Biol* 2015, 35:49-59.

Trabucchi, M., and R. Mategot: **Subcellular Heterogeneity of the microRNA Machinery**. *Trends Genet* 2019, 35(1):15-28.

VandenDriessche, T., D. Collen, and M. K. Chuah: **Biosafety of onco-retroviral vectors**. *Curr Gene Ther* 2003, 3(6):501-515.

Venter, J. C., M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, *et al.*: **The sequence of the human genome**. *Science* 2001, 291(5507):1304-1351.

Venturini, G., E. Capanna, and B. Fontana: **Size and structure of the bird genome. II. Repetitive DNA and sequence organization**. *Comp Biochem Physiol B* 1987, 87(4):975-979.

Volff, J. N.: **Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes**. *Bioessays* 2006, 28(9):913-922.

Wachtl, G., E. Schad, K. Huszar, A. Palazzo, Z. Ivics, A. Tantos, and T. I. Orban: **Functional Characterization of the N-Terminal Disordered Region of the piggyBac Transposase**. *Int J Mol Sci* 2022, 23(18).

Wang, D., P. S. Chang, Z. Wang, L. Sutherland, J. A. Richardson, E. Small, P. A. Krieg, and E. N. Olson: **Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor**. *Cell* 2001, 105(7):851-862.

Wang, J., T. Bing, N. Zhang, X. Liu, and D. Shangguan: **FnCas12a/crRNA assisted dumbbell-PCR detection of IsomiRs with terminal and inner sequence variants**. *Chem Commun (Camb)* 2020, 56(69):10038-10041.

Wang, X., A. Ramat, M. Simonelig, and M. F. Liu: **Emerging roles and functional mechanisms of PIWI-interacting RNAs**. *Nat Rev Mol Cell Biol* 2022.

Wang, Y., R. Medvid, C. Melton, R. Jaenisch, and R. Blelloch: **DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal**. *Nat Genet* 2007, 39(3):380-385.

Weiner, A. M., and L. T. Gray: **What role (if any) does the highly conserved CSB-PGBD3 fusion protein play in Cockayne syndrome?** *Mech Ageing Dev* 2013, 134(5-6):225-233.

Wells, J. N., and C. Feschotte: **A Field Guide to Eukaryotic Transposable Elements**. *Annu Rev Genet* 2020, 54:539-561.

Wen, J., E. Ladewig, S. Shenker, J. Mohammed, and E. C. Lai: **Analysis of Nearly One Thousand Mammalian Mirtrons Reveals Novel Features of Dicer Substrates**. *PLoS Comput Biol* 2015, 11(9):e1004441.

Westholm, J. O., E. Ladewig, K. Okamura, N. Robine, and E. C. Lai: **Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs**. *RNA* 2012, 18(2):177-192.

Wicker, T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman: **A unified classification system for eukaryotic transposable elements**. *Nat Rev Genet* 2007, 8(12):973-982.

Wicks, S. R., C. J. de Vries, H. G. van Luenen, and R. H. Plasterk: **CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in Caenorhabditis elegans**. *Dev Biol* 2000, 221(2):295-307.

Wightman, B., I. Ha, and G. Ruvkun: **Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans**. *Cell* 1993, 75(5):855-862.

Wilson, M. H., J. M. Kaminski, and A. L. George, Jr.: **Functional zinc finger/sleeping beauty transposase chimeras exhibit attenuated overproduction inhibition**. *FEBS Lett* 2005, 579(27):6205-6209.

Wilson, R. C., A. Tambe, M. A. Kidwell, C. L. Noland, C. P. Schneider, and J. A. Doudna: **Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis**. *Mol Cell* 2015, 57(3):397-407.

Winter, J., and S. Diederichs: **Argonaute-3 activates the let-7a passenger strand microRNA**. *RNA Biol* 2013, 10(10):1631-1643.

Winter, J., S. Jung, S. Keller, R. I. Gregory, and S. Diederichs: **Many roads to maturity: microRNA biogenesis pathways and their regulation**. *Nat Cell Biol* 2009, 11(3):228-234.

Woodard, L. E., X. Li, N. Malani, A. Kaja, R. H. Hice, P. W. Atkinson, F. D. Bushman, N. L. Craig, and M. H. Wilson: **Comparative analysis of the recently discovered hAT transposon TcBuster in human cells**. *PLoS One* 2012, 7(11):e42666.

Woods, N. B., A. Muessig, M. Schmidt, J. Flygare, K. Olsson, P. Salmon, D. Trono, C. von Kalle, and S. Karlsson: **Lentiviral vector transduction of NOD/SCID repopulating cells results in multiple vector integrations per transduced cell: risk of insertional mutagenesis**. *Blood* 2003, 101(4):1284-1289.

Wyman, S. K., E. C. Knouf, R. K. Parkin, B. R. Fritz, D. W. Lin, L. M. Dennis, M. A. Krouse, P. J. Webster, and M. Tewari: **Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity**. *Genome Res* 2011, 21(9):1450-1461.

Xia, X., Y. Zhang, C. R. Zieth, and S. C. Zhang: **Transgenes delivered by lentiviral vector are suppressed in human embryonic stem cells in a promoter-dependent manner**. *Stem Cells Dev* 2007, 16(1):167-176.

Xiong, C., S. Sun, W. Jiang, L. Ma, and J. Zhang: **ASDmiR: A Stepwise Method to Uncover miRNA Regulation Related to Autism Spectrum Disorder**. *Front Genet* 2020, 11:562971.

Xu, Z. L., H. Mizuguchi, A. Ishii-Watabe, E. Uchida, T. Mayumi, and T. Hayakawa: **Optimization of transcriptional regulatory elements for constructing plasmid vectors**. *Gene* 2001, 272(1-2):149-156.

Yang, J. S., M. D. Phillips, D. Betel, P. Mu, A. Ventura, A. C. Siepel, K. C. Chen, and E. C. Lai: **Widespread regulatory activity of vertebrate microRNA* species**. *RNA* 2011, 17(2):312-326.

Yue, W. H., H. F. Wang, L. D. Sun, F. L. Tang, Z. H. Liu, H. X. Zhang, W. Q. Li, Y. L. Zhang, Y. Zhang, C. C. Ma, B. Du, L. F. Wang, Y. Q. Ren, Y. F. Yang, X. F. Hu, Y. Wang, W. Deng, L. W. Tan, Y. L. Tan, Q. Chen, *et al.*: **Genome-wide association study identifies a susceptibility locus for schizophrenia in Han Chinese at 11p11.2**. *Nat Genet* 2011, 43(12):1228-1231.

Yusa, K.: **piggyBac Transposon**. *Microbiol Spectr* 2015, 3(2):MDNA3-0028-2014.

Yusa, K., L. Zhou, M. A. Li, A. Bradley, and N. L. Craig: **A hyperactive piggyBac transposase for mammalian applications**. *Proc Natl Acad Sci U S A* 2011, 108(4):1531-1536.

Zamore, P. D., and B. Haley: **Ribo-gnome: the big world of small RNAs**. *Science* 2005, 309(5740):1519-1524.

Zayed, H., Z. Izsvak, O. Walisko, and Z. Ivics: **Development of hyperactive sleeping beauty transposon vectors by mutational analysis**. *Mol Ther* 2004, 9(2):292-304.

Zelli, V., C. Compagnoni, R. Capelli, A. Corrente, J. Cornice, D. Vecchiotti, M. Di Padova, F. Zazzeroni, E. Alesse, and A. Tessitore: **Emerging Role of isomiRs in Cancer: State of the Art and Recent Advances**. *Genes (Basel)* 2021, 12(9).

Zeng, X., J. Chen, J. F. Sanchez, M. Coggiano, O. Dillon-Carter, J. Petersen, and W. J. Freed: **Stable expression of hrGFP by mouse embryonic stem cells: promoter activity in the undifferentiated state and during dopaminergic neural differentiation**. *Stem Cells* 2003, 21(6):647-653.

Zerbino, D. R., A. Frankish, and P. Flicek: **Progress, Challenges, and Surprises in Annotating the Human Genome**. *Annu Rev Genomics Hum Genet* 2020, 21:55-79.

Zhang, R., Y. Q. Wang, and B. Su: **Molecular evolution of a primate-specific microRNA family**. *Mol Biol Evol* 2008, 25(7):1493-1502.

Zhou, H., M. L. Arcila, Z. Li, E. J. Lee, C. Henzler, J. Liu, T. M. Rana, and K. S. Kosik: **Deep annotation of mouse iso-miR and iso-moR variation**. *Nucleic Acids Res* 2012, 40(13):5864-5875.

A DISSZERTÁCIÓ ALAPJÁT KÉPEZŐ KÖZLEMÉNYEK EREDETI VERZIÓI

(csak az elektronikus verzióban szerepelnek)

Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome

TAMAS I. ORBAN and ELISA IZAURRALDE
EMBL, D-69117 Heidelberg, Germany

ABSTRACT

RNA interference (RNAi) is a conserved RNA silencing pathway that leads to sequence-specific mRNA decay in response to the presence of double-stranded RNA (dsRNA). Long dsRNA molecules are first processed by Dicer into 21–22-nucleotide small interfering RNAs (siRNAs). The siRNAs are incorporated into a multimeric RNA-induced silencing complex (RISC) that cleaves mRNAs at a site determined by complementarity with the siRNAs. Following this initial endonucleolytic cleavage, the mRNA is degraded by a mechanism that is not completely understood. We investigated the decay pathway of mRNAs targeted by RISC in *Drosophila* cells. We show that 5′ mRNA fragments generated by RISC cleavage are rapidly degraded from their 3′ ends by the exosome, whereas the 3′ fragments are degraded from their 5′ ends by XRN1. Exosome-mediated decay of the 5′ fragments requires the *Drosophila* homologs of yeast Ski2p, Ski3p, and Ski8p, suggesting that their role as regulators of exosome activity is conserved. Our findings indicate that mRNAs targeted by siRNAs are degraded from the ends generated by RISC cleavage, without undergoing decapping or deadenylation.

Keywords: argonaute proteins; exosome; mRNA decay; RNA interference; RISC; Ski complex; XRN1

INTRODUCTION

RNA interference (RNAi) belongs to a family of related processes that lead to the silencing of gene expression by double-stranded RNA (dsRNA) molecules. These pathways act not only at the post-transcriptional level by eliciting mRNA decay or repressing translation, but also at the transcriptional level by inducing heterochromatin formation (for review, see Ambros 2004; Lippman and Martienssen 2004; Meister and Tuschl 2004).

The process of RNAi is triggered by the presence of long dsRNA molecules in the cell. During the initiation step of RNAi, these dsRNA molecules are cleaved into small interfering RNA duplexes (siRNAs) by the RNase III-like enzyme Dicer (Bernstein et al. 2001; Elbashir et al. 2001; Grishok et al. 2001; Hutvágner et al. 2001). The siRNAs are characterized by a 19–20-base-pair (bp) duplex region and a 2-nucleotide (nt) 3′ overhang on each strand. During the effector step of RNAi, the siRNAs are incorporated into a multimeric protein complex called RNA-induced silencing complex (RISC), where they serve as guides to select complementary mRNA substrates for degradation (Hammond et al. 2001; Schwarz et al. 2002).

dsRNA molecules may originate from viral replication, transcription of pseudogenes and repetitive sequence elements, or during transposition of mobile genetic elements. Consequently, gene silencing pathways are thought to contribute to the stability of the genome by silencing the expression of viruses, transposons, and repetitive sequences (for review, see Plasterk 2002; Ambros 2004; Meister and Tuschl 2004). Another endogenous source of dsRNAs results from the transcription of microRNA (miRNA) genes. miRNAs are 19–25-nt-long RNAs produced by cleavage of dsRNA hairpins encoded in the genome (for review, see Ambros 2004). Plant miRNAs are often fully complementary to their targets and elicit mRNA decay (for review, see Baulcombe 2004). In contrast, animal miRNAs are only partially complementary to their targets and do not generally elicit decay but repress translation (for review, see Ambros 2004).

Additionally, long dsRNA molecules can be introduced exogenously. In *Drosophila* and *Caenorhabditis elegans*, exogenous dsRNAs are processed by Dicer and enter the RNAi pathway (for review, see Meister and Tuschl 2004). In mammalian cells, exogenous dsRNA molecules elicit an interferon response and lead to a nonspecific degradation of cellular mRNAs and a general inhibition of translation (for

Reprint requests to: Elisa Izaurralde, EMBL, Meyerhofstrasse 1, 69117 Heidelberg, Germany; e-mail: izaurralde@embl-heidelberg.de; fax: +49 6221 387 306.

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.7231505.

review, see Williams 1999). However, chemically synthesized siRNAs artificially introduced in the cell enter the RNAi pathway and specifically target complementary mRNAs for degradation (Elbashir et al. 2001).

Degradation of mRNAs targeted by the RNAi pathway is initiated by endonucleolytic cleavage of the mRNA within the region complementary to the siRNA (Elbashir et al. 2001; Nykänen et al. 2001; Martinez et al. 2002). Structural and biochemical studies have provided compelling evidence that the endonucleolytic activity of RISC resides in the Argonaute proteins (AGO proteins), which constitute core components of RISC (Tabara et al. 1999; Hammond et al. 2001; Liu et al. 2004; Meister et al. 2004; Rand et al. 2004). Argonautes form a conserved family of highly basic proteins, characterized by a central PAZ domain and a C-terminal Piwi domain (Cerutti et al. 2000; Carmell et al. 2002). The PAZ domain binds to single-stranded 3′ ends of nucleic acids (Lingel et al. 2003, 2004; Song et al. 2003; Yan et al. 2003; Ma et al. 2004), and is involved in the specific recognition of the 2-nt 3′ overhangs of siRNAs. The Piwi domain adopts an RNase H-like fold, and is thought to catalyze the endonucleolytic cleavage (Liu et al. 2004; Song et al. 2004).

Metazoan genomes encode more than one Argonaute protein (Carmell et al. 2002). There are four annotated paralogs in *Drosophila*, eight in human, and more than 20 in *C. elegans*. Despite their similar domain organization, not all Argonaute proteins are capable of mediating endonucleolytic cleavage. Indeed, in human and *Drosophila* cells, only AGO2-containing RISC is able to catalyze mRNA cleavage, even though all Argonaute proteins bind siRNAs (Liu et al. 2004; Meister et al. 2004; Okamura et al. 2004; Rand et al. 2004).

Following endonucleolytic cleavage by RISC, the mRNA is degraded by a mechanism that has not been elucidated. In *Arabidopsis*, mutations in the gene encoding one of the major cytoplasmic 5′-to-3′ exonucleases (XRN4) leads to the stabilization of 3′ mRNA fragments derived from transcripts that are targets of endogenous miRNAs (Souret et al. 2004). That study suggested that mRNAs targeted by RISC are not degraded by a dedicated pathway, but are handed over to the general RNA decay machinery in the cell.

There are two major cytoplasmic pathways for degrading bulk mRNA in eukaryotes (for review, see Parker and Song 2004). Both pathways are initiated by shortening of the poly(A)-tail by deadenylases. In one pathway, deadenylation triggers decapping, and this exposes the mRNA body for 5′-to-3′ digestion by the major cytoplasmic 5′-to-3′ exonuclease XRN1. In yeast and mammalian cells, decay of mRNAs through this pathway is thought to occur in specialized cytoplasmic bodies or mRNA decay foci (also known as P-bodies or GW-bodies) where XRN1 and the decapping enzyme DCP2 colocalize (Ingelfinger et al. 2002; van Dijk et al. 2002; Eystathioy et al. 2003; Sheth and Parker 2003; Cougot et al. 2004).

In the second mRNA decay pathway, deadenylation is followed by 3′-to-5′ degradation of the transcript. This requires the exosome, a multimeric assembly of 3′-to-5′ exonucleases, and the Ski complex. The Ski complex was identified in yeast as a heterotrimeric protein complex (consisting of Ski2p, Ski3p, and Ski8p) that regulates exosome activity (Anderson and Parker 1998; Brown et al. 2000).

In addition to their role in general RNA decay, XRN1, the exosome, and the Ski complex function in mRNA surveillance pathways by which aberrant mRNAs are detected and degraded. These include the nonsense-mediated mRNA decay pathway (NMD), a pathway that eliminates mRNAs harboring premature translation termination codons (PTCs), and the nonstop decay (NSD) pathway that degrades mRNAs lacking a stop codon (for review, see Parker and Song 2004).

In this study we investigated the decay pathway of mRNAs targeted by RISC in *Drosophila* cells. In agreement with results reported by Okamura et al. (2004) and Rand et al. (2004), we show that endonucleolytic cleavage requires *Drosophila* AGO2, but not the additional *Drosophila* Argonaute paralogs (i.e., AGO1, PIWI, or Aubergine). Following the first AGO2-dependent cleavage, the resulting 5′ fragment is degraded by the exosome, whereas the 3′ fragment is degraded by XRN1. Degradation of the 5′ fragment is prevented in cells depleted of *Drosophila* homologs of the yeast proteins Ski2p, Ski3p, and Ski8p, suggesting that the role of these proteins in 3′-to-5′ mRNA decay is conserved.

RESULTS

Generation of RNAi reporters for *Drosophila* cells

To elucidate the mechanism by which mRNAs targeted by RISC are degraded, we made use of two previously described reporter constructs in which the coding region of the *Drosophila* alcohol dehydrogenase (adh) gene is placed downstream of the constitutive actin 5C promoter (constitutive adh reporter, 5C-adh, Fig. 1A), or downstream of a inducible metallothionein (Mtn) promoter (inducible adh reporter, Mtn-adh, Fig. 1A; Gatfield et al. 2003; Gatfield and Izaurralde 2004). We also designed an additional reporter in which the cDNA encoding human NXF1 is placed under the control of the Mtn promoter (Fig. 1A, Mtn-NXF1).

The reporter constructs were cotransfected with a plasmid encoding puromycin acetyl transferase into *Drosophila* Schneider cells (S2 cells) to generate polyclonal cell lines constitutively expressing adh, or expressing adh or human NXF1 after induction with copper sulfate. Cells were treated with dsRNAs targeting a central region of ~300–450 nt of adh or NXF1 mRNAs (Fig. 1A, black boxes). The steady-state levels of the transcripts were analyzed by Northern blot and normalized to those of the endogenous rp49 mRNA (encoding ribosomal protein L32). In cells treated with the specific dsRNAs, the normalized levels of the reporter tran-

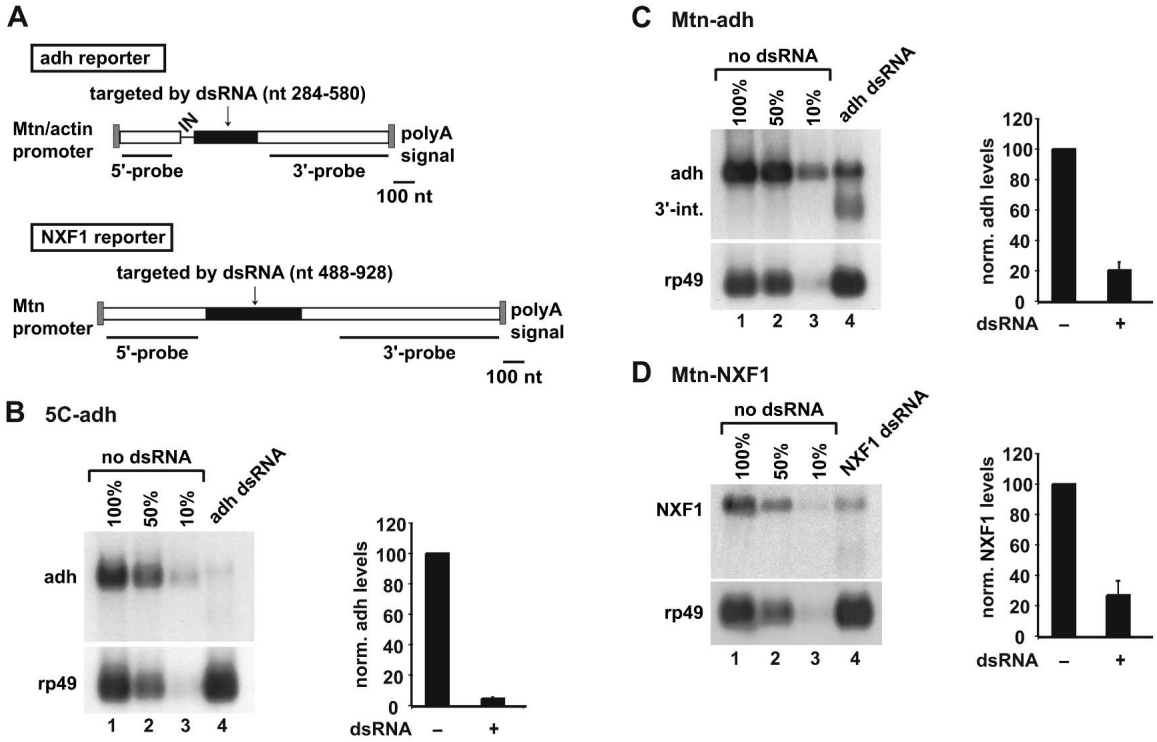


FIGURE 1. Generation of RNAi reporters for *Drosophila* cells. (*A*) Schematic representation of the reporters. White boxes, exons; gray boxes, sequences derived from vector pAc5.1b (Invitrogen) or pRmHa, respectively; black boxes, sequences complementary to the dsRNAs; IN, introns. The fragments of the transcript detected by the 5′ or 3′ probes used in this study are indicated. (*B–D*) S2 cells expressing the indicated reporter constructs were treated with the corresponding dsRNAs. Total RNA samples were isolated and analyzed by Northern blot using probes specific for adh (3′ probe), NXF1 (3′ probe), and rp49 mRNAs. In lanes *1–3*, dilutions of total RNA samples isolated from untreated cells were loaded to assess the efficiency of RNAi. The levels of the reporters were quantitated in at least three independent experiments and normalized (norm.) to those of rp49 mRNA. These values were set to 100% in untreated cells. Mean values ± standard deviations (SD) are shown.

scripts were reduced to 5%–25% of the levels detected in untreated cells (Fig. 1B–D), indicating that the reporter transcripts were effectively targeted by the dsRNAs.

We next investigated whether the reduction of mRNA levels caused by the dsRNAs was a consequence of increased mRNA turnover. In the experiment shown in Figure 2A–C, the half-life of the inducible adh reporter was determined after a short induction pulse followed by the inhibition of transcription by actinomycin D. In untreated cells, the half-life of Mtn-adh mRNA was longer than 4 h, while in cells treated with adh dsRNA the half-life was < 15 min (Fig. 2A–C). Thus, the overall reduction in steady-state levels of the reporter transcripts targeted by dsRNAs is a direct consequence of a higher turnover rate.

To assess whether the reporters are degraded through the RNAi pathway, we investigated whether depletion of Argonaute proteins prevents dsRNA-mediated decay of the transcripts. S2 cells expressing 5C-adh reporter mRNA were treated with dsRNAs specific for AGO1, AGO2, PIWI, Aubergine (AUB), or GFP as a control. Four days after the first treatment, cells were treated again with the same dsRNAs and, in addition, adh dsRNA. The expression levels of adh mRNA were analyzed 3 d after the second transfection. In cells depleted of AGO2, adh dsRNA failed to reduce the expression levels of its cognate transcript (Fig. 2D, lane 4). In contrast, in cells depleted of AGO1, PIWI, or AUB, addition of adh dsRNA triggered a reduction of adh mRNA levels as efficiently as in control cells (Fig. 2D). Note that AGO1, AGO2, PIWI, and Aubergine are likely to be expressed in S2 cells, as the corresponding mRNAs are detectable by microarray analysis and RT-PCR (data not shown). Together, these observations suggest that only AGO2, but not AGO1, PIWI, or AUB, mediates siRNA-guided target cleavage in *Drosophila*, in agreement with Okamura et al. (2004) and Rand et al. (2004). These results also indicate that the mRNA reporters are degraded through the RNAi pathway in cells treated with the corresponding dsRNAs.

3′ mRNA fragments generated by RISC-mediated endonucleolytic cleavage are degraded by XRN1

RISC initiates decay by cleaving the mRNA within a region complementary to the siRNA. This cleavage should generate a 5′ and a 3′ mRNA decay intermediate. In *Arabidopsis*, the 3′ decay intermediates are degraded by XRN4, a cytoplasmic 5′-to-3′ exonuclease (Souret et al. 2004). The *Drosophila* genome encodes a single cytoplasmic 5′-to-3′ exo-

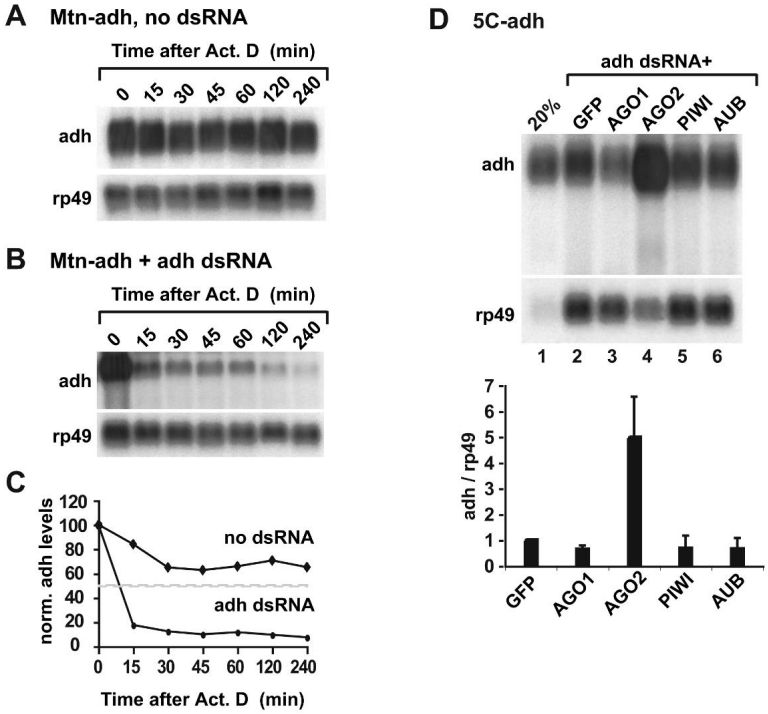


FIGURE 2. Reporter mRNAs are degraded through the RNAi pathway. (*A*,*B*) S2 cells expressing Mtn-adh were treated with adh dsRNA. Expression of adh was induced for 45 min in treated and untreated cells. Following induction, transcription was inhibited by actinomycin D (5 μg/mL) for the times indicated above the lanes. Total RNA samples were isolated and analyzed as described in Figure 1. (*C*) The levels of adh mRNA normalized to those of rp49 mRNA shown in panels *A* and *B* are plotted as a function of time. (*D*) S2 cells constitutively expressing adh mRNA were treated with the dsRNAs indicated above the lanes. Total RNA samples were isolated and analyzed as described in Figure 1. The levels of the reporters were quantitated in at least three independent experiments and normalized (norm.) to those of rp49 mRNA. These values were set to 1 in cells treated with GFP and adh dsRNAs. Mean values ± SD are shown.

nuclease known as XRN1 or Pacman. To investigate whether *Drosophila* XRN1 is responsible for the decay of the 3′ mRNA fragment generated by RISC cleavage, we depleted the protein in cells constitutively expressing adh mRNA. When cells were also treated with adh dsRNA, depletion of XRN1 led to the accumulation of mRNA species with higher electrophoretic mobility (Fig. 3A, lane 3, 3′-int.). These shorter RNA species were detected with a probe complementary to a region downstream of the dsRNA-target site, but not with a probe hybridizing upstream of the targeted region (see Figs. 3A, 5A below). The shorter RNA species were not observed when adh dsRNA was omitted (data not shown), and depletion of AGO2 prevented their appearance (Fig. 3A, lane 7), indicating that they represent 3′ decay intermediates that arose from the messages undergoing RNAi. Similar results were obtained for the Mtn-adh and Mtn-NXF1 reporters (Fig. 3B,C).

To quantify the effect of the XRN1 depletion, the steady-state levels of the full-length transcripts and of the 3′ decay intermediates were quantitated in at least three independent experiments. The values obtained for the intermediate were divided by those obtained for the full-length transcript, in order to correct for potential nonspecific effects of the depletions and this ratio was set to a value of 1 in control cells (i.e., cells treated with GFP and adh dsRNAs). Although the 3′ fragments were also detected in control cells treated with the specific dsRNA (with or without addition of GFP dsRNA; Fig. 1B–D, lanes 4; Fig. 3A–C), their levels increase three- to eightfold in cells depleted of XRN1 (Fig. 3A–C). Depletion of the exosome components RRP4 or CSL4, or of SKI2 did not increase the levels of the 3′ fragments, indicating that these fragments are not degraded by deadenylation and 3′-to-5′ exonucleolytic digestion by the exosome.

If the 3′ fragments were degraded by XRN1, one would expect the half-life of these fragments to be increased in cells depleted of XRN1 relative to cells treated with GFP dsRNA. Indeed, in cells treated with GFP and adh dsRNAs, the adh 3′ intermediate had a half-life shorter than 15 min, but was strongly stabilized in cells depleted of XRN1 (Fig. 4A–C). We conclude that the 3′ mRNA fragments generated by RISC activity are degraded from their 5′ ends by XRN1.

Decay of the 5′ fragment is mediated by the exosome

RISC-mediated endonucleolytic cleavage produces 5′ decay intermediates that have been observed both in vivo and in vitro (Elbashir et al. 2001; Nykänen et al. 2001; Martinez and Tuschl 2004; Schwarz et al. 2004), and could potentially be degraded by the exosome. In order to determine whether the exosome is indeed involved in this process, we silenced the expression of the exosome component RRP4 or CSL4 or the exosome cofactor SKI2 in cells expressing 5C-adh and treated with adh dsRNA. The effect of these depletions on the adh mRNA levels was analyzed by Northern blot and compared to the effects observed when GFP dsRNA was added. Depletion of RRP4, CSL4, or SKI2 led to the appearance of mRNA fragments that are detected using a probe spanning sequences 5′ to the region targeted by the dsRNA (Fig. 5, 5′ int.). These fragments were not observed in the absence of adh dsRNA (data not shown) or in cells treated with AGO2 dsRNA (Fig. 5A, lane 7), indicating that they represent 5′ decay intermediates arising from the reporter targeted by RNAi. Similar results were seen with the Mtn-adh and NXF1 reporters (Fig. 5B,C).

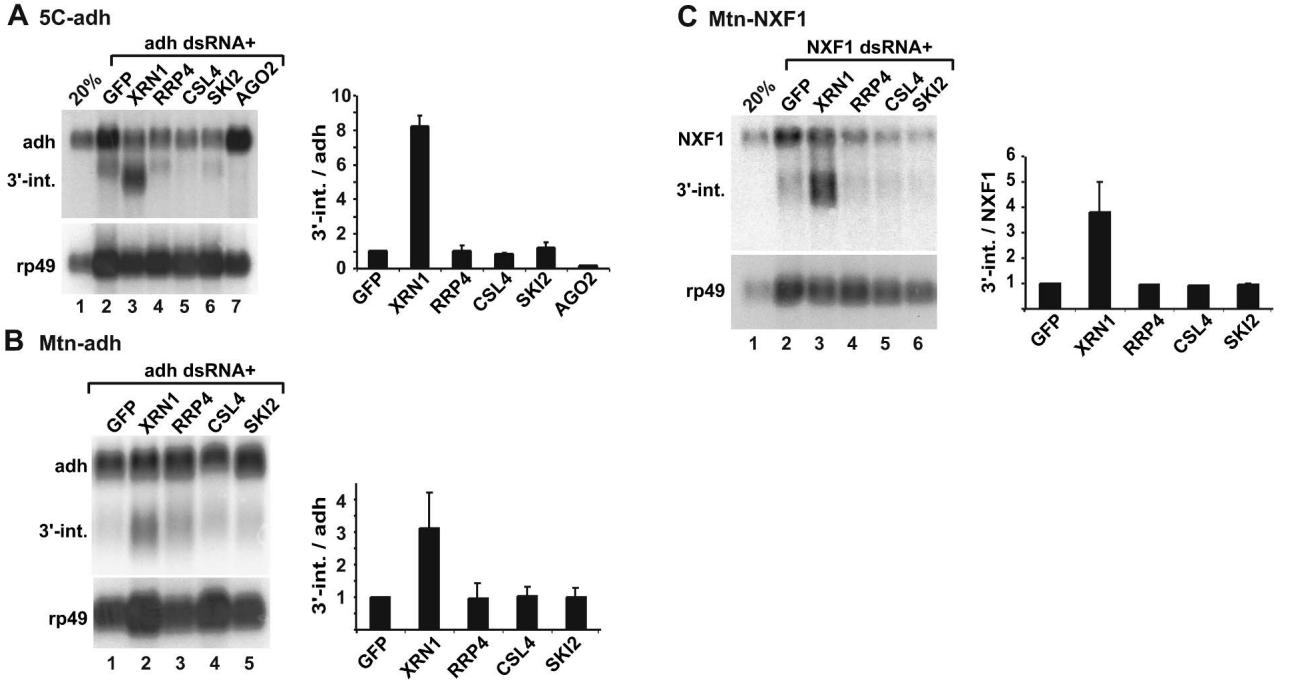


FIGURE 3. The 3′ fragments are degraded by XRN1. (*A–C*) S2 cells expressing 5C-adh, Mtn-adh, or Mtn-NXF1 were treated with the dsRNAs indicated above the lanes. Total RNA samples were analyzed as described in Figure 1. Northern blots were hybridized with probes complementary to the 3′ ends of the reporters as shown in Figure 1A. In lane *1* of panels *A* and *C*, 20% of an RNA sample isolated from untreated cells was loaded to assess the efficiency of RNAi. The levels of the 3′-intermediate (3′-int.) were normalized to those of the full-length mRNA in at least three independent experiments. These values were set to 1 in control cells treated with GFP and the specific (adh or NXF1) dsRNAs. Mean values ± SD are shown.

Relative to the full-length transcript, the levels of the 5′ fragments increased three- to ninefold in cells depleted of RRP4, CSL4, or SKI2, but remained unchanged in cells depleted of XRN1 (Fig. 5A–C). Moreover, the half-life of the 5′ fragments was < 15 min in control cells, but increased to 60 min in cells depleted of RRP4 (Fig. 6A–C). We conclude that the 5′ mRNA fragments generated by RISC activity are degraded from their 3′ ends by the exosome.

Drosophila homologs of Ski proteins are required for exosome-mediated decay of the 5′ intermediates

In yeast, the Ski complex consists of Ski2p, Ski3p, and Ski8p (Brown et al. 2000). Although putative homologs of the Ski proteins are encoded by the higher eukaryotic genomes, it is unclear whether these proteins assemble in a complex with similar functions as in yeast. The observation that depletion of the *Drosophila* homolog of Ski2p [encoded by the gene *twister* (*tst*), also known as CG10210] leads to a strong stabilization of the 5′ decay intermediates (Fig. 5) prompted us to test whether the putative *Drosophila* homologs of Ski3p and Ski8p are also involved in this process. The closest homologs of Ski3p and Ski8p are encoded by the uncharacterized *Drosophila* genes CG8777 and CG3909, respectively.

In yeast, the interaction between the exosome and the Ski complex is thought to be mediated by the GTP-binding protein Ski7p (van Hoof et al. 2000; Araki et al. 2001). Ski7p is structurally related to eRF3 and Hbs1p. Surprisingly, higher eukaryotic genomes encode orthologs of eRF3 and Hbs1, but no obvious orthologs of Ski7p. We therefore also tested whether *Drosophila* HBS1 could be the functional homolog of Ski7p.

In cells depleted of CG8777 or CG3909, but not of HBS1, the adh 5′ decay intermediate increased in abundance relative to the full-length transcript (Fig. 6D,E). These results argue that *tst* (or CG10210), CG8777, and CG3909 encode bona fide components of the Ski complex. Therefore we propose to call CG10210, CG8777, and CG3909 *Drosophila ski2*, *ski3*, and *ski8*, respectively.

RISC-mediated decay of mRNA occurs independently of decapping and deadenylation

The results described above suggest that the decay intermediates are degraded from the ends generated by RISC cleavage, without prior decapping or deadenylation. We therefore tested whether the 3′ fragments that accumulate in cells depleted of XRN1 are polyadenylated. If this were the case, we would expect these fragments to change their electrophoretic mobility upon removal of the poly(A)-tail by oligo(dT)-targeted ribonuclease H (RNase H) cleavage. Both

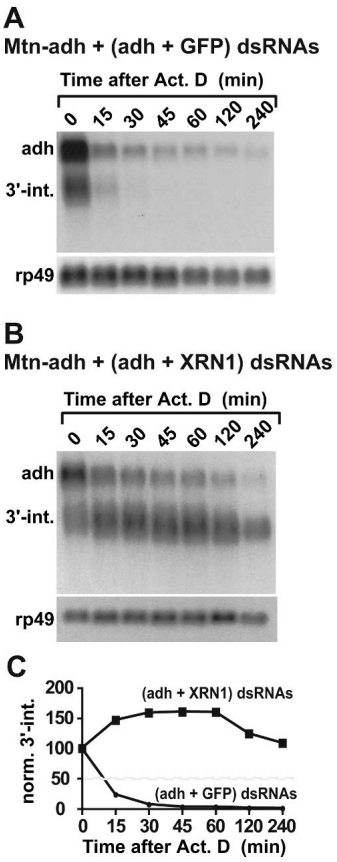


FIGURE 4. Decay rates of the 3′ fragments in cells depleted of XRN1. (*A*,*B*) S2 cells expressing Mtn-adh were treated with the indicated dsRNAs. Expression of adh was induced for 45 min. Transcription was inhibited by actinomycin D (5 μg/mL) for the times indicated above the lanes. Total RNA samples were isolated and analyzed as described in Figure 1. Northern blots were hybridized with a 3′ probe (see Fig. 1A). (*C*) The levels of the 3′ fragments normalized to those of rp49 mRNA shown in panels *A* and *B* are plotted as a function of time.

the full-length transcript and the 3′ fragments change mobility following removal of the poly(A)-tail, indicating that these fragments are polyadenylated (Fig. 7A). In contrast, the 5′ fragments stabilized by depletion of SKI2 did not change mobility following RNase H treatment in the presence of oligo(dT) (Fig. 7B).

To test for the presence of a cap structure, we performed co-immunoprecipitations with antibodies recognizing the 5′ cap of mRNAs. We observed that full-length adh mRNA and the 5′ decay intermediates, but not the 3′ fragments, were preferentially co-immunoprecipitated with these antibodies (Fig. 7C,D). These results provide further evidence for a model in which the decay of mRNA fragments generated by RISC is independent of deadenylation and decapping.

Efficient decay of the 5′ fragments requires ongoing translation

The 5′ fragments generated by RISC cleavage harbor a cap structure but lack a stop codon and a poly(A)-tail. As such, these mRNA fragments could be substrates for the nonstop decay (NSD) pathway. In yeast, this pathway requires the exosome, the Ski complex, and Ski7p (Frischmeyer et al. 2002; van Hoof et al. 2002). Moreover, mRNAs degraded through the NSD pathway are stabilized following inhibition of translation (Frischmeyer et al. 2002; van Hoof et al. 2002). To determine whether decay of the intermediates requires translation, cells constitutively expressing adh mRNA were treated with adh and GFP dsRNAs. As shown in Figures 3A and 5A (lane 2), this treatment leads to the appearance of the intermediates. Cells were then treated

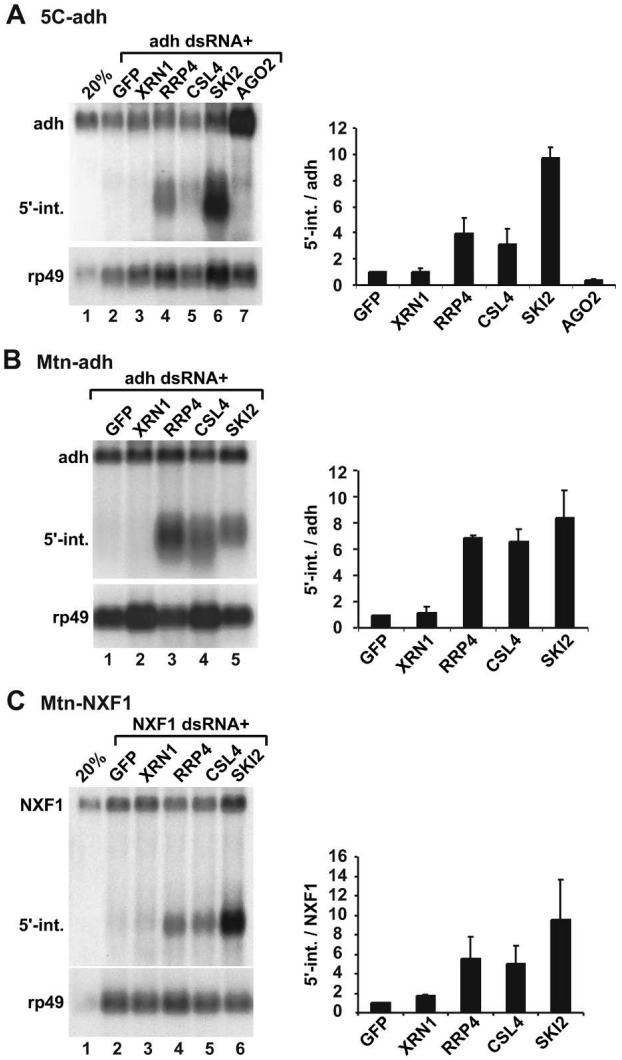


FIGURE 5. The 5′ fragments are degraded by the exosome. (*A–C*) S2 cell-lines expressing 5C-adh, Mtn-adh, or Mtn-NXF1 were treated with the dsRNAs indicated above the lanes. Total RNA samples were analyzed by Northern blot using probes detecting the 5′ decay intermediates (see Fig. 1A). The levels of the 5′ fragments were normalized to those of the full-length mRNA in at least three independent experiments. These values were set to 1 in control cells treated with GFP dsRNA and dsRNAs targeting the reporters. Mean values ± SD are shown. In lane *1* of panels *A* and *C*, a dilution of an RNA sample isolated from untreated cells was loaded to assess the efficiency of RNAi.

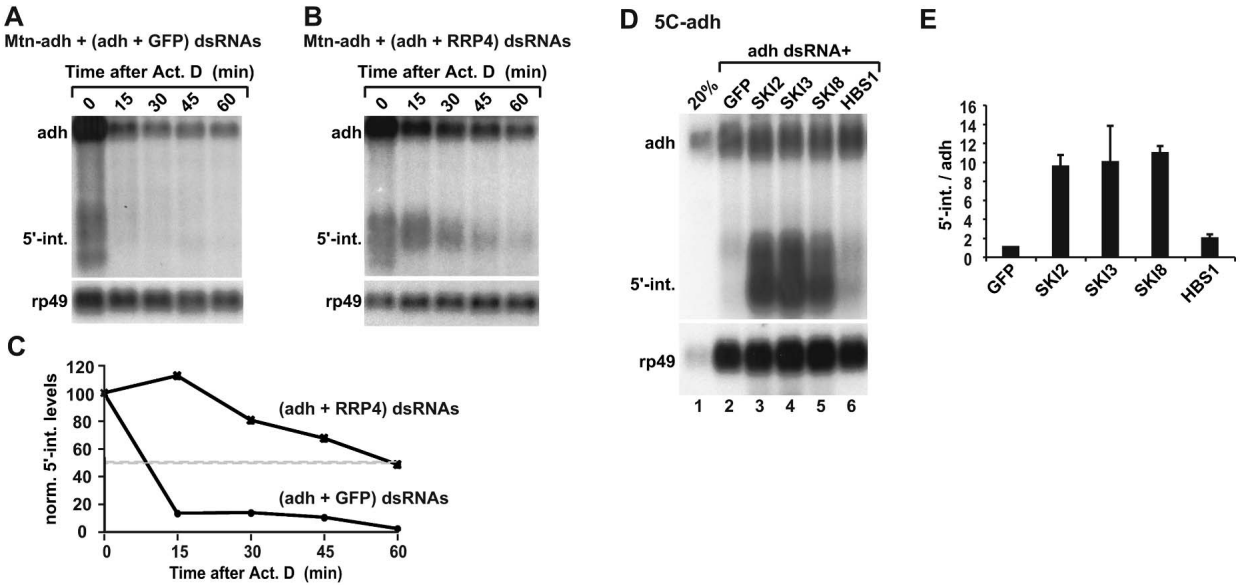


FIGURE 6. Degradation of the 5′ decay intermediates. (*A*,*B*) S2 cell-lines expressing Mtn-adh were treated with adh and RRP4 dsRNAs. Expression of adh mRNA was induced for 45 min. Next, transcription was inhibited with actinomycin D (5 μg/mL) for the times indicated above the lanes. Total RNA samples were isolated and analyzed by Northern blot with a 5′ probe. (*C*) The levels of the 5′ fragments normalized to those of rp49 mRNA shown in panels *A*,*B* are plotted as a function of time. (*D*) S2 cell lines expressing 5C-adh were transfected with the dsRNAs indicated above the lanes. RNA samples were analyzed by Northern blot, as described in Figure 1. The steady-state levels of the 5′ decay intermediates were quantitated and normalized to those of the full-length transcript.

with cycloheximide for 45 min, and the steady-state levels of the intermediates were quantitated and normalized to those of the full-length mRNA. In the presence of cycloheximide, the levels of the 5′ fragment increased three- to fivefold relative to the levels of rp49 (or adh) mRNAs, whereas the levels of the 3′ fragments remained unchanged (Fig. 8A–C). This suggests that efficient decay of the 5′ fragment requires ongoing translation.

DISCUSSION

In this study we analyzed the pathway by which mRNAs targeted by RNAi are degraded. We show that following siRNA-guided endonucleolytic cleavage by RISC, the resulting 5′ and 3′ mRNA fragments are degraded from the ends generated by RISC activity without decapping or deadenylation. Mechanistic studies of RISC-mediated cleavage have revealed that the reaction requires Mg^{2+} ions and leaves 3′-OH and 5′ phosphate termini (Martinez and Tuschl 2004; Schwarz et al. 2004). This is compatible with the observation that the 5′ and 3′ RNA fragments generated by RISC-cleavage are degraded by the exosome and XRN1, respectively. Indeed, the exosome has a strong preference for RNAs with 3′ hydroxyl groups (Mitchell et al. 1997), and RNAs having 5′ phosphate groups are the preferred substrates of XRN1 (Stevens 2001).

The decay pathway of mRNAs targeted by RISC parallels nonsense-mediated mRNA decay (NMD) in *Drosophila*. Indeed, in this organism, degradation of PTC-containing messages is initiated by endonucleolytic cleavage in the vicinity of the nonsense codon (Gatfield and Izaurralde 2004). The resulting 5′ fragments are rapidly degraded by the exosome, and this requires at least SKI2, while the 3′ fragments are degraded by XRN1 (Gatfield and Izaurralde 2004). Notably, the 5′ fragments derived from mRNAs undergoing NMD or RNAi may also enter the nonstop decay (NSD) pathway, as ribosomes engaged in translating these fragments would be stalled at their 3′ ends. This raises the question of how the mRNA surveillance and RNAi machineries interact with general RNA decay enzymes and recruit them specifically to their targeted mRNAs.

Finally, the accumulation of the mRNA fragments flanking the region targeted by dsRNAs in cells depleted of the exosome or XRN1 confirms the observation that in *Drosophila* there is no amplification of the silencing triggers, and only the region complementary to the dsRNA is cleaved by RISC (Schwarz et al. 2002; Roignant et al. 2003).

Drosophila SKI2, SKI3, and SKI8 are required for exosome-mediated decay of the 5′ intermediates

The yeast proteins Ski2p, Ski3p, and Ski8p form a complex required for exosome-mediated 3′-to-5′ mRNA decay in yeast (Anderson and Parker 1998; Brown et al. 2000). The precise mechanism by which these proteins regulate exosome activity is unknown. Ski2p is a putative RNA helicase of the DEVH family (Widner and Wickner 1993). Ski3p is characterized by the presence of 10 tetratricopeptide repeats

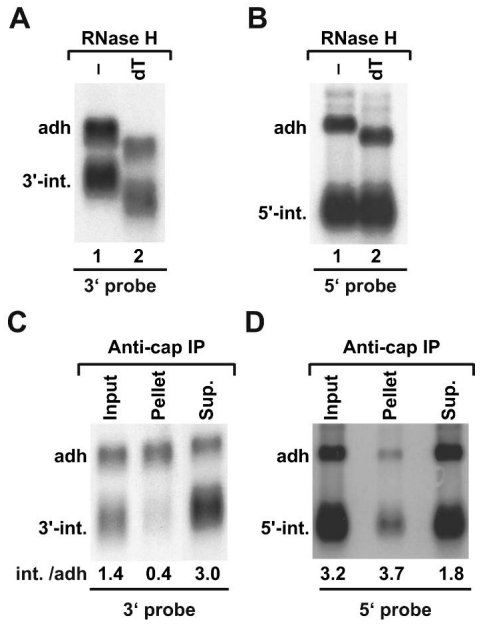


FIGURE 7. The 5′ intermediates are capped, whereas the 3′ intermediates are polyadenylated. Cells expressing the inducible adh reporter were codepleted of SKI2 and XRN1. In (*A*,*B*), total RNA samples were subjected to oligo(dT)-targeted RNase H cleavage. In (*C*,*D*), total RNA samples were co-immunoprecipitated with anti-cap antibodies. Northern blots were hybridized with 5′ or 3′ probes as indicated. The positions of the 3′ and 5′ intermediates are shown. In *C*, one-tenth of the inputs and supernatants and 100% of the immunoprecipitates were analyzed by Northern blot. In *D*, 100% of the inputs, supernatants, and pellets was loaded. The levels of the intermediates relative to those of the full-length transcript are indicated below the lanes.

(TPRs; Rhee et al. 1989), 34-amino-acid α-helical repeats known to mediate protein:protein interactions (D'Andrea and Regan 2003). Ski8p (also known as Rec14) folds into a seven-bladed propeller, a fold thought to serve as a scaffold for the assembly of multimeric protein complexes (Cheng et al. 2004; Madrona and Wilson 2004). In addition to its role in mRNA decay, Ski8p plays a role in the formation of meiotic DNA double-strand breaks and associates with chromosomes during meiosis (Arora et al. 2004). These two roles of Ski8p are conserved in fungi, but can be genetically separated (Evans et al. 1997; Tessé et al. 2003; Arora et al. 2004).

Putative homologs of the yeast Ski proteins have been identified in metazoa, but whether they have similar functions as in yeast remains unresolved. In this regard, human Ski2 (known as Ski2w) is mainly nucleolar (Qu et al. 1998) while yeast Ski2p is predominantly cytoplasmic (Brown et al. 2000), suggesting divergent functions. In this study we characterized the components of the putative Ski complex in *Drosophila*. We show that SKI2, SKI3, and SKI8 are all required for exosome-mediated decay of 5′ mRNA fragments generated by RISC cleavage.

The observation that degradation of the 5′ intermediates is prevented in cells depleted of SKI2, SKI3, or SKI8, despite the presence of the exosome, suggests that these proteins are essential for exosome activity or for its recruitment to the mRNA, as has been suggested for their yeast homologs (Anderson and Parker 1998; van Hoof et al. 2000; Araki et al. 2001). Further biochemical and functional studies are required to determine the precise composition and role of the Ski complex in metazoa.

In addition to the exosome and the Ski complex, the yeast protein Ski7p is also required for 3′-to-5′ mRNA decay and has been implicated in both the NMD and the NSD pathways (van Hoof et al. 2000, 2002; Araki et al. 2001; Taka-

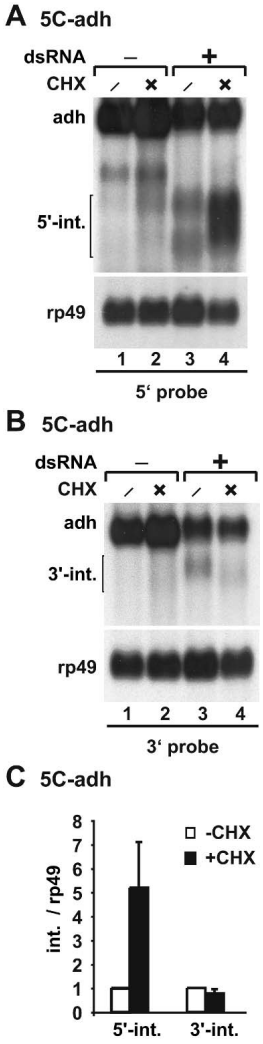


FIGURE 8. Degradation of the 5′ fragments requires ongoing translation. (*A*,*B*) Cells expressing 5C-adh were transfected with adh (and GFP) dsRNA as indicated (lanes *3*,*4*). Translation was inhibited with cycloheximide (CHX, 100 μg/mL) for 45 min (lanes *2*,*4*). Total RNA samples were isolated and analyzed by Northern blot with probes detecting the 5′ or the 3′ decay intermediates. Lanes *1* and *2* show that the intermediates are not observed in treated (CHX+) or untreated (CHX−) cells in the absence of adh dsRNA. (*C*) The levels of the intermediates were normalized to those of the rp49 mRNA in at least three independent experiments. These values were set to 1 in untreated cells (−CHX). Mean values ± SD are shown.

hashi et al. 2003). Ski7p interacts with the exosome and the Ski complex via its unique N-terminal domain (Araki et al. 2001). This domain is sufficient for general 3′-to-5′ mRNA degradation (Araki et al. 2001). In contrast, the C-terminal domain of Ski7p, which is structurally related to the GTPase domains of EF1A, eRF3 and Hbs1p, is dispensable for 3′-to-5′ mRNA decay but is necessary for decay of transcripts lacking a stop codon (Araki et al. 2001; van Hoof et al. 2002). These observations lead to a model in which the C-terminal domain of Ski7p binds to the A site of the ribosomes stalled at the 3′-end of nonstop mRNAs, while the N-terminal domain would recruit the exosome and the Ski complex to target the aberrant transcript for decay.

Surprisingly, there are no obvious orthologs of Ski7p in higher eukaryotes. This has raised the possibility that HBS1, the closest homolog of Ski7p, could have evolved additional functions in metazoa. However, we show that *Drosophila* HBS1 is not involved in decay of mRNA fragments generated by RISC. Similarly, *Drosophila* HBS1 is not implicated in the decay of mRNAs degraded by the NMD pathway (D. Gatfield and E. Izaurralde, unpubl.). It would be of interest to determine how ribosomes translating mRNA fragments lacking termination codons are released from the 3′-end of the transcript in higher eukaryotes.

Recruitment of decay enzymes and recycling of RISC

RISC can catalyze multiple rounds of mRNA cleavage (Hutvágner and Zamore 2002). How RISC dissociates from the mRNA after cleavage is unknown. Presumably, this process involves unwinding of the siRNA/mRNA duplexes. Unwinding would also be required to release the mRNA fragments, so that their ends are exposed to the exonucleolytic activities of the exosome and XRN1. Several RNA helicases have been implicated in RNAi, but with the exception of Armitage, their precise role in this process remains unresolved (for review, see Meister and Tuschl 2004). Alternatively, RISC release could be facilitated or may require degradation of the mRNA products. In this regard, RISC components may interact with factors involved in mRNA decay and promote their recruitment to the targeted mRNA. It is also possible that RISC components remain bound to the 3′ decay intermediates and escort them to P-bodies where the intermediates decay rapidly and the RISC complex would dissociate and be recycled. Whether decay of the 3′ mRNA fragments generated by RISC occurs in the cytoplasm or in mRNA decay bodies, however, remains to be established.

MATERIALS AND METHODS

Construction of RNAi reporters

The adh reporters have been described (Gatfield et al. 2003; Gatfield and Izaurralde 2004). To generate the NXF1 reporters, a cDNA fragment encoding human NXF1 was cloned into the EcoRI-SalI sites of vector pRmHa containing the metallothionein promoter (Bunch et al. 1988).

Cell culture and RNA interference

Establishment of stable cell lines expressing the RNAi reporters and RNAi were performed as described (Gatfield et al. 2003; Gatfield and Izaurralde 2004); dsRNAs corresponding to *Drosophila* XRN1 (*pcm*, CG3291), RRP4 (CG3931), CSL4 (CG6249), SKI2 (*tst*, CG10210), SKI3 (CG8777), SKI8 (CG3909), and HBS1 (CG1898) were transcribed in vitro from ~700-bp cDNA fragments amplified from a Schneider cell cDNA library.

RNA analysis

RNA preparation and Northern blot analysis were performed essentially as described by Gatfield et al. (2003). RNAs from inducible cell lines were harvested 45 min after induction with 0.5 mM copper sulfate. For determination of RNA half-lives, control or depleted cells were incubated with actinomycin D (5 μg/mL) for the times indicated in the figures. Total RNA samples were analyzed by Northern blot. The levels of the reporters and intermediates were quantitated and normalized to the levels of rp49 mRNA (whose levels relative to 18S rRNA do not change within the time-frame of the experiments).

Body-labeled DNA probes were generated by unidirectional PCR or random priming according to standard protocols. RNase H (USB) digestion using a $(dT)_{15}$ oligonucleotide was performed according to the manufacturer's instructions.

Immunoprecipitation of capped mRNAs was performed from total RNA preparations using the monoclonal antibody H20 in NET buffer (50 mM Tris-HCl pH 7.5, 150 mM NaCl, 0.1% NP40, 1 mM EDTA).

ACKNOWLEDGMENTS

We thank David Gatfield and Kevin Czaplinski for helpful discussions and Reinhard Lührmann for the anti-cap antibodies. This study was supported by the European Molecular Biology Organisation (EMBO) and the Human Frontier Science Program Organization (HFSPO). T.I.O. is a recipient of the Hungarian State Eötvös fellowship.

Received November 9, 2004; accepted December 20, 2004.

REFERENCES

Ambros, V. 2004. The functions of animal microRNAs. *Nature* **431:** 350–355.

Anderson, J.S. and Parker, R.P. 1998. The 3′ to 5′ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3′ to 5′ exonucleases of the exosome complex. *EMBO J.* **17:** 1497–1506.

Araki, Y., Takahashi, S., Kobayashi, T., Kajiho, H., Hoshino, S., and Katada, T. 2001. Ski7p G protein interacts with the exosome and the Ski complex for 3′-to-5′ mRNA decay in yeast. *EMBO J.* **20:** 4684–4693.

Arora, C., Kee, K., Maleki, S., and Keeney, S. 2004. Antiviral protein Ski8 is a direct partner of Spo11 in meiotic DNA break formation, independent of its cytoplasmic role in RNA metabolism. *Mol. Cell*

13: 549–559.
Baulcombe D. 2004. RNA silencing in plants. *Nature* **409:** 356–363.
Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409:** 363–366.
Brown, J.T., Bai, X., and Johnson, A.W. 2000. The yeast antiviral proteins Ski2p, Ski3p, and Ski8p exist as a complex in vivo. *RNA* **6:** 449–457.
Bunch, T. A., Grinblat, Y., and Goldstein, L.S. 1988. Characterization and use of the *Drosophila* metallothionein promoter in cultured *Drosophila melanogaster* cells. *Nucleic Acids Res.* **16:** 1043–1061.
Carmell, M.A., Xuan, Z., Zhang, M.Q., and Hannon, G.J. 2002. The Argonaute family: Tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis. *Genes & Dev.* **16:** 2733–2742.
Cerutti, L., Mian, N., and Bateman, A. 2000. Domains in gene silencing and cell differentiation proteins: The novel PAZ domain and redefinition of the Piwi domain. *Trends Biochem. Sci.* **25:** 481–482.
Cheng, Z., Liu, Y., Wang, C., Parker, R., and Song, H. 2004. Crystal structure of Ski8p, a WD-repeat protein with dual roles in mRNA metabolism and meiotic recombination. *Protein Sci.* **13:** 2673–2684.
Cougot, N., Babajko, S., and Seraphin, B. 2004. Cytoplasmic foci are sites of mRNA decay in human cells. *J. Cell Biol.* **165:** 31–40.
D'Andrea, L.D. and Regan, L. 2003. TPR proteins: The versatile helix. *Trends Biochem. Sci.* **28:** 655–662.
Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Dev.* **15:** 188–200.
Evans, D.H., Li, Y.F., Fox, M.E., and Smith, G.R. 1997. A WD repeat protein, Rec14, essential for meiotic recombination in *Schizosaccharomyces pombe*. *Genetics* **146:** 1253–1264.
Eystathioy, T., Jakymiw, A., Chan, E.K., Seraphin, B., Cougot, N., and Fritzler, M.J. 2003. The GW182 protein colocalizes with mRNA degradation associated proteins hDcp1 and hLSm4 in cytoplasmic GW bodies. *RNA* **9:** 1171–1173.
Frischmeyer, P.A., van Hoof, A., O'Donnell, K., Guerrerio, A.L., Parker, R., and Dietz, H.C. 2002. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. *Science* **295:** 2258–2261.
Gatfield, D. and Izaurralde, E. 2004. Nonsense-mediated messenger RNA decay is initiated by endonucleolytic cleavage in *Drosophila*. *Nature* **429:** 575–578.
Gatfield, D., Unterholzner, L., Ciccarelli, F.D., Bork, P., and Izaurralde, E. 2003. Nonsense-mediated mRNA decay in *Drosophila*: At the intersection of the yeast and mammalian pathways. *EMBO J.* **22:** 3960–3970.
Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106:** 23–34.
Hammond, S.M., Boettcher, S., Caudy, A.A., Kobayashi, R., and Hannon, G.J. 2001. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science* **293:** 1146–1150.
Hutvágner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297:** 2056–2060.
Hutvágner, G., McLachlan, J., Pasquinelli, A.E., Balint, E., Tuschl, T., and Zamore, P.D. 2001. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293:** 834–838.
Ingelfinger, D., Arndt-Jovin, D.J., Luhrmann, R., and Achsel, T. 2002. The human LSm1–7 proteins colocalize with the mRNA-degrading enzymes Dcp1/2 and Xrnl in distinct cytoplasmic foci. *RNA* **8:** 1489–1501.
Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. 2003. Structure and nucleic-acid binding of the *Drosophila* Argonaute 2 PAZ domain. *Nature* **426:** 465–469.
———. 2004. Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. *Nat. Struct. & Mol. Biol.* **11:** 576–577.
Lippman, Z. and Martienssen, R. 2004. The role of RNA interference in heterochromatic silencing. *Nature* **431:** 364–370.
Liu, J., Carmell, M.A., Rivas, F.V., Marsden, C.G., Thomson, J.M., Song, J.J., Hammond, S.M., Joshua-Tor, L., and Hannon, G.J. 2004. Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305:** 1437–1441.
Ma, J.B., Ye, K., and Patel, D.J. 2004. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429:** 318–322.
Madrona, A.Y. and Wilson, D.K. 2004. The structure of Ski8p, a protein regulating mRNA degradation: Implications for WD protein structure. *Protein Sci.* **13:** 1557–1565.
Martinez, J. and Tuschl, T. 2004. RISC is a 5′ phosphomonoester-producing RNA endonuclease. *Genes & Dev.* **18:** 975–980.
Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110:** 563–574.
Meister, G. and Tuschl, T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431:** 343–349.
Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* **15:** 185–197.
Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M., and Tollervey, D. 1997. The exosome: A conserved eukaryotic RNA processing complex containing multiple 3′→5′ exoribonucleases. *Cell* **91:** 457–466.
Nykänen, A., Haley, B., and Zamore, P.D. 2001. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107:** 309–321.
Okamura, K., Ishizuka, A., Siomi, H., and Siomi, M.C. 2004. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes & Dev.* **18:** 1655–1666.
Parker, R. and Song, H. 2004. The enzymes and control of eukaryotic mRNA turnover. *Nat. Struct.& Mol. Biol.* **11:** 121–127.
Plasterk, R.H. 2002. RNA silencing: The genome's immune system. *Science* **296:** 1263–1265.
Qu, X., Yang, Z., Zhang, S., Shen, L., Dangel, A.W., Hughes, J.H., Redman, K.L., Wu, L.C., and Yu, C.Y. 1998. The human DEVH-box protein Ski2w from the HLA is localized in nucleoli and ribosomes. *Nucleic Acids Res.* **26:** 4068–4077.
Rand, T.A., Ginalski, K., Grishin, N.V., and Wang, X. 2004. Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity. *Proc. Natl. Acad. Sci.* **101:** 14385–14389.
Rhee, S.K., Icho, T., and Wickner, R.B. 1989. Structure and nuclear localization signal of the SKI3 antiviral protein of *Saccharomyces cerevisiae*. *Yeast* **5:** 149–158.
Roignant, J.Y., Carre, C., Mugat, B., Szymczak, D., Lepesant, J.A., and Antoniewski, C. 2003. Absence of transitive and systemic pathways allows cell-specific and isoform-specific RNAi in *Drosophila*. *RNA* **9:** 299–308.
Schwarz, D.S., Hutvagner, G., Haley, B., and Zamore, P.D. 2002. Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways. *Mol. Cell* **10:** 537–548.
Schwarz, D.S., Tomari, Y., and Zamore, P.D. 2004. The RNA-induced silencing complex is a Mg2+-dependent endonuclease. *Curr. Biol.* **14:** 787–791.
Sheth, U. and Parker, R. 2003. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300:** 805–808.
Song, J.J., Liu, J., Tolia, N.H., Schneiderman, J., Smith, S.K., Martienssen, R.A., Hannon, G.J., and Joshua-Tor, L. 2003. The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat. Struct. Biol.* **10:** 1026–1032.
Song, J.J., Smith, S.K., Hannon, G.J., and Joshua-Tor, L. 2004. Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* **305:** 1434–1437.
Souret, F.F., Kastenmayer, J.P., and Green, P.J. 2004. AtXRN4 degrades mRNA in Arabidopsis and its substrates include selected

miRNA targets. *Mol. Cell* **15:** 173–183.
Stevens, A. 2001. 5′-exoribonuclease 1: Xrn1. *Methods Enzymol.* **342:** 251–259.
Tabara, H., Sarkissian, M., Kelly, W.G., Fleenor, J., Grishok, A., Timmons, L., Fire, A., and Mello, C.C. 1999. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* **99:** 123–132.
Takahashi, S., Araki, Y., Sakuno, T., and Katada, T. 2003. Interaction between Ski7p and Upf1p is required for nonsense-mediated 3′-to-5′ mRNA decay in yeast. *EMBO J.* **22:** 3951–3959.
Tessé, S., Storlazzi, A., Kleckner, N., Gargano, S., and Zickler, D. 2003. Localization and roles of Ski8p protein in Sordaria meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. *Proc. Natl. Acad. Sci.* **100:** 12865–12870.
van Dijk, E., Cougot, N., Meyer, S., Babajko, S., Wahle, E., and Seraphin, B. 2002. Human Dcp2: A catalytically active mRNA decapping enzyme located in specific cytoplasmic structures. *EMBO J.* **21:** 6915–6924.
van Hoof, A., Staples, R.R., Baker, R.E., and Parker, R. 2000. Function of the ski4p (Csl4p) and Ski7p proteins in 3′-to-5′ degradation of mRNA. *Mol. Cell. Biol.* **20:** 8230–8243.
van Hoof, A., Frischmeyer, P.A., Dietz, H.C., and Parker, R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* **295:** 2262–2264.
Widner, W.R. and Wickner, R.B. 1993. Evidence that the SKI antiviral system of *Saccharomyces cerevisiae* acts by blocking expression of viral mRNA. *Mol. Cell. Biol.* **13:** 4331–4341.
Williams, B.R. 1999. PKR; a sentinel kinase for cellular stress. *Oncogene* **18:** 6112–6120.
Yan, K.S., Yan, S., Farooq, A., Han, A., Zeng, L., and Zhou, M.M. 2003. Structure and conserved RNA binding of the PAZ domain. *Nature* **426:** 468–474.

STEM CELLS®

orban.tamas_188_24

Applying a "Double-Feature" Promoter to Identify Cardiomyocytes Differentiated from Human Embryonic Stem Cells Following Transposon-Based Gene Delivery

TAMÁS I. ORBÁN,[a] ÁGOTA APÁTI,[a] ANDREA NÉMETH,[a] NÓRA VARGA,[a] VIRÁG KRIZSIK,[a] ANITA SCHAMBERGER,[a] KORNÉLIA SZEBÉNYI,[a] ZSUZSA ERDEI,[a] GYÖRGY VÁRADY,[a] ÉVA KARÁSZI,[a] LÁSZLÓ HOMOLYA,[a] KATALIN NÉMET,[a] ELEN GÓCZA,[b] CSABA MISKEY,[c] LAJOS MÁTÉS,[c] ZOLTÁN IVICS,[c] ZSUZSANNA IZSVÁK,[c] BALÁZS SARKADI[a]

[a]Membrane Research Group of the Hungarian Academy of Sciences, Semmelweis University and National Blood Center, Budapest, Hungary; [b]Genetic Modification Program Group, Agricultural Biotechnology Center, Gödöllő, Hungary; [c]Mobile DNA Group, Max-Delbrück Center for Molecular Medicine, Berlin, Germany

Key Words. Sleeping Beauty transposon • Human embryonic stem cells • CAG promoter • ''Double-feature'' promoter • Cardiomyocytes • Lentiviral gene delivery

ABSTRACT

Human embryonic stem (HuES) cells represent a new potential tool for cell-therapy and gene-therapy applications. However, these approaches require the development of efficient, stable gene delivery, and proper progenitor cell and tissue separation methods. In HuES cell lines, we have generated stable, enhanced green fluorescent protein (EGFP)-expressing clones using a transposon-based (*Sleeping Beauty*) system. This method yielded high percentage of transgene integration and expression. Similarly to a lentiviral expression system, both the undifferentiated state and the differentiation pattern of the HuES cells were preserved. By using the CAG promoter, in contrast to several other constitutive promoter sequences (such as CMV, elongation factor 1α, or phosphoglycerate kinase), an exceptionally high EGFP expression was observed in differentiated cardiomyocytes. This phenomenon was independent of the transgene sequence, methods of gene delivery, copy number, and the integration sites. This "double-feature" promoter behavior, that is providing a selectable marker for transgene expressing undifferentiated stem cells, and also specifically labeling differentiated cardiomyocytes, was assessed by transcriptional profiling. We found a positive correlation between CAG promoter-driven EGFP transcription and expression of cardiomyocyte-specific genes. Our experiments indicate an efficient applicability of transposon-based gene delivery into HuES cells and provide a novel approach to identify differentiated tissues by exploiting a nontypical behavior of a constitutively active promoter, thereby avoiding invasive drug selection methods. STEM CELLS *2009;27:1077–1087*

Disclosure of potential conflicts of interest is found at the end of this article.

INTRODUCTION

The application of human embryonic stem (HuES) cells provides new hopes in the clinical treatment of a number of diseases and, at the same time, these stem cells are excellent models for studying tissue development and physiological cell differentiation. The so called "regenerative medicine" makes use of cells that can grow and differentiate to replace a damaged tissue. Also, efficient and stable gene delivery into stem cells should form the basis of successful gene therapy applications. Currently, the most widely applied methods for gene delivery into various stem/progenitor cells are based on the use of viral vector constructs. By now, there are numerous efficient retrovirus- or lentivirus-based methods which allow stable genomic incorporation of the foreign DNAs with high gene product expression levels. However, virus-based gene therapy technologies also have serious draw-backs, including safety concerns of virus production, and the preferential incorporation of foreign genes into active host gene loci, which may cause uncontrolled proliferation of the gene-modified stem cells [1, 2]. On the other hand, although non-viral gene delivery techniques are usually considered to be less efficient, with the emergence and refinement of the transposon-based

Author contributions: T.I.O.: conception and design; collection and assembly of data; data analysis and interpretation; manuscript writing; Á.A.: conception and design; collection and assembly of data; data analysis and interpretation; manuscript writing; A.N.: collection and assembly of data; N.V.: collection and assembly of data; V.K.: collection and assembly of data; A.S.: collection and assembly of data; K.S.: collection and assembly of data; Z.E.: collection and assembly of data; G.V.: collection and interpretation of data; É.K.: collection of data; L.H.: interpretation of data; K.N.: collection and interpretation of data; E.G.: interpretation of data; C.M.: data analysis; L.M.: interpretation of data; Z. Ivics: data analysis; manuscript writing; Z. Izsvák: data analysis and interpretation of data; B.S.: conception and design; financial support; collection and assembly of data; data analysis and interpretation; manuscript writing; final approval of manuscript; T.I.O. and Á.A.: contributed equally to this article.

Correspondence: Balázs Sarkadi, MD, PhD, Membrane Research Group of the Hungarian Academy of Sciences, Semmelweis University and National Blood Center, Dioszegi u. 64., Budapest, H-1113, Hungary. Telephone: +36-1-372-4353; Fax: +36-1-372-4353; e-mail: sarkadi@biomembrane.hu Received December 17, 2008; accepted for publication February 11, 2009; first published online in STEM CELLS *EXPRESS* February 20, 2009. 1066-5099/2009/$30.00/0 doi: 10.1002/stem.45

methods, they certainly represent a plausible alternative to viral applications [3–6].

Current methods for directed tissue differentiation from various HuES cells often apply endogenous morphogenic proteins or invasive chemicals to obtain the tissue(s) of interest [7–9]. However, the use of such artificial cocktail of drugs may affect natural differentiation and, among other problems, may induce undesired gene expression profiles. In addition, it is often advisable to examine whether tissues derived after such invasive attempts reliably represent naturally occurring cell types [10]. A common way of overcoming this problem is to use a combination of tissue-specific promoters and certain drug resistance genes as selection markers. This method can provide the solution for obtaining the desired cell types, as it is based on "spontaneous" differentiation of HuES cells (see [11]) and selecting for the differentiated cell types in a later phase, e.g., by certain antibiotics [12, 13]. In principle, this could result in the enrichment of desired tissues in large quantities and most likely provide less interference with the gene expression profiles of the tissues of interest. However, it still requires drug selection which may induce an altered way of differentiation, again asking for the need to prove that the obtained tissues represent naturally occurring cell types [14].

The use of tissue specific promoters generates another technical problem, as in this case transgene expression is not detectable in the undifferentiated HuES cells [15]. Especially for non-viral applications, where the efficiency of gene delivery is generally low, this represents a serious disadvantage. To avoid this, one can use another, constitutive promoter to drive the expression of an additional selection marker gene, and first select for transgene delivery, then, after spontaneous differentiation, for specific tissues. Apart from being cumbersome and time-consuming, this also creates another technically challenging problem, requiring either the codelivery of different transgenes or the use of a larger cargo vector, often resulting in lower delivery efficiency.

Delivering transgenes into HuES cells also requires a careful design as to which promoters are to be used for yielding high expression levels. Previous studies are controversial concerning the applicability of widely used transgene promoters. The cytomegalovirus immediate early 1 (CMV) promoter has been claimed to be a useful system for expressing transgenes in undifferentiated mouse embryonic stem (ES) cells [16]. Some studies, however, found that the CMV promoter has only a moderate activity when compared with the elongation factor 1α (EF1α) promoter [17]. Other investigations revealed that the CMV promoter is strongly silenced in ES cells by means of epigenetic modification via DNA methylation [18, 19]. Chung et al. [20] carried out a systematic study in comparing three promoters in undifferentiated and differentiated mouse ES cells. They demonstrated that the EF1α and the CMV-chicken β-actin fusion (a version of CAG-CMV-Actin-Globin) promoters yielded robust gene expression in undifferentiated mouse ES cells and effective transgene expression both in embryoid bodies and differentiated neuronal precursors. On the other hand, the CMV promoter was only active in the neuronal precursors [20]. Similar systematic study was carried out in human ES cells, using a lentiviral transduction method [21]. It was demonstrated that the CAG promoter drove gene expression in the ES cells more efficiently than the (strongly silenced) CMV promoter, whereas the EF1α and the phosphoglycerate kinase (PGK) promoters showed much stronger transgene expression and were thereby claimed to be the promoters of choice for HuES cells. Nevertheless, it is still unsettled which promoter is to be used in ES cells, especially if one needs to have strong transgene expression in certain types of tissues, originated from the genetically modified ES cells.

In this study, we have developed a non-viral methodology for efficient and stable gene delivery into HuES cells. We applied *Sleeping Beauty* (SB) transposition [4, 22] for non-viral gene delivery and expression, examined the efficiency of stable gene delivery, as well as the expression level and the effect of a reporter gene construct on early stem cell differentiation. In this system, enhanced green fluorescent protein (EGFP)-expressing SB transposon-based vectors were delivered into HuES cells together with a hyperactive version of the SB transposase that shows ~32-fold higher activity than the originally reported first-generation transposase [3]. By comparing this system to a lentiviral gene delivery into the HuES cells, we provide a detailed description of the two stable gene modification methods.

In addition, we have developed an efficient and relatively simple genetic system to selectively enrich transgene-expressing cardiomyocytes differentiated from HuES cells. We found that one version of the widely used CAG promoter exhibits a previously unknown "double-feature" characteristic: it can be used to mark genetically modified undifferentiated HuES cells and it also provides a selection platform specifically for cardiomyocytes during subsequent spontaneous differentiation. We documented that this unusual behavior is independent of gene delivery methods, and the sequence, copy number, and chromosomal integration site of the transgene.

Materials and Methods

Embryonic Stem Cell Culturing and Differentiation

The HuES cell lines HUES9 (originally provided by Dr. Douglas Melton, Harvard University) and BG01V (from ATCC) were maintained essentially as described earlier [23]; cells from passage no. 35 were used for these analyses. Differentiation of the HuES cells were initiated spontaneously via the embryoid body (EB) formation pathway [23]. Cell types were identified by morphological signs under phase contrast light microscopy and by immunostaining for various protein markers.

Promoter and Vector Constructs

Plasmids used in this project contained the cDNA of the highly fluorescent marker amaxaGFP or EGFP (Fig. 1A; the original CAG-amaxaGFP transposon plasmid was kindly provided by Evelyn Zeira via collaboration in the EU FP6 INTHER project). By a Hind-III and NheI restriction digestion followed by ligation, we cloned the CMV viral promoter, the CAG promoter, the human PGK promoter or the short version of the human EF1α promoter upstream of the transgene in the transposon vectors (Fig. 1B). For the lentiviral constructs, the used transgene cassettes were exactly the same, as in all cases the entire transcription unit was removed from the transposon vector by restriction digestion and inserted into the viral vector by blunt end ligation (Fig. 1C).

Transposon-Based Gene Delivery

For the SB transposase, we used an enhanced version of the enzyme having 32 times higher activity than the originally reconstructed transposase ([3] and unpublished results); for transposition control, we applied a DDE inactive mutant form of the enzyme [22]. For transfection, the FuGENE® 6 (Roche Applied Science, Rotkreuz, Switzerland, http://www.roche-applied-science.com) reagent was used according to the manufacturer's instruction. Before transfection, HuES cells were

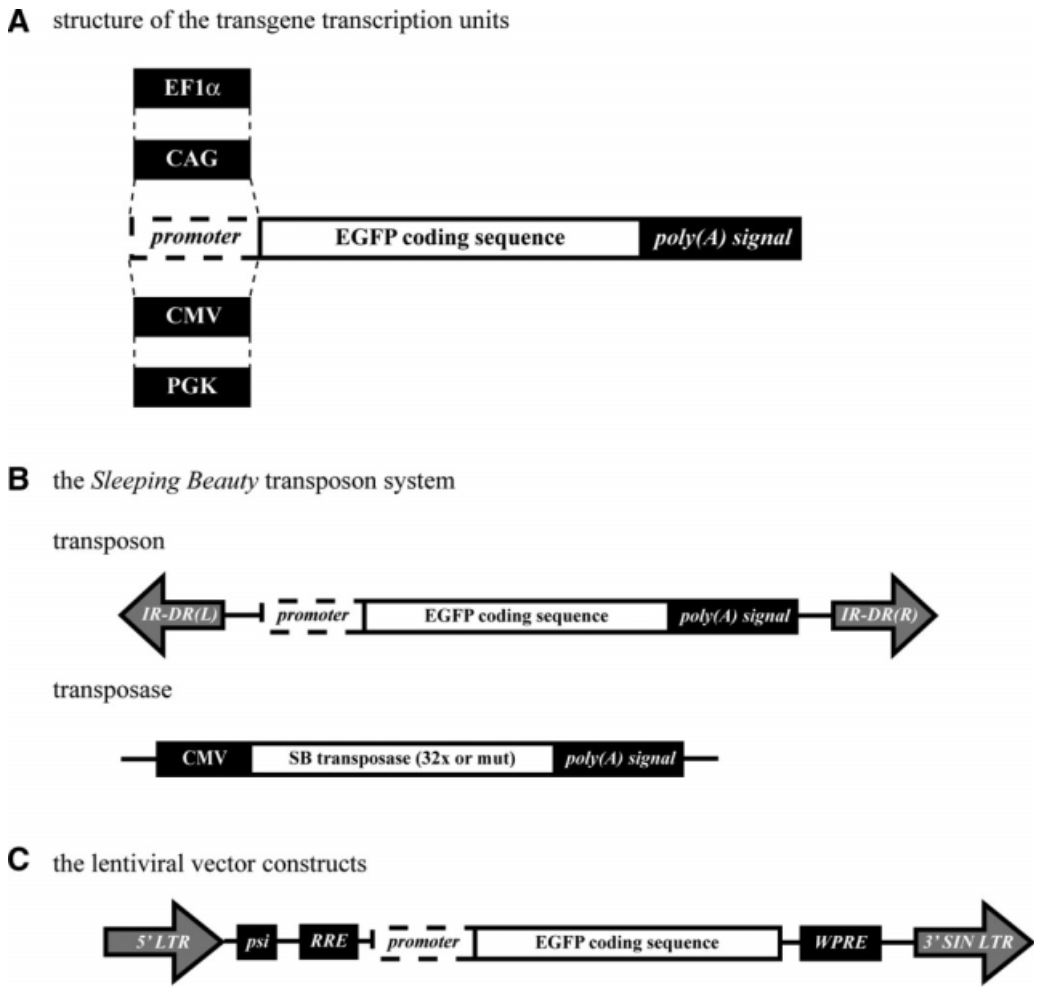


Figure 1. Vector constructs. **(A):** Four different promoters were used to drive transgene expression. **(B):** Structure of the cotransfected plasmids for the *Sleeping Beauty* transposon system. IR-DR(L) and (R) stand for left and right inverted repeat–direct repeat transposon sequence. **(C):** Structure of the lentiviral constructs. Psi and RRE are elements required for viral packaging. Sizes of the different elements are not drawn to scale. Abbreviations: CAG, CMV enhancer-chicken β-actin-rabbit β1-globin fusion; CMV, cytomegalovirus immediate early 1; EF1α, elongation factor 1α; EGFP, enhanced green fluorescent protein; IR-DR(L), inverted repeat-direct repeat, left; IR-DR(R), inverted repeat-direct repeat, right; LTR, (viral) long terminal repeat; PGK, phosphoglycerate kinase; SIN, self-inactivating (virus); SB, Sleeping Beauty; WPRE, woodchuck hepatitis B virus posttranscriptional regulatory element.

separated from the mouse feeder cells and placed on gelatin-coated plates. On the subsequent day, the cells were cotransfected with transposon and transposase plasmids in a 10:1 ratio to avoid overproduction inhibition of the transposase [4, 5]. Next day, the transfected HuES cells were placed back onto the mouse feeder layer to keep them in an undifferentiated state. Although some differentiation could start during the 2 days without feeder cells, in our hands it did not significantly jeopardize the preservation of pluripotency (The current transfection protocol was not efficiently working in a Matrigel system.) To provide evidence that the transposon constructs are capable of transposition, a nested excision PCR method was applied, as described previously. This method amplifies the "footprint" sequence left in the plasmid after transposon excision, whereas no PCR product is obtained if no excision occurs [24]. To determine the integration sites of the transgenes in human genomic DNA, splinkerette PCR and inverse PCR methods were applied, essentially as described earlier [22, 25]. Copy number determination was carried out by real-time PCR using the relative standard curve method with a TaqMan® assay designed specifically for the right inverted repeat sequences of the SB transposon vector. As a single-copy gene control, a predeveloped TaqMan® assays for the RNA subunit of the RNaseP enzyme, was purchased from the manufacturer; analyses were carried out using the StepOne Real-Time PCR System (Applied Biosystems, Foster City, CA, http://www.appliedbiosystems.com).

Lentiviral Transduction Procedure

For the viral-based gene delivery, the third generation lentiviral vector system was used, as described in [26] (Fig. 1C). Determination of virus titers and transduction procedures were performed essentially as described previously [27]. The HUES cells were transduced by a MOI of 2–5, and the further handling of undifferentiated cells and cell differentiation protocols were the same as for the transposon-based gene delivery system.

Flow Cytometry

EGFP-expressing undifferentiated HuES cell colonies were manually selected by using a fluorescent microscope under sterile conditions, to enrich the marker gene expressing population. Such heterogeneous colonies (Fig. 2A) were harvested from the mouse feeder cells, washed with PBS and were

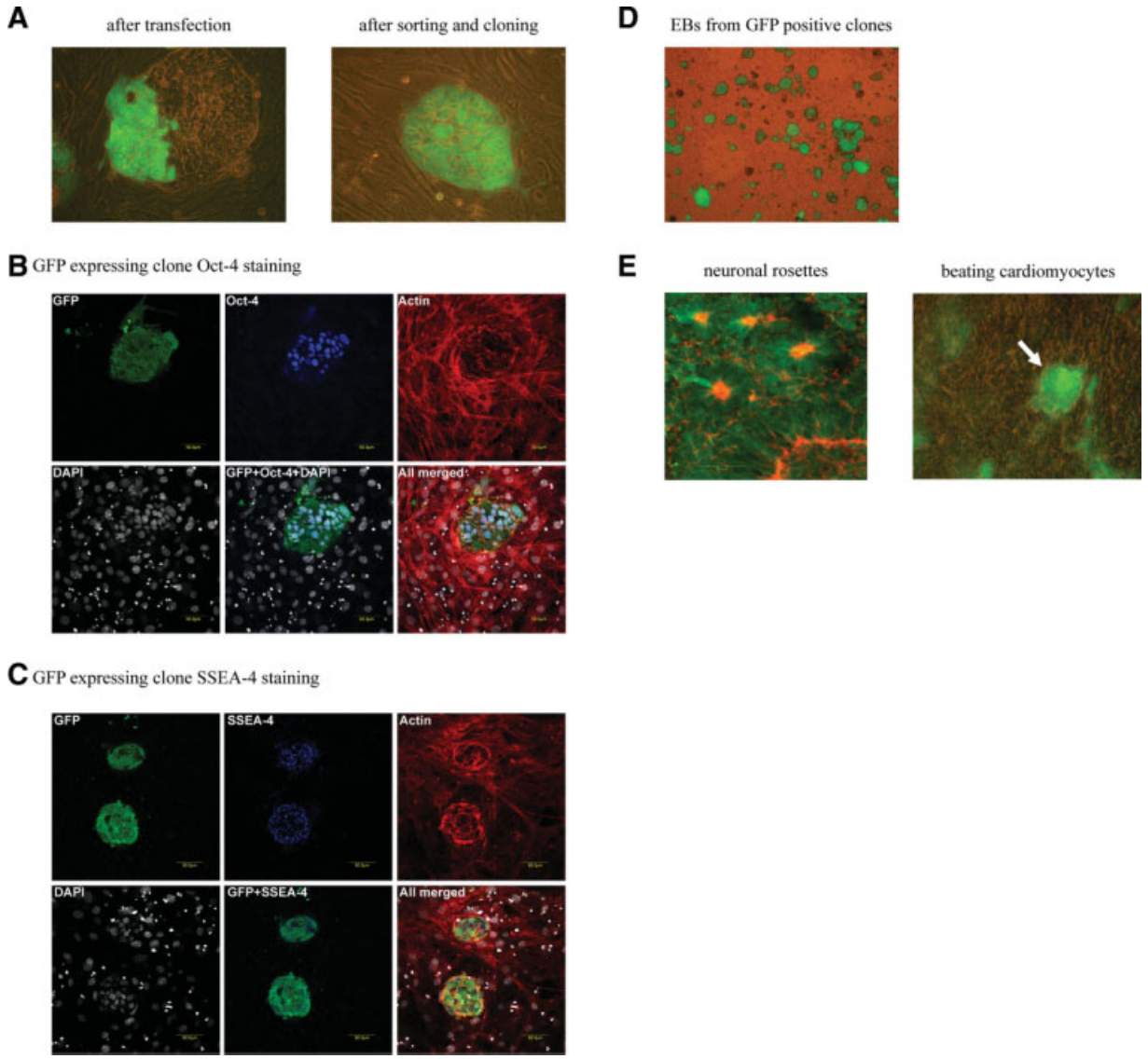


Figure 2. Human embryonic stem cells stably expressing the transgene, exemplified by the CAG-amaxaGFP transgene, delivered by the SB transposon system. **(A):** Morphology and GFP expression of human embryonic stem cell (HUES9) clumps. GFP expression was detected by fluorescence microscopy following transfection (left panel) or after subsequent sorting and cloning of the transgene expressing cells (right panel, magnification: ×200). **(B):** GFP expressing HUES9 clone on mouse feeder cells stained for the Oct-4 embryonic stem cell marker. This transcription factor is localized in the nucleus and it is only expressed in the cells of the HuES clump and not in the surrounding feeder cells. Green: amaxaGFP; blue: Oct-4; red: TRITC-phalloidin, representing actin filaments; white: DAPI staining for the nucleus. Confocal images, scale bars represent 50 μm. **(C):** GFP expressing HUES9 clone on mouse feeders stained for the SSEA-4 embryonic stem cell marker. This plasma membrane protein could only be detected in HuES clump cells and not in the surrounding embryonic feeders. Green: amaxaGFP; blue: SSEA-4; red: TRITC-phalloidin, representing actin filaments; white: DAPI staining for the nucleus. Confocal images, scale bars represent 50 μm. **(D):** Six-days-old EBs formed from GFP expressing HUES9 clones observed by fluorescence microscopy (magnification: ×40). **(E):** Neuronal and myocardial cells differentiated from GFP expressing HUES9 clones. The left panel represents a confocal image of neuronal rosettes (scale bar represents 50 μm). The right panel depicts a typical area with beating cardiomyocytes (white arrow), indicating a prominent difference of amaxaGFP expression between the cardiomyocytes and the other surrounding tissues (phase contrast fluorescence microscopy image, ×40 magnification). Green: GFP fluorescence, red: TRITC-phalloidin, representing actin filaments. Abbreviations: DAPI, 4′,6-diamidino-2-phenylindole; EB, embryoid body; GFP, green fluorescent protein.

sorted based on GFP fluorescence using the fluorescence-based cell sorting (FACS) Aria High Speed Cell Sorter (Beckton-Dickinson, San Jose, CA, http://www.bdbiosciences.com). Mock-transfected HuES cells were measured to set the level for EGFP-positivity with the FACS Diva analysis software; propidium iodide staining was used to gate out the nonviable cells. Sorted EGFP positive single cells were placed onto the mouse feeder cells and monitored until the formation of surviving clones. For further gene expression analysis, differentiated cells from EGFP-expressing HuES clones were sorted into four different artificial fractions based on increasing EGFP fluorescent signal intensity. Cells obtained from different fractions were washed with PBS and immediately resuspended in Trizol (Invitrogen, Carlsbad, CA, http://www.invitrogen.com) for further RNA analysis.

Immunohistochemical Assays

Immunostaining procedures were performed essentially as described previously [23]. For stem cell markers, monoclonal antibodies against Oct-4 (Santa Cruz Biotechnology,

Santa Cruz, CA, http://www.scbt.com), SSEA-4, or podocalyxin (R&D Systems Inc., Minneapolis, MN, http://www.rndsystems.com) were used. TRITC-phalloidin (Sigma-Aldrich, St. Louis, MO, http://www.sigmaaldrich.com) staining was used for the detection of actin filaments, anti-myosin IIA (Santa Cruz Biotechnology) was used for nonmuscle cell detection, DAPI (Invitrogen) was used for nuclear staining, and GFP was evaluated by direct fluorescence. All documented measurements were carried out on cells from four independent GFP-expressing clones, at least in triplicate stainings.

RNA Analysis

RNA isolation was carried out from cells collected in Trizol reagent (Invitrogen) according to the manufacturer's instruction. cDNA samples were prepared from 0.1 μg total RNA using the Promega Reverse Transcription System Kit, as specified by the manufacturer. Tissue specific genes were selected using the TiProD tissue-specific promoter database (tiprod.cbi.pku.edu.cn:8080/index.html). The following markers were selected: OCT-4 and NANOG transcripts as undifferentiated stem cell markers; ACTC (cardiac-specific α-actin), NPPA (natriuretic peptide precursor A) and PLN (phospholamban) genes as cardiac specific markers; PAX6 (paired box gene six) gene as an early marker for neuronal differentiation; CAPG (capping protein (actin filament), gelsolin-like) gene as a skin differentiation marker; P0 ribosomal protein as endogenous controls. Predeveloped real-time TaqMan® assays for the listed genes were purchased from Applied Biosystems. For quantifying EFGP mRNA, specific TaqMan® assay were designed for the cDNA sequence. Real-time PCR analyses were carried out using the StepOne Real-Time PCR System (Applied Biosystems), according to the manufacturer's instructions.

RESULTS

Comparison of the Transposon-Based and Lentiviral Gene Delivery Systems

An important goal in ES cell biology is to create transgenic stem cells with stable transgene expression to provide a source for potentially any type of tissues. The first difficulty to overcome during this process is to achieve a high rate of stable gene delivery into HuES cells. For this purpose, we applied the SB transposon system and tested several constitutive promoters previously shown to drive marker gene expression in ES cell lines, namely, the cytomegalovirus immediate early 1 (CMV) promoter, the EF1α promoter, the PGK promoter, and the artificial CMV enhancer—chicken β-actin–rabbit β1-globin fusion (CAG) promoter. To compare the efficiency and the applicability of viral and non-viral gene delivery, we have also applied a lentiviral-based vector system to express the EGFP marker gene with the promoters mentioned earlier (Fig. 1).

In the case of the transposon system, the lipid-based transfection was less efficient (5%–10%) than viral transduction (in most cases above 20%), but the selected transgene expressing cells were found to be stable and useful for later cloning and differentiation studies in both cases. As expected from the data in the literature, the CMV promoter was not applicable in our HuES cell lines, as transgene expression rapidly faded away soon after transfection/transduction, most likely due to the silencing of this viral promoter [18, 19]. The other constitutive promoters (PGK, EF1α, and CAG) expressed the transgene in undifferentiated and in differentiating HuES cells. To avoid the problem of copy number variation when comparing expression levels among different tissue types after differentiation, we established clones from transgenic stem cells expressing EGFP by all these three different promoters. For each case, a manual enrichment of the EGFP-expressing HuES cells was followed by a flow cytometry sorting and the selected cells were placed back onto mouse feeders at a density to allow individual clones to grow separately (Fig. 2A). These clones preserved their ES cell characteristics, as they expressed the ES cell markers Oct-4, SSEA-4, and podocalyxin (Fig. 2B, 2C, and data not shown). The uniform EGFP expression in these clones, grown in an undifferentiated state for 3–20 passages, was preserved; thus, at least in this period, a stability of foreign gene expression was observed (supporting information Fig. 2A).

Subsequently, we followed differentiation of these HuES clones by embryoid body (EB) formation in vitro. The resulting EBs and the various tissue types differentiated from them were similar to those observed in the untransfected HuES cells, and there was no observable change in the general pattern or timing of these differentiating clones (Figs. 2D, 2E). Concerning cardiomyocytes, 19–33 beating areas/100 plated EBs were detected in different transgenic clones which was not significantly different from the average number of 28 beating areas observed during differentiation of control HuES cells. A detailed analysis of the gene expression patterns in these differentiated colonies is underway in our laboratory (see discussion). Collectively, these experiments indicated that neither the transduction nor the combined transfection/transposition procedure altered the cardiomyocyte differentiation potential of HuES cells per se and stable transgene expression could be achieved in these pluripotent cells by our gene delivery methods.

Robust CAG Promoter-Driven Gene Expression in Cardiomyocytes

HuES cells form EBs when detached from the plate surface, and under well-defined conditions the EBs spontaneously differentiated into various tissue types (such as neuronal, epithelial, myocardial, and other progenitors [7, 23, 28, 29]). Cardiomyocytes that differentiate from HuES cells can easily be recognized by the spontaneous contractions of small tissue patches, by the immunochemical recognition of cardiac-specific markers [30], and by a clear response to pharmacological modulations (increased contraction frequency after adrenaline, decreased frequency, and arrest after calcium channel blockers [31, 32]). Although we do not include a detailed documentation here, the cardiomyocytes differentiated from the parental or gene-modified HuES cells showed all these parameters, without observable differences caused by the actual methods of gene transfer (supporting information Video 1). As the BG01V cell line has a lower potential to generate beating cardiomyocytes, most of the detailed experiments were carried out in the HUES9 cell line.

When comparing various tissue types differentiated from such EGFP-expressing HuES clones, an intriguing observation in CAG promoter-driven EGFP cells was that the fluorescent signal intensity in cardiomyocytes was always significantly higher than in the other surrounding tissue types. As the phenomenon was first seen for SB transposon-derived clones, this gene delivery experiment was repeated several times and the outcome was always the same: CAG-driven EGFP expression was extremely strong in differentiated cardiomyocytes as compared with cells with PGK promoter- or EF1α promoter-driven transgenes (Figs. 2E and 3). In fact, in the case of the CAG promoter, all beating cardiomyocytes could easily be

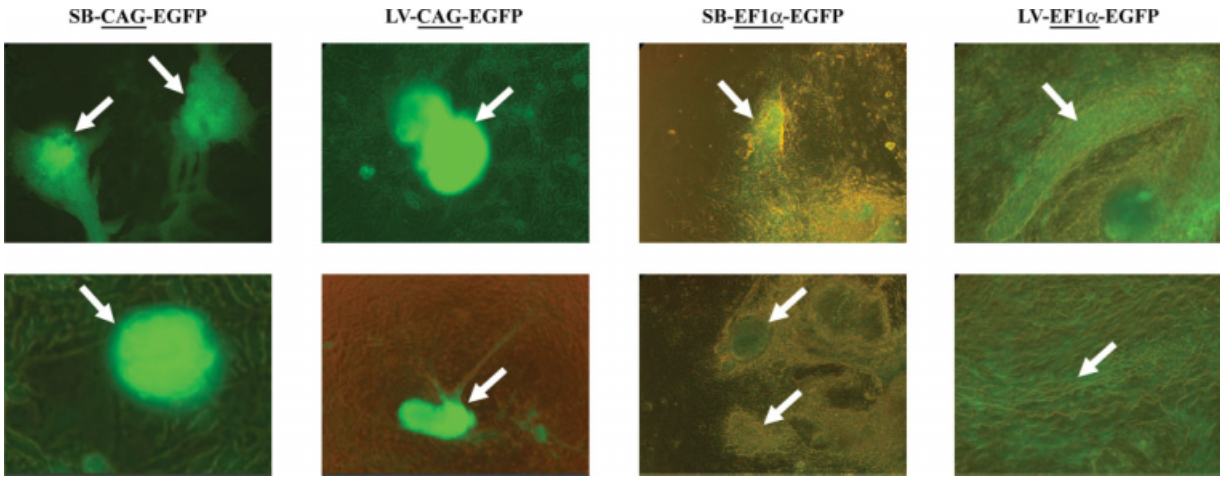


Figure 3. The "double-feature" phenomenon of the CAG promoter is independent of the gene delivery method. Several independent differentiations from transposon (SB) or lentiviral (LV) clones expressing EGFP resulted in numerous spontaneously contracting cardiomyocytes (white arrows, see also supporting information Videos 2 and 3). When the CAG promoter was used, the EGFP fluorescence in cardiomyocytes was always extremely high so that it masked the green fluorescence in the surrounding tissues (left "CAG" panels). In cells expressing the EF1α-driven transgene, cells were expressing the EGFP almost uniformly and did not show such an expression bias toward cardiomyocytes (right "EF1α" panels). Abbreviations: CAG, CMV enhancer-chicken β-actin-rabbit β1-globin fusion; EF1α, elongation factor 1α; EGFP, enhanced green fluorescent protein; LV, lentiviral; SB, Sleeping Beauty.

identified by the extremely high transgene expression. This unexpected behavior of a constitutive promoter allowed solving two problems at the same time: following transgene expression in undifferentiated cell types and selecting for a particular tissue type after differentiation.

The "Double-Feature" Behavior of the CAG Promoter Is Independent of the Gene Delivery Method and the Sequence, Copy Number, and the Integration Sites of the Transgenes

To prove that the phenomenon is a characteristic of the CAG promoter, we first compared the results of two different CAG-driven fluorescent transgenes (EGFP and amaxaGFP) delivered into the HuES cells by the SB transposon system. EGFP (e.g., www.clontech.com) and the amaxaGFP (www.amaxa.com) are sequentially unrelated proteins, isolated from different organisms. Despite this sequence difference, the same phenomenon was observed, that is the differentiated cardiomyocytes showed a markedly higher fluorescent signal as compared with other differentiated cell types (Figs. 2E, 3, and supporting information Video 2, 3). Lentiviral gene delivery provided an additional proof that the CAG promoter, but not the EF1α or the PGK promoters, has this behavior of being transcriptionally extremely active selectively in cardiomyocytes (Fig. 3 and data not shown). Thus, this feature of the CAG promoter is independent of the gene delivery method.

To further characterize the phenomenon, we compared clones with different copy numbers of the CAG-EGFP transgene for both transposon-derived and lentiviral-derived HuES clones. The transgene copy numbers determined by real-time PCR were found to be 3–6 in our experiments. The differentiation-linked EGFP expression behavior was detected in all cases, clearly showing that copy number differences do not influence the "double-feature" behavior.

To document the independence of this phenomenon from transgene integration into specific host loci, we carried out integration site determination for transposon-based amaxaGFP-expressing HuES cells (Fig. 4). We identified several different genetic loci (being on different chromosomes) in which SB transposition occurred. In all cases, although these sites were apparently not related to any cardiomyocyte-related protein coding genes or miRNA clusters, an increased expression of EGFP in the cardiomyocytes was observed. All these data indicated that it is not the site of integration, but the CAG promoter itself that is responsible for the observed effect.

Selection for High, CAG Promoter-Mediated Transcription Enriches Differentiated Cells Expressing Cardiomyocyte Markers

In principle, higher EGFP fluorescence in particular cell types can be the result of various factors, including higher transcription or translation rate, or certain posttranslational modifications. Alternatively, the size or the higher cytoplasm/nucleus ratio in the observed cell types could also account for the higher signal intensity. To distinguish among these possibilities, we separated cells based on EGFP fluorescence by FACS at the 30th day of differentiation and isolated total cellular RNA from each separated fraction. Four different EGFP fractions were analyzed from cells differentiated from either CAG-EGFP or EF1α-EGFP-expressing HuES cell clones, created by either lentiviral-based or transposon-based gene delivery (Fig. 5B and supporting information Fig. 1B). By real-time quantitative PCR analysis, we measured the transcriptional profiles of cardiac-specific, early neuron-specific, and skin-specific marker genes, as well as stem cell-specific (pluripotency marker) genes, and also the EGFP mRNA levels in the different cell populations. As expected, the expression levels of pluripotency marker genes, OCT-4 and NANOG, were high in the undifferentiated HuES clones, whereas practically undetectable in the differentiated populations (Fig. 5A and supporting information Fig. 1A). In the differentiated cells, the EGFP transcription level closely correlated with the fluorescent signal intensity, thereby proving that high EGFP intensity was the result of increased transcription, as a consequence of increased promoter activity (Fig. 5C and supporting information Fig. 1C).

When analyzing cardiac-specific transcripts, we measured the mRNA levels of the ACTC, PLN, and NPPA genes (see Materials and Methods). For all cells originated from CAG-EGFP clones (irrespective of the gene delivery method applied), these cardiac-specific transcripts exhibited a

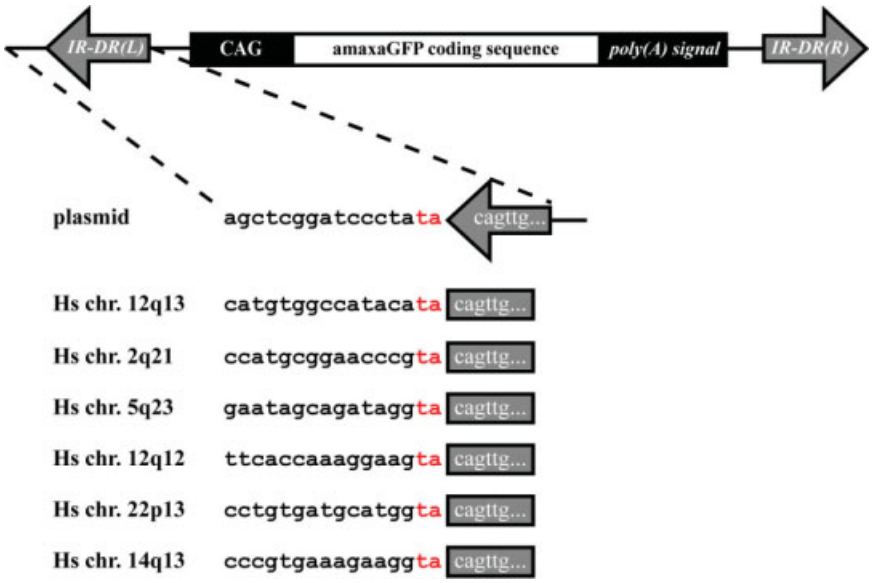


Figure 4. Examples of transposon integration sites. Splinkerette and inverse PCR techniques were used to determine the SB-CAG-amaxaGFP integration sites in HuES clones. In all cases, a "ta" sequence (marked by red) was found in the genomic locus next to the transposon IR-DR sequence which is a footprint of *Sleeping Beauty* transposition [25]. None of the detected integration sites were apparently related to any cardiomyocyte-associated protein coding genes or miRNA clusters.

significant, several-fold higher level in high EGFP expressing populations, indicating that in differentiated cells, cardiomyocytes were located in the fractions with high EGFP signal intensity (as observed earlier by fluorescence and confocal microscopy). On the other hand, in fractions from cells expressing the EF1α-EGFP transgene there was no such correlation: the examined cardiac specific genes showed a random distribution among the different EGFP expressing populations (Fig. 5C and supporting information Fig. 1C). When analyzing early neuron-specific (PAX6) or skin-specific (CAPG) transcripts, we did not find any correlation between their mRNA levels and EGFP transcriptional levels, and these genes had random expression profiles among different fractions for both CAG-EGFP- and EF1α-EGFP-expressing cells (Fig. 5C and supporting information Fig. 1C). Taken together, these expression studies provide evidence that the transcriptional activity of the CAG promoter, in contrast to other constitutive promoters examined, is significantly higher, selectively in cardiomyocytes. In addition, it also raises the possibility of flow cytometric separation of cardiomyocytes, based on high CAG promoter-driven transgene expression, especially because the currently available cardiac-specific cell surface antibodies require fixation and/or permeabilization of the cells (supporting information Fig. 2B).

Deciphering the "Double-Feature" Behavior by Bioinformatics Approaches

The artificial CAG promoter used in this study contains the CMV enhancer region, two sequences from the chicken β-actin promoter and one short part of the rabbit β1-globin promoter (Fig. 6A). To understand the sequential reasons behind the "double-feature" phenomenon, we conducted a bioinformatics analysis of the CAG promoter for potential tissue-specific transcription factor binding sites, by using the program MatInspector 7.7.3. When the analysis was narrowed down to heart-specific transcription factors, several potential binding sites were revealed (Fig. 6B). An especially intriguing feature of the CAG promoter is that it contains a consensus Nkx2.5 binding site (nt 234–248), three serum response factor (SRF.01, SRF.02, and SRF.03) binding sites (nt 317–335, 55–73, 595–613), as well as a MEF2-SL1.01 binding site (nt 624–646). These are all located within the CMV enhancer and the first chicken β-actin promoter region (Fig. 6B).

Several studies have documented that Nkx2.5 expression is specifically required for cardiogenesis, and the expression of this transcription factor is present from the early, neonatal development, to the fully matured form of cardiomyocytes [35–38]. SRF proteins were documented to cooperate with MEF2 and Nkx2.5 to promote cardiogenesis, and the activation of SRF by its interaction with myocardin provides a strict tissue-specificity for SRF binding promoters in the cardiac tissues [39]. Based on these data, it is not unexpected that the CAG promoter, most probably through multiple interactions with cardiac-specific transcription factors, shows an exquisite overexpression in cardiomyocytes.

Discussion

To establish stable transgene integration into ES cells, the careful design of the gene delivery method and the choice of promoters are crucial points. Among non-viral gene deliveries, transposon-based methods are becoming increasing popular, because their advantages outweigh their potential technological drawbacks. They certainly require less sophisticated and expensive facilities, their use is simpler than viral applications, and their use can also overcome safety and ethical issues currently associated with viral-based vector systems. One of the most effective transposon-based platforms is the SB system. One particular advantage of the SB system in the context of potential human applications is undoubtedly the random genomic integration profile, which lowers the chance of insertional mutagenesis or induced oncogenesis. These problems represent a serious drawback for viral-based gene delivery methods [1, 2]. A further advantage of the SB system is the availability of hyperactive transposases, providing highly efficient gene delivery tools [3, 24]. There is one report in which the SB system was successfully applied in HuES cells [33], but this study used an older version of the transposase (SB11), and a nucleofection technology in a given HuES cell line, which may not be directly applicable for other stem cell cultures. Therefore, we aimed to examine the use of a hyperactive SB transposase, combined with a lipid-based transfection methodology, and compared this with the efficient lentiviral gene delivery system.

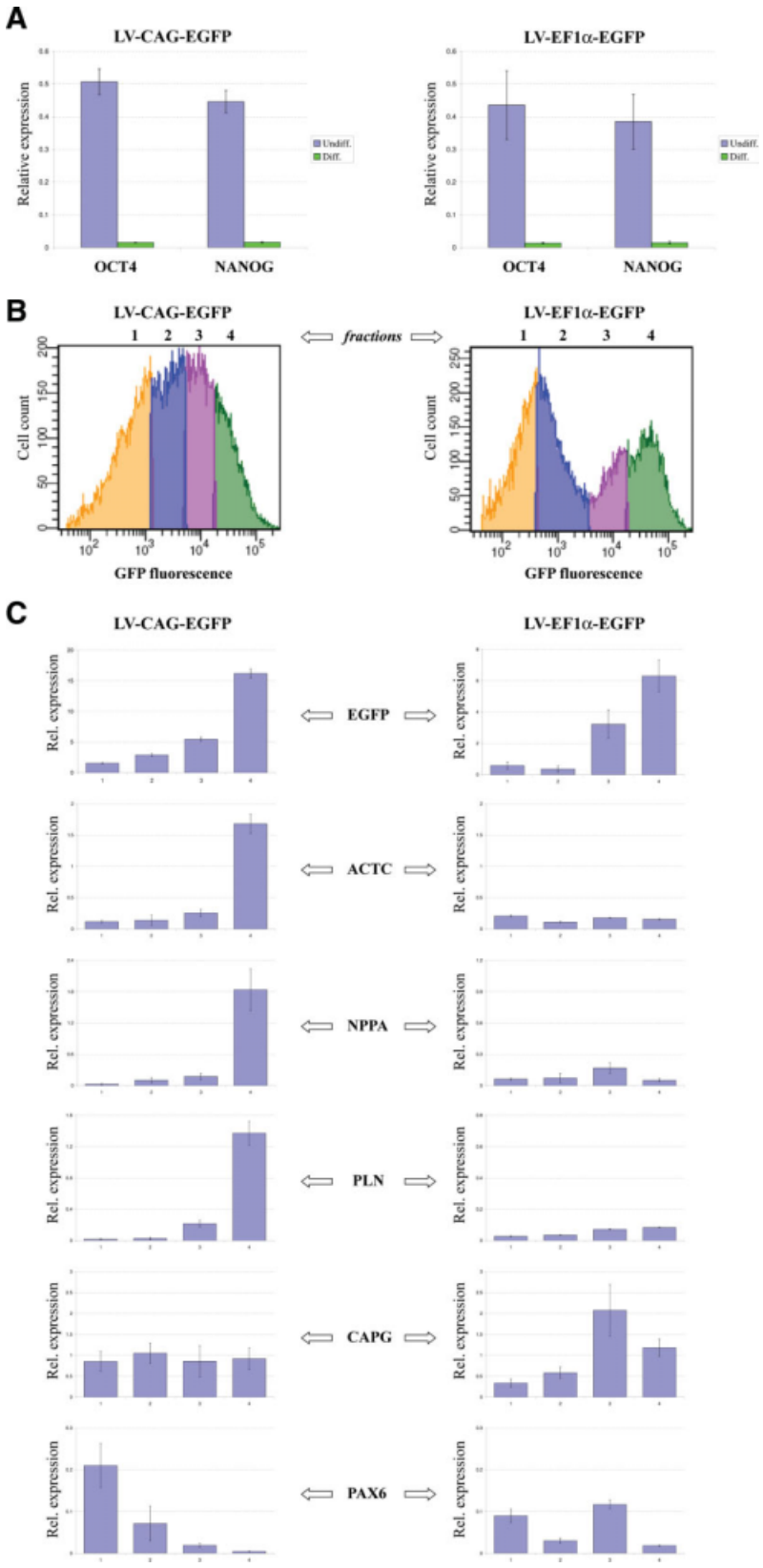


Figure 5. Expression profiles of HuES clones expressing either CAG-EGFP or EF1α-EGFP transgenes after lentiviral gene delivery. **(A):** Comparing the expression levels of stem cell markers in undifferentiated (Undiff.) and in differentiated (Diff.) HuES cells. The transcription levels of OCT-4 and NANOG were almost undetectable and at least 40-times lower in differentiated cells. **(B):** Fluorescence-based cell sorting diagrams indicating the sorted populations of differentiated EGFP-expressing cells. For unknown reasons, two major EGFP peaks were detected in this particular experiment after differentiating EF1α-EGFP expressing HuES clones. Total RNAs were isolated from cells collected from the four indicated fractions in both cases, representing cells with increasing EGFP signal intensity. **(C):** Transcription levels of EGFP and differentiation specific endogenous genes in the four different fractions. Increasing EGFP transcript level positively correlated with increasing fluorescence signal intensity. mRNA levels of cardiac specific genes (ACTC, NPPA, and PLN) showed a several fold increase in the cell population with the highest EGFP intensity for the CAG promoter-driven transgene, whereas no such tendency was observed for the EIF1α promoter-driven transgene. No correlation between EGFP transcription level and skin-specific (CAPG) or early neuron-specific (PAX6) gene transcripts could be identified for either of the promoters. In all real-time PCR experiments, results represent mean values ± S.E.M. of at least three independent measurements; transcription levels were normalized to the ribosomal P0 endogenous control. Abbreviations: ACTC, cardiac-specific alpha-actin; CAG, CMV enhancer-chicken β-actin-rabbit β1-globin fusion; CAPG, capping protein (actin filament), gelsolin-like; EF1α, elongation factor 1α; (E)GFP, (enhanced) green fluorescent protein; LV, lentiviral; NPPA, natriuretic peptide precursor A; PLN, phospholamban.

Recent, as yet unpublished experiments clearly indicate the higher specific integration efficiency of the hyperactive transposase system in tissue-derived stem cells (Ivics and Izsvak, personal communication). Our experiments demonstrate that the same SB transposon system is an effective gene delivery method that can be applied to create stable cell lines for a given transgene even if the transfection efficiency is low and a drug-based selection method is not applied. These

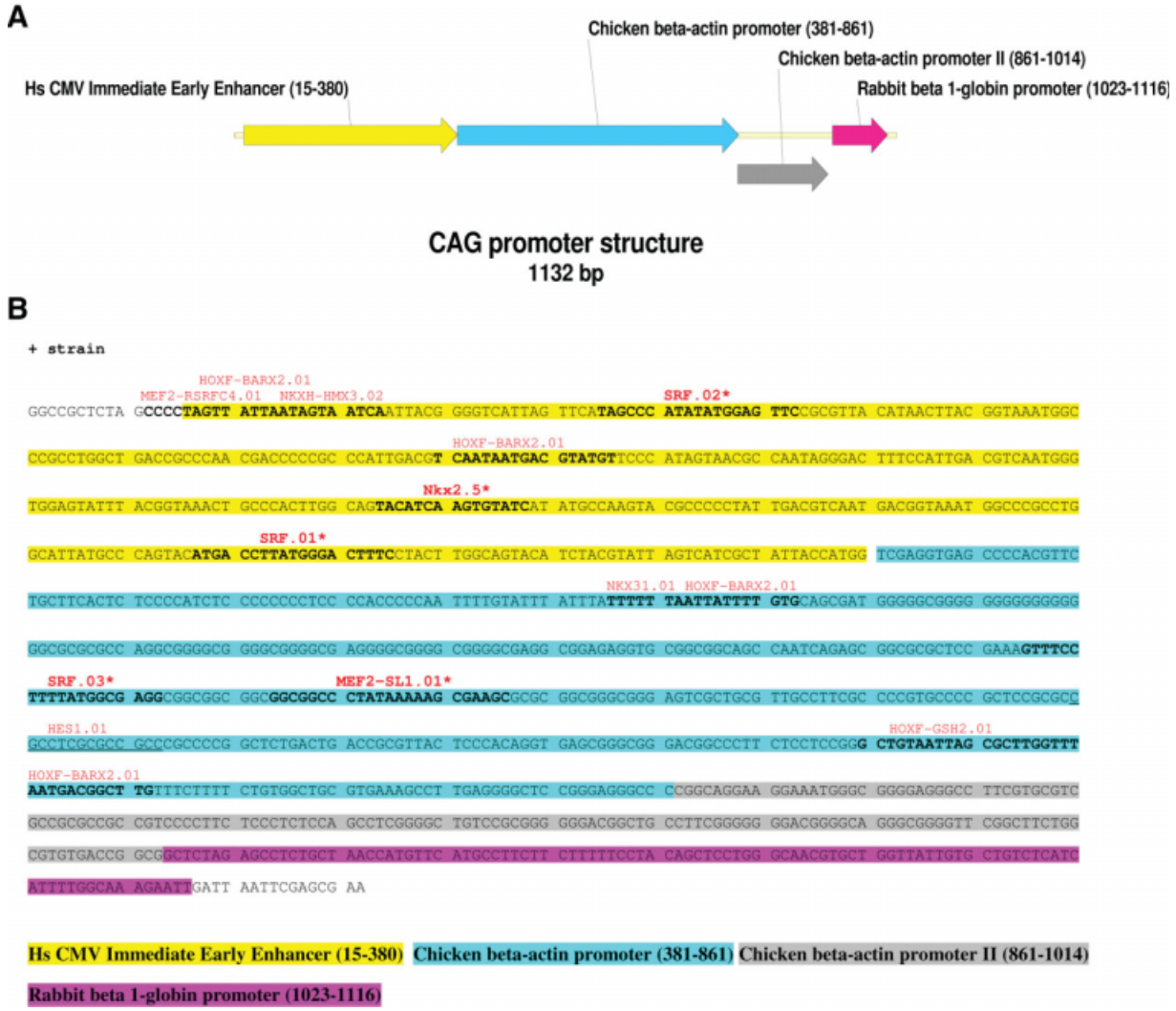


Figure 6. Structure of the applied CAG promoter. **(A):** Proportional representation of the distinct elements of the fusion promoter. **(B):** The predicted cardiac-specific transcription factor binding sites. The most prominent one is the complete consensus site of the Nkx2.5 homeodomain factor, located in the CMV Immediate Early Enhancer element (nt 234–248). See text for further details. Color coding is intentionally comparable between the two panels. Abbreviations: CAG, CMV enhancer-chicken β-actin-rabbit β1-globin fusion; CMV, cytomegalovirus.

conditions especially apply to HuES cells, which are difficult subjects for transfection, and may not be selectable by commonly used antibiotics without jeopardizing their pluripotency status (Fig. 2). We propose that the transposon protocol presented here could be a method of choice for gene delivery into HuES cells.

To assure the similarity of the transgene-expressing HuES cells and their differentiation to the normal parental cells, we have analyzed a number of gene expression patterns, including the expression of the ABCG2 protein [23], and found no apparent differences in these parameters (Figs. 2, 5 and data to be reported elsewhere). For a systematic comparison of the differentiation patterns in gene-modified cells, a detailed analysis of a full gene expression profile by a TaqMan based real-time quantitative PCR assay system is underway in our laboratory.

The use of suitable promoter(s) for foreign gene expression in stem cells is still an unresolved issue, although several studies described the potential silencing or the variable transcriptional efficiency of commonly used promoters [19–21]. It is widely accepted, however, that some housekeeping promoters, such as the PGK or the EF1α promoters, as well as the constitutive artificial CAG promoter most likely represent the best choices for stem cell applications, as these promoters also remain active in a wide variety of differentiated tissues, although with variable intensities [19–21, 34]. On the other hand, the need to express transgenes specifically in a given tissue raises the problem of controlling transgene delivery and tissue-specific transcription at the same time. Current solutions to this problem offer the use of two different transcription units separately (with the problem of the delivery of two plasmids or viruses), or in one cargo (with the problem of larger plasmid or virus vector size), both of which decrease delivery efficiency. Based on the present study, we propose that the use of the "double-feature" CAG promoter solves both problems: it provides a means for transgene detection in undifferentiated cells, as well as a later separation platform for cardiomyocytes, based on its selectively high transcription rate in these cell types (Figs. 3 and 5). We provided evidence that this feature is independent of the gene delivery method, transgene sequence, copy number, or integration sites (Figs. 3 and 4). We are currently investigating the applicability of cardiomyocytes selected by this method for potential therapeutic purposes. As an important first criterion, these cells

show negligible OCT-4 and NANOG expression (Fig. 5A), indicating the absence of undifferentiated cells that might represent a risk of tumor formation [30–32].

It should be noted that the term "CAG promoter" is inconsistently used in the scientific literature. The version applied in our experiments is probably very close to the originally created artificial promoter (Fig. 6A), as it contains the three crucial elements (*C*MV enhancer, *β-a*ctin sequence, rabbit *β*1-*g*lobin sequence), although it has two shorter parts of chicken actin promoter rather than the original long sequence [40]. It is often impossible to decipher the exact "CAG" sequence used by a particular research group, but from the size of the construct in several cases it is likely different from the "canonical" promoter. A common mistake is that a construct without the rabbit *β*1-*g*lobin sequence is also called a "CAG" promoter [41, 42], or the CMV enhancer sequence is sometimes shortened, e.g., in the commercially available pDRIVE-CAG plasmid (www.invivogen.com). If our bioinformatics analysis correctly identifies the Nkx2.5 and the serum response element binding sites as responsible sequences for the cardiac-specific upregulation, than many of the generally used "CAG" promoters may not show the "double-feature" characteristics. Currently, we have embarked on a detailed experimental analysis of these questions; deletions and binding factor consensus site mutations are hoped to reveal the background of this phenomenon. Moreover, as this phenomenon holds true for two human-derived ES cell lines, HUES9 and BG01V, we have started similar experiments in mouse ES cells, and preliminary results indicate that the same CAG promoter also drives cardiac-specific marker expression in this species (data not shown).

An important issue is whether it is possible to design "double-feature" promoters with different tissue specificities. By understanding the exact sequence characteristics from the current constellation of the CAG promoter, we hope to be able to create other "double-feature" promoters. These could become important tools in ES cell research and in future gene therapy applications.

Conclusion

In this study, we show that hyperactive SB transposons represent a powerful gene delivery methodology in HuES cell-based applications. By random transgene integration, this method provides an attractive alternative of virus-based applications. We also document that the CAG promoter shows unexpected characteristics when applied in HuES cells. This promoter has a long-term activity in undifferentiated cells, thus enabling the investigator to control the efficiency of transgene delivery and selecting for transgene-expressing cells. At the same time, the CAG promoter also provides a possibility in later differentiation stages for cardiomyocyte selection, via its unique, tissue-specific transcriptional upregulation. The application of this "double-feature" promoter provides a tool for human cardiomyocyte recognition and separation, and the system may be developed to follow various cell differentiation pathways.

Acknowledgments

The authors are grateful to Zsuzsanna Sebestyén for the excellent technical assistance. The authors appreciate the gift of HUES9 cell line by Dr. Douglas Melton, HHMI, and the original CAG-amaxaGFP transposon plasmid by Evelyn Zeira, Hadassah Medical Organization, Israel. This work was supported by EU FP6-INTHER (LSHB-CT-2005018961), OTKA (AT 048986, and NK72057), NKFP-1A-060/2004, ETT 405/2006, and KKK grants.

Disclosure of Potential Conflicts of Interest

The authors indicate no potential conflicts of interest.

References

1 Schroder AR, Shinn P, Chen H et al. HIV-1 integration in the human genome favors active genes and local hotspots. Cell 2002;110:521–529.
2 VandenDriessche T, Collen D, Chuah MK. Biosafety of onco-retroviral vectors. Curr Gene Ther 2003;3:501–515.
3 Ivics Z, Izsvak Z. Transposons for gene therapy! Curr Gene Ther 2006;6:593–607.
4 Izsvak Z, Ivics Z. Sleeping beauty transposition: Biology and applications for molecular therapy. Mol Ther 2004;9:147–156.
5 Izsvak Z, Ivics Z, Plasterk RH. Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. J Mol Biol 2000;302:93–102.
6 Ivics Z, Izsvak Z. Transposable elements for transgenesis and insertional mutagenesis in vertebrates: A contemporary review of experimental strategies. Methods Mol Biol 2004;260:255–276.
7 Lev S, Kehat I, Gepstein L. Differentiation pathways in human embryonic stem cell-derived cardiomyocytes. Ann N Y Acad Sci 2005; 1047:50–65.
8 Karner E, Unger C, Sloan AJ et al. Bone matrix formation in osteogenic cultures derived from human embryonic stem cells in vitro. Stem Cells Dev 2007;16:39–52.
9 Chen K, Wu L, Wang ZZ. Extrinsic regulation of cardiomyocyte differentiation of embryonic stem cells. J Cell Biochem 2008;104:119–128.
10 Tomescot A, Leschik J, Bellamy V et al. Differentiation in vivo of cardiac committed human embryonic stem cells in postmyocardial infarcted rats. Stem Cells 2007;25:2200–2205.
11 Passier R, Oostwaard DW, Snapper J et al. Increased cardiomyocyte differentiation from human embryonic stem cells in serum-free cultures. Stem Cells 2005;23:772–780.
12 Kolossov E, Bostani T, Roell W et al. Engraftment of engineered ES cell-derived cardiomyocytes but not BM cells restores contractile function to the infarcted myocardium. J Exp Med 2006;203: 2315–2327.
13 Huber I, Itzhaki I, Caspi O et al. Identification and selection of cardiomyocytes during human embryonic stem cell differentiation. FASEB J 2007;21:2551–2563.
14 Duan Y, Catana A, Meng Y et al. Differentiation and enrichment of hepatocyte-like cells from human embryonic stem cells in vitro and in vivo. Stem Cells 2007;25:3058–3068.
15 Gallo P, Grimaldi S, Latronico MV et al. A lentiviral vector with a short troponin-I promoter for tracking cardiomyocyte differentiation of human embryonic stem cells. Gene Ther 2008;15:161–170.
16 Ward CM, Stern PL. The human cytomegalovirus immediate-early promoter is transcriptionally active in undifferentiated mouse embryonic stem cells. Stem Cells 2002;20:472–475.
17 Zeng X, Chen J, Sanchez JF et al. Stable expression of hrGFP by mouse embryonic stem cells: Promoter activity in the undifferentiated state and during dopaminergic neural differentiation. Stem Cells 2003; 21:647–653.
18 Krishnan M, Park JM, Cao F et al. Effects of epigenetic modulation on reporter gene expression: Implications for stem cell imaging. FASEB J 2006;20:106–108.
19 Liew CG, Draper JS, Walsh J et al. Transient and stable transgene expression in human embryonic stem cells. Stem Cells 2007;25: 1521–1528.
20 Chung S, Andersson T, Sonntag KC et al. Analysis of different promoter systems for efficient transgene expression in mouse embryonic stem cell lines. Stem Cells 2002;20:139–145.
21 Xia X, Zhang Y, Zieth CR et al. Transgenes delivered by lentiviral vector are suppressed in human embryonic stem cells in a promoter-dependent manner. Stem Cells Dev 2007;16:167–176.

22 Ivics Z, Hackett PB, Plasterk RH et al. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 1997;91:501–510.
23 Apati A, Orban TI, Varga N et al. High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells. Biochim Biophys Acta 2008;1778:2700–2709.
24 Ivics Z, Katzer A, Stuwe EE et al. Targeted Sleeping Beauty transposition in human cells. Mol Ther 2007;15:1137–1144.
25 Vigdal TJ, Kaufman CD, Izsvak Z et al. Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. J Mol Biol 2002;323:441–452.
26 Schambach A, Galla M, Maetzig T et al. Improving transcriptional termination of self-inactivating γ-retroviral and lentiviral vectors. Mol Ther 2007;15:1167–1173.
27 Ujhelly O, Ozvegy C, Varady G et al. Application of a human multidrug transporter (ABCG2) variant as selectable marker in gene transfer to progenitor cells. Hum Gene Ther 2003;14:403–412.
28 Schulz TC, Palmarini GM, Noggle SA et al. Directed neuronal differentiation of human embryonic stem cells. BMC Neurosci 2003;4:27.
29 Thomson JA, Itskovitz-Eldor J, Shapiro SS et al. Embryonic stem cell lines derived from human blastocysts. Science 1998;282:1145–1147.
30 Capi O, Gepstein L. Myocardial regeneration strategies using human embryonic stem cell-derived cardiomyocytes. J Control Release 2006; 116:211–218.
31 Kehat I, Kenyagin-Karsenti D, Snir M et al. Human embryonic stem cells can differentiate into myocytes with structural and functional properties of cardiomyocytes. J Clin Invest 2001;108:407–414.
32 Xu C, Police S, Rao N et al. Characterization and enrichment of cardiomyocytes derived from human embryonic stem cells. Circ Res 2002;91:501–508.
33 Wilber A, Linehan JL, Tian X et al. Efficient and stable transgene expression in human embryonic stem cells using transposon-mediated gene transfer. Stem Cells 2007;25:2919–2927.
34 Chan KK, Wu SM, Nissom PM et al. Generation of high-level stable transgene expressing human embryonic stem cell lines using Chinese hamster elongation factor-1 α promoter system. Stem Cells Dev 2008; 17:825–836.
35 Chen CY, Schwartz RJ. Recruitment of the tinman homolog Nkx-2.5 by serum response factor activates cardiac α-actin gene transcription. Mol Cell Biol 1996;16:6372–6384.
36 Sepulveda JL, Belaguli N, Nigam V et al. GATA-4 and Nkx-2.5 coactivate Nkx-2 DNA binding targets: Role for regulating early cardiac gene expression. Mol Cell Biol 1998;18:3405–3415.
37 Christoforou N, Miller RA, Hill CM et al. Mouse ES cell-derived cardiac precursor cells are multipotent and facilitate identification of novel cardiac genes. J Clin Invest 2008;118:894–903.
38 Linhares VL, Almeida NA, Menezes DC et al. Transcriptional regulation of the murine Connexin40 promoter by cardiac factors Nkx2-5, GATA4 and Tbx5. Cardiovasc Res 2004;64:402–411.
39 Wang D, Chang PS, Wang Z et al. Activation of cardiac gene expression by myocardin, a transcriptional cofactor for serum response factor. Cell 2001;105:851–862.
40 Niwa H, Yamamura K, Miyazaki J. Efficient selection for high-expression transfectants with a novel eukaryotic vector. Gene 1991;108:193–199.
41 Sawicki JA, Morris RJ, Monks B et al. A composite CMV-IE enhancer/β-actin promoter is ubiquitously expressed in mouse cutaneous epithelium. Exp Cell Res 1998;244:367–369.
42 Xu ZL, Mizuguchi H, Ishii-Watabe A et al. Optimization of transcriptional regulatory elements for constructing plasmid vectors. Gene 2001;272:149–156.


See www.StemCells.com for supporting information available online.

orban.tamas_188_24

Mobile DNA

METHODOLOGY Open Access

Reliable transgene-independent method for determining *Sleeping Beauty* transposon copy numbers

Orsolya Kolacsek[1], Virág Krízsik[1], Anita Schamberger[1], Zsuzsa Erdei[1], Ágota Apáti[1], György Várady[1], Lajos Mátés[2], Zsuzsanna Izsvák[2,3], Zoltán Ivics[2,3], Balázs Sarkadi[1], Tamás I Orbán[1*]

Abstract

Background: The transposon-based gene delivery technique is emerging as a method of choice for gene therapy. The *Sleeping Beauty* (SB) system has become one of the most favored methods, because of its efficiency and its random integration profile. Copy-number determination of the delivered transgene is a crucial task, but a universal method for measuring this is lacking. In this paper, we show that a real-time quantitative PCR-based, transgene-independent (qPCR-TI) method is able to determine SB transposon copy numbers regardless of the genetic cargo.

Results: We designed a specific PCR assay to amplify the left inverted repeat-direct repeat region of SB, and used it together with the single-copy control gene *RPPH1* and a reference genomic DNA of known copy number. The qPCR-TI method allowed rapid and accurate determination of SB transposon copy numbers in various cell types, including human embryonic stem cells. We also found that this sensitive, rapid, highly reproducible and non-radioactive method is just as accurate and reliable as the widely used blotting techniques or the transposon display method. Because the assay is specific for the inverted repeat region of the transposon, it could be used in any system where the SB transposon is the genetic vehicle.

Conclusions: We have developed a transgene-independent method to determine copy numbers of transgenes delivered by the SB transposon system. The technique is based on a quantitative real-time PCR detection method, offering a sensitive, non-radioactive, rapid and accurate approach, which has a potential to be used for gene therapy.

Background

Transposon-based systems have become the method of choice for gene delivery, and their applications as potential genetic vehicles are receiving great interest [1-3]. In recent years, the *Sleeping Beauty* (SB) transposon has been emerging as the most favorable delivery system, because of its random integration profile and the lack of similar transposon-like elements in the human genome, which significantly minimizes the risk often represented by viral-based methods [4-6]. Owing to its advantageous characteristics, SB is the first transposon-based system to be used in a clinical trial for a hematologic malignancy [7]. Recently, a novel hyperactive version of the originally reconstituted SB transposase was developed [8], which, apart from making the system more favorable than other widely used non-viral methods, further substantiates its applicability as a mutagenic tool to perform genetic analyses, similar to the transposon-based systems in *D. melanogaster* and *C. elegans* [9,10]. Although already possessing clear advantages, rigorous characterization of the SB system still remains to be carried out to set up standard methods concerning its applicability. One of the important issues in setting up gene-therapy guidelines or genome-wide mutagenesis protocols is that of copy-number determination in stable clones [11-13].

Various technical methods have been developed to determine transgene copy numbers after gene delivery, including Southern blotting and the specific PCR-based transposon display method [14,15]. In most cases, these

* Correspondence: orbant@biomembrane.hu
[1]Membrane Research Group of the Hungarian Academy of Sciences, Semmelweis University and National Blood Center, Budapest, Hungary
Full list of author information is available at the end of the article

are performed using radioactively labeled probes; although fluorescent labeling can also be used, its threshold detection levels are generally lower. Depending on the transgene used, other techniques such as *in situ* hybridization quantification of fluorescent marker proteins such as green fluorescent protein (GFP) can also be employed [16]. Although these methods are widely accepted and used, they are usually laborious and require specific chemicals and equipment. In addition, these detection methods are often limited to the measurement of a specific transgene, and lengthy pilot experiments are often required to determine the exact measurements needed to accurately quantify a newly arising gene of interest within a particular delivery system [17-19].

During this study, we aimed to develop an accurate method for quantifying SB transposon copy numbers, independent of the transgene sequence. We term this the real-time quantitative PCR-based, transgene-independent (qPCR-TI) method. It can be used for any SB-based gene delivery experiments without *a priori* optimization of the protocol.

To establish this method, we used specific probe sets designed for the left and right inverted repeat-direct repeat (IRDR) regions, which are the recognition motifs of the transposase and therefore required for any SB transposition reaction [20]. As an internal control for normalization, a probe for the *RPPH1* gene, the H1 RNA subunit of the RNaseP enzyme complex, was used. This gene is a widely accepted one-copy gene of the haploid human genome [21]. Comparing this system with the radioactive transposon display and Southern/dot blotting techniques, we provide evidence that using the IRDR-L specific probe set in comparative $2^{-\Delta\Delta Ct}$ measurements can reliably and accurately quantify SB transposon copy numbers in various cell lines, regardless of the transgene used. Apart from being sensitive, accurate and rapid, this real-time PCR-based quantification method also offers a powerful non-radioactive technique as an alternative against other standard methods.

Results and Discussion

The exact and rapid quantification of transgene copy numbers is often required for gene-delivery experiments. As we generally use the SB transposon system in our laboratory, we aimed to develop a real-time PCR-based technique that would be transgene-independent, specific for the transposon regions and therefore widely applicable. To optimize the qPCR-TI method, we began with clones of HEK-293 cells with SB transposons carrying two transcription units expressing GFP and the puromycin-resistance gene, which are both under the control of the CAG promoter (Figure 1A). This transgene setup allowed generation of clones with various copy numbers by either fluorescence-activated cell sorting (FACS) or antibiotic selection. Specific TaqMan® (Applied Biosystems, Foster City, CA, USA) assays were designed for the two IRDR motifs of the SB transposon and for the GFP sequence (Figure 1A). The widely applicable SB transposon version used throughout this study has two asymmetric IRDR regions ('left' and 'right' [22]). In most transposon flanking sequences, the two IRDR regions are repeat-rich DNA sequences, which makes PCR primer design relatively difficult. Moreover, the left and the right IRDRs are very similar to each other, which further increases the difficulty of designing specific assays for them. Nevertheless, we could still develop specific assays for each; neither of the IRDR-L nor the IRDR-R probe set gave signals in the exclusive presence of the other template (data not shown).

As the first (and simplest) approach, absolute quantification of DNA samples was performed using plasmid dilution series complemented with transposon-free non-specific genomic (g)DNA. However, the difficulties of determining the exact nucleic-acid concentration of very dilute samples and the differences in purity between samples made it necessary to abandon absolute quantification, and to include an internal copy control to overcome these problems with relative quantification. The *RPPH1* gene, the H1 RNA subunit of the RNaseP enzyme complex, was chosen as this is a widely-accepted one copy gene of the haploid human genome [21] (http://www.ncbi.nlm.nih.gov/ieb/research/acembly/index.html). However, the assay efficiency for the IRDR-R region differed significantly from that of the others, including the *RPPH1* endogenous control assay. Various conditions for the IRDR-R set were tried, and although template concentration seemed to be a crucial factor, the widely accepted template range of 10 to 40 ng still produced efficiency values that were significantly lower than those of the other assays (<90%) (Figure 1B). Sequence constraints originating from the similarity to IRDR-L hindered us designing other specific assays with different combinations of primers and probes in this short (228 bp) and repeat-rich region. Therefore, if this assay were to be included for measurement, the relative standard curve method would be the only acceptable quantification method, as it is the most suitable to compare reactions with suboptimal PCR efficiency. Apart from the setting up of standard curves (for both the transposon-specific assays and the *RPPH1* endogenous control), relative quantification also requires the use of a calibrator (a reference sample with a known copy number, preferably '1') to ensure the precision of quantification.

In the search for a potential calibrator sample, generated clones were screened by FACS for the lowest possible GFP signal, assuming that clones with one copy number should be among those samples (the signal

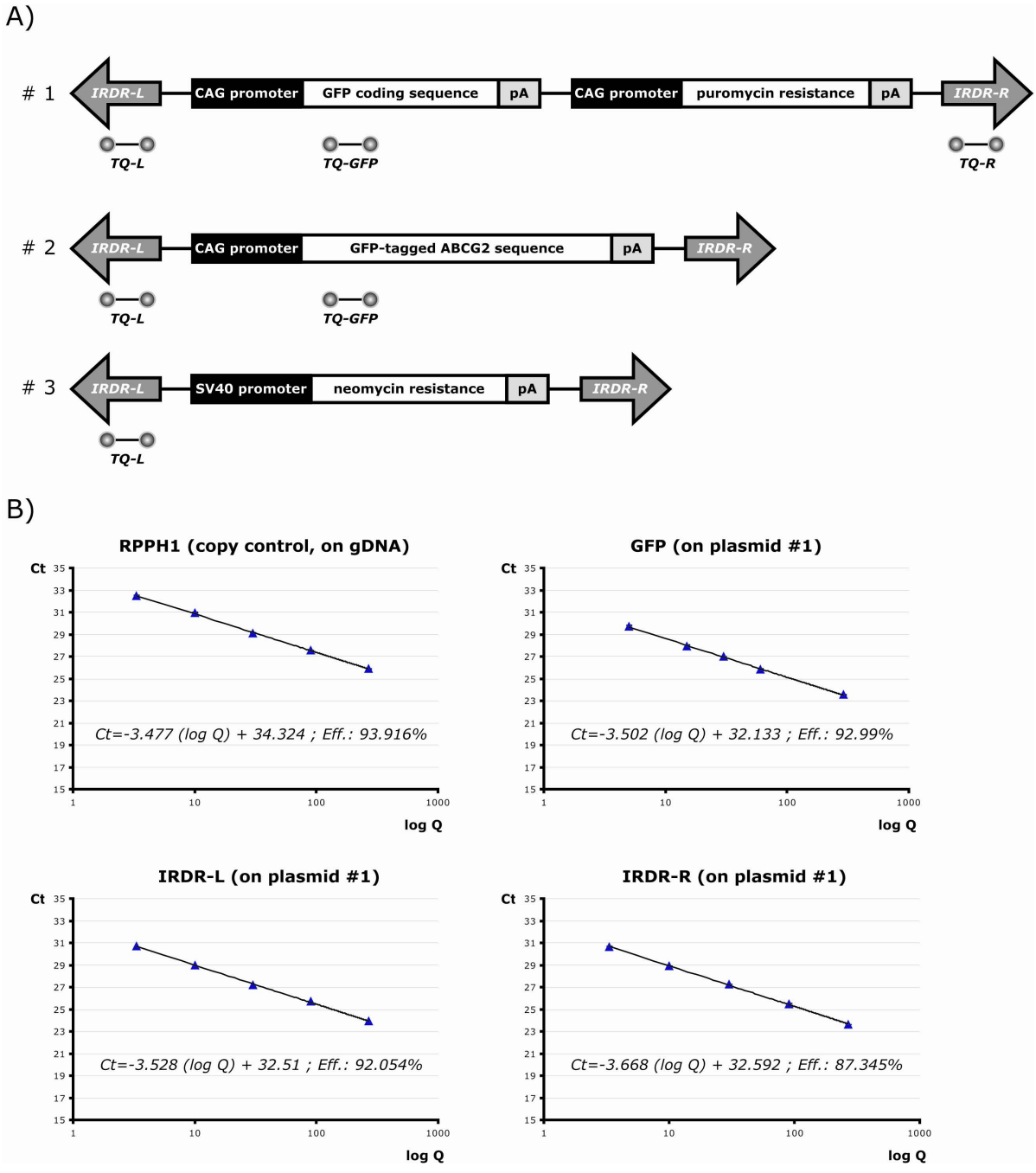


Figure 1 Real-time PCR assay designed for different transposon and transgene regions. **(A)** Structure of the used SB transposons with asymmetric IRDRs [22]. For each construct, the TaqMan® assays (TQ) used for copy-number determination are indicated. Sequences are not drawn to scale. IRDR-L/-R = inverted repeat-direct repeat left/right regions; pA = SV40 polyadenylation signals. **(B)** Efficiencies of the real-time assays determined by standard curves. For all assays, a dilution series was prepared from pooled genomic DNA samples from clones containing integrated transposon 1. The efficiency of the IRDR-R TaqMan® assay was notably lower than that of the others (<90%).

could also vary because of positional effects of different integration sites). Although the CAG promoter we used is known to be less prone to silencing [23-25], we had to make sure the lowest fluorescent signals were also associated with the lowest real-time signals when normalized to the *RPPH1* level, in order to exclude the potential presence of silenced copies. Using the GFP TaqMan® assay, several clones with one integrated transposon copy and numerous others with three or four copies were found (Figure 2A,B). Using the IRDR-L

orban.tamas_188_24

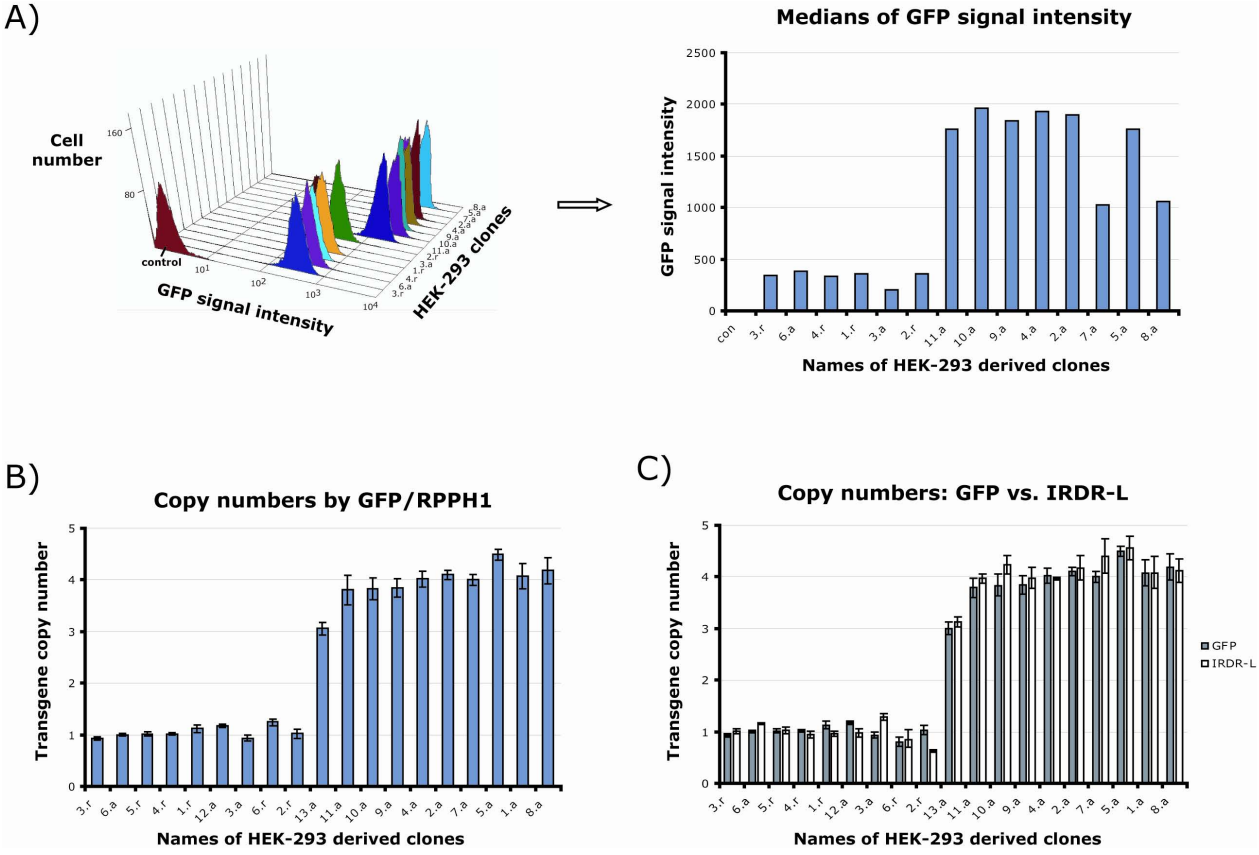


Figure 2 Comparing copy-number determination by green fluorescent protein (GFP) or transposon-specific real-time PCR. **(A)** Fluorescence-activated cell sorting (FACS) analysis of different HEK-293 derived clones expressing GFP. Higher fluorescent intensities indicate higher copy numbers, although signals can vary because of integration position effects and/or transgene silencing. The control sample shows the autofluorescence detected in non-transfected HEK-293 cells. **(B)** Copy numbers determined by transgene (GFP) specific real-time PCR assay normalized to the level of one copy control *RPPH1*; clones analyzed by FACS (A) and other clones established subsequently were examined. Various clones with low GFP expression level were determined to have one integrated transposon copy, whereas the majority with higher GFP fluorescence was found to have four transposon copies. In the case of clone 5.a, further analysis revealed that it was not a clone but rather a mixture of clones with an average copy number around 4.5. **(C)** Comparison of two techniques. The copy values determined by the transgene independent TaqMan® assay for the IRDR-L sequence correlated well with the GFP-based copy numbers. Clone 2.r originated from random integration, so the transposon repeat sequence might not be intact, and the partial presence of IRDR-L could result in a lower signal, therefore this clone was not included among the controls for later experiments. a = clones obtained from active transposition; r = clones obtained from random integration (from transfection with the mutant transposase). For copy numbers, values are means ± SEM of at least three independent measurements.

set, very similar copy numbers could be calculated using the relative standard curve method (Figure 2C), whereas the IRDR-R TaqMan® set gave unreliable results, mainly due to the problems discussed earlier (Additional file 1). Because the assays for *RPPH1*, GFP and the transposon IRDR-L had very similar efficiency values (Figure 1B), we also tried another approach, calculating the copy numbers in the examined clones by the comparative Ct ($2^{-\Delta\Delta Ct}$) method in the same experiments. The results based on GFP or IRDR-L were in agreement with each other and with the results of the relative standard curve method. Moreover, technical errors could be further decreased by using a pool of gDNA samples with known copy number as a reference. We therefore concluded that once we left out the specific but less efficient assay for the IRDR-R region, the comparative Ct method could be used for reliable and precise transposon copy-number determination using the IRDR-L TaqMan® assay. Abandoning the relative standard curve method also allowed inclusion of more samples in one reaction plate, as no more dilution series with several parallels were required.

To test the qPCR-TI method on other samples, we examined clones of the HUES9 human embryonic stem cell line expressing the GFP-tagged ABCG2 transporter [26] generated by the SB transposon system. Again, the GFP and the IRDR-L TaqMan® assays could be compared with each other (Figure 1A, transposon 2). As a general assay setup,

the *RPPH1* control and reference samples (pool of clones with known copy numbers) were used. As shown in Figure 3, the $2^{-\Delta\Delta Ct}$ method produced the same copy numbers, using either of the probe sets. These experiments therefore supported the use of the IRDR-L repeat specific assay for transposon copy-number determination, as it gave the same results as the assay specific for the carried internal transgene.

To compare our transgene-independent quantification approach with other techniques, we measured copy numbers of clones generated from HeLa cells by transposons containing a neomycin-resistance (neoR) gene (Figure 1A, transposon 3). Such clones were ideal for comparison because of the different transgene sequences and because their copy numbers were also determined by the Southern/dot blotting techniques or the transposon display method [5]. Several clones were tested, and the qPCR-TI method gave the same copy numbers as determined by the other radioactive methods (Table 1). For higher (>5) copy-number clones, the qPCR-TI method was also reasonably accurate, with occasional low relative-error margins (≤9%). The slight differences in some cases could be due to the inaccuracy of the standard methods for this range [14,15]. In addition, it has been suggested that precise values of very high copy numbers are more reliably measured by dot blot rather than transposon display methods. We found that the copy number of clone 4 determined by the dot-blot technique correlated well with data produced by the qPCR-TI. For low copy-number clones, only one clone (2/2 of neoR; see Table 1) did not give identical results with the different techniques. A difference of one copy number here clearly represents a higher percentage error margin, but this error might be related to the difference in integration sites in that particular clone (see discussion below). Taken together, the results of the neoR transposon clones indicated that the qPCR-TI technique is just as sensitive and accurate as the other widely used methods.

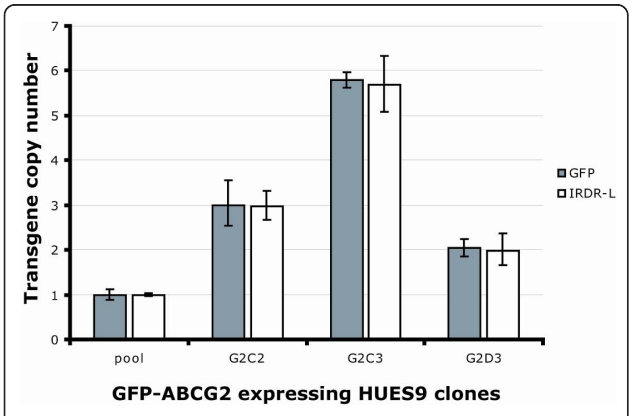


Figure 3 Copy-number determinations of green fluorescent protein (GFP)-ABCG2 expressing HUES9 clones. The sample 'pool' indicates the equimolar mixture of gDNA samples from the first four single-copy clones on Figure 2C. Later examination of the G2C3 line indicates that it is not derived from a clone but rather from a mixture of cells with five and six transposon copies. Values are means ± SEM of at least three independent measurements.

A further proof of principle was given by the determination of the transposon copy numbers in HUES9 clones previously generated using another sequentially distinct transgene. In those experiments, the amaxaGFP (a special fluorescent protein from a *Pontellina* copepod species, http://www.lonzabio.com) was carried by the transposon to generate clones of an embryonic stem-cell line, and the transposon integration sites were determined by the splinkerette PCR and the inverse PCR methods [27]. Based on these integration assays, copy numbers were estimated to be one to six in various clones, although all integrated copies may not be reliably detected by these methods because of the different flanking genomic sequences. When using the qPCR-TI method for several clones using the IRDR-L assay, the measured transposon copies were almost always the same as those previously claimed on the basis of the different proven integration sites (Table 1). One exception here was clone B1, where qPCR-TI gave a result one copy higher, similarly to the 2/2 neoR clone. A difference of one copy number here again undoubtedly represents a higher discrepancy with higher percentage error margin. However, because all the other low copy-number clones gave identical results with the various techniques, the two outliers might represent the lower sensitivity of the standard methods due to the dependence of transgene-integration sites [15]. These comparisons lead us to the conclusion that the qPCR-TI method provides reliable results for different SB transposon constructs, thereby being a consistent transgene-independent copy-number quantification method.

Using the experiments described above, the newly developed transgene-independent method for determining SB transposon copy numbers was validated: (i) it provided the same results as the assays specific for the carried transgene sequence and (ii) it could also reliably replace widely used standard radioactive techniques. The TaqMan® assay designed for the IRDR-L region of the transposon provides the basis for transgene independence as it is present in all SB constructs. In fact, 'symmetric' SB transposons with two IRDR-L (but not two IRDR-R) flanking sequences are functional [28], and the qPCR-TI method is also applicable to such constructs (with an obvious correction factor of 0.5). We found evidence that the PCR efficiency of this probe set is similar to the *RPPH1* single-copy control, so reliable quantification can be performed using the comparative $2^{-\Delta\Delta Ct}$ method. To ensure precise and rapid quantification,

Table 1 Comparing the qPCR-TI method with other standard techniques

Clone name	Methods	Copy numbers	
		By standard methods	By qPCR-TI[a]
Transposons carrying the neomycin resistance gene			
2/1	Transposon display/Southern blotting	8 to 10	8
2/2	Transposon display/Southern blotting	3	4
2/3	Transposon display/Southern blotting	10 to 12	10
2/9	Transposon display/Southern blotting	1	1
1	Transposon display/Southern blotting	12 to 13	13
4	Dot blot	52	50
5	Transposon display/Southern blotting	15	15
6	Transposon display/Southern blotting	12	11
7	Transposon display/Southern blotting	1	1
8	Transposon display/Southern blotting	2	2
9	Transposon display/Southern blotting	1	1
Transposons carrying the amaxaGFP transgene			
A3	Splinkerette PCR/inverse PCR	2	2
A4	Splinkerette PCR/inverse PCR	4	4
A5	Splinkerette PCR/inverse PCR	4	4.5[b]
A6	Splinkerette PCR/inverse PCR	2	2
B1	Splinkerette PCR/inverse PCR	1	2
B3	Splinkerette PCR/inverse PCR	3	3
B5	Splinkerette PCR/inverse PCR	2	2

[a]Quantitative PCR, transgene independent.

[b]For sample A5 from the amaxa green fluorescent clones, real-time PCR measurement indicated that it is more likely to be a mixed population of cells rather than a single clone.

reference samples (calibrators) with known copy numbers are also included, preferably a pool of gDNAs from different clones, to minimize discrepancies resulting from different transgenic sampling techniques and purities. The method could also be extended to other non-human gDNA samples; however, a suitable and validated single-copy reference gene control must always be used.

Another technical point that should be considered is the transposition-independent, random integration of the transgene. Because this is a stochastic process, it could possibly lead to the integration of the carried transcription unit without the transposon IRDR sequences. In such cases, the qPCR-TI method clearly underestimates transgene copy numbers, as it only detects copies resulted from *bona fide* transposition. As a general rule, we always include control experiments with gene delivery using the mutant transposase to estimate the level of random integration [20]. According to previous experiments, this phenomenon is generally very rare when using the new hyperactive SB100x transposase, but its extent can vary between different cell lines. Nevertheless, if such random background integration increases significantly, it may be necessary to measure the copy numbers of the transgene itself in the samples generated with the active transposase.

Conclusions

We have developed a sensitive and reliable real-time PCR-based (qPCR-TI) method for measuring SB transposon copy numbers. When compared with widely used standard methods, such as various blotting techniques or transposon display, it proved to be just as accurate as those other methods, while also offering a faster and non-radioactive method. However, the real advantage of this method is the transgene independence, which makes it applicable for any scientists working with *Sleeping Beauty* transposon constructs. Therefore, we believe that qPCR-TI could become the method of choice for gene therapy and general gene-delivery applications.

Methods

Cell-culture maintenance and creation of clones

Human embryonic kidney cells (HEK-293) were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum, 1% L-glutamine and 1% penicillin/streptomycin (Invitrogen, Carlsbad, CA, USA). Transfected cell populations were first enriched for transgene expression by flow cytometry (see below). Subsequently, cell clones were created by serial dilutions in 96-well plates. Selected clones were further analyzed by flow cytometry and harvested for

genomic DNA isolation (see below). The HUES9 embryonic stem-cell line (originally provided by Dr. Douglas Melton, Harvard University, USA) was maintained essentially as described previously [29], using cells from passage 35. To create transgene-expressing HUES9 clones, we used our previously developed method for human embryonic stem-cell lines [27].

Transfection and transposition

HEK-293 and HUES9 cells were transfected using a transfection reagent (FuGENE® 6; Roche Applied Science, Rotkreuz, Switzerland) in accordance with the manufacturer's instructions. The transfection mix contained 1 μg of a given transposon plasmid (Figure 1A) and 100 ng of the hyperactive SB100x *Sleeping Beauty* transposase, in a 10:1 ratio to minimize the overproduction inhibition phenomenon [5,8]. To visualize the random integration background, a control transfection with the inactive DDE motif mutant of the transposase was carried out, using the same experimental setup [20].

Flow cytometry

GFP-expressing cells were analyzed by a flow cytometer (FACSCalibur; Becton-Dickinson, San Jose, CA, USA) with Cellquest-Pro analysis software (Becton-Dickinson). Mock-transfected cells were used as labeling controls, and propidium iodide or 7-aminoactinomycin D staining was used to exclude non-viable cells. To select and clone cells expressing GFP, a fluorescence based cell sorter (FACSAria High Speed Cell Sorter; Becton-Dickinson) was used in accordance with the manufacturer's instructions.

Genomic DNA isolation, transposon display and Southern/dot blotting

After treatment with trypsin, cells were separated by centrifugation and washed with 1 × phosphate-buffered saline. After careful removal of the liquid supernatant, the dry cell pellets were stored at -80°C until further processing. Genomic DNAs were isolated from the cells by standard phenol-chloroform extraction after cell lysis and proteinase K digestion. DNA samples were quantified with a spectrophotometer (GeneQuant II; Pharmacia Biotech, Piscataway, NJ, USA). Transposon display and Southern-/dot-blotting techniques were performed essentially as described previously [5,14].

Quantitative real-time PCR

Reactions were performed on a real-time PCR platform (StepOne™ or StepOnePlus™; Applied Biosystems, Foster City, CA, USA) in accordance with the manufacturer's instructions. The gDNA samples (30 ng each) were run in triplicate, in singleplex reactions with a final volume of 20 μl using TaqMan® chemistry. All primers and probes were designed by Primer Express software (version 3.0; Applied Biosystems), and probes were labeled with 5'-FAM and 3'-nonfluorescent (minor groove binding) quencher molecules. Sequences for the TaqMan® assays are given in Table 2. Final concentrations of primers and probes were 250 and 900 nM, respectively. Data were analyzed by StepOne software (version 2.1; Applied Biosystems).

Table 2 Primers and probes used for quantitative real-time PCR

Primer/probe name	Sequence 5'→3'
RPPH1	
Forward	AGCTGAGTGCGTCCTGTCACT
Reverse	TCTGGCCCTAGTCTCAGACCTT
Probe	CACTCCCATGTCCC
GFP	
Forward	GAGCGCACCATCTTCTTCAAG
Reverse	TGTCGCCCTCGAACTTCAC
Probe	ACGACGGCAACTACA
IRDR-L	
Forward	CTCGTTTTTCAACTACTCCACAAATTTCT
Reverse	GTGTCATGCACAAAGTAGATGTCCTA
Probe	CTGACTTGCCAAAACT
IRDR-R	
Forward	GCTGAAATGAATCATTCTCTCTACTATTATTCTGA
Reverse	AATTCCCTGTCTTAGGTCAGTTAGGA
Probe	TCACCACTTTATTTTAAGAATGTG

Additional material

Additional file 1: Supplementary Figure 1: Comparison of the IRDR-R assay with the GFP specific real-time PCR method. Selected HEK-293 clones were examined for transposon copy numbers in parallel by the accepted green fluorescent protein (GFP) specific assay and the assay specific for *Sleeping Beauty* (SB) inverse repeat-direct repeat, right (IRDR)-R. In contrast to the IRDR, left (IRDR-L) real-time assay, the IRDR-R specific assay failed to reproduce previously determined copy numbers consistently (see Figure 2C). For this particular experiment, 30 ng genomic (g)DNA was used for the reaction. Although different starting gDNA concentrations (higher than the recommended range of 10 to 40 ng) improved the reproducibility of the IRDR-R assay, it still did not reach the reliability level of the GFP or the IRDR-L assays.

Acknowledgements

We thank Dr Douglas Melton for the gift of the HUES9 cell line. T I O is a recipient of the János Bolyai Scholarship from the Hungarian Academy of Sciences. This work was supported by grants from OTKA (NK72057), ETT (213-09), ES2Heart Jedlik (OM00203/2007), STEMKILL Jedlik (OM00108/2008) and National Development Agency grant KMOP-1.1.2-07/1-2008-0003.

Author details

[1]Membrane Research Group of the Hungarian Academy of Sciences, Semmelweis University and National Blood Center, Budapest, Hungary. [2]Mobile DNA Group, Max-Delbrück Center for Molecular Medicine, Berlin,

Germany. [3]Department of Human Genetics, University of Debrecen, Debrecen, Hungary.

Authors' contributions

OK established the HEK clone; OK and VK optimized the real-time PCR and performed copy-number measurements; AS, ZE and ÁA established the HUES9 clones; GV helped in FACS measurements; LM measured copy numbers in HeLa clones; ZsI and ZI gave technical help and advices with the SB transposon work; BS provided financial support and discussed the data; and TIO designed the overall strategy, analyzed the data and wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Received: 5 November 2010 Accepted: 3 March 2011
Published: 3 March 2011

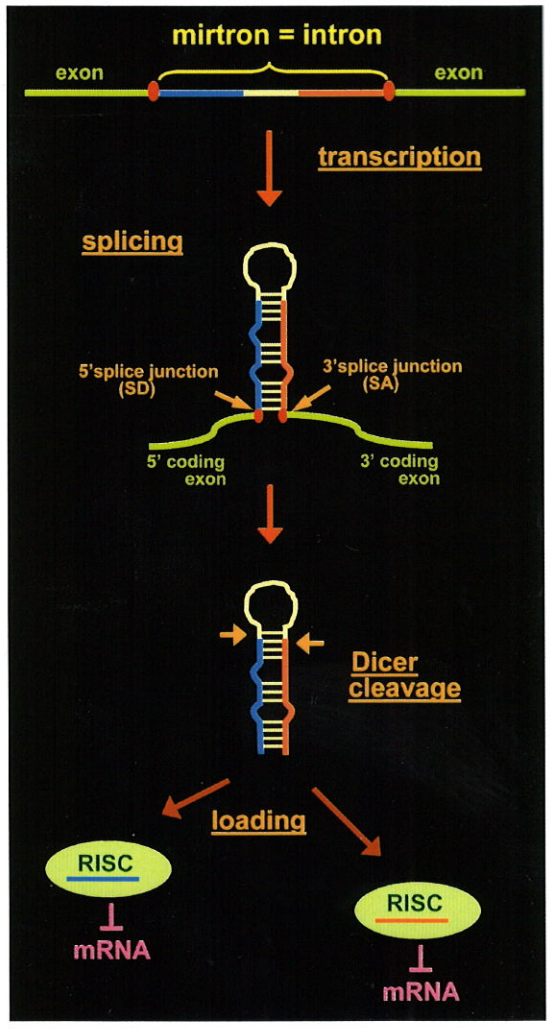
References

1. Ivics Z, Izsvak Z: **Transposons for gene therapy!** *Curr Gene Ther* 2006, **6**:593-607.
2. VandenDriessche T, Ivics Z, Izsvak Z, Chuah MK: **Emerging potential of transposons for gene therapy and generation of induced pluripotent stem cells.** *Blood* 2009, **114**:1461-1468.
3. Claeys Bouuaert C, Chalmers RM: **Gene therapy vectors: the prospects and potentials of the cut-and-paste transposons.** *Genetica* 2010, **138**:473-484.
4. Izsvak Z, Ivics Z: **Sleeping beauty transposition: biology and applications for molecular therapy.** *Mol Ther* 2004, **9**:147-156.
5. Grabundzija I, Irgang M, Mates L, Belay E, Matrai J, Gogol-Doring A, Kawakami K, Chen W, Ruiz P, Chuah MK, VandenDriessche T, Izsvák Z, Ivics Z: **Comparative analysis of transposable element vector systems in human cells.** *Mol Ther* 2010, **18**:1200-1209.
6. Hackett PB, Largaespada DA, Cooper LJ: **A transposon and transposase system for human application.** *Mol Ther* 2010, **18**:674-683.
7. Williams DA: **Sleeping beauty vector system moves toward human trials in the United States.** *Mol Ther* 2008, **16**:1515-1516.
8. Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, Grzela DP, Schmitt A, Becker K, Matrai J, Ma L, Samara-Kuko E, Gysemans C, Pryputniewicz D, Miskey C, Fletcher B, VandenDriessche T, Ivics Z, Izsvák Z: **Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates.** *Nat Genet* 2009, **41**:753-761.
9. Ryder E, Russell S: **Transposable elements as tools for genomics and genetics in Drosophila.** *Brief Funct Genomic Proteomic* 2003, **2**:57-71.
10. Mates L, Izsvak Z, Ivics Z: **Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives.** *Genome Biol* 2007, **8(Suppl 1)**:S1.
11. Bian Q, Belmont AS: **BAC TG-EMBED: one-step method for high-level, copy-number-dependent, position-independent transgene expression.** *Nucleic Acids Res* 2010, **38**:e127.
12. Sivalingam J, Krishnan S, Ng WH, Lee SS, Phan TT, Kon OL: **Biosafety assessment of site-directed transgene integration in human umbilical cord-lining cells.** *Mol Ther* 2010, **18**:1346-1356.
13. Huang X, Haley K, Wong M, Guo H, Lu C, Wilber A, Zhou X: **Unexpectedly high copy number of random integration but low frequency of persistent expression of the Sleeping Beauty transposase after trans delivery in primary human T cells.** *Hum Gene Ther* 2010, **21**:1577-1590.
14. Wicks SR, de Vries CJ, van Luenen HG, Plasterk RH: **CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in Caenorhabditis elegans.** *Dev Biol* 2000, **221**:295-307.
15. Devon RS, Porteous DJ, Brookes AJ: **Splinkerettes–improved vectorettes for greater efficiency in PCR walking.** *Nucleic Acids Res* 1995, **23**:1644-1645.
16. Moeller F, Nielsen FC, Nielsen LB: **New tools for quantifying and visualizing adoptively transferred cells in recipient mice.** *J Immunol Methods* 2003, **282**:73-82.
17. Wang LJ, Chen YM, George D, Smets F, Sokal EM, Bremer EG, Soriano HE: **Engraftment assessment in human and mouse liver tissue after sex-mismatched liver cell transplantation by real-time quantitative PCR for Y chromosome sequences.** *Liver Transpl* 2002, **8**:822-828.
18. Ballester M, Castello A, Ibanez E, Sanchez A, Folch JM: **Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals.** *Biotechniques* 2004, **37**:610-613.
19. Joshi M, Keith Pittman H, Haisch C, Verbanac K: **Real-time PCR to determine transgene copy number and to quantitate the biolocalization of adoptively transferred cells from EGFP-transgenic mice.** *Biotechniques* 2008, **45**:247-258.
20. Ivics Z, Hackett PB, Plasterk RH, Izsvak Z: **Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells.** *Cell* 1997, **91**:501-510.
21. Baer M, Nilsen TW, Costigan C, Altman S: **Structure and transcription of a human gene for H1 RNA, the RNA component of human RNase P.** *Nucleic Acids Res* 1990, **18**:97-103.
22. Cui Z, Geurts AM, Liu G, Kaufman CD, Hackett PB: **Structure-function analysis of the inverted terminal repeats of the sleeping beauty transposon.** *J Mol Biol* 2002, **318**:1221-1235.
23. Chung S, Andersson T, Sonntag KC, Bjorklund L, Isacson O, Kim KS: **Analysis of different promoter systems for efficient transgene expression in mouse embryonic stem cell lines.** *Stem Cells* 2002, **20**:139-145.
24. Liew CG, Draper JS, Walsh J, Moore H, Andrews PW: **Transient and stable transgene expression in human embryonic stem cells.** *Stem Cells* 2007, **25**:1521-1528.
25. Xia X, Zhang Y, Zieth CR, Zhang SC: **Transgenes delivered by lentiviral vector are suppressed in human embryonic stem cells in a promoter-dependent manner.** *Stem Cells Dev* 2007, **16**:167-176.
26. Orban TI, Seres L, Ozvegy-Laczka C, Elkind NB, Sarkadi B, Homolya L: **Combined localization and real-time functional studies using a GFP-tagged ABCG2 multidrug transporter.** *Biochem Biophys Res Commun* 2008, **367**:667-673.
27. Orban TI, Apati A, Nemeth A, Varga N, Krizsik V, Schamberger A, Szebenyi K, Erdei Z, Varady G, Karaszi E, Homolya L, Német K, Gócza E, Miskey C, Mátés L, Ivics Z, Izsvák Z, Sarkadi B: **Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery.** *Stem Cells* 2009, **27**:1077-1087.
28. Izsvak Z, Khare D, Behlke J, Heinemann U, Plasterk RH, Ivics Z: **Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition.** *J Biol Chem* 2002, **277**:34581-34588.
29. Apati A, Orban TI, Varga N, Nemeth A, Schamberger A, Krizsik V, Erdelyi-Belle B, Homolya L, Varady G, Padanyi R, Karászi E, Kemna EW, Német K, Sarkadi B: **High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells.** *Biochim Biophys Acta* 2008, **1778**:2700-2709.

doi:10.1186/1759-8753-2-5
Cite this article as: Kolacsek *et al.*: Reliable transgene-independent method for determining *Sleeping Beauty* transposon copy numbers. *Mobile DNA* 2011 **2**:5.

RNA biology

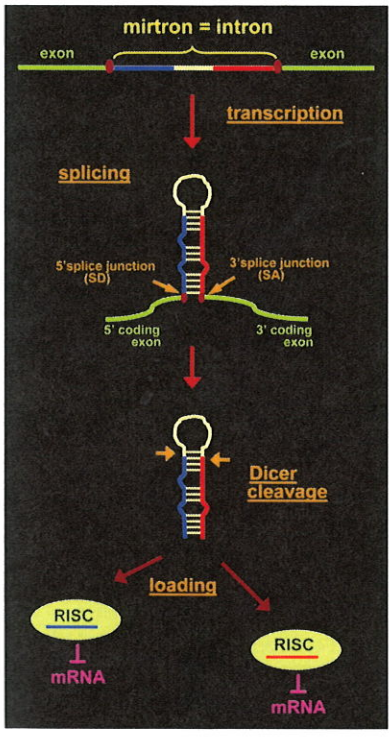
Volume 9 • Issue 9 • September 2012



LANDES BIOSCIENCE

RNA biology

Volume 9 Issue 9
September 2012



About the cover.
Schematic representation of the mammalian mirtron pathway. This special group of microRNAs mature via an alternative processing pathway where the pre-miRNA represents a full intron and instead of the Drosha/DGCR8 complex, the splicing machinery carries out the first cleavage step during processing. In the case of the human mir-877 mirtron, miRNAs can be matured from both arms simultaneously. For more information see Schamberger, et al. pp. 1177–85.

RESEARCH PAPER

RNA Biology 9:9, 1–9; September 2012; © 2012 Landes Bioscience

orban.tamas_188_24

Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner

Anita Schamberger,[1] Balázs Sarkadi[1,2] and Tamás I. Orbán[1,*]

[1]Institute of Molecular Pharmacology; Research Centre for Natural Sciences; Hungarian Academy of Sciences; Budapest, Hungary; [2]Membrane Research Group of the Hungarian Academy of Sciences; Department of Biophysics; Semmelweis University and National Blood Center; Budapest, Hungary

Keywords: mirtron, miRNA, miRNA*, mir-877, mir-1226, mir-1233, RNA interference

Abbreviations: miRNA, microRNA; nt, nucleotide; RISC, RNA-induced silencing complex; 3'-UTR, 3'-untranslated region; MEF, mouse embryonic fibroblast

Mirtrons are short intronic microRNA (miRNA) precursors representing an alternative, Drosha/DGCR8-independent miRNA biogenesis pathway. In this study we characterized three predicted human mirtrons. Their expression was proven to be context-independent, since functional mirtrons could be derived either from their endogenous or from a heterologous coding environment. Systematic testing revealed that both 5'- and 3'-arms of mir-877 are capable of producing functional miRNA simultaneously in the various cell types examined. On the other hand, experimental validations revealed that the predicted mir-1233 is not a *bona fide* mirtron. For functional mirtrons, we were able to detect mature mirtron-derived miRNAs for the first time by qRT-PCR or northern blot analysis, when silencing activity was proven by functional assays. Our results emphasize the need for functional testing of both arms of miRNAs and the importance of experimentally validating human mirtrons since, in spite of being localized in a short intron, predicted species could mature via other miRNA processing pathways.

Introduction

Since the discovery of the first small regulatory RNA molecules,[1,2] microRNAs (miRNAs) have been proven to represent an important level of gene expression control.[3-6] Most of these 20–24 nucleotide (nt) long molecules are formed via a canonical pathway, beginning with the transcription of a long primary precursor (pri-miRNA) with hairpin structure, followed by the cleavage of the nuclear RNase III-like enzyme Drosha, mediated by its partner protein, DGCR8. The so formed pre-miRNA is then recognized and transported by the Exportin-5 shuttle system to the cytoplasm where Dicer, another RNase III-type enzyme, produces the double-stranded miRNA:miRNA* duplex. Further processing involves the incorporation of this species into an Argonaute protein containing RNA-induced silencing complex (RISC) where one strand is removed and the mature RISC is formed. The regulatory effect is then manifested by the destabilization and/or translational inhibition of the target mRNA molecule via partial base pairing of the miRNA and the 3'-untranslated region (3'-UTR) of the mRNA.[7]

Recent discoveries revealed that various alternative miRNA biogenesis pathways exist which could bypass certain steps of the canonical process (reviewed in refs. 8,9). The most prominent Drosha/DGCR8-independent pathway is the mirtron pathway which relies on mRNA splicing and lariat debranching for the formation of pre-miRNA. Mirtrons were first described and experimentally validated in *Drosophila melanogaster* and in *Caenorhabditis elegans*.[10,11] Previous small RNA data analyses[12-17] also suggested the existence of mammalian mirtrons and two contemporary works also investigated the mirtron biogenesis pathway in mammalian cells.[18,19] However, further experimental validation still remained an important issue since experimental data on functional mirtrons are still incomplete, and the presence of mirtron derived mature miRNA species could not be detected directly.[18] In addition, some predicted mirtrons termed "simtrons" were proven to mature via a different pathway, further demanding for experimental confirmation of predicted mirtrons in higher eukaryotes.[19]

In this study we systematically characterized three predicted human mirtrons and confirmed that two of them are indeed bona fide mirtrons. On the other hand, mir-1233 did not show any mirtron characteristic features and it is presumed to be processed via a different miRNA biogenesis pathway. We demonstrate that functional mirtrons can be produced in a coding context-independent manner. Moreover, we show that mir-877 can liberate functional mature miRNAs from both of its 5'- and

*Correspondence to: Tamás I. Orbán; Email: orbant@biomembrane.hu
Submitted: 04/30/12; Revised: 07/02/12; Accepted: 07/04/12
http://dx.doi.org/10.4161/rna.21359

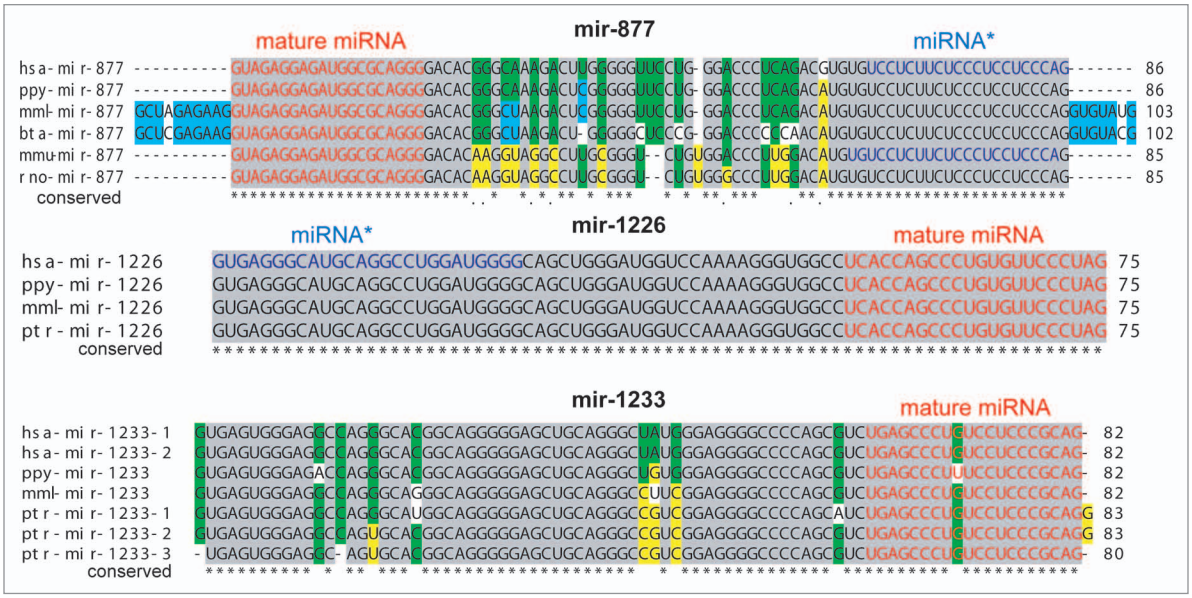


Figure 1. Sequences and evolutionary conservation of three predicted mirtrons. Alignments of mir-877, mir-1226 and mir-1233 were performed using the online program available at www.ebi.ac.uk/Tools/msa/clustalw2/. Mature miRNAs are indicated by red, whereas predicted miRNA* species by blue characters. Asterisks under the alignments represent conserved nucleotides among species. Identical nucleotides are indicated by the same background colors. Hsa, *Homo sapiens*; ppy, *Pongo pygmaeus*; mml, *Macaca mulatta*; bta, *Bos Taurus*; mmu, *Mus musculus*; rno, *Rattus norvegicus*; ptr, *Pan troglodytes*.

3'-arms simultaneously, and we can detect both mature miRNA forms directly in various cell lines examined.

Results and Discussion

Processing of predicted human mirtrons from their original genomic environment. To test whether human mirtrons are indeed functional, we subjected three predicted human mirtron sequences for detailed investigations (mir-877, mir-1226 and mir-1233). These miRNAs are strongly conserved in various mammals. For mir-877, the miRNA and miRNA*, while for mir-1226, even the loop sequence shows a very strong conservation among species (**Fig. 1**). To examine the expression of these mirtrons from their original genomic context, we cloned ~300–400 nt genomic pri-mirtron fragments (mirtrons with their flanking exons) downstream of EGFP (pEGFP_mir-877eie, pEGFP_mir-1226eie, pEGFP_mir-1233eie, **Fig. 2A**). This cloning strategy was previously proven to generate mature miRNAs from similarly sized pri-miRNAs.[11,20,21] After expressing the constructs in HEK-293 cells, splicing was analyzed by RT-PCR using primers flanking the mirtronic cassettes. In all cases, splicing with moderate efficiency was observed (**Fig. 2B**) and splicing accuracy was confirmed by sequencing the shorter RT-PCR products (data not shown).

Having established that the examined mirtrons can be spliced out from their original genomic environment, we assessed their capabilities to silence gene expression. Given that mature miRNAs could originate from both arms of a pre-miRNA in a tissue specific manner,[22] we tested both 5'- and 3'-arms of all examined mirtrons in luciferase sensor assays. Two antisense copies of either the 5'- or the 3'-arm of a particular mirtron were cloned downstream of renilla luciferase in the psiCHECK2 vector. Co-transfection of the mir-877 construct alongside with its 5'- or 3'-sensors resulted in ~60% and ~40% knockdown efficiency, respectively (**Fig. 2C**). For mir-1226, we could detect a moderate but significant effect (~20% downregulation) only with its 3'-sensor construct, whereas mir-1233 exhibited no repression on its sensors. To rule out technical pitfalls for the latter case, we examined the silencing ability of the predicted mature mir-1233 sequence as an shRNA. Expressed from a psiSTRIKE vector, the mature miRNA was capable of potently silence its sensor (~80% repression, **Fig. S1A**), suggesting that impaired processing and/or RISC incorporation may account for its inability to affect gene expression when produced from a mirtron.

A previous study[12] suggested that the 3'-arms of mir-1226 and mir-1233 harbor the dominant miRNA species. In the case of mir-1226, our data indeed confirmed this prediction. As for mir-877, small RNA deep sequencing data suggested the predominant usage of the 5'-arm and we indeed observed a more pronounced effect on the 5'-sensor. However, our results clearly demonstrated that both the miRNA and the miRNA* species are functional at the same time. Repeating the above experiments in various cell lines (e.g., MCF7, HeLa) provided similar results, ruling out cell type specific effects (**Fig. 2D**). In summary, despite the low representation of the 3'-arm of mir-877 in deep sequencing data (two orders of magnitude less than that of the 5'-arm, see mirbase.org), we could provide evidence that it is also incorporated into mature RISCs alongside with the 5'-derived miRNA species, and they have simultaneous silencing ability in various cell lines examined.

Processing of human mirtrons from a heterologous coding context. To investigate whether the mirtron biogenesis pathway

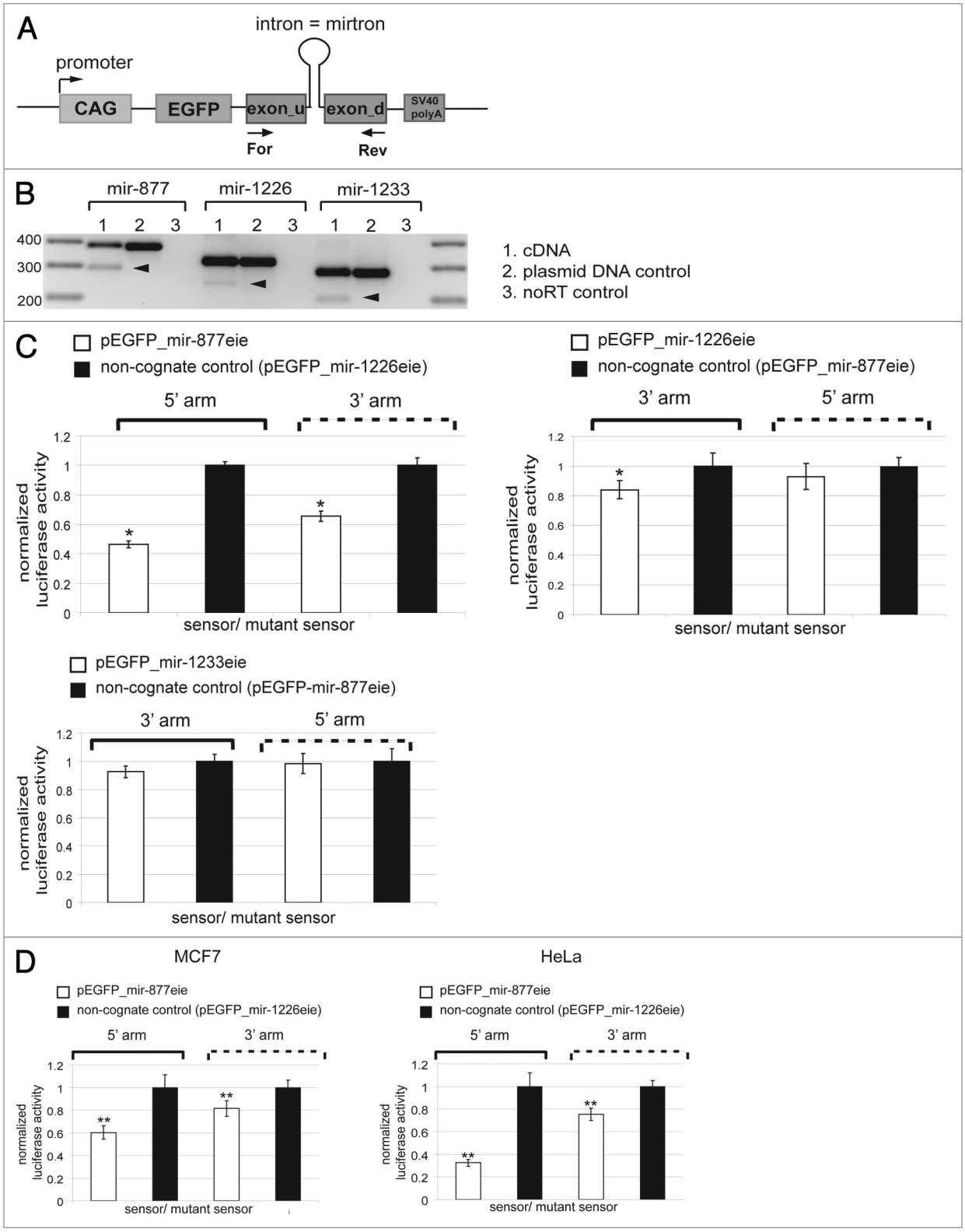


Figure 2. For figure legend, see page 4.

can function in a context-independent manner, we tested the expression of human mirtrons from a heterologous coding environment. We modified the EGFP sequence to separate its coding region into two exons.[23] Proper splicing and expression of the modified EGFP (EGFPm) were confirmed by using an 81 nt long mouse intron and its 5' splice site mutated control in the coding

Figure 2. Analysis of predicted human mirtrons in their original genomic context. (A) Schematic representation of vectors expressing mirtrons from their original genomic context. Mirtron sequences and their flanking host exons were cloned downstream of EGFP (upstream exon = exon_u, downstream exon = exon_d). Arrows indicate primers used for RT-PCR. (B) Analysis of splicing by RT-PCR. Black arrowheads show the spliced forms of RNAs. (C) Luciferase sensor assays in HEK-293 cell line. Sensor constructs containing two antisense copies of the 5'- or 3'-arm of the particular mirtron (mir-877, mir-1226 or mir-1233) were cloned downstream of renilla luciferase in psiCHECK2. Sensor plasmids were cotransfected with the indicated mirtron expression constructs. Plotted is the relative repression of the respective sensor vs. mutant sensor normalized to the non-cognate control. Black lines indicate experiments for the predicted mature miRNA, whereas dashed lines for the predicted miRNA* species. Experiments with three parallels were repeated at least twice, one representative experiment is shown. Error bars represent standard deviations; *$p < 0.05$. (D) Luciferase assays for mir-877 in MCF7 and HeLa cell lines. Experiments with three parallels were repeated at least twice, one representative experiment is shown. Error bars represent standard deviations; **$p < 0.025$.

region of EGFPm (data not shown). Subsequently, the studied mirtrons were cloned between the artificial exons of EGFPm as intronic sequences (**Fig. 3A**). To include a control for the canonical miRNA biogenesis pathway, we also inserted the mir-33b pre-miRNA with its relatively short host intron (475 nt) into the EGFPm construct. Expression in all cases resulted EGFP fluorescence (**Fig. 4B**, **Fig. S2A**) indicating successful splicing, although the efficiency varied among constructs (**Fig. 3B**, **Fig. S1B**). It was low for mir-1226 and higher for mir-877, while in the case of mir-1233, splicing efficiency was nearly 100%.

We also analyzed the ability of the above miRNAs to regulate gene expression when expressed from this heterologous environment (**Fig. 3C**). In luciferase sensor assays, the canonical positive control mir-33b exhibited a strong repression on its sensor (~80% downregulation, **Fig. S1C**). In the same test system, mir-877 and mir-1226 provided essentially the same results to those observed when they were expressed from their endogenous sequence context either in HEK-293 or HeLa cells (**Fig. 3C**). mir-1233 did not silence its sensor, on the contrary, it produced a modest but reproducible increase in luciferase activity, even when compared with various non-cognate controls (**Fig. 3C**, **Fig. S2B**). The mechanism behind the detected mir-1233 activity is currently unknown. There are specific examples when miRNAs can enhance translation,[24-26] but further experiments are needed to better understand the unusual behavior of mir-1233. However, it is important to note that this stimulatory effect was not detectable in DGCR8 deficient cells (**Fig. S2C**) which suggests a DGCR8-dependent mechanism and therefore argues against the maturation of this miRNA via even the simtron pathway.[19]

Nevertheless, the above data strongly support that bona fide human mirtrons (mir-877 and mir-1226) can enter the miRNA biogenesis pathway and function properly when expressed either from their original or from a heterologous intronic locus.

Maturation of bona fide mirtrons depends on splicing but not on Drosha/DGCR8 processing. To examine the mirtronic features of these sequences we first tested their dependency on splicing. The consensus 5' donor splice site of each mirtron was mutated (sm) in order to prevent intron removal (GT →TG, **Fig. 4A**). Upon transfection into mammalian cells, no EGFP expression was detected by fluorescence microscopy (**Fig. 4B**, **Fig. S2A**) and RT-PCR analysis confirmed that the introduced mutations indeed abolished splicing (**Fig. 4C**).

Next, luciferase assays were performed to investigate how the inhibition of splicing affected the ability of the different constructs to silence gene expression. The dominant arms were tested for each mirtron: 5'-sensor for mir-877sm and 3'-sensor for mir-1226sm. In both cases the silencing effect was abolished when the splicing-deficient mutants were expressed (**Fig. 4D**). Impaired splicing had no effect on the ability of the canonical mir-33b miRNA to silence gene expression (**Fig. S2B**).

Unlike canonical miRNAs, the maturation process of mirtrons is suggested to be independent of the DGCR8/Drosha complex. To confirm this, we performed luciferase sensor assays in wild type and DGCR8 deficient mouse embryonic fibroblasts (MEFs)[27] (**Fig. 4E**). As expected, the canonical mir-33b miRNA strongly repressed its sensor in wild type cells (~90% knockdown), which was essentially abolished in DGCR8 deficient MEFs. In contrast, the inhibitory effect of mir-877 on its sensor observed in wild type MEFs was increased even further in DGCR8 deficient MEFs (~50% repression).

Our results confirm the findings of a contemporary study[18] with a more direct method showing that the maturation process of mammalian mirtrons is indeed independent of the Drosha/DGCR8 complex while it requires splicing. The observed increased inhibitory effect of mir-877 could be the result of reduced substrate competition from endogenous pre-miRNAs in the DGCR8 deficient cell line, since disruption of the main endogenous miRNA processing machinery may lead to an increased accessibility of the miRNA pathway downstream of Drosha.

Detection of the mature miRNA forms of human mirtrons. Until now, mammalian mirtrons could only be detected indirectly by deep sequencing and 3'RACE.[12,18] In contrast, we were able to detect the mir-1226 mirtron directly by northern blot analysis. The pre- and mature forms of the mirtron can be identified when expressed from its endogenous or a heterologous environment but not from the 5' splice site mutated vector. As expected, the pre- and mature forms of the canonical mir-33b were also detectable by expressing it from both wild type and splice site mutant constructs (**Fig. 5A**). Northern analysis for the mature mir-1233 revealed several bands even in the case of the splicing control, making it impossible to unambiguously identify the pre- and the mature forms of the mir-1233 and further suggesting some unknown processing mechanisms (data not shown). As for mir-877 we were not able to detect either the pre- or the mature forms, most likely due to its low abundance, which finding is in agreement with previous observations.[18]

To circumvent this problem, we applied the more sensitive real-time PCR detection method to verify the existence of mirtron derived miRNAs. We used the *Sleeping Beauty* transposon technology[28] to establish HeLa cell lines stably expressing

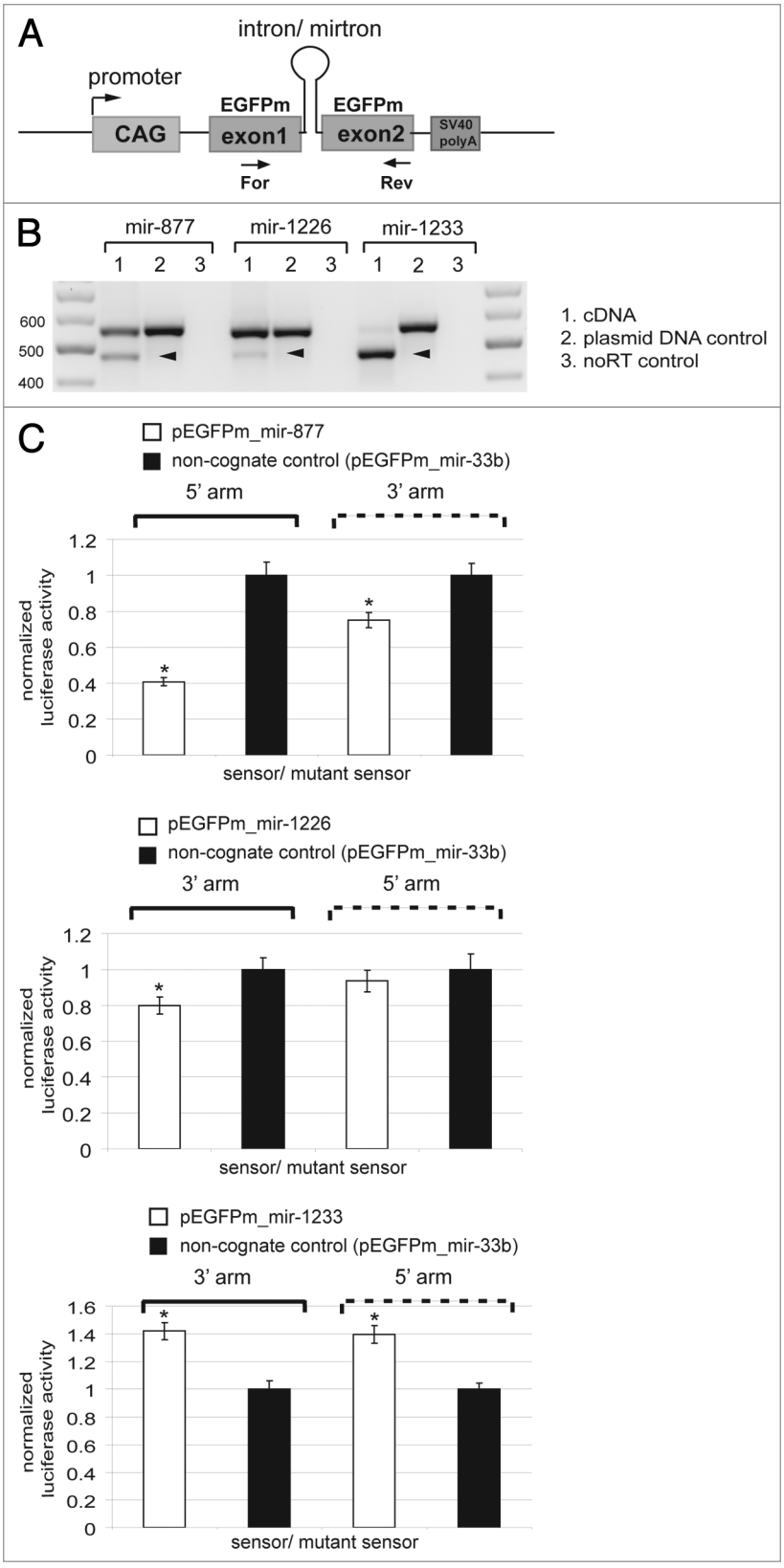


Figure 3. Analysis of human mirtrons in heterologous coding environment. (A) Construct used for expressing mirtron as an artificial EGFPm intron. Arrows indicate primers used for RT-PCR. (B) RT-PCR results for the analysis of splicing. The spliced forms of RNAs are shown by black arrowheads. (C) Luciferase sensor assays of the three mirtrons in HeLa cells. Experimental details are the same as in Figure 2C; error bars represent standard deviations, *p < 0.025.

the examined mirtrons or the canonical mir-33b miRNA as EGFP introns. In agreement with our luciferase measurements, we were able to detect mature miRNA species originating from both the 5'- and the 3'-arms of mir-877 (**Fig. 5B**). Mature form of mir-1226 from a heterologous coding environment was also detectable at a level of one order of magnitude higher than that of mir-877, being consistent with the northern blot analysis data. In addition, miRNA species of these mirtrons could also be detected by real-time PCR when expressed transiently from their endogenous genomic context (**Fig. S3**). Concerning mir-1233, however, we were not able to detect mature miRNA molecules (**Fig. 5B**) which is in accordance with our luciferase assays and northern blot experiments.

There are emerging data on the mirtron biogenesis pathway, indicating the importance of this alternative miRNA maturation pathway in various organisms, including plant species.[29,30] Our results underline the importance of experimental validation of predicted mirtrons, especially in mammalian systems. In summary, we have confirmed that mir-877 and mir-1226 are typical mirtrons while mir-1233 might represent a different maturation pathway. Our results suggest that mir-1233 is neither a bona fide mirtron nor a simtron. In spite of its unusual features, mir-1233 miRNA was claimed to have medical relevance in certain leukemia samples[31] and as a biomarker for renal cell carcinoma.[32] In another study, potential function for mir-1226 as a canonical miRNA in tumor suppression was also described.[33] Our study confirmed that mir-877 and mir-1226 are typical mirtrons and showed that they function properly when expressed either from their endogenous or from a heterologous coding context. In a previous study, higher splicing efficiencies were detected for some predicted mirtrons when longer flanking genomic sequences were included in the reporter constructs which may be due to the presence of other splicing enhancer elements.[19] However, our results suggest that neither good splicing efficiency nor high overall miRNA expression level is a strict prerequisite for mirtron functionality. We have provided the first experimental evidence that mature miRNAs can be derived simultaneously from both arms of the mir-877 mirtron and both are able to repress their respective sensors. Concerning target predictions for mir-877, there is a predicted target for the 5'-arm derived miRNA species, although the effect could not be reassuringly validated.[18] Until now, overwhelming amount of deep sequencing data suggest

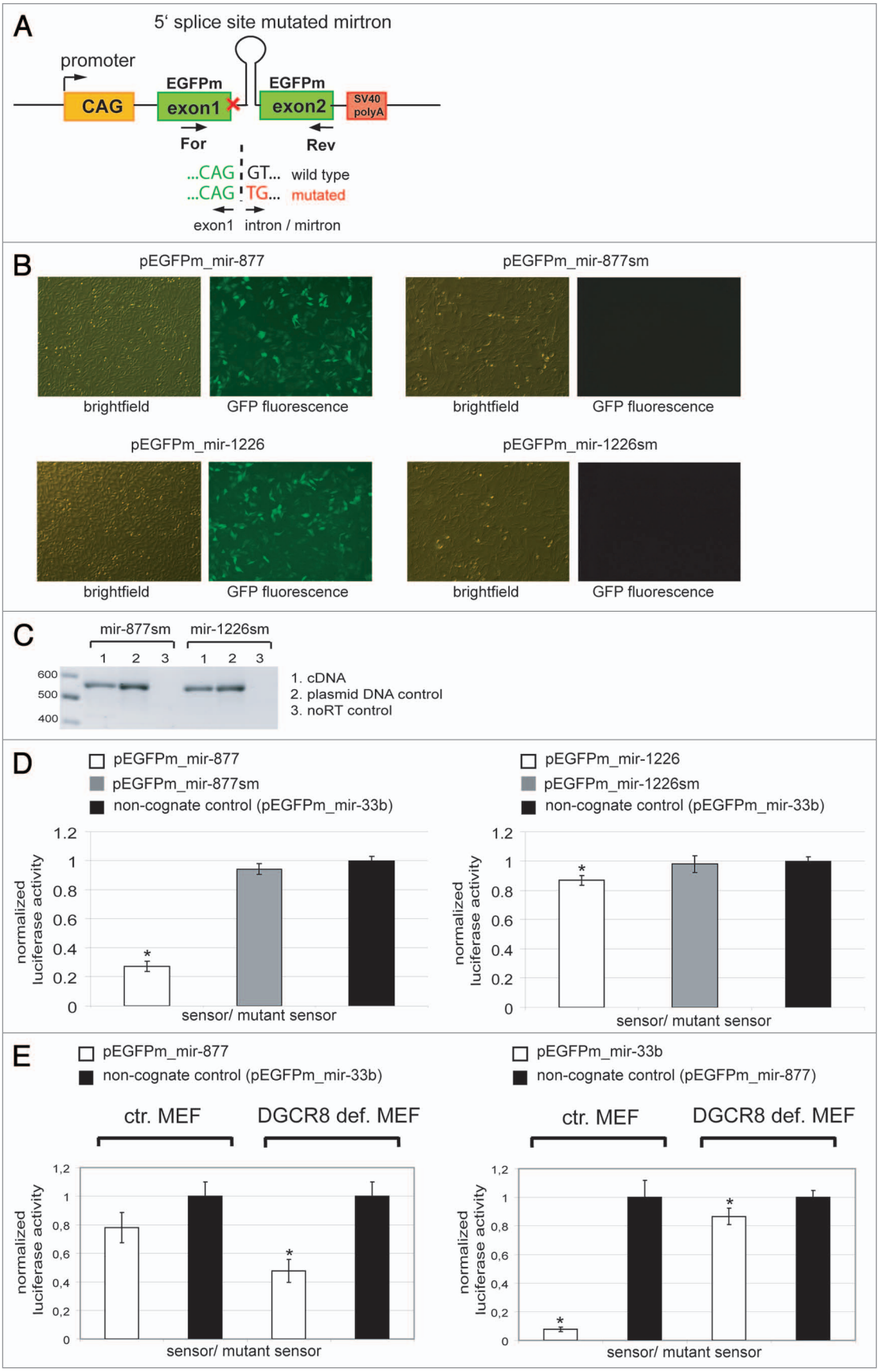


Figure 4. For figure legend, see page 7.

Figure 4. The human mirtron biogenesis pathway is splicing dependent and DGCR8 (Drosha) independent. (A) Schematic representation of vectors expressing 5' splice site mutated mirtrons. Arrows under the exons indicate primers used for RT-PCR. (B) Fluorescence microscopy photos of HeLa cells expressing non-mutated and mutated (sm) mirtrons from heterologous coding context. EGFP positive cells represent successful splicing. (C) RT-PCR results for analysis of splicing from mirtrons with splice site mutations; splicing is not detectable. (D) Luciferase sensor assays of the wild type vs. splice site mutant mirtrons in HeLa cells. For technical details, see Figure 2C and the methods section; *p < 0.05. (E) Luciferase sensor assays in MEF cell lines. The canonical pathway (mir-33b) requires DGCR8 (Drosha), whereas the repression effect of a mirtron (mir-877, 5'-arm derived miRNA) on the luciferase activity is more pronounced in the DGCR8 deficient background. *p < 0.05.

that mature miRNA species could be derived from both arms of a given miRNA. Based on functional assays, here we provided experimental evidence that both arm-derived miRNA species can have regulatory function simultaneously in a particular cell line. In the light of our results, it seems to be important to also examine the "passenger" strand when exploring the relevant physiological functions of different miRNAs since the simultaneously expressed small RNA species might bear separate functions at the same time.

Materials and Methods

Plasmid constructs. Human mirtrons with their flanking host gene exons were amplified and blunt end ligated into our transposon based EGFP expression vector behind the coding region into the BsaBI restriction site.[28] For expressing mirtrons from a heterologous sequence environment, first a PvuII restriction site was introduced into the EGFP sequence by site directed mutagenesis.[23] Oligonucleotides corresponding to the sense and antisense sequence of the specific mirtrons were hybridized to form a double-stranded DNA, then inserted as an artificial intron into the PvuII site of EGFP by blunt end ligation. To generate splicing controls, oligonucleotides with mutated 5' splicing sites (GT > TG) were used. For non-mirtronic controls, the third intron of mouse IgCε gene[23] and the human SREBF1 gene intron containing mir-33b were PCR amplified from genomic DNA samples and inserted into the PvuII site of EGFP.

For the luciferase constructs, hybridized double stranded oligonucleotides for the wild type and the seed region mutated target sequences of corresponding mirtrons were ligated between the XhoI/NotI restriction sites of renilla luciferase 3'-UTR in the psiCHECK2 vector (Promega). To express mir-1233 as shRNA, double stranded oligonucleotides corresponding to the mature mirtron sequence were inserted into the linearized psiSTRIKE vector (Promega).

The sequences of the used oligonucleotides are shown in **Table S1**. All constructs were verified by DNA sequencing using an ABI310 Genetic Analyzer (Applied Biosystems).

Cell cultures and manipulation. HEK-293, MCF-7 and HeLa cell lines were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% of fetal calf serum, 1% of L-glutamine, and 1% of penicillin/streptomycin (Invitrogen) using standard cell culture methodology. Wild type and DGCR8 deficient mouse embryonic fibroblast cell lines[34]—kindly provided by Robert Blelloch—were also maintained in DMEM except that the medium did not contain antibiotics.

For the highest transfection efficiency, cell lines were transfected with reagents providing the most optimal results. Namely, HEK-293 was transfected with FuGENE® 6 (Roche Applied Science), HeLa and MCF7 cell lines were transfected with FuGENE® HD (Roche Applied Science), whereas MEF cell lines were transfected with Lipofectamine® 2000 (Invitrogen). EGFP fluorescence was detected by a IX51 fluorescence microscope (Olympus).

To establish cell lines stably expressing EGFPm-mirtron/miRNA constructs, we applied the *Sleeping Beauty* transposon based gene delivery technology as described earlier.[28] Following transfection, cells were sorted for GFP positivity at day 8 and subsequently at day 17 using a FACS Aria High Speed Cell Sorter (Beckton-Dickinson), to obtain homogenously expressing cell populations.

RNA analysis. Total RNA was isolated from cultured cells using Trizol reagent (Invitrogen). To remove genomic DNA contaminations, RNA samples were treated with DNaseI (New England Biolabs) at 37°C for 1 h. For cDNA preparations, 1 μg of total RNA was reverse transcribed with random primers using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems). Polymerase chain reaction was performed on cDNA samples using primers indicated in **Table S1**.

For northern blot analysis of miRNAs, we essentially followed the protocol published recently.[35] Briefly, 20–30 μg total RNA samples were run on a 12% polyacrylamide gel for 2 h by constant 400 V. RNAs were blotted and crosslinked to a Hybond N+ membrane (Amersham), then hybridized with ^{32}P-labeled LNA probes (Exiqon) specific for the studied mirtrons, and ^{32}P-labeled DNA probe for the U6 snRNA as endogenous control (**Table S1**). Radioactively labeled RNA markers (Decade Markers, Ambion) were used as size markers. Labeled membranes were exposed to X-ray films by standard methodology.

To carry out real-time quantitative PCR, we used small RNA specific TaqMan® assays and reagents, and performed the reactions on StepOne™ or StepOnePlus™ platforms, according to the manufacturer's instructions (Applied Biosystems). For mir-877*, mir-1226, mir-1233 and mir-33b, pre-developed miRNA TaqMan® assays were purchased, whereas a custom made small RNA TaqMan® assay was ordered for mir-877.

Luciferase assays. For each construct, 300 ng of the mirtron/miRNA expressing plasmids were co-transfected with 15 ng of sensor or mutant sensor luciferase plasmids into cells seeded on a 24-well plate. Sensors containing two fully complementary antisense copies of the predicted mature miRNA species were cloned downstream of renilla luciferase in psiCHECK2 vector. Mutant sensors differ in 3 mismatched nucleotides in the predicted miRNA seed region (every second nucleotide). Luciferase activity was measured at 48 h posttransfection by a 2030 Multilabel Reader luminometer (PerkinElmer) using the Dual-Luciferase Reporter Assay System (Promega). Signal specific for firefly luciferase

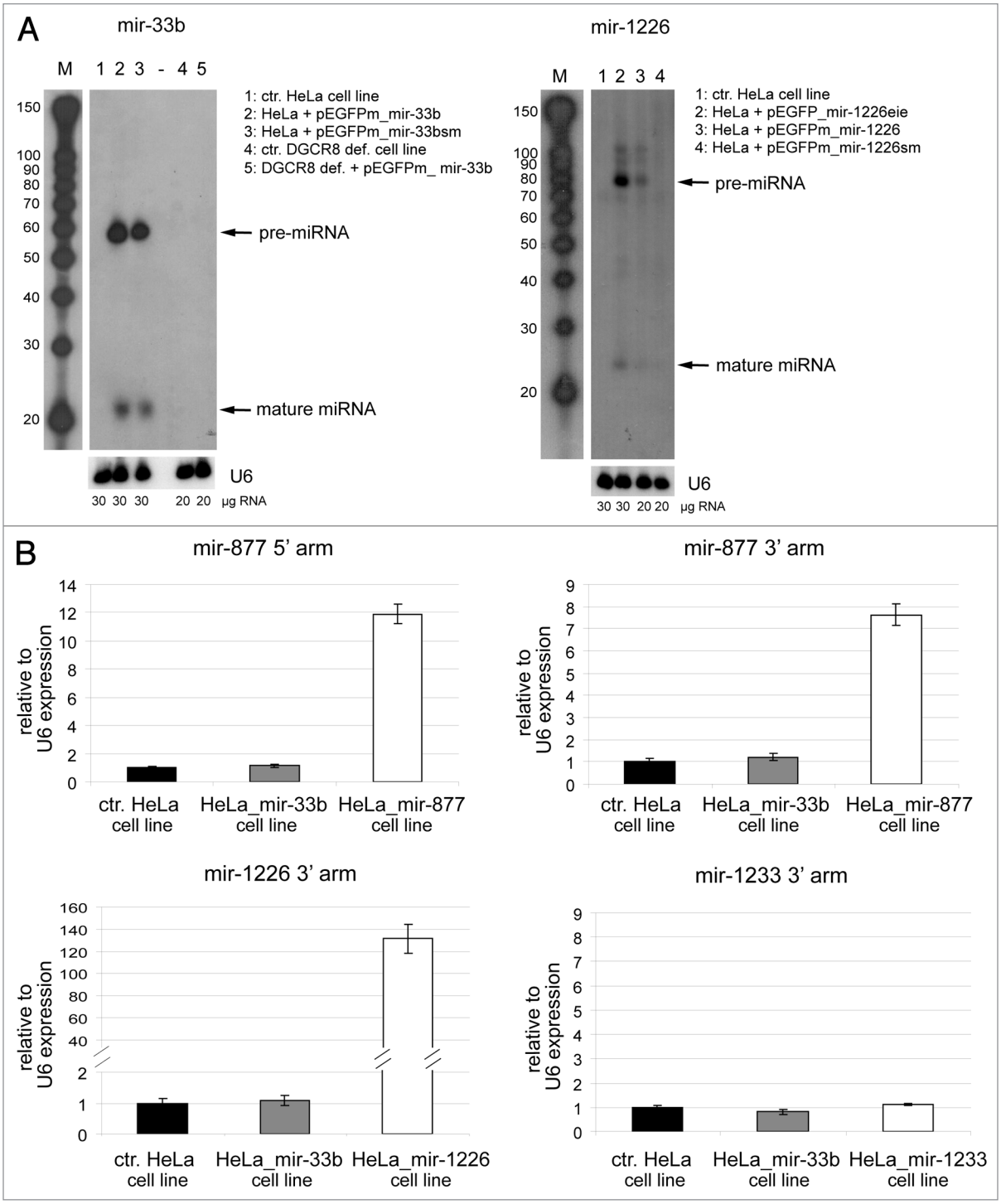


Figure 5. Detection of mature miRNA forms. (A) Northern blot analysis of mir-33b canonical (left) and mir-1226 mirtronic (right) miRNAs in the indicated cell lines expressing the different constructs. Numbers above the lanes correspond to RNA samples from given cell lines transfected with the indicated miRNA/mirtron expression plasmid. Different amounts of total RNA loaded on the gel indicated below the corresponding wells. Signals for the U6 snRNA served as loading control, M stands for labeled RNA markers. (B) Real-time PCR results using TaqMan® assays for mature miRNA detection in different cell lines indicated. The assay for U6 snRNA served as the endogenous control, error bars represent standard error of the means.

expressed from the same psiCHECK2 plasmid was used to normalize for transfection efficiency. To exclude any non-specific effects, a given sensor/mutant sensor ratio was always normalized to the signal ratio measured in the presence of an unrelated miRNA expression construct (non-cognate control, its value was arbitrarily set to 1). Experiments with three parallels were repeated at least twice. For statistical analysis, two-sided Student's t-test was performed.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We would like to thank Károly Fátyol for helpful discussions and critical reading of the manuscript, and Robert Blelloch's laboratory for providing the wild type and the DGCR8 deficient MEF cell lines. We are indebted to Éva Várallyay and Zoltán Havelda for the technical help with the small RNA specific Northern analysis. Tamás I. Orbán is a recipient of the János Bolyai Scholarship of the Hungarian Academy of Sciences. This work was supported by grants from OTKA (NK83533), STEMKILL (OM00108/2008), and KMOP-1.1.2–07/1–2008–0003.

Supplemental Material

Supplemental material may be found here: www.landesbioscience.com/journals/rnabiology/article/21359

References

1. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75:843-54; PMID:8252621; http://dx.doi.org/10.1016/0092-8674(93)90529-Y.
2. Wightman B, Ha I, Ruvkun G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 1993; 75:855-62; PMID:8252622; http://dx.doi.org/10.1016/0092-8674(93)90530-4.
3. Carthew RW, Sontheimer EJ. Origins and Mechanisms of miRNAs and siRNAs. Cell 2009; 136:642-55; PMID:19239886; http://dx.doi.org/10.1016/j.cell.2009.01.035.
4. Ghildiyal M, Zamore PD. Small silencing RNAs: an expanding universe. Nat Rev Genet 2009; 10:94-108; PMID:19148191; http://dx.doi.org/10.1038/nrg2504.
5. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 2008; 9:102-14; PMID:18197166; http://dx.doi.org/10.1038/nrg2290.
6. Slezak-Prochazka I, Durmus S, Kroesen BJ, van den Berg A. MicroRNAs, macrocontrol: regulation of miRNA processing. RNA 2010; 16:1087-95; PMID:20423980; http://dx.doi.org/10.1261/rna.1804410.
7. Huntzinger E, Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 2011; 12:99-110; PMID:21245828; http://dx.doi.org/10.1038/nrg2936.
8. Miyoshi K, Miyoshi T, Siomi H. Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. Mol Genet Genomics 2010; 284:95-103; PMID:20596726; http://dx.doi.org/10.1007/s00438-010-0556-1.
9. Yang JS, Lai EC. Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. Mol Cell 2011; 43:892-903; PMID:21925378; http://dx.doi.org/10.1016/j.molcel.2011.07.024.
10. Ruby JG, Jan CH, Bartel DP. Intronic microRNA precursors that bypass Drosha processing. Nature 2007; 448:83-6; PMID:17589500; http://dx.doi.org/10.1038/nature05983.
11. Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 2007; 130:89-100; PMID:17599402; http://dx.doi.org/10.1016/j.cell.2007.06.028.
12. Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC. Mammalian mirtron genes. Mol Cell 2007; 28:328-36; PMID:17964270; http://dx.doi.org/10.1016/j.molcel.2007.09.028.
13. Babiarz JE, Ruby JG, Wang Y, Bartel DP, Blelloch R. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. Genes Dev 2008; 22:2773-85; PMID:18923076; http://dx.doi.org/10.1101/gad.1705308.
14. Babiarz JE, Hsu R, Melton C, Thomas M, Ullian EM, Blelloch R. A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in Dgcr8 versus Dicer1 knockouts and small RNA sequencing. RNA 2011; 17:1489-501; PMID:21712401; http://dx.doi.org/10.1261/rna.2442211.
15. Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, et al. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. Genes Dev 2010; 24:992-1009; PMID:20413612; http://dx.doi.org/10.1101/gad.1884710.
16. Chong MM, Zhang G, Cheloufi S, Neubert TA, Hannon GJ, Littman DR. Canonical and alternate functions of the microRNA biogenesis machinery. Genes Dev 2010; 24:1951-60; PMID:20713509; http://dx.doi.org/10.1101/gad.1953310.
17. Glazov EA, Kongsuwan K, Assavalapsakul W, Horwood PF, Mitter N, Mahony TJ. Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. PLoS One 2009; 4:e6349; PMID:19633723; http://dx.doi.org/10.1371/journal.pone.0006349.
18. Sibley CR, Seow Y, Saayman S, Dijkstra KK, El Andaloussi S, Weinberg MS, et al. The biogenesis and characterization of mammalian microRNAs of mirtron origin. Nucleic Acids Res 2012; 40:438-48; PMID:21914725; http://dx.doi.org/10.1093/nar/gkr722.
19. Havens MA, Reich AA, Duelli DM, Hastings ML. Biogenesis of mammalian microRNAs by a non-canonical processing pathway. Nucleic Acids Res 2012; 40:4626-40; PMID:22270084; http://dx.doi.org/10.1093/nar/gks026.
20. Lai EC, Tam B, Rubin GM. Pervasive regulation of Drosophila Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. Genes Dev 2005; 19:1067-80; PMID:15833912; http://dx.doi.org/10.1101/gad.1291905.
21. Stark A, Brennecke J, Russell RB, Cohen SM. Identification of Drosophila MicroRNA targets. PLoS Biol 2003; 1:E60; PMID:14691535; http://dx.doi.org/10.1371/journal.pbio.0000060.
22. Mah SM, Buske C, Humphries RK, Kuchenbauer F. miRNA*: a passenger stranded in RNA-induced silencing complex? Crit Rev Eukaryot Gene Expr 2010; 20:141-8; PMID:21133843; http://dx.doi.org/10.1615/CritRevEukarGeneExpr.v20.i2.40.
23. Lacy-Hulbert A, Thomas R, Li XP, Lilley CE, Coffin RS, Roes J. Interruption of coding sequences by heterologous introns can enhance the functional expression of recombinant genes. Gene Ther 2001; 8:649-53; PMID:11320412; http://dx.doi.org/10.1038/sj.gt.3301440.
24. Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. Science 2007; 318:1931-4; PMID:18048652; http://dx.doi.org/10.1126/science.1149460.
25. Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 2008; 30:460-71; PMID:18498749; http://dx.doi.org/10.1016/j.molcel.2008.05.001.
26. Henke JI, Goergen D, Zheng J, Song Y, Schüttler CG, Fehr C, et al. microRNA-122 stimulates translation of hepatitis C virus RNA. EMBO J 2008; 27:3300-10; PMID:19020517; http://dx.doi.org/10.1038/emboj.2008.244.
27. Han J, Pedersen JS, Kwon SC, Belair CD, Kim YK, Yeom KH, et al. Posttranscriptional crossregulation between Drosha and DGCR8. Cell 2009; 136:75-84; PMID:19135890; http://dx.doi.org/10.1016/j.cell.2008.10.053.
28. Orbán TI, Apáti A, Németh A, Varga N, Krizsik V, Schamberger A, et al. Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery. Stem Cells 2009; 27:1077-87; PMID:19415778; http://dx.doi.org/10.1002/stem.45.
29. Meng Y, Shao C. Large-scale identification of mirtrons in Arabidopsis and rice. PLoS One 2012; 7:e31163; PMID:22348048; http://dx.doi.org/10.1371/journal.pone.0031163.
30. Joshi PK, Gupta D, Nandal UK, Khan Y, Mukherjee SK, Sanan-Mishra N. Identification of mirtrons in rice using MirtronPred: a tool for predicting plant mirtrons. Genomics 2012; 99:370-5; PMID:22546559; http://dx.doi.org/10.1016/j.ygeno.2012.04.002.
31. Rainer J, Ploner C, Jesacher S, Ploner A, Eduardoff M, Mansha M, et al. Glucocorticoid-regulated microRNAs and mirtrons in acute lymphoblastic leukemia. Leukemia 2009; 23:746-52; PMID:19148136; http://dx.doi.org/10.1038/leu.2008.370.
32. Wulfken LM, Moritz R, Ohlmann C, Holdenrieder S, Jung V, Becker F, et al. MicroRNAs in renal cell carcinoma: diagnostic implications of serum miR-1233 levels. PLoS One 2011; 6:e25787; PMID:21984948; http://dx.doi.org/10.1371/journal.pone.0025787.
33. Jin C, Rajabi H, Kufe D. miR-1226 targets expression of the mucin 1 oncoprotein and induces cell death. Int J Oncol 2010; 37:61-9; PMID:20514397.
34. Wang Y, Medvid R, Melton C, Jaenisch R, Blelloch R. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet 2007; 39:380-5; PMID:17259983; http://dx.doi.org/10.1038/ng1969.
35. Várallyay E, Burgyán J, Havelda Z. MicroRNA detection by northern blotting using locked nucleic acid probes. Nat Protoc 2008; 3:190-6; PMID:18274520; http://dx.doi.org/10.1038/nprot.2007.528.

OPEN ACCESS Freely available online

PLOS ONE

3′ IsomiR Species and DNA Contamination Influence Reliable Quantification of MicroRNAs by Stem-Loop Quantitative PCR

Anita Schamberger[1], **Tamás I. Orbán**[1,2]*

1 Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary, **2** Chemical Technology Transfer Ltd., Budapest, Hungary

Abstract

MicroRNAs (miRNAs) are ~20–24 nucleotide-long regulatory RNAs that have been proven to play important roles in many cellular processes. Since their discovery, a number of different techniques have been developed to detect and accurately quantify them. For individual mature miRNA measurements, quantitative stem-loop real-time PCR represents a widely used method. Although there are some data on optimization of this technique, there are still many factors that have not been investigated yet. In this study, we have thoroughly optimized this technique and pointed out several important factors that influence reliable quantification. First, we found that total RNA input can affect the measurements. Second, our data showed that carryover DNA contamination could also mislead the detection in a sequence-specific manner. Additionally, we provided evidence that different 3′ isomiR species of a particular miRNA can be reverse transcribed and cross-detected even by specifically targeted assays. Besides these, we have investigated the measurement of reaction efficiencies from total RNA samples and the accuracy of simultaneous reverse transcription reactions for increasing reliability and cost effectiveness without the loss of sensitivity and specificity. In summary, we provide a detailed, refined protocol for reliable detection of microRNA species by quantitative stem-loop PCR.

Citation: Schamberger A, Orbán TI (2014) 3′ IsomiR Species and DNA Contamination Influence Reliable Quantification of MicroRNAs by Stem-Loop Quantitative PCR. PLoS ONE 9(8): e106315. doi:10.1371/journal.pone.0106315

Editor: Pedro Gonzalez, Duke University, United States of America

Received November 26, 2013; **Accepted** August 5, 2014; **Published** August 29, 2014

Funding: Anita Schamberger is a recipient of the Jedlik Ányos predoc fellowship. This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 'National Excellence Program' and also supported by the TransRat grant (KMR_12-2012-0112) from the Hungarian National Development Agency given to Tamás I. Orbán. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Author Tamás I. Orbán is employed by Chemical Technology Transfer Ltd. Chemical Technology Transfer Ltd provided support in the form of salary for author TIO, but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The specific role of this author is articulated in the ''author contributions'' section.

Competing Interests: Author Tamás I. Orbán is employed by Chemical Technology Transfer Ltd. There are no patents, products in development or marketed Cover Letter products to declare. This does not alter the authors' adherence to all the PLoS ONE policies on sharing data and materials.

* Email: orban.tamas@ttk.mta.hu

Introduction

MicroRNAs (miRNAs) are short, non-coding regulatory RNA molecules that control mRNA stability and translation by targeting the 3′ untranslated region of given mRNA species [1,2]. They influence various cellular functions and now are believed to form a crucial and extensive regulatory network similar to that of transcription factors [3]. The biogenesis of miRNAs consists of different, subsequent processing steps during which mature miRNA is liberated from longer precursor RNA forms [4–6]. In order to understand proper regulation and function, the different RNA forms can be studied and measured by various techniques. In the general laboratory practice, however, it is often sufficient to measure individual mature miRNA steady state levels. Nevertheless, measurements are challenging due to their short size, and sequence specific detection methods are more limited than in the case of mRNA molecules. Traditional hybridization techniques using radioactively or fluorescently labeled nucleic acids are generally applied, including in situ hybridization [7,8] or Northern blotting [9–11]. Their sensitivity can be strongly increased by using specifically modified artificial nucleotides, such as locked nucleic acids (LNAs) [12–15], but miRNAs with low abundance can still be beyond the sensitivity of these methods [16,17].

Similarly to mRNA detection and quantification, measuring the expression level of miRNA species by real-time PCR represents one of the most sensitive and accurate methods developed so far for such purposes. However, due to the short nature of miRNAs, a specific stem-loop real-time PCR technique has been developed among other methodologies [18–20]. The detection of mature miRNAs by this technique is composed of two main steps (Figure 1). The first step is a specifically targeted cDNA synthesis when a sequence specific stem-loop primer is hybridized to the mature miRNA and used to initiate the reverse transcription reaction. The second step is the real-time PCR during which the extended and transcribed miRNA is quantified using oligos specific for the miRNA and the primer loop sequences. This technique is fast and could be standardized for high-throughput purposes. However, this method has the *a priori* assumption that the miRNA in question has a well-defined 3′ end. Conversely, based on deep sequencing results, recent reports described significant sequence length heterogeneity of miRNAs originating from a given locus, often having significant variability of their 5′

and/or 3′ ends [21,22]. Moreover, the distribution of such isomiRs seems to vary among cell types or physiological statuses of the cells [23,24]. Therefore, such 3′ end variability could seriously influence miRNA detection by stem-loop PCR by interfering with the very first step, the sequence specific reverse transcription. There are several data on optimization of miRNA detection from discussing RNA isolation techniques to comparing various platforms [25–30]. Nevertheless, there are many other factors during individual mature miRNA detection by the widely used stem-loop quantitative PCR that are not discussed yet, although they play important roles in the accuracy and reproducibility of the measurements.

In this study, we intended to systematically investigate the stem-loop real-time PCR detection method of small RNA molecules. Careful optimization of this technique pointed to a previously underestimated aspect, that total RNA input and DNA contamination could severely influence the accurate detection. Moreover, we provide evidence that 3′ isomiR species are not exclusively measured by the stem-loop qRT-PCR methodology, and thereby can be cross-detected. This latter problem could not be overcome even by using the poly(A)-tailing-based qRT-PCR methodology. On the other hand, simultaneous reverse transcription of the target miRNA and the endogenous control does not necessarily influence the outcome of the results and may be a more accurate and cost effective approach for miRNA level quantitation. Based on our experiments, we suggest a refined protocol of miRNA detection by stem-loop real-time PCR technology.

Results

Relative quantification, reaction efficiency and the amount of reverse transcribed RNA

In quantitative RT-PCR applications, determination of the target is based on absolute or relative quantification. For individual miRNA measurements, relative quantification is the commonly used method, when the amount of the target is determined relative to an endogenous control. Since the target is compared to the control, they must be amplified with similar efficiencies. The accurate amplification efficiency in practice is calculated from the slope of a standard curve made by at least 5 points, encompassing the relevant concentration range of the application. Making an accurate and reproducible standard curve for miRNAs from total RNA samples (which is physiologically more relevant than using synthetic oligos) is challenging, since many miRNAs are present in low abundance. A sensitive balance has to be found between the sufficient dilution of the reverse transcription reaction (e.g.: for mRNA detection, it is a minimum of 1:10) and an optimal C_t value (delayed by the dilution of the reverse transcription reaction; Figure S1). Therefore, we recommend the employment of small dilution steps (e.g. 1.5×) with only 3 or 4 points in the strict range of the measurement. If there is appropriate correlation between the control and target, the relative quantification method can be used at the particular dilution range for the analysis of the measurements.

The next question is about the optimal amount of total RNA used for the reverse transcription reaction. As mentioned above, mature miRNA levels are often low in certain samples. Therefore, one could speculate to increase the amount of total RNA to increase the input of mature miRNAs in the reverse transcription reaction. To investigate this question, we measured more and less abundant miRNAs (abundance was estimated based on previous data: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_089374.pdf and [17,30]) from different total RNA input, relative to the widely used U6 small nuclear RNA (snRNA) or to the miR-21-5p endogenous miRNA. Target and endogenous control samples were prepared simultaneously and measured in the same plate during the real-time PCR reaction. The increase of the total RNA amount resulted in a decrease of mature miRNA detection when applying the U6 endogenous snRNA control (Figure 2A). It has dropped significantly above 20 ng in general and the effect did not seem to depend on the abundance of the miRNA target. Considering the miR-21-5p endogenous control, the effect of the total RNA input on the measurements was less pronounced (Figure 2B). Finally, the lower range of RNA input was measured less accurately probably because the low template concentration leads to delayed C_t values. Summarizing the results, the optimal range of RNA input varies depending on endogenous controls and targets, therefore, should be optimized. Based on our data, however, 10 ng total RNA input can be appropriate when using U6 endogenous snRNA control and 20 ng with the application of miR-21-5p endogenous miRNA control.

Different targets can be reverse transcribed in the same reaction

For cDNA synthesis of miRNAs, the different small RNA targets have unique, sequence specific stem-loop primers to assist their reverse transcription. Although numerous miRNAs are reverse transcribed together in array experiments [31–33], it is indicated in the general protocol that for individual miRNA measurements, the endogenous control and the target have to be reverse transcribed in separate reactions (http://tools.lifetechnologies.com/content/sfs/manuals/cms_042167.pdf). To examine the feasibility of the simultaneous reactions, we compared real-time PCR measurements of simultaneously and separately reverse transcribed samples. We measured the level of several miRNAs including miR-1226-3p and miR-33b-5p in stably overexpressing HeLa cell lines, and the endogenous miR-21-5p in normal HeLa cell line. We found that there is no significant difference in the results when the reverse transcription was done separately or together with the endogenous control for the investigated assays (Figure 3). However, it is important to note that the long term storage of different hairpin primers mixed together in the same solution is not recommended as it may lead to a false positive detection of mature miRNAs (data not shown). In summary, the level of an individual miRNA can be determined by using cDNA samples in which the given target and the endogenous control are reverse transcribed together, thereby reducing potential pipetting errors and making the measurements more cost effective.

DNA contamination significantly influences the measurement of mature miRNAs

Next, we investigated the effect of genomic and plasmid DNA on miRNA measurements. Based on our previous data from transient transfections, we had indications that contaminating DNA might interfere with mature miRNA detection. Thus, we examined whether stem-loop qRT-PCR is specific to the present mature miRNA or it has false positive signal from samples which do not contain the particular target. We tested miR-1226-3p and miR-33b-5p assays on genomic DNA (gDNA), total RNA and plasmid DNA (encoding the corresponding miRNA) samples. The investigated plasmids differ only in their pre-miRNA coding sequence. Total RNA and gDNA samples were derived from mir-1226 and mir-33b overexpressing or parental HeLa cell lines. We compared DNase treated and non-treated parallels for each sample. After reverse transcription, we measured the mature

orban.tamas_188_24

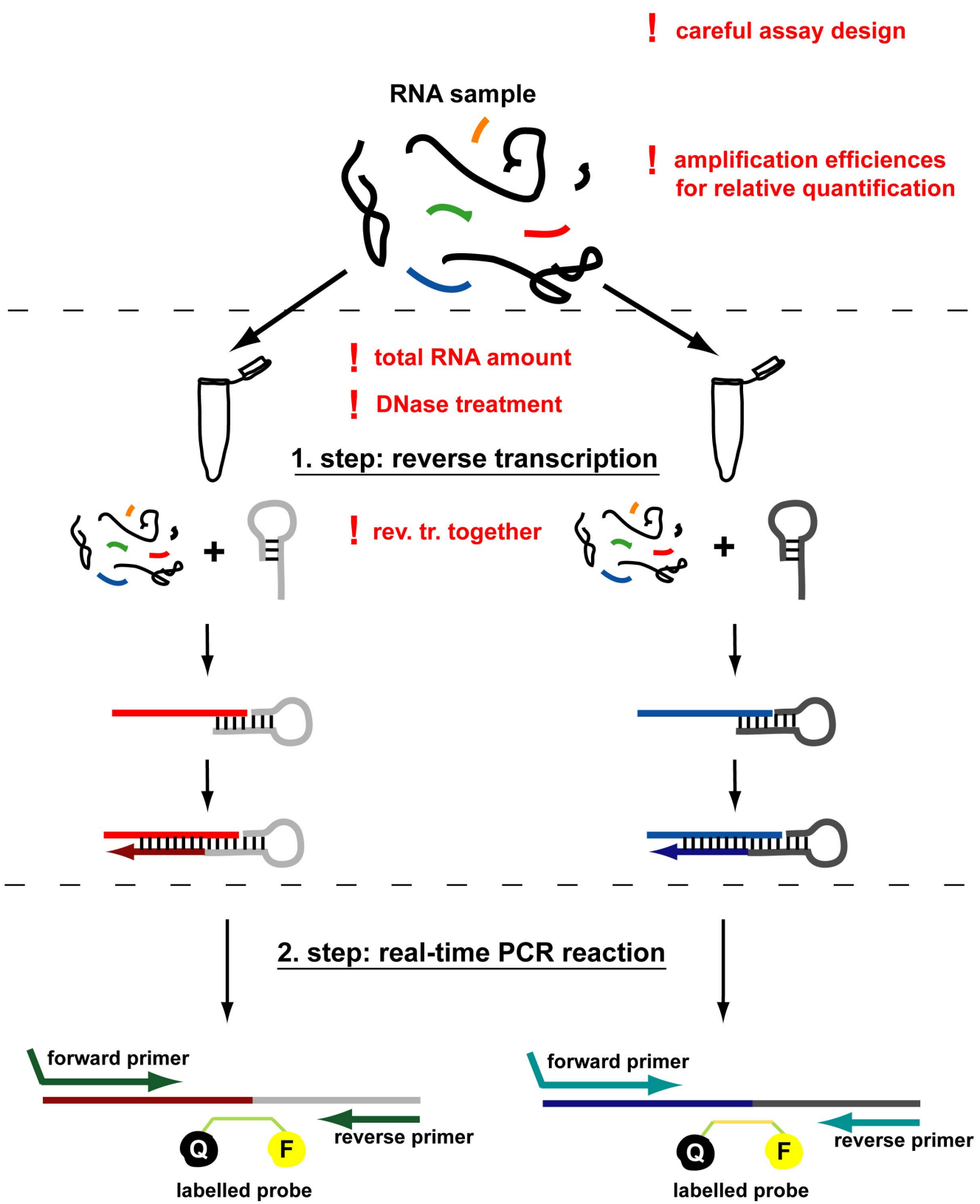
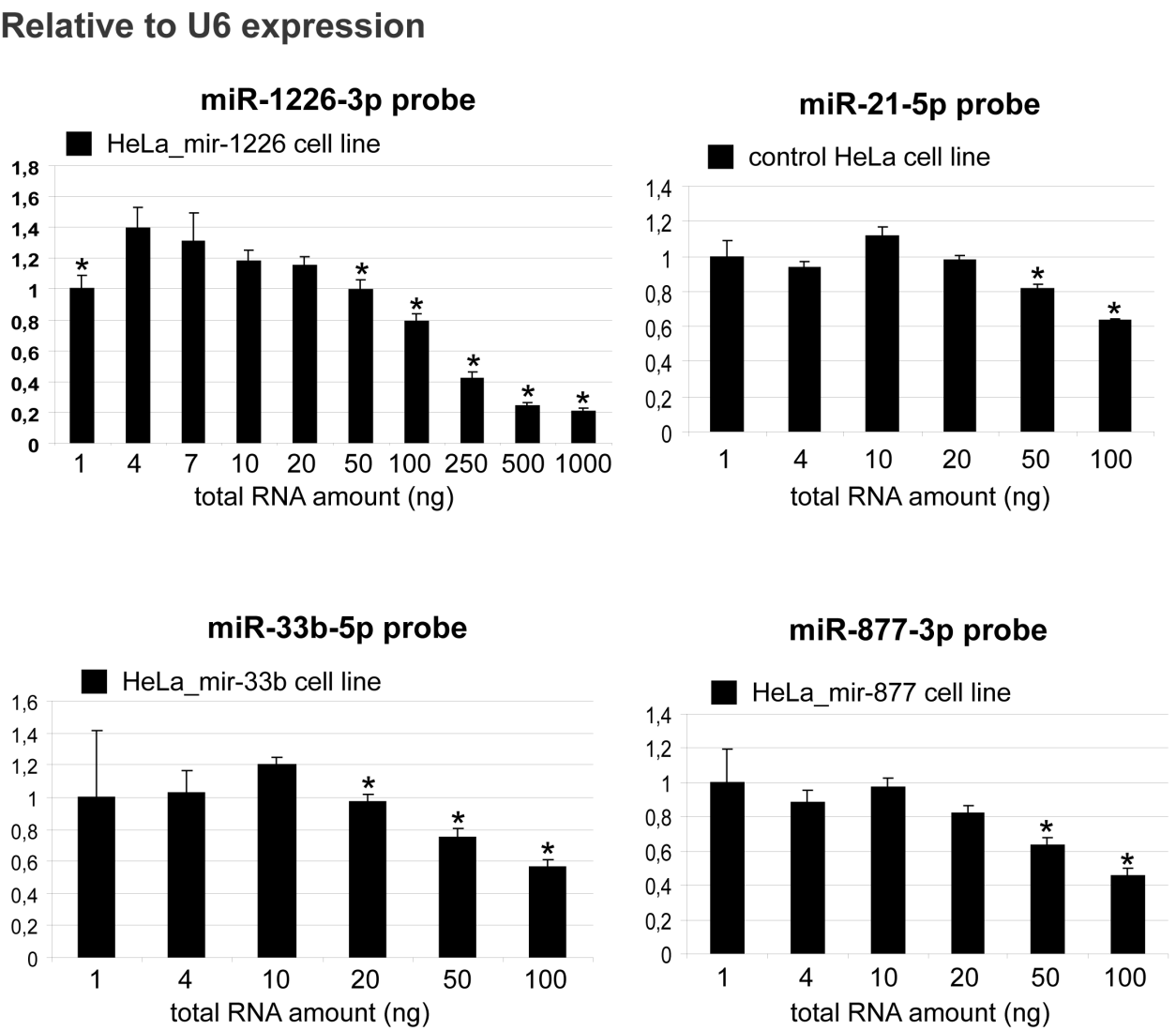


Figure 1. Schematic representation of stem-loop microRNA quantitative RT-PCR. The two main steps are reverse transcription and real-time PCR. In the first step, mature miRNA is extended and reverse transcribed by a sequence specific stem-loop primer. In the second step, the reverse transcribed miRNA is quantified by a fluorescently labeled hybridization probe using the strand replacement reaction. According to the previous protocol, all targets (e.g. endogenous control and target) should be reverse transcribed separately. In the dual-labeled probe based detection systems Q stands for quencher, F for fluorophore. Red exclamation marks indicate crucial points of the procedure that are discussed in this paper.
doi:10.1371/journal.pone.0106315.g001

A



B

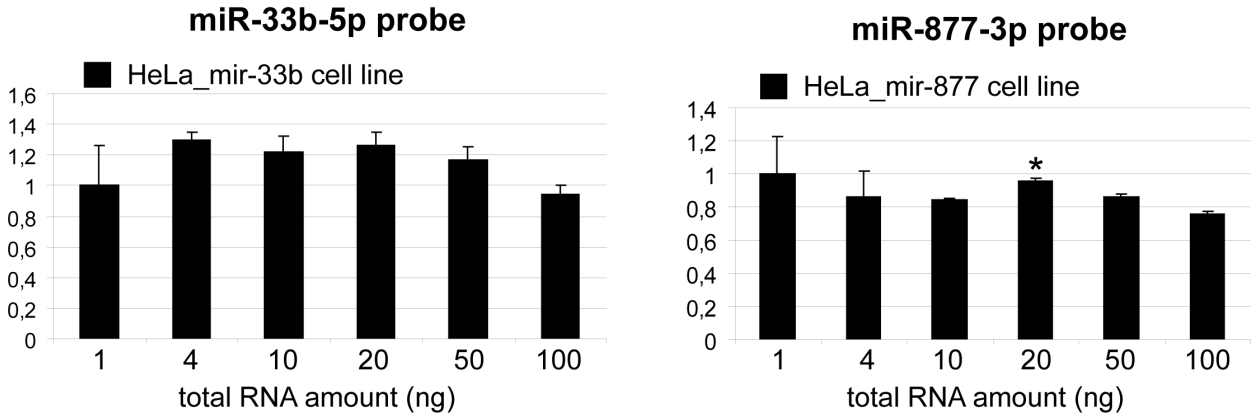


Figure 2. Mature miRNA detection from different amount of total RNA input. Different amount of total RNA samples were reverse transcribed for the detection of a particular mature miRNA by real-time PCR. MiRNAs with various abundance were measured, using relative quantification. The more abundant miR-1226-3p, miR-33b-5p and the less abundant miR-877-3p (mir-877*) were measured from cell lines stably overexpressing the corresponding miRNA, while the abundant endogenous miR-21-5p from parental HeLa cell line. Concerning the endogenous controls, the U6 snRNA (**A**) and the endogenous miR-21-5p (**B**) were applied. The optimal range of RNA input varies depending on endogenous controls and targets. The corresponding concentration series (controls and targets) were prepared and measured simultaneously. Mean values of three independent experiments (three biological parallels with three technical replicates) are shown. Error bars represent S.E.M.; samples are compared to a chosen optimal condition (10 ng of total RNA). *: $p<0.05$.
doi:10.1371/journal.pone.0106315.g002

miRNA levels in the above samples (Figure 4). The miR-1226-3p probe detected 4 fold higher amounts (2 C_t difference) of mature miRNA from the gDNA of mir-1226 overexpressing cell line compared to the two "non-relevant" gDNA samples (from HeLa and HeLa_mir-33b cell lines). In the case of the miR-33b-5p assay, the measured miRNA levels were similar among all gDNA samples. For RNA samples, as it was expected, both miR-1226-3p and miR-33b-5p probes resulted in significantly higher detected

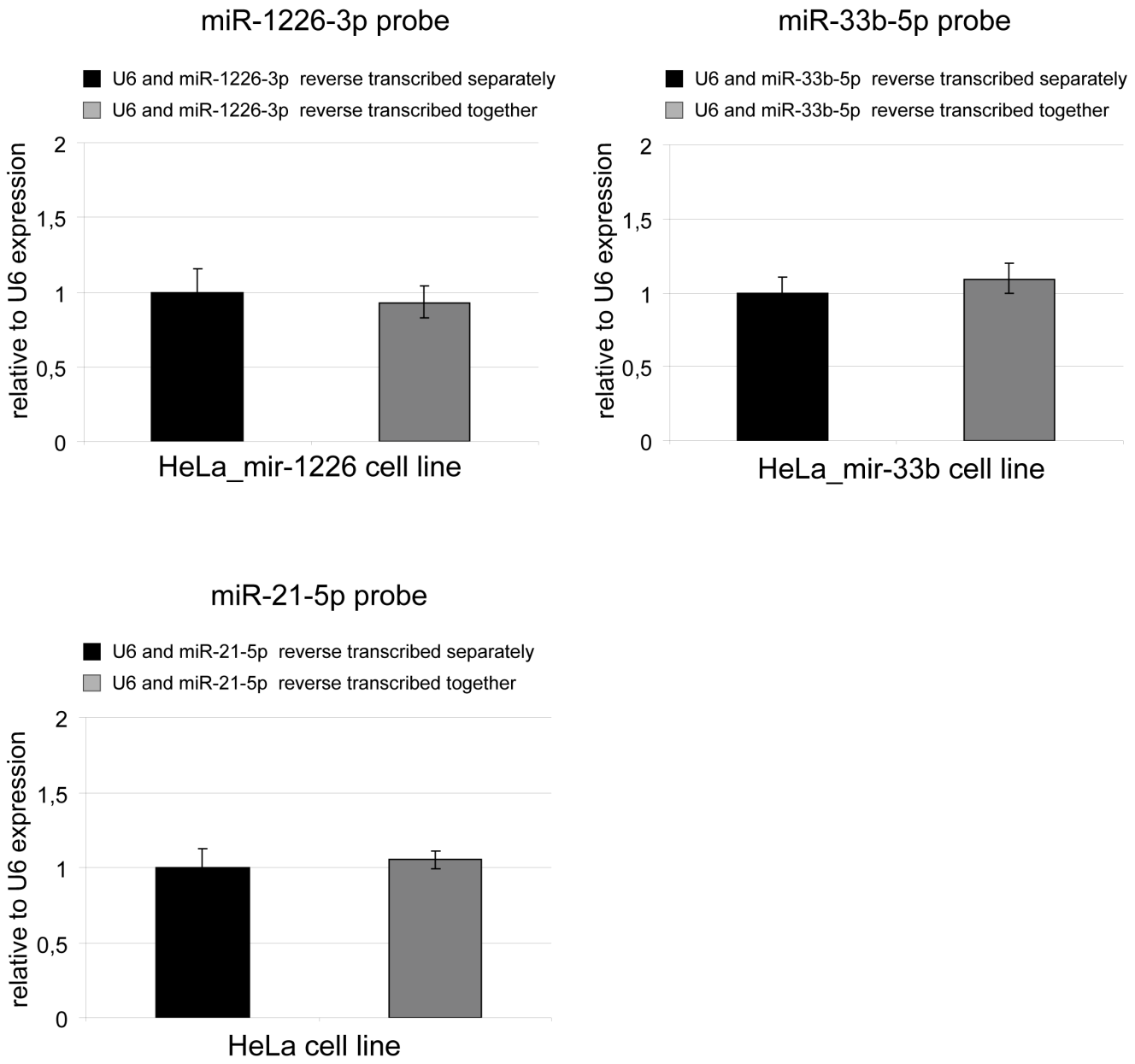


Figure 3. Reverse transcription of target and control can be done simultaneously. Mature miRNA levels of miR-1226-3p and miR-33b-5p were detected in stably overexpressing HeLa cell lines, whereas the endogenous miR-21-5p in parental HeLa cell line. cDNA samples were used from simultaneous or separate reverse transcription reactions of the endogenous control and the target. Experiments were carried out with three RT parallels and three technical replicates, error bars represent standard deviations.
doi:10.1371/journal.pone.0106315.g003

mature miRNA levels from the corresponding miRNA overexpressing cell line than in the controls. Concerning plasmid DNAs, apparently similar amount of miR-33b-5p was detected from mir-33b encoding plasmid as from mir-33b overexpressing cell line derived RNA. This striking false effect was even more pronounced in the case of miR-1226-3p when the mir-1226 expression plasmid served as a template. There was 9 C_t difference compared to the mir-1226 overexpressing cell line derived RNA sample, and about 14 C_t difference compared to the RNA backgrounds, representing an apparent 512× and 16384× higher miRNA amount, respectively. From plasmids encoding other "non-relevant" miRNA, very low signals were detected both for miR-1226-3p and miR-33b-5p. The above data indicate that the false positive signals from the relevant plasmid samples are miRNA sequence specific. Therefore, although mature miRNA molecules are not present, signals can be apparently detected from DNA containing the coding sequence of the corresponding pre-miRNA form. These results were also confirmed by experiments using miR-877-3p and miR-877-5p assays (data not shown).

Next, we intended to address the question that which part of the measurement (reverse transcription or real-time PCR) misleads the mature miRNA detection. To answer this question, first we made quantitative real-time PCR for miR-1226-3p from reverse transcribed and non-transcribed samples. We tested gDNA and RNA samples from mir-1226 overexpressing cell line and we also used mir-1226 encoding plasmid samples. Mir-33b overexpressing cell line and mir-33b encoding plasmid samples served as non-relevant controls. As shown in Figure 5A, there is a slight detection during the real-time PCR reaction from the relevant plasmid DNA without reverse transcription, but the majority of the false positive signal is detected only when the reverse transcription reaction is performed. We obtained similar results with the miR-33b-5p assay (Figure S2). In summary, these data reveal the unexpected fact that DNA may serve as a template during the reverse transcription reaction in a (stem-loop primer) sequence specific manner.

In further experiments, we tested whether the above phenomenon has a real relevance, for example when investigating miRNAs in transiently transfected cells. In such cases, RNA samples are prepared from cells containing the transfected plasmids. For these measurements we used RNA samples from HeLa cells transiently transfected with different amounts of a mir-1226 encoding plasmid (Figure 5B). The data showed that when samples were not treated with DNase, a significantly higher amount of miRNA was detected as compared to the DNase treated samples. This problem occurred not only by using the Trizol based total RNA isolation method, but also when applying a column-based isolation protocol such as the *mir*Vana Kit (Figure S3). These results indicate that there is plasmid DNA contamination in the total RNA samples which indeed misleads the accurate detection of mature miRNAs.

3′ isomiR forms of miRNAs are cross-detected

Emerging data strengthen the existence of isomiRs which are the results of the heterogeneous nature of miRNA processing, leading to variation in the length and/or sequence of mature miRNAs [23]. Since the exact 3′ end sequence seems to be crucial for stem-loop quantitative PCR, we investigated whether the different 3′ end variants of miRNAs can be exclusively detected by this technique. We applied different assays, designed for different 3′ isomiR species of a particular miRNA and tested the detection on various synthetic RNA oligonucleotides (~10^5 molecules/reaction).

First, among numerous 3′ isomiRs of miR-877-5p, the three most abundant species were analyzed by specific assays (http://www.mirbase.org/cgi-bin/get_read.pl?acc = MI0005561, at date of November, 2013). Since there were no commercially available pre-designed assays for them, we used custom made TaqMan assays (Life Technologies, CA, USA). They were tested on synthetic RNA oligonucleotides, identical to the miR-877-5p isomiR sequences. Each assay was tested for each isomiR species, bearing nucleotide differences in their 3′ ends (Figure 6A). Assays specific for the "GACA" and "GAC" 3′ends detected both "GACA" and "GAC" ended RNAs similarly, while "GA" ending was detected with ~3 Ct delay. On the other hand, "GA" specific assay detected all three isoforms similarly. As concerning non-reverse transcribed (no RT) controls, signals were detected in the case of all three probes, indicating that these real-time PCR assays are somehow able to detect their synthetic RNA targets without reverse transcription. However, there were at least 10 Ct differences between the values of no RT controls and the reverse transcribed target containing samples.

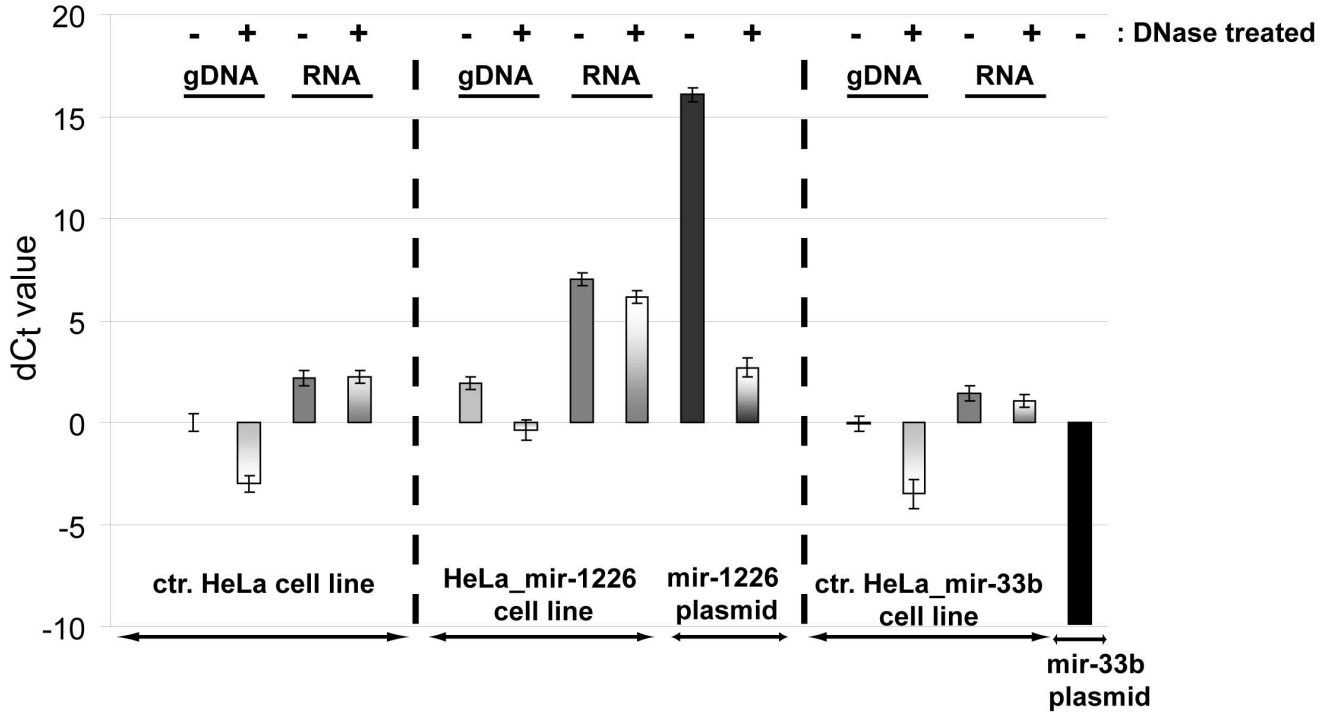
In other experiments, we tested miR-33b-5p isomiRs. There are two indicated 3′ isomiR forms of miR-33b-5p in the miRBase database, with 1 nt difference in their 3′ ends. The shorter mature miRNA is marked as the reference sequence but the longer form seems to be more abundant in investigated cell lines based on deep sequencing data (http://www.mirbase.org/cgi-bin/get_read.pl?acc = MI0003646, at date of November, 2013). We tested these variants by commercially available pre-designed TaqMan assays (Life Technologies, CA, USA). Both "GCA" and "GC" 3′ end specific assays detected the corresponding template better than the other isomiR, but the cross-detections were still considerable (2–4 C_t delayed; Figure 6B). To examine if a different miRNA detecting qRT-PCR approach might overcome this problem, we analyzed the above miR-33b-5p isomiRs by using the poly(A)-tailing-based method [19]. However, the isomiRs were also strongly cross-detected in those experiments (Figure S4), and even the melting curve analysis could not make reliable indication that more isomiRs are present when applying mixed isomiR population as a template for the different assays (data not shown).

The above data indicate that the examined quantitative real-time PCR methods for miRNA detection are not exclusively specific for a given isomiR, consequently 3′ isomiR species can be cross-detected in various extent. These results underline that careful selection of the assay is essential, since the accurate measurement of the given mature miRNA species strongly relies on the selected assay.

Discussion

In this study, we examined several factors in detail influencing accuracy and reliability of the miRNA quantitative stem-loop PCR. Considering the reverse transcription step of this methodology, our data indicate that the increase of the total RNA amount can result in a lower apparent miRNA expression level. This phenomenon could occur due to dissimilar reaction efficiencies of the target and the control in certain ranges of total RNA amount. Thus, it may lead to elevated detection of the endogenous control compared to the target (at higher concentration ranges, as suggested by the analysis of the raw data), therefore resulting in an apparent decrease in the level of the target. For a particular endogenous control/target pair the optimal amount of the total RNA for the reverse transcription reaction can vary, therefore pilot investigations are advisable prior to the real experiments. However, based on our experiments, 10–20 ng of total RNA might be adequate. In addition to these data, we provided evidence that the target of interest can be reverse transcribed

orban.tamas_188_24



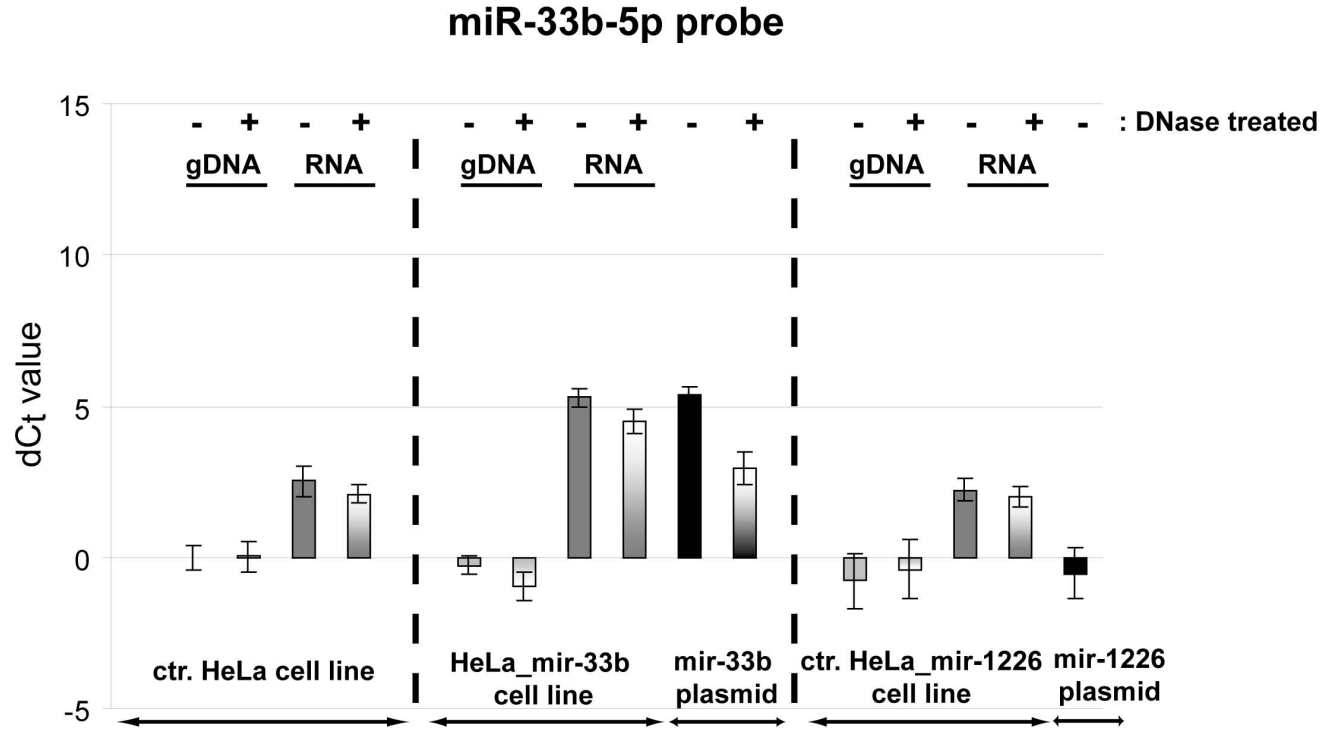


Figure 4. The effect of DNA contamination on mature miRNA measurements. Mature miR-1226-3p and miR-33b-5p detection were tested on the indicated samples, with or without DNase treatment. On the y-axis, dC_t value is represented, calculated as the Ct difference between the examined samples and the gDNA of control HeLa cell line (Ct = 33,8 for miR-1226-3p and Ct = 34,9 for miR-33b-5p experiments). Signal was not detected up to 40 reaction cycles for the mir-33b plasmid control by the miR-1226-3p assays. One C_t difference represents about 2× higher detected mature miRNA level. The effect of DNA is probe specific and plasmid DNAs have more pronounced effect on the measurements than gDNA contaminations. Experiments were carried out with three replicates at least twice, one representative experiment is shown. Error bars represent standard deviations.
doi:10.1371/journal.pone.0106315.g004

together in one reaction with the appropriate endogenous control. Apart from lowering the costs of experiments, it has the advantage of reducing pipetting errors and thereby making the measurements more accurate.

Next, we found the unexpected result that contrary to the claims of the original protocol [18], DNA could serve as a template during mature miRNA measurements, mostly during the reverse transcription reaction. Our data suggest that the corresponding pre-miRNA coding sequence is detected by the stem-loop primer. There are data that reverse transcriptases can use (single stranded) DNA as a template, therefore this DNA dependent DNA polymerase activity might be an explanation for our observation. However, the extent of the DNA-derived false detection varied among different miRNA targeting assays, and plasmid DNAs had more pronounced effects on the detection than gDNA contaminations. The significance of DNA contamination is further underlined by the fact that miRNA expression studies are often carried out on transiently transfected cells, which contain a significant amount of plasmid DNA originating from the used expression vector. Additionally to this, DNA and RNA molecules are both detected at 260 nm by spectrophotometry, therefore DNA contamination also disturbs the accurate measurement of RNA concentration. All these factors imply that extensive DNase treatment is a critical part of this miRNA quantification protocol which cannot be omitted when using certain RNA isolation methods. Our data show, that in contrast to total RNA isolation using either Trizol reagent or *mir*Vana Kit, no significant DNA contamination present in the RNA samples when applying small RNA isolation by the *mir*Vana Kit (Figure S3).

In addition to the technical issues described above, the recently discovered isomiR species impose another challenge on miRNA detection by the stem-loop qPCR technique, as the sequence diversity of miRNA species could be quite extensive both at the 5′ and the 3′ ends. Although there are emerging data on the existence of this variability, neither all mechanisms responsible for the generation of isomiRs nor their potential functional differences are clear as yet. Even if 3′ variability appears to be redundant in function at present [23], it represents a problematic issue not only for stem-loop qRT-PCR, but also for miRNA detection by the poly(A)-tailing based methodology ([21]; Figure 6 and Figure S4). Therefore, the *a priori* knowledge of the exact 3′ sequence is a prerequisite for designing an accurate, specific assay for any particular small RNA species, and examining miRNA databases and available online deep sequencing data is strongly recommended. Additionally, we would like to point out that the indicated reference sequences in databases often represent only a small proportion of isomiRs, therefore it could mislead researchers in assay design. Thus, as it was shown in the case of miR-33b-5p and miR-877-5p, cautious selection or design of the assay is essential, since only one nucleotide difference in the 3′ end terminus can cause inaccurate detection, leading to false representation of a mature miRNA form.

In addition to our findings, we would like to note that besides the factors investigated here, there are other issues reported to influence reliable miRNA detection. For example, when applying the widely used Trizol reagent based RNA isolation method also for miRNAs, it is important to keep in mind that the extraction efficiency of miRNAs with low GC content or stable secondary structure is sensitive for the initial number of the cells [34].

Summarizing our results, we provide a detailed and improved protocol for proper application of quantitative stem-loop RT-PCR for the accurate detection of mature miRNA species (see Figure 1 and Materials and Methods).

Materials and Methods

Refined, detailed protocol for stem-loop quantitative RT-PCR of individual miRNAs

Assay design. Careful assay selection for the proper isomiR species is crucial to evade misleading data. If the desired isomiR species is not known for a given miRNA, several previously annotated abundant isoforms should be tested in parallel.

RNA isolation. If total RNA isolation is done by Trizol reagent, the usage of minimum $1–2\times10^6$ cell/ml Trizol is strongly recommended (see ref [34]). The assessment of the quality of the isolated RNA sample (e.g.: by BioAnalyzer, Agilent Technologies) is also advisable.

DNase treatment (strongly recommended for total RNA samples).

e.g. 5 μg of total RNA,
2 μl (4 unit) of DNase (New England Biolabs),
2 μl of 10× DNase buffer,
1 μl (40 unit) of RNasin (Life Technologies), in total volume of 20 μl.

Incubate at 37°C for 1 hour, inactivate at 75°C for 10 minutes, then put on ice. Quantification of RNA by spectrophotometry (e.g.: NanoDrop 2000 Spectrophotometer, Thermo Scientific).

cDNA preparation (TaqMan MicroRNA Reverse Transcription Kit, Life Technologies). For one reaction:

0.15 μl of 100 mM dNTP Mix,
1 μl of Reverse Transcriptase,
1.5 μl of 10× buffer,
0.19 μl of RNase inhibitor,
1.16 μl of H_2O.

Mix gently, then add 5 μl of total RNA (2 ng/μl).

Mix gently and add 3 μl of endogenous control specific RT primer and 3 μl of target specific RT primer.

Reverse transcribe the RNA according to the manufacturer's instructions (16°C for 30′, 42°C for 30′, 85°C 5′).

Important note: reverse transcription efficiency may vary among samples in different type of PCR tubes.

cDNA dilution for quantitative PCR. Dilute the total 15 μl of cDNA volume 5 times by adding 60 μl of H_2O.

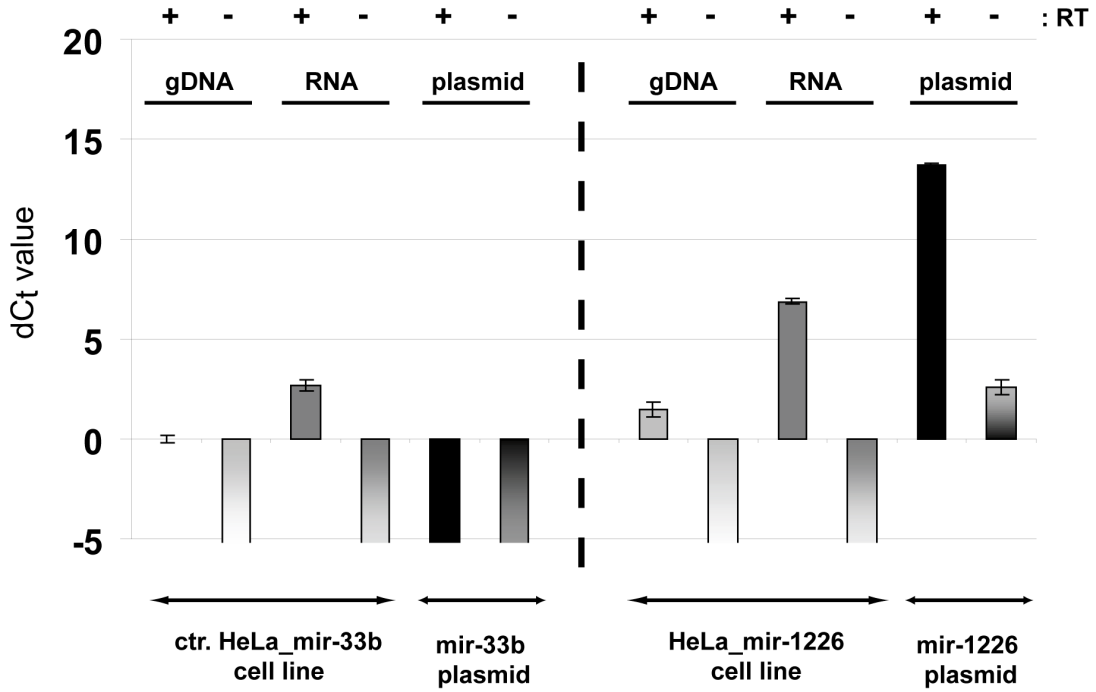
Quantitative real-time PCR (using TaqMan MicroRNA Assays, Life Technologies). Perform the samples in triplicate in singleplex reactions, in a final volume of 20 μl.

for one reaction:

10 μl of 2× Mix (TaqMan Universal Master Mix II with UNG, Life Technologies),
1 μl of 20× probe,
9 μl of diluted cDNA.

orban.tamas_188_24

A



B

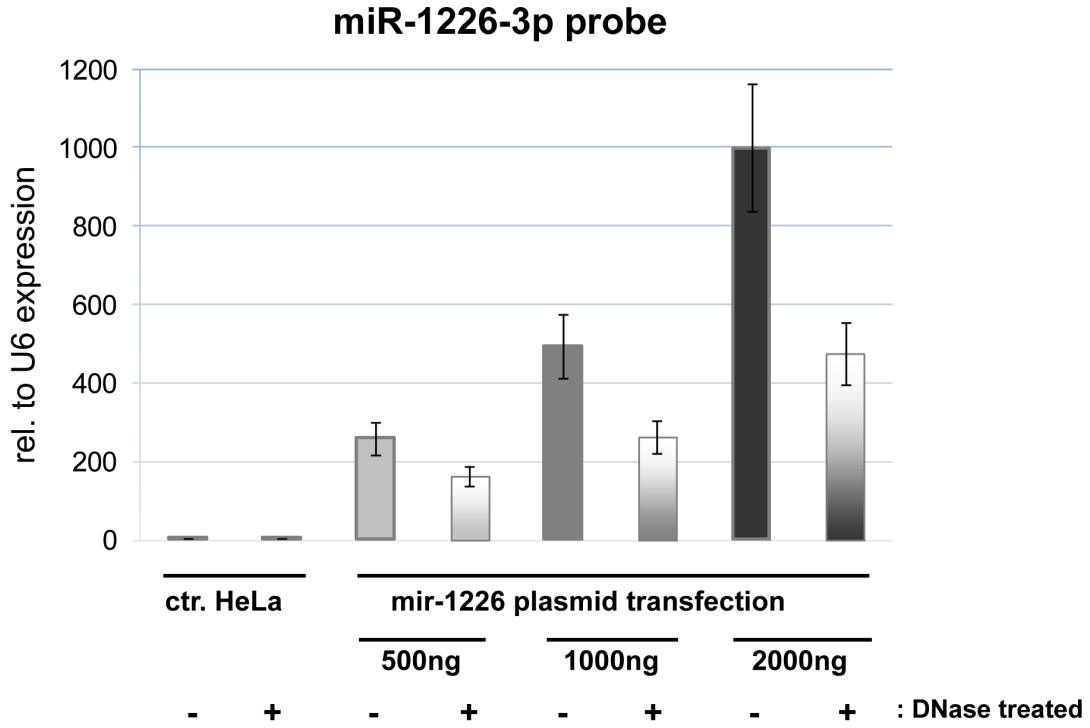


Figure 5. DNA can serve as a template during miRNA detection. (**A**) False positive signal of DNA derives mainly from the reverse transcription reaction. Mature miR-1226-3p was tested in the indicated samples, with or without reverse transcription (RT). On the y-axis, dC_t value is represented, calculated as the Ct difference between the examined samples and the gDNA of control HeLa_mir-33b cell line (Ct = 35,9). One C_t difference represents about 2× higher detected mature miRNA level. (**B**) DNA contamination remains in total RNA samples during isolation by the widely used Trizol reagent. Total RNA samples were isolated from transiently transfected HeLa cells; the transfected plasmid DNA amounts are indicated. Samples were DNase treated and non-treated, then reverse transcribed and subjected to real-time PCR. Expression values relative to U6 snRNA are shown on the y-axis. Experiments were carried out with three replicates at least from three independent experiments; one representative experiment is shown, error bars represent standard deviations.
doi:10.1371/journal.pone.0106315.g005

The final dilution of the cDNA in the reaction is 11×. Always apply non-template controls for the different assays. Perform the PCR reaction according to the manufacturer's instructions (50°C for 2′, 95°C for 10′, in 40 cycles: 95°C for 15″, 60°C 1′).

Data analysis. If relative quantification is to be applied, make sure by standard curve analysis that it is indeed applicable for comparison of the particular assays. Always check the baseline and threshold values since big differences in Ct values of the samples or little contamination in the non-template control might cause false auto fit by the program.

Plasmid constructs and isolation

EGFP embedded mir-1226, mir-33b and mir-877 expression plasmids were cloned as described earlier [17]. Plasmid DNAs were isolated by QIAGEN Plasmid Midi Kit using EndoFree Plasmid Buffer Set.

Cell cultures and manipulation

Parental HeLa cell line [35] was kindly provided by Zsuzsanna Izsvák (Mobile DNA Group, Max-Delbrück Center, Berlin, Germany). Cells were maintained in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% of fetal calf serum, 1% of L-glutamine, and 1% of penicillin/streptomycin (Life Technologies) using standard cell culture methodology. Mir-1226, mir-33b and mir-877 stably expressing cell lines were established by the *Sleeping Beauty* transposon based gene delivery technology as described earlier [17]. For transient transfections, 3×10^5 HeLa cells per wells were seeded onto a 6-well plate for transfection on the next day by FuGENE HD reagent (Life Technologies) using plasmid DNAs as indicated (DNA:lipid reagent = 1 μg:3 μl). Transfection efficiencies were followed by EGFP fluorescence, detected by a IX51 fluorescence microscope (Olympus). Cells were collected for total RNA isolation 48 h after transfection.

Genomic DNA isolation

After trypsinization, cells were centrifuged and washed with 1× phosphate-buffered saline. Then, after careful removal of the liquid supernatant, cell pellets were stored at −80°C until further processing. Genomic DNAs were isolated from the cells by standard phenol-chloroform extraction after cell lysis and proteinase K digestion. To remove RNA contamination from genomic DNA, samples were RNaseA treated at 37°C for 1 hour before proteinase K treatment.

miRNA analysis

Total RNA was isolated from cultured cells using either the Trizol reagent or the *mir*Vana miRNA Isolation Kit (Life Technologies); small RNA samples were isolated using the *mir*Vana miRNA Isolation Kit (Life Technologies). $\sim2\times10^6$ of cells were harvested and prepared according to the manufacturer's instructions. To remove DNA contaminations, RNA samples were treated with DNaseI (New England Biolabs) at 37°C for 1 hour. When applying the stem-loop qRT-PCR for cDNA preparations, if not indicated otherwise, 10 ng of total RNA (or gDNA, or plasmid) was reverse transcribed with miRNA specific stem-loop primers using TaqMan MicroRNA Reverse Transcription Kit (Life Technologies). For the poly(A)-tailing based qRT-PCR method, the miRCURY LNA Universal RT microRNA PCR Starter Kit from Exiqon was used following the manufacturer's protocol. "no RT" controls were prepared by inactivating the reverse transcriptase at 98°C for 20 minutes prior to adding it to the cDNA Mix (stem-loop qRT-PCR) or leaving out the enzyme mix from the reaction (Exiqon Kit). Related sample series were prepared simultaneously and measured in the same plate during the real-time PCR reaction. The general "cutoff" value was 40 cycles, as a standard used by the program of the used instruments. We always used three technical replicates for the real-time PCR measurements, and biological and RT parallels as indicated in the figure legends. When measuring miR-33b-5p and miR-877-3p with different total RNA input (shown in Figure2A and B), we used the same three independent RNA samples (isolated from the respective miRNA overexpressing HeLa cell line) for the experiments with the U6 and miR-21-5p endogenous controls. For isomiR detections, synthetic 5′-phosphate RNA oligos were purchased from Sigma. Assuming an average of 10^5 miRNA copies per 10 ng of total RNA [30], diluted RNA oligos were used either alone or supplemented with 10 ng of control HeLa total RNA samples for reverse transcription. Quantification was performed by quantitative real-time PCR using either TaqMan MicroRNA Assays and TaqMan Universal Master Mix II with UNG (Life Technologies) or LNA PCR primers sets and ExiLENT SYBR Green master mix (Exiqon). The real-time quantification reactions were performed on StepOne™ or StepOnePlus™ platforms (Life Technologies), according to the manufacturer's instructions; the data was analyzed by StepOne software (version 2.1; Life Technologies). Our data were represented as by the StepOne program (relative to an endogenous control, delta delta Ct values) or when it was not relevant, as delta Ct values (comparing to a control sample). For statistical analysis, two-sided Student's t-test was performed. The following TaqMan MicroRNA Assays were used in miRNA quantification, catalog numbers are in brackets: U6 small nuclear RNA [001973], hsa-miR-1226-3p [245467_mat], hsa-miR-33b-5p = hsa-miR-33b-5p "GC" [002085], hsa-miR-33b-5p "GCA" [001565], hsa-miR-21-5p [000397] and hsa-miR-877-3p = hsa-miR-877* [241029_mat]. For custom made miRNA assays, the following RNA sequences were used for assay design: 5′-GUAGAGGAGAUGGCG-CAGGGGACA for the hsa-miR-877-5p "GACA" isomiR (= hsa-miR-877-5p), 5′-GUAGAGGAGAUGGCGCAGGG-GAC for the hsa-miR-877-5p "GAC" isomiR and 5′-GUAGAG-GAGAUGGCGCAGGGGA for the hsa-miR-877-5p "GA" isomiR. For Exiqon LNA PCR primers sets, the following assays were used, catalog numbers are in brackets: has-miR-33b-5p "GC" [205860] and custom designed mir-33b-5p-GCA_1 "GCA" [206999].

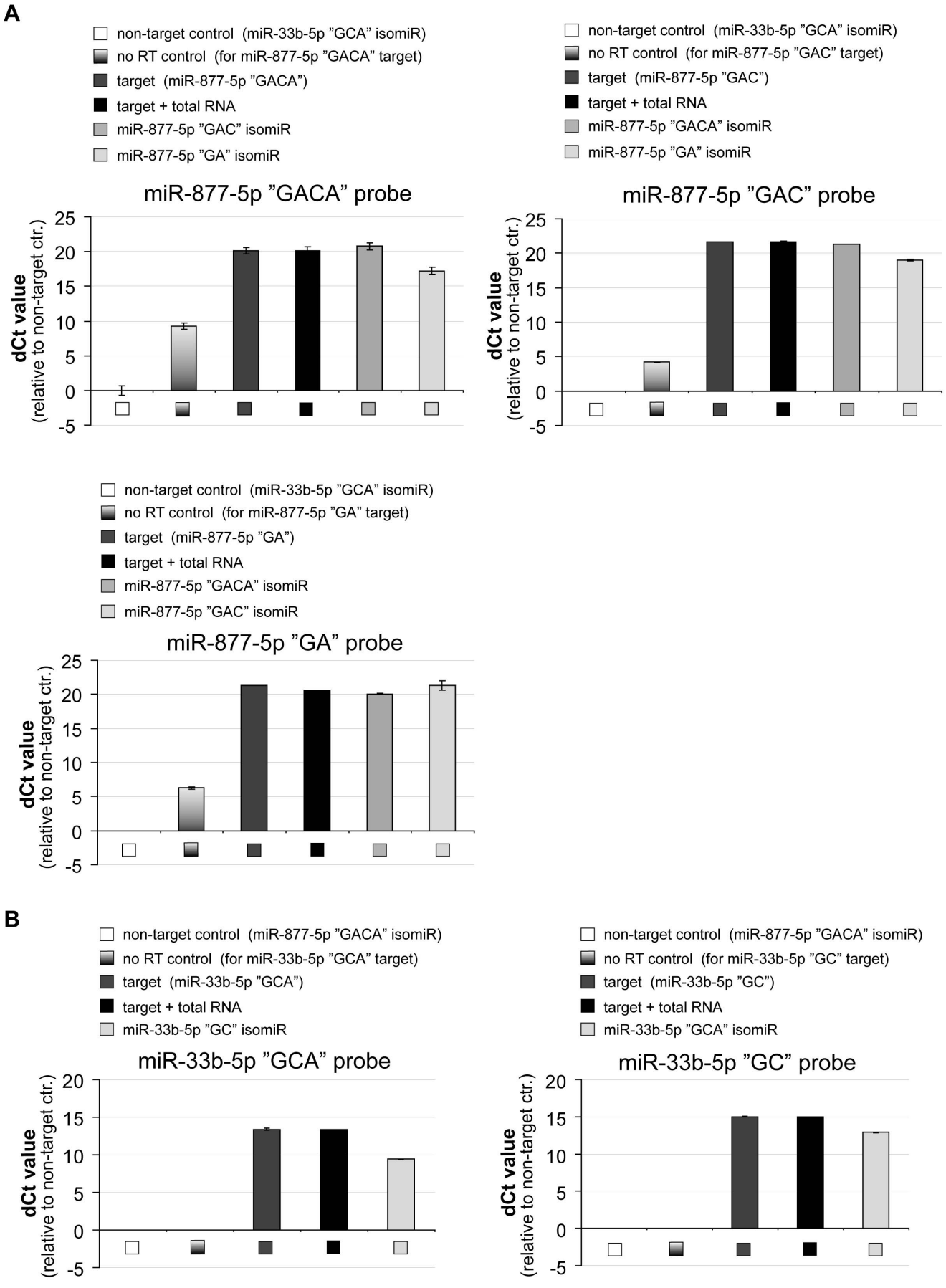

A
non-target control (miR-33b-5p "GCA" isomiR)
no RT control (for miR-877-5p "GACA" target)
target (miR-877-5p "GACA")
target + total RNA
miR-877-5p "GAC" isomiR
miR-877-5p "GA" isomiR
miR-877-5p "GACA" probe
dCt value (relative to non-target ctr.)
non-target control (miR-33b-5p "GCA" isomiR)
no RT control (for miR-877-5p "GAC" target)
target (miR-877-5p "GAC")
target + total RNA
miR-877-5p "GACA" isomiR
miR-877-5p "GA" isomiR
miR-877-5p "GAC" probe
dCt value (relative to non-target ctr.)
non-target control (miR-33b-5p "GCA" isomiR)
no RT control (for miR-877-5p "GA" target)
target (miR-877-5p "GA")
target + total RNA
miR-877-5p "GACA" isomiR
miR-877-5p "GAC" isomiR
miR-877-5p "GA" probe
dCt value (relative to non-target ctr.)
B
non-target control (miR-877-5p "GACA" isomiR)
no RT control (for miR-33b-5p "GCA" target)
target (miR-33b-5p "GCA")
target + total RNA
miR-33b-5p "GC" isomiR
miR-33b-5p "GCA" probe
dCt value (relative to non-target ctr.)
non-target control (miR-877-5p "GACA" isomiR)
no RT control (for miR-33b-5p "GC" target)
target (miR-33b-5p "GC")
target + total RNA
miR-33b-5p "GCA" isomiR
miR-33b-5p "GC" probe
dCt value (relative to non-target ctr.)

Figure 6. 3′ isomiR species may not be distinguished by the stem-loop qRT-PCR. Detection of the 3′ isomiRs of miR-877-5p (**A**) and miR-33b-5p (**B**) by the corresponding stem-loop assays using synthetic RNA oligonucleotides as templates. Non-target controls were chosen as reference samples. (Ct values were 36/37/39,6 for the "GACA"/"GAC"/"GA" miR-877-5p isomiRs respectively, while signals were not detected up to 40 reaction cycles for the miR-33b isomiR assays.) The 3′ isomiRs of a particular miRNA locus are cross-detected using stem-loop qRT-PCR, although the extent varies among the different probes. Experiments were carried out at least in four independent measurements. One representative experiment is shown, error bars represent standard deviations.
doi:10.1371/journal.pone.0106315.g006

Supporting Information

Figure S1 Determination of reaction efficiencies of different targets. (**A**) Standard curves with 1.5× dilution series and 5 points. Template concentrations are presented in a logarithmic scale; R^2 values represent the correlation coefficients of the fitted lines. (**B**) Amplification efficiencies calculated from different ranges of the curves. 1–5 for five points; 2–5 for four points, omitting the obvious outlier of the measurement from the most concentrated template. (It is below the recommended minimum of 1:10 dilution of the cDNA sample in the qPCR reaction).
(TIF)

Figure S2 DNA can serve as a template for the reverse transcription reaction. False positive signal of DNA derives from the reverse transcription reaction. Mature miR-33b-5p assay was measured in the indicated samples, with or without reverse transcription (RT). On the y-axis, dC_t value is represented (calculated as the Ct difference between the examined samples and the gDNA of control HeLa_mir-1226 cell line). Control gDNA data are above C_t of 35; one C_t difference represents about 2× higher detected mature miRNA level. Experiments were carried out in three replicates; one representative experiment is shown, error bars represent standard deviations.
(TIF)

Figure S3 Residing DNA contamination in RNA samples prepared by different RNA isolation procedures by *mir*Vana miRNA Isolation Kit. RNA samples were isolated from parental (control) and transiently transfected HeLa cells. Samples were DNase treated and non-treated, then reverse transcribed and subjected to real-time PCR. Expression values relative to U6 snRNA are shown on the y-axis. Experiments were carried out with three technical replicates from three independent experiments (biol. repl.), error bars represent standard deviations. There is remaining DNA contamination in the total RNA samples (**A**), but not in the small RNA enriched samples (**B**) when prepared by the *mir*Vana Kit. The expression level of miR-1226-3p from total RNA (with DNase treatment) and from small RNA samples (with or without DNase treatment) is similar.
(TIF)

Figure S4 3′ isomiRs of miR-33b-5p are cross-detected using the poly(A)-tailing based quantitative RT-PCR method. 3′ isomiRs of miR-33b-5p were detected by isomiR-specific primer sets using synthetic RNA oligonucleotides as templates. Non-target controls served as references samples (Ct values >33). In the no RT control reactions, particularly no signals (Ct >39) were detected. The two different 3′ isomiRs are significantly cross-detected by the specific primer sets and even the post-PCR SYBR Green-based melting curve analysis could not make reliable distinction between the different isomiR-specific PCR products when applying mixed isomiR population as a template (data not shown). Experiments were carried out at least twice, one representative experiment is shown, error bars represent standard deviations.
(TIF)

Acknowledgments

We would like to thank Károly Fátyol for helpful discussions.

Author Contributions

Conceived and designed the experiments: AS TIO. Performed the experiments: AS TIO. Analyzed the data: AS TIO. Contributed reagents/materials/analysis tools: AS TIO. Wrote the paper: AS TIO.

References

1. Carthew RW, Sontheimer EJ (2009) Origins and Mechanisms of miRNAs and siRNAs. Cell 136: 642–655.
2. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10: 94–108.
3. Arora S, Rana R, Chhabra A, Jaiswal A, Rani V (2013) miRNA-transcription factor interactions: a combinatorial regulation of gene expression. Mol Genet Genomics 288: 77–87.
4. Slezak-Prochazka I, Durmus S, Kroesen BJ, van den Berg A (2010) MicroRNAs, macrocontrol: regulation of miRNA processing. RNA 16: 1087–1095.
5. Miyoshi K, Miyoshi T, Siomi H (2010) Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. Mol Genet Genomics 284: 95–103.
6. Yang JS, Lai EC (2011) Alternative miRNA Biogenesis Pathways and the Interpretation of Core miRNA Pathway Mutants. Mol Cell 43: 892–903.
7. de Planell-Saguer M, Rodicio MC, Mourelatos Z (2010) Rapid in situ codetection of noncoding RNAs and proteins in cells and formalin-fixed paraffin-embedded tissue sections without protease treatment. Nat Protoc 5: 1061–1073.
8. Nuovo G, Lee EJ, Lawler S, Godlewski J, Schmittgen T (2009) In situ detection of mature microRNAs by labeled extension on ultramer templates. Biotechniques 46: 115–126.
9. Sempere LF, Freemantle S, Pitha-Rowe I, Moss E, Dmitrovsky E, et al. (2004) Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol 5: R13.
10. Valoczi A, Hornyik C, Varga N, Burgyan J, Kauppinen S, et al. (2004) Sensitive and specific detection of microRNAs by northern blot analysis using LNA-modified oligonucleotide probes. Nucleic Acids Res 32: e175.
11. Pall GS, Codony-Servat C, Byrne J, Ritchie L, Hamilton A (2007) Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. Nucleic Acids Res 35: e60.
12. Silahtaroglu AN, Nolting D, Dyrskjot L, Berezikov E, Moller M, et al. (2007) Detection of microRNAs in frozen tissue sections by fluorescence in situ hybridization using locked nucleic acid probes and tyramide signal amplification. Nat Protoc 2: 2520–2528.
13. Kloosterman WP, Wienholds E, de Bruijn E, Kauppinen S, Plasterk RH (2006) In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. Nat Methods 3: 27–29.
14. Varallyay E, Burgyan J, Havelda Z (2008) MicroRNA detection by northern blotting using locked nucleic acid probes. Nat Protoc 3: 190–196.
15. Obernosterer G, Martinez J, Alenius M (2007) Locked nucleic acid-based in situ detection of microRNAs in mouse tissue sections. Nat Protoc 2: 1508–1514.
16. Sibley CR, Seow Y, Saayman S, Dijkstra KK, El Andaloussi S, et al. (2012) The biogenesis and characterization of mammalian microRNAs of mirtron origin. Nucleic Acids Res 40: 438–448.
17. Schamberger A, Sarkadi B, Orbán TI (2012) Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. RNA Biol 9: 1177–1185.
18. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, et al. (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res 33: e179.
19. Shi R, Chiang VL (2005) Facile means for quantifying microRNA expression by real-time PCR. Biotechniques 39: 519–525.

20. Raymond CK, Roberts BS, Garrett-Engele P, Lim LP, Johnson JM (2005) Simple, quantitative primer-extension PCR assay for direct monitoring of microRNAs and short-interfering RNAs. RNA 11: 1737–1744.
21. Lee LW, Zhang S, Etheridge A, Ma L, Martin D, et al. (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA 16: 2170–2180.
22. Westholm JO, Ladewig E, Okamura K, Robine N, Lai EC (2012) Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. RNA 18: 177–192.
23. Neilsen CT, Goodall GJ, Bracken CP (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. Trends Genet 28: 544–549.
24. Li SC, Liao YL, Ho MR, Tsai KW, Lai CH, et al. (2012) miRNA arm selection and isomiR distribution in gastric cancer. BMC Genomics 13 Suppl 1: S13.
25. Wang WX, Wilfred BR, Baldwin DA, Isett RB, Ren N, et al. (2008) Focus on RNA isolation: obtaining RNA for microRNA (miRNA) expression profiling analyses of neural tissue. Biochim Biophys Acta 1779: 749–757.
26. Ach RA, Wang H, Curry B (2008) Measuring microRNAs: comparisons of microarray and quantitative PCR measurements, and of different total RNA prep methods. BMC Biotechnol 8: 69.
27. Becker C, Hammerle-Fickinger A, Riedmaier I, Pfaffl MW (2010) mRNA and microRNA quality control for RT-qPCR analysis. Methods 50: 237–243.
28. Benes V, Castoldi M (2010) Expression profiling of microRNA using real-time quantitative PCR, how to use it and what is available. Methods 50: 244–249.
29. Chugh P, Dittmer DP (2012) Potential pitfalls in microRNA profiling. Wiley Interdiscip Rev RNA 3: 601–616.
30. Redshaw N, Wilkes T, Whale A, Cowen S, Huggett J, et al. (2013) A comparison of miRNA isolation and RT-qPCR technologies and their effects on quantification accuracy and repeatability. Biotechniques 54: 155–164.
31. Tang F, Hajkova P, Barton SC, Lao K, Surani MA (2006) MicroRNA expression profiling of single whole embryonic stem cells. Nucleic Acids Res 34: e9.
32. Lao K, Xu NL, Sun YA, Livak KJ, Straus NA (2007) Real time PCR profiling of 330 human micro-RNAs. Biotechnol J 2: 33–35.
33. Mestdagh P, Feys T, Bernard N, Guenther S, Chen C, et al. (2008) High-throughput stem-loop RT-qPCR miRNA expression profiling using minute amounts of input RNA. Nucleic Acids Res 36: e143.
34. Kim YK, Yeo J, Kim B, Ha M, Kim VN (2012) Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. Mol Cell 46: 893–895.
35. Grabundzija I, Irgang M, Mates L, Belay E, Matrai J, et al. (2010) Comparative analysis of transposable element vector systems in human cells. Mol Ther 18: 1200–1209.

Chapter 22

Experimental Validation of Predicted Mammalian MicroRNAs of Mirtron Origin

Anita Schamberger and Tamás I. Orbán

Abstract

MicroRNAs (miRNAs) are ~22 nucleotide-long noncoding RNAs influencing many cellular processes by their regulatory functions on gene expression. MiRNAs of mirtron origin represent the most prominent group of the alternatively processed miRNAs. They reside in short introns, which are essentially equivalent to the precursor form of the given miRNA. Consequently, their maturation is independent of the Drosha/DGCR8 complex, while depends on the mechanism of mRNA splicing. The number of predicted human mirtron sequences increases as a consequence of the growing deep sequencing data and refined bioinformatics tools. However, experimental validations of particular sequences are also essential. In this chapter, we intend to provide detailed protocols for the investigation of predicted mirtron sequences. First, we use the *Sleeping Beauty* transposon-based gene-delivery system for the development of cell lines stably overexpressing mirtrons. The processing of functional mature miRNAs is then detected by a luciferase assay using a very strict "triple control" system. In addition, *bona fide* mirtron features are confirmed by demonstrating splicing dependency through splice site mutations, while Drosha/DGCR8 independency is assessed in DGCR8 deficient cell line. Finally, the presence of mirtron-derived mature miRNAs is detected by quantitative real-time PCR.

Key words miRNA, Mirtron, *Sleeping Beauty* transposon, Stable miRNA expression, Luciferase assay, qRT-PCR, Real-time PCR, Splicing, DGCR8, Drosha

1 Introduction

MicroRNAs (miRNAs) are short, single-stranded regulatory RNAs about 20–24 nucleotides in length. These small noncoding RNA molecules form a gene regulatory network comparable to that of transcription factors by targeting mRNA molecules via sequence complementarity [1–4]. MiRNAs often control several mRNAs simultaneously, and have an impact on most cellular processes, e.g., on the regulation of ontogenesis, cell proliferation, cell differentiation, and apoptosis, and could play a role in the pathomechanisms of cancer and metabolic diseases [2, 5, 6]. Most of these small miRNA molecules are generated via a canonical pathway,

M. Lucrecia Alvarez and Mahtab Nourbakhsh (eds.), *RNA Mapping: Methods and Protocols*, Methods in Molecular Biology, vol. 1182, DOI 10.1007/978-1-4939-1062-5_22,

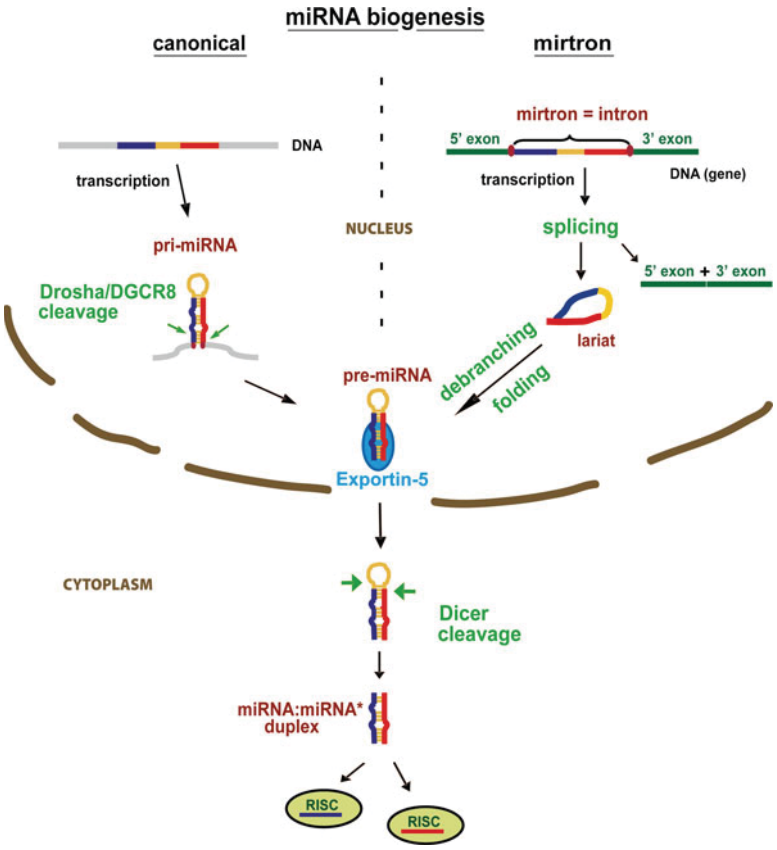


Fig. 1 Comparison of the canonical and mirtron biogenesis pathways. These two miRNA maturation pathways differ in their first cleavage step in which a pre-miRNA is liberated from the primary transcript. In the canonical pathway, the Drosha/DGCR8 microprocessor complex carries out the first cleavage reaction, whereas the splicing machinery is used in the mirtron-biogenesis pathway

but there are also emerging data on various alternative miRNA maturation routes.

In the canonical miRNA biogenesis pathway (Fig. 1), after the transcription of the primary miRNA transcript (pri-miRNA), an RNaseIII-type enzyme called Drosha cuts out the hairpin structured precursor miRNA form (pre-miRNA). This cleavage step is mediated by DGCR8, a partner protein of Drosha. The pre-miRNA is then transported from the nucleus to the cytoplasm, where Dicer, another RNaseIII-type enzyme cuts the precursor. This second cleavage step liberates a small RNA (miRNA:miRNA*) duplex, from which one strand is incorporated into an Argonaute (Ago) protein containing RNA-induced silencing complex (RISC), while the other strand is degraded. The miRNA-mediated gene silencing effect can be manifested by translation inhibition and/or mRNA decay [7, 8]. The alternative biogenesis pathways could bypass

certain steps of the canonical process, typically one of the two cleavage steps. Therefore, there are Drosha-independent and Dicer-independent alternative maturation pathways [9, 10].

The most prominent Drosha-independent miRNA biogenesis is the mirtron pathway (Fig. 1). It was first described in *Drosophila melanogaster* and *Caenorhabditis elegans* [11, 12], and experimentally demonstrated also in mammals [13–15]. MiRNAs of mirtron origin are localized in short introns where the whole intron is essentially equivalent to the pre-miRNA form. Thus, the first step of the mirtron processing is different from the canonical one: the pre-miRNA is cleaved out from the primary transcript by the splicing machinery instead of the Drosha/DGCR8 complex [16]. Apart from the originally described type of mirtrons, there are other closely related groups: the 3′- and 5′-tailed mirtrons. Their ends are also defined by the intronic splicing donor and acceptor sites, but they have extended ends in the pre-miRNA form either in 3′ or in 5′ direction. These "overhangs" are further processed by exonucleases [17–19].

There are emerging data on mirtrons indicating the importance of this alternative miRNA maturation pathway in various organisms, including plant species [20, 21]. However, beside small RNA data analyses and bioinformatics predictions, experimental validation of predicted mirtron sequences is also an important issue. In this chapter, we provide detailed protocols for the experimental validation of mammalian mirtrons. First, we describe an efficient way for the development of mirtron-expressing stable cell lines by using the *Sleeping Beauty* transposon-based gene-delivery system. Next, to test the presence of mirtron-derived functional miRNA species, we describe a protocol for luciferase assay using a very strict "triple control" analysis. To confirm mature miRNA processing by the mirtron pathway, details of testing splicing-dependent and Drosha/DGCR8-independent maturation are provided. Finally, we describe a refined protocol of the real-time quantitative stem-loop PCR to accurately detect and quantify mature miRNAs.

2 Materials

2.1 Generation of Mirtron-Encoding Vectors

2.1.1 Annealing of Oligonucleotides and Ligation

1. DNA oligonucleotides (sense and antisense strands of mirtron coding sequence) can be purchased from commercial source (Sigma; *see* **Note 1**).
2. NEBuffer2 (New England BioLabs): 50 mM NaCl, 10 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM dithiothreitol (pH = 7.9).
3. Heating block.
4. GFP-encoding plasmid suitable for cloning (*see* **Note 2**).
5. T4 DNA Ligase (New England BioLabs).

2.1.2 Transformation

1. Competent *Escherichia coli* cells.
2. Sterile Luria-Bertani (LB) medium: 10 g tryptone, 5 g yeast extract, 10 g NaCl in 1 l distilled H_2O.
3. LB-agar (1 l LB + 15 g agar) plated in Petri dishes.
4. Antibiotics matching with the transformed plasmid (ampicillin, 100 mg/ml).
5. Water bath.
6. 37 °C incubator/shaker.

2.1.3 Colony PCR, Gel Electrophoresis, and Plasmid Isolation

1. PCR Master Mix: 50 U/ml Taq DNA polymerase in reaction buffer of pH = 8.5, 400 μM dNTP Mix, 3 mM $MgCl_2$ (Promega).
2. Primers flanking the intronic mirtron sequences: forward (5′– TTCTTCAAGTCCGCCATGCC) and reverse (5–ACT TGTACAGCTCGTCCATGCCG) (*see* **Note 3**).
3. PCR machine.
4. Agarose.
5. 10× TBE buffer: 890 mM Tris–HCl, 890 mM boric acid, 20 mM EDTA (pH = 8).
6. Ethidium bromide.
7. DNA ladder and loading dye.
8. Gel electrophoresis equipment.
9. Plasmid isolation kit (QIAprep®Spin Miniprep Kit, Qiagen).

2.2 Development of Stable Mirtron-Expressing HeLa Cell Lines

2.2.1 Cell Line Maintenance and Transfection

1. HeLa cell line (ATCC).
2. Fetal bovine serum.
3. Dulbecco's Modified Eagle Medium (DMEM).
4. Phosphate-buffered saline (PBS).
5. Trypsin.
6. *Sleeping Beauty* transposon and transposase vectors.
7. FuGENE® HD reagent (Promega).

2.2.2 Cell Sorting

1. Tubes suitable for flow cytometry.
2. Flow cytometry cell sorter (FACS Aria High Speed Cell Sorter, Beckton-Dickinson).

2.3 Detection of Mirtron Function by Luciferase Assay

1. FuGENE® HD reagent (Promega).
2. psiCHECK2 vector (Promega; *see* **Note 4**).
3. Dual Luciferase® Reporter Assay System Kit (Promega), containing special reagents such as the "Passive Lysis Buffer" and "LARII" and "Stop & Glo" reagents.

4. Rocking platform or orbital shaker.
5. Opaque plate (OptiPlate-96, PerkinElmer).
6. Glass tube.
7. Luminometer (2030 Multilabel Reader luminometer, PerkinElmer).

2.4 Investigation of Mirtron Features

1. Splice site mutant mirtron oligonucleotides (*see* **Note 5**).
2. Materials and equipment for the generation of vectors (*see* Subheading 2.1).
3. Transfection reagent (*see* Subheading 2.2.1).
4. Luciferase assay (*see* Subheading 2.3).

2.4.1 Splicing Dependency

2.4.2 DGCR8 Independency

1. Wild type and DGCR8-deficient mouse embryonic fibroblast (MEF) cell lines.
2. Fetal bovine serum.
3. Dulbecco's Modified Eagle Medium (DMEM).
4. Phosphate-buffered saline (PBS).
5. Trypsin.
6. Lipofectamine® 2000 (Invitrogen).
7. Opti-MEM (Gibco®, Invitrogen).
8. Luciferase assay (*see* Subheading 2.3).

2.5 Detection of Mature miRNAs by qRT-PCR

Use RNase-free reagents and equipment.

2.5.1 Total RNA Isolation

1. Trizol reagent (Invitrogen; *see* **Note 6**).
2. Chloroform.
3. Isopropyl alcohol.
4. 75 % ethanol.
5. Formamide.
6. Ethidium bromide.
7. Spectrophotometer (NanoDrop 2000 Spectrophotometer, Thermo Scientific).
8. Gel electrophoresis equipment.

2.5.2 miRNA-Specific Reverse Transcription

1. DNase (2 U/μl) and 10× DNase buffer (New England BioLabs; *see* **Note 7**).
2. RNase inhibitor (20 U/μl, Life Technologies).
3. Spectrophotometer (NanoDrop 2000 Spectrophotometer, Thermo Scientific).

4. miRNA Reverse Transcription Kit (TaqMan® MicroRNA, Life Technologies): 100 mM dNTP Mix, Reverse Transcriptase (50 U/μl), Reverse Transcriptase buffer, RNase inhibitor (20 U/μl).
5. miRNA-specific primer (provided in TaqMan® MicroRNA Assays, Life Technologies or can be synthesized).
6. PCR machine.

2.5.3 Quantitative Real-Time PCR

1. Master Mix (TaqMan® Universal Master Mix II with UNG, Life Technologies).
2. miRNA-specific probe (20× probe provided in TaqMan® MicroRNA Assays, Life Technologies or can be synthesized).
3. cDNA.
4. Real-time PCR machine (StepOne™ or StepOnePlus™ platforms, Life Technologies).
5. Equipments for real-time PCR (MicroAmp Fast Optical 48 or 96-well Reaction plate, MicroAmp Optical Adhesive Film or MicroAmp Fast Reaction Tubes and Cap Strip, Life Technologies).
6. Software for data analysis (StepOne version 2.1, Life Technologies).

3 Methods

3.1 Generation of Mirtron-Encoding Vectors

Because of its ability to make stable cell lines, we use the *Sleeping Beauty* transposon-based vector system for the expression of mirtrons (Fig. 2a; [22, 23]). We clone mirtron sequences in the previously modified EGFP coding sequence (a PvuII site was introduced) as artificial introns [15, 24].

3.1.1 Mirtron Sequence

Mirtron sequence can be designed based on miRNA databases, e.g., miRBase (http://www.mirbase.org). Subsequently, mirtron-encoding (intronic) sequences can be obtained by annealing DNA oligonucleotides (sense and its respective antisense counterpart).

3.1.2 Annealing of Sense and Antisense Oligonucleotides and Ligation

1. Mix 5 μl of mirtron-encoding sense oligonucleotide [100 μM] and 5 μl of antisense oligonucleotide [100 μM]. Add 5 μl of NEB2 buffer and 85 μl of dH_2O.
2. Mix and incubate for 3 min at 96 °C, then turn off heating block, and let the tube cool down slowly in block to room temperature (typically overnight). The annealed double-stranded oligos can be used next day for ligation or stored at −20 °C.

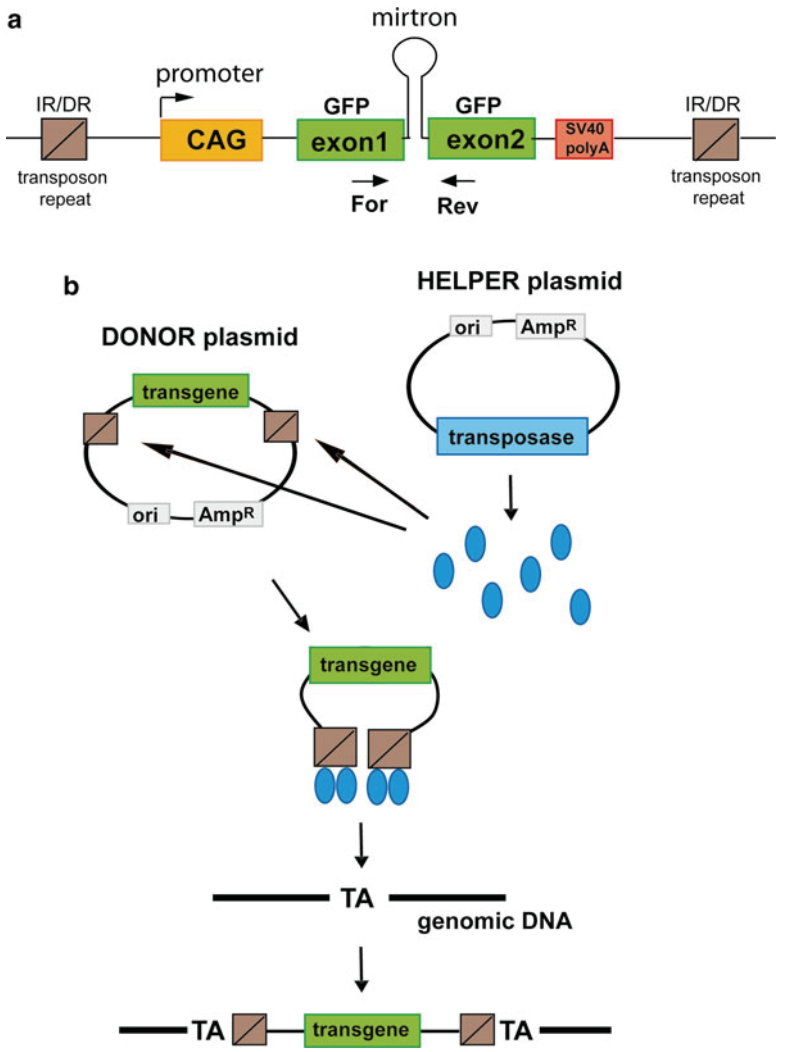


Fig. 2 Overview of the *Sleeping Beauty* transposon-based gene-delivery system. (**a**) Schematic illustration of a transposon vector, suitable for stable mirtron expression. *IR/DR* stands for inverted repeat-direct repeat sequences, *For* and *Rev* indicate binding sites for the PCR primers used for subsequent analysis (colony PCR, detection of splicing). (**b**) The "cut-and-paste" transposition mechanism of the *Sleeping Beauty* system. Following cotransfection, the transposase protein (*blue ellipses*) is expressed from the helper plasmid. Subsequently, the transposase recognizes the IR/DR repeats on the donor plasmid, cleaves the DNA, and liberates the transgene carrying transposon. Then, the transposase/transgene complex scans the genomic DNA for a target sequence, and integrates the transposon at a "TA" dinucleotide. During the integration process, the "TA" motif is duplicated, leaving a typical "footprint" of the *Sleeping Beauty* system. *ori* bacterial replication origin sequence, *Amp*^{R} ampicillin resistance gene

3. For ligation, mix 1 μl of 10× ligase buffer, 0.5 μl (~10 ng) of plasmid backbone, 0.5 μl of hybridized oligos (previously diluted in 1:6), 7 μl of H_2O, and 1 μl of ligase. Incubate at room temperature for ~2–5 h.

3.1.3 Transformation

1. Prechill on ice the tube containing the ligation reaction, and add 100 μl of competent cells.
2. Incubate on ice for 30 min.
3. Heat shock competent cells for 40 s at 42 °C.
4. Return cells to ice for 1–2 min.
5. Add 890 μl of LB medium and incubate at 37 °C for 20–60 min in a shaking incubator (200–300 rpm; *see* **Note 8**).
6. Centrifuge the tube at ~8,000 × *g* for 1 min and take off 900 μl of supernatant.
7. Resuspend the pellet and spread onto antibiotic-containing LB-agar plates (e.g.: 100 mg/ml ampicillin, corresponding to the backbone resistance).
8. Incubate at 37 °C for 16–20 h for colony formation.

3.1.4 Colony PCR, Gel Electrophoresis, and Plasmid Isolation

Colony PCR is a suitable method for screening successful ligations. Colonies are picked one by one, their plasmids are liberated by cell lysis, and then they can be used as template for PCR (*see* also **Note 3**).

1. Aliquot 10–10 μl of H_2O into PCR tubes.
2. Prepare the PCR mix: 12.5 μl of PCR 2× Master Mix (Promega), 1 μl of forward primer [10 μM], 1 μl of reverse primer [10 μM], 9.5 μl of H_2O. Aliquot 24–24 μl of PCR mix into PCR tubes.
3. Pick a single colony and suspend it in 10 μl of H_2O. Pick as many colonies as you wish (approx. 5–10 per ligation).
4. Lyse bacteria at 95 °C for 3 min, then cool down to 4 °C or put on ice.
5. Add 1 μl of lysate (from supernatant) to an aliquoted PCR mix (24 μl).
6. Perform PCR reaction: 95 °C for 3 min; 35 cycles of: 95 °C for 30 s (denaturation), 59 °C for 30 s (annealing), 72 °C for 45 s (extension); 72 °C for 5 min (final extension).
7. Run 3 μl of PCR samples on 3 % agarose gel (*see* **Note 9**).
8. PCR amplicons in the appropriate size can be purified and sequenced.
9. Selected colonies can be grown overnight in 3–5 ml of LB (containing the appropriate antibiotics) in a shaking incubator.
10. Plasmid DNA can be prepared by any commercially available method (e.g.: QIAprep®Spin Miniprep Kit).

3.2 Development of Stable Mirtron-Expressing HeLa Cell Lines

The *Sleeping Beauty* transposon-based gene-delivery system is a simple and effective method for the generation of stable transgene-expressing cell lines. It requires two plasmids: a donor plasmid (encoding the gene of interest flanked by the transposon repeats) and a helper plasmid (encoding the transposase; Fig. 2b). The transposase inserts the transposon from the plasmid to a genomic location by a "cut-and-paste" mechanism [25, 26].

3.2.1 Transfection

1. Plate 3×10^6 HeLa cells/well in a 6-well plate in 2 ml of medium.
2. Next day mix 1 μg of transposon and 100 ng of transposase-encoding plasmid with 100 μl of serum- and antibiotics-free DMEM.
3. Add 3.3 μl of FuGENE® HD reagent (avoid the contact between the plastic wall of the tube with undiluted transfection reagent), mix gently by tapping the tube, and incubate for 45–60 min at room temperature (*see* **Note 10**).
4. Add the transfection complex to the cells in a drop-wise manner, swirl the plate gently to mix.
5. After 48 h, check transfection efficiency by fluorescence microscope (EGFP fluorescence). Ideally, about 70–100 % of cells should be positive for EGFP.

3.2.2 Cell Sorting

1. 2 days after transfection, transfer the cells into a 25 cm^2 cell culture flask.
2. When cells are confluent, passage them to a 75 cm^2 flask.
3. When cells are confluent (normally at day 8 after transfection; *see* **Note 11**), collect 3/4 of the cell population in 0.5–1 ml of medium for sorting, and save the rest as a backup cell culture.
4. Sort the cells based on marker gene expression (GFP fluorescence), gating the cells rigorously (*see* **Note 12**).
5. Put the sorted cells back into a 6-well plate or into a 25 cm^2 cell culture flask (*see* **Note 13**).
6. To achieve 100 % positive cell population, sort the cells again for marker gene (GFP) expression at day 17 after transfection.

3.3 Detection of Mirtron Function by Luciferase Assay

The assay can be measured either in stably or transiently transfected mirtron-expressing HeLa cell line. We test every mirtron for both 5′- and 3′- arms with sensor and mutant sensor constructs. Sensor constructs contain two fully complementary antisense copies of either the 5′- or the 3′-arm of a particular mirtron, cloned into the 3′-untranslated region (3′-UTR) of Renilla luciferase in the psiCHECK2 vector. Mutant sensors differ in three mismatched nucleotides in the predicted miRNA seed region (*see* also **Note 4**). For luciferase assay, prepare three parallels for each setup (*see* **Note 14**). To exclude nonspecific effects, we compare mirtron-derived data

with a negative control (non-cognate miRNA) using the particular mirtron-specific sensors (*see* also Subheading 3.3.3). We describe here a single-sample measurement in a plate-reading luminometer; however, luciferase assays can be measured also in a multi-sample manner (*see* also **Note 15**).

3.3.1 Transfection

1. Plate 0.5×10^6 stable mirtron-expressing HeLa cells or 0.5×10^6 "parental" HeLa cells per well in a 24-well plate (0.5 ml of medium per well).
2. Next day, transfect cells using the FuGENE® HD reagent, (*see* Subheading 3.2.1). For stable cell lines, use 15 ng of luciferase sensor or mutant sensor construct (*see* **Notes 4** and **14**) and 0.045 μl of transfection reagent per well (typically, prepare a mastermix for several reactions because of the small volumes). For transient expression, use 300 ng of mirtron-encoding plasmid and 15 ng of luciferase sensor (or mutant sensor) construct with 0.945 μl of transfection reagent per well.
3. For transient transfection, check transfection efficiency after 24 h by fluorescence microscope (EGFP fluorescence).

3.3.2 Luciferase Assay

1. 48 h after transfection, remove the culture medium, and rinse the cells carefully with PBS. Remove the PBS completely.
2. Lyse the cells by passive lysis: add 100 μl of 1× Passive Lysis Buffer (1:10 dilution of the 10× Passive Lysis Buffer provided in the Dual Luciferase® Reporter Assay System).
3. Place the culture plate on an orbital shaker or rocking platform for 15–30 min to ensure complete lysis.
4. Prepare LARII and Stop & Glo reagents (provided in the Dual Luciferase® Reporter Assay System) according to the manufacturer's instructions (*see* **Note 16**).
5. Pipette 100 μl of LARII (provided in the Dual Luciferase® Reporter Assay System) into a well of a 96-well OptiPlate (*see* also **Note 15**).
6. Add 20 μl of lysate, mix gently by pipetting.
7. Measure luminescence immediately for 10 s, and record the data (firefly activity).
8. Add 100 μl of Stop & Glo reagent (Dual Luciferase® Reporter Assay System), mix thoroughly to quench luminescence from firefly luciferase reaction.
9. Immediately measure luminescence (Renilla luciferase activity) for 10 s, record the data.

3.3.3 Evaluation of the Data

The psiCHECK2 vector encodes two luciferases. The sensor element for the mirtron is cloned in the 3′ UTR of the Renilla luciferase (the level of this protein can be altered by silencing), while the

firefly luciferase serves as transfection control. Consequently, the raw data of Renilla activity is always normalized with the data of the firefly activity.

To ensure that the silencing effect is specific for a given mirtron, we use two additional controls: a non-cognate control and a sensor/seed-mutant sensor pair (*see* also **Note 14**). After normalization of the data with the transfection control (firefly luciferase), we calculate the sensor/mutant sensor ratio for the mirtron, and compare it to the sensor/mutant sensor ratio of a non-cognate miRNA control. Using this "triple control" system, we maximize the exclusion of nonspecific effects, which can result either from the different plasmid constructs and/or from endogenous small RNAs. In addition, we use three replicates for each experimental setup.

3.4 Investigation of Mirtron Features: Splicing Dependency and DGCR8 Independency

Splicing dependency of mature mirtron generation can be investigated by abolishing the splicing of mirtrons. The generation of splice site-mutated mirtron-encoding vectors is similar to that of the wild type construct (*see* Subheading 3.1), except that a 5′ dinucleotide mutated oligonucleotide (corresponding to the mutant splicing donor site) is ligated into the marker gene as an artificial intron (*see* **Note 17**).

3.4.1 Generation of 5′ Splice Site-Mutated Mirtron-Encoding Vectors

3.4.2 Functional Testing of Mutant Mirtrons by Luciferase Assay

A luciferase assay is used to examine if splicing influences the biogenesis of functional miRNAs of mirtron origin. In this assay, we compare the capability of a wild type and a 5′ splice (donor site) mutated mirtron to produce a functional mature miRNA. For mutated mirtrons, we use the same setup described earlier, including a non-cognate control and sensor/mutant-sensor pairs (*see* Subheading 3.3).

3.4.3 Functional Testing of the Generation of a Mature miRNA of Mirtron Origin in a DGCR8-Deficient Cell Line

To prove that a mature miRNA is processed by the mirtron biogenesis pathway, a Drosha/DGCR8-independent processing has to be demonstrated. We investigate Drosha/DGCR8-independent mature miRNA generation using a luciferase assay in a DGCR8-deficient mouse embryonic fibroblast (MEF) cell line (*see* **Note 18**). For this measurement, we recommend to use a canonical miRNA as a control [15]. In wild type MEFs, functional mature miRNAs are processed normally from both the mirtron and the canonical pathway. However, in DGCR8-deficient MEFs, the detection of a functional mature miRNA is expected only in the case of *bona fide* mirtrons, since functional mature miRNA generation by the canonical pathway is abolished.

1. Plate 0.3×10^5 DGCR8-deficient MEF cells or 0.6×10^5 wild type MEF cells in 0.4 ml of antibiotic-free DMEM per well in a 24-well plate.

2. Next day prepare the transfection mix by adding 0.33 µl of Lipofectamine® 2000 to 50 µl of serum- and antibiotics-free Opti-MEM. Mix gently. Incubate for 15 min.
3. Mix 50 µl of serum- and antibiotics-free Opti-MEM with 300 ng of mirtron-encoding plasmid and 15 ng of sensor (or mutant-sensor) construct.
4. After incubation, combine transfection mix generated in **step 2** with the DNA mix prepared in **step 3**. Mix gently and incubate for 20 min.
5. Add the transfection complex to the cells in a drop-wise manner, swirl the plate gently to mix.
6. After 48 h, check transfection efficiency by fluorescence microscope (EGFP fluorescence).

3.5 Detection of Mature miRNAs by qRT-PCR

Stem-loop quantitative PCR is a good method of choice for directly detecting low levels of mature miRNAs (*see* **Note 19**; Fig. 3).

3.5.1 Total RNA Isolation

1. Add 1 ml of Trizol per well to confluent cells (100 % confluency) in a 6-well plate, mix thoroughly by pipetting. A minimum of $1–2 \times 10^6$ cells/ml Trizol is strongly recommended (*see* **Note 6**). Samples can be processed immediately or stored at –80 °C.
2. Incubate for 2–3 min at room temperature before adding 0.2 ml of chloroform.
3. Shake for 15 s by hand and incubate for additional 2–3 min at room temperature.
4. Centrifuge at 12,000 × *g* for 15 min at 4 °C.
5. Transfer about 3 × 150 µl of the aqueous phase to a fresh tube (*see* **Note 20**).
6. Add 0.5 ml of isopropanol, mix by inverting the tubes, and incubate at room temperature for 10 min.
7. Centrifuge at 12,000 × *g* for 10 min at 4 °C.
8. Remove the supernatant, and wash the pellet by adding 1 ml of 75 % ethanol. Vortex the samples.
9. Centrifuge at 7,500 × *g* for 10 min at 4 °C.
10. Remove the supernatant, and let the pellet dry (do not let it dry completely because it would decrease its solubility).
11. Dissolve the RNA in 50 µl of preheated (60–70 °C) RNase-free H_2O.
12. Measure purity and concentration by spectrophotometry (e.g.: NanoDrop). Ideally, RNA concentration should be 1–1.5 µg/µl.

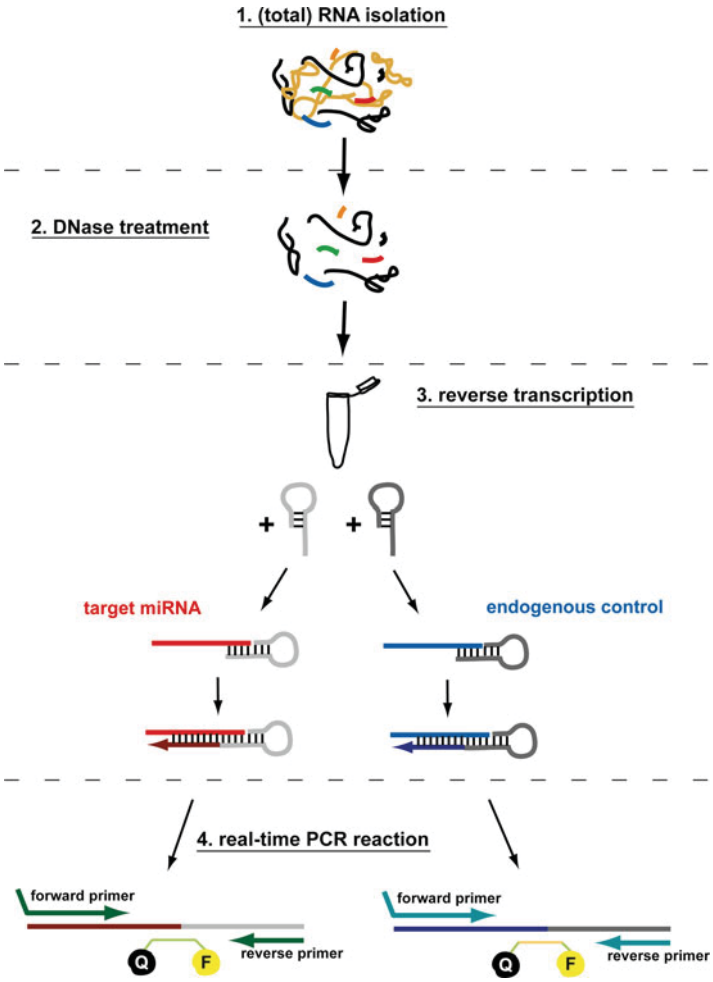


Fig. 3 Overview of methods used for the detection of mature miRNAs by qRT-PCR. *F* fluorophore, *Q* quencher

13. Check integrity of the samples by gel electrophoresis:
Mix 1 µg (~1–2 µl) of RNA and 10 µl of formamide, incubate at 70 °C for 10 min, and chill on ice. Add 1 µl of ethidium bromide [10 mg/ml] and 2.5 µl of 6× loading dye. Analyze samples by electrophoresis in a 1 % agarose gel. The ratio of the intensity of the bands corresponding to the 28S and 18S ribosomal RNA reflects the integrity of the RNA sample (ideally it should be about 2:1).

3.5.2 *miRNA-Specific Reverse Transcription*

1. DNase treatment of the total RNA is strongly recommended before reverse transcription (*see* **Note 7**):
Mix ~5 µg of total RNA, 2 µl (4 U) of DNase, 2 µl of 10× DNase buffer, 1 µl (20 U) of RNase inhibitor in the total volume of 20 µl. Incubate at 37 °C for 1 h, inactivate the enzyme at 75 °C

for 10 min, and place it on ice. Quantify RNA concentration by spectrophotometry.

2. For the reverse transcription reaction, mix 0.15 μl of 100 mM dNTP Mix, 1.5 μl of 10× buffer, and 1.16 μl of H_2O. Add 0.19 μl of RNase inhibitor [20 U/μl] and 1 μl of Reverse Transcriptase [50 U/μl]. Mix gently, do not vortex. These amounts are per sample but a preparation of a mastermix is recommended.
3. Add 5 μl of DNase treated total RNA [2 ng/μl].
4. Mix gently and add 3 μl of endogenous control-specific RT primer and 3 μl of target-specific RT primer.
5. Reverse transcribe the RNA according to the manufacturer's instructions (16 °C for 30 min, 42 °C for 30 min, 85 °C for 5 min; *see* **Note 21**).

3.5.3 Quantitative Real-Time PCR

1. Dilute five times the total volume of cDNA (15 μl) by adding 60 μl of H_2O (*see* **Note 22**).
2. Prepare real-time reactions by mixing 10 μl of 2× Reaction Mix, 1 μl of 20× probe, and 9 μl of diluted cDNA. (Perform each sample in triplicate in singleplex reactions.)
3. Cover and centrifuge the plate or strip at 500 × *g* for 1 min.
4. Perform the PCR reaction according to the manufacturer's instructions (50 °C for 2 min, 95 °C for 10 min followed by 40 cycles of 95 °C for 15 s, 60 °C for 1 min).
5. Analyze the data (StepOne version 2.1, *see* **Note 23**).

4 Notes

1. Mirtron-sized long oligonucleotides must be purified by at least HPLC although the most suitable method is PAGE separation. To get sufficient yield after purification, the starting amount for synthesis should be a minimum of 100 nmol. Depending on the cloning plan, oligonucleotides can be designed to be complementary to each other for blunt end ligation, or with 5′ or 3′ overhangs for the formation of sticky ends. In the case of blunt end ligation, 5′ phosphorylated oligos should be purchased if the removal of phosphate groups of the backbone's ends is planned.
2. Annealed mirtron sequences can be inserted into a desired marker gene (e.g., EGFP) as artificial introns [12, 13, 15]. Alternatively, mirtrons can be amplified from genomic DNA with their flanking exons and then this minigene can be cloned behind a marker gene [11, 12, 14, 15].
3. To screen for successful ligations by colony PCR, usually one primer is designed to anneal in the region of the insert and the

other one on the backbone sequence. In the case of numerous different mirtron ligations, employment of a backbone-specific primer pair is more practical. However, the primers should anneal as close to the insert as possible for easy distinction between the plasmids containing the insert and the empty vector during gel electrophoresis.

4. To test the mirtron function using a luciferase assay, make sensor and mutant-sensor plasmids. Sensor constructs contain two fully complementary antisense copies of the particular predicted mature mirtron species, which is cloned downstream of Renilla luciferase in the psiCHECK2 vector. The mutant-sensor plasmid differs from the sensor plasmid in the three mismatched nucleotides in the predicted miRNA seed region (every second nucleotide). Annealed double-stranded oligonucleotides are ligated between the XhoI/NotI restriction sites of Renilla luciferase 3′-UTR according to the manufacturer's instructions.
5. To investigate splicing dependency, generation of a splicing-mutant mirtron construct is suitable. In such constructs, intronic splicing-signal sequences are mutated to block the removal of the intron by splicing, thereby also inhibiting the first step of mirtron biogenesis. Mutations can be introduced at both 5′ (donor) and 3′ (acceptor) splice sites of the intron; however, the mutation of only the 5′ splice site is usually sufficient (e.g., change GT to TG but see also refs. [11–15]). Mutant oligonucleotides can be purchased from a commercial source (*see* also **Note 1**).
6. There are several total and miRNA isolation methods and reagents; however, total RNA extraction by Trizol reagent is still the most commonly used method for miRNA studies. On the other hand, it is important to keep in mind that the efficient isolation of miRNAs with low GC content or stable secondary structure is sensitive to the initial number of cells [27]. Therefore, for reproducible measurements, we recommend to use a relatively high cell number (a minimum of $1–2 \times 10^6$ cells/ml) for this type of isolation.
7. Before reverse transcription, we strongly recommend DNase treatment of RNA samples. First, contamination of RNA samples with genomic and transiently transfected DNA plasmid might interfere with the accurate determination of RNA concentration, since both DNA and RNA are measured by absorbance at 260 nm. Second, DNA contamination might influences mature miRNA detection by stem-loop qRT-PCR by giving false positive signals.
8. Alternatively, depending on the nature of antibiotics, competent cells can be spread onto LB-agar plates immediately after heat shock. Typically, this is possible for ampicillin, which is an inhibitor of cell wall synthesis. However, for antibiotics inhibiting

translation such as kanamycin, a short time of growth in a nonselective medium is required to obtain resistant cells.

9. Amplicon length differs depending on the localization of the primer pair. To distinguish an empty vector from the longer (70–100 nucleotides) mirtron-containing amplicon, gel electrophoresis using different agarose concentrations can be applied. As an example, we use 3 % agarose gels to distinguish between 470 bp and 550 bp amplicons.
10. Although 0–15 min is recommended to obtain DNA molecules efficiently packaged in lipid complexes, longer incubation of DNA with FuGENE® HD reagent (45–60 min) usually improves transfection efficiency.
11. Sorting the cells about a week after transfection is preferable because marker gene expression detected at that time originates mostly from the inserted genomic copies; transiently residing plasmids are usually lost by then. Another advantage is that the higher number of sorted cells obtained about a week after transfection improves cell viability.
12. It is possible to separate cells with different fluorescent intensity corresponding to different levels of transgene expression, but it depends on the type of the cell sorter. After the sorting procedure, there is a heterogeneous cell population with different marker gene copy numbers. If it is desired, it can be cloned further to get a homogenous one-cell-derived population.
13. Generally, a significant proportion of the cells (approximately 30–50 % of the total) die during the sorting process. Keep this in mind to select the appropriate cell culture dish size for plating the sorted cells, since cells generally do not grow well in low cell density.
14. A typical experimental setup for testing the function of a potential miRNA derived from a mirtron (either of its 5′ or 3′ arm):
    - mirtron + sensor (three parallel transfections)
    - mirtron + seed-mutant sensor (three parallel transfections)
    - non-cognate miRNA (negative control, having unrelated sequence) + mirtron sensor (three parallel transfections)
    - non-cognate miRNA + mirtron seed-mutant sensor (three parallel transfections).

    You have to perform 12 transfections for testing one arm of the mirtron, and 12 transfections for testing the other arm of potential miRNA species. In summary, there are 24 transfections in one experiment for a particular mirtron.
15. Samples can be measured one by one in a single sample or in a plate-reading luminometer. Alternatively, more samples can be measured simultaneously in a plate-reading luminometer by

pipetting no more than 2–4 samples in the plate at the same time. Another option is to culture, transfect, and measure cells directly in a specific opaque plate. In this case, a plate-reading luminometer has to be equipped with two injectors. When measuring single samples in multiwell opaque plate, we advise using only every second well to prevent cross talk between them. It is important to keep the samples away from bright lights and sunlight.

16. LARII is provided in lyophilized form and, according to the manufacturer's instructions, should be resuspended in 10 ml of Luciferase Assay Buffer II. However, the supplied volume of buffer is usually about 11 ml, and all can be added directly to the LARII. It is also important to note that the Stop & Glo reagent has to be prepared in a glass vial or in a siliconized polypropylene tube.
17. Splicing deficiency of the splice site-mutated construct can be verified by RT-PCR with the same pair of primers used for colony detection by PCR (localized outside of the mirtron insert, in the region of the marker gene). Moreover, in case of a fluorescent reporter (EGFP), mutant constructs can also be tested for splicing deficiency based on fluorescence, since no fluorescent signals should be detected after transfection.
18. The independency of mirtron processing from the Drosha/DGCR8 complex can be examined in Drosha- or DGCR8-deficient cells (knocking down Drosha or DGCR8 with siRNA or shRNA, or in genetically knocked out cell lines). Although in HeLa cells the knockdown of Drosha with siRNAs was demonstrated at both mRNA and protein levels, this was not sufficient to abolish the canonical miRNA biogenesis pathway. Alternatively, a transdominant-negative Drosha (TN-Drosha) [14], or as we did, DGCR8-deficient mouse embryonic fibroblasts can be used [15, 28]. Moreover, a DGCR8-knockout mouse embryonic stem cell line can also be purchased from commercial source.
19. There are several methods for direct detection of mature miRNAs such as Northern blot and qRT-PCR. The former method is more traditional and can detect the different processed miRNA forms. Although its sensitivity can be strongly increased by locked nucleic acids (LNAs; [29, 30]), the detection of ~20–22-nucleotide-long mature miRNAs by qRT-PCR is a more sensitive technique [15]. For this measurement, we use stable mirtron-expressing cell lines and two negative controls to check the specificity of the reactions. One negative control is a sample from the original, genetically non-modified (parental) cell line and the other one is from another, unrelated miRNA-overexpressing cell line-derived sample.

20. When removing the aqueous phase during the phase separation step of RNA isolation, approximately 500–600 μl can be recovered from 1 ml Trizol. However, to avoid touching the interphase and thereby contaminating with DNA, we recommend removing a smaller volume.
21. Reverse transcription efficiency may vary between samples when using different type of PCR tubes.
22. The final dilution of cDNA in the qPCR reaction is 11× (it has to be a minimum of 10×). Always use non-template controls in all the assays.
23. Always check the baseline and threshold values since big differences in Ct values between samples as well as a little contamination in the non-template control might cause a wrong autofit by the program.

Acknowledgements

We would like to thank Károly Fátyol for helpful discussions. Anita Schamberger is a recipient of the Jedlik Ányos predoc fellowship. This research was supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 "National Excellence Program" and also supported by the TransRat grant KMR_12-2012-0112 given to Tamás I. Orbán.

References

1. Filipowicz W, Bhattacharyya SN, Sonenberg N (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet 9:102–114
2. Carthew RW, Sontheimer EJ (2009) Origins and mechanisms of miRNAs and siRNAs. Cell 136:642–655
3. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10:94–108
4. Slezak-Prochazka I, Durmus S, Kroesen BJ et al (2010) MicroRNAs, macrocontrol: regulation of miRNA processing. RNA 16:1087–1095
5. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. Cell 136:215–233
6. Krol J, Loedige I, Filipowicz W (2010) The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet 11:597–610
7. Pasquinelli AE (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet 13: 271–282
8. Yates LA, Norbury CJ, Gilbert RJ (2013) The long and short of microRNA. Cell 153:516–519
9. Miyoshi K, Miyoshi T, Siomi H (2010) Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. Mol Genet Genomics 284:95–103
10. Yang JS, Lai EC (2011) Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. Mol Cell 43:892–903
11. Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. Nature 448:83–86
12. Okamura K, Hagen JW, Duan H et al (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 130: 89–100
13. Sibley CR, Seow Y, Saayman S et al (2012) The biogenesis and characterization of mammalian

microRNAs of mirtron origin. Nucleic Acids Res 40:438–448

14. Havens MA, Reich AA, Duelli DM et al (2012) Biogenesis of mammalian microRNAs by a non-canonical processing pathway. Nucleic Acids Res 40:4626–4640
15. Schamberger A, Sarkadi B, Orban TI (2012) Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. RNA Biol 9:1177–1185
16. Westholm JO, Lai EC (2011) Mirtrons: microRNA biogenesis via splicing. Biochimie 93:1897–1904
17. Flynt AS, Greimann JC, Chung WJ et al (2010) MicroRNA biogenesis via splicing and exosome-mediated trimming in Drosophila. Mol Cell 38:900–907
18. Ladewig E, Okamura K, Flynt AS et al (2012) Discovery of hundreds of mirtrons in mouse and human small RNA data. Genome Res 22:1634–1645
19. Curtis HJ, Sibley CR, Wood MJ (2012) Mirtrons, an emerging class of atypical miRNA. Wiley Interdiscip Rev RNA 3: 617–632
20. Meng Y, Shao C (2012) Large-scale identification of mirtrons in Arabidopsis and rice. PLoS One 7:e31163
21. Joshi PK, Gupta D, Nandal UK et al (2012) Identification of mirtrons in rice using MirtronPred: a tool for predicting plant mirtrons. Genomics 99:370–375
22. Ivics Z, Hackett PB, Plasterk RH et al (1997) Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell 91:501–510
23. Ammar I, Izsvak Z, Ivics Z (2012) The Sleeping Beauty transposon toolbox. Methods Mol Biol 859:229–240
24. Lacy-Hulbert A, Thomas R, Li XP et al (2001) Interruption of coding sequences by heterologous introns can enhance the functional expression of recombinant genes. Gene Ther 8: 649–653
25. Feschotte C, Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet 41:331–368
26. Izsvak Z, Ivics Z (2004) Sleeping beauty transposition: biology and applications for molecular therapy. Mol Ther 9:147–156
27. Kim YK, Yeo J, Kim B et al (2012) Short structured RNAs with low GC content are selectively lost during extraction from a small number of cells. Mol Cell 46:893–895
28. Wang Y, Medvid R, Melton C et al (2007) DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. Nat Genet 39:380–385
29. Varallyay E, Burgyan J, Havelda Z (2008) MicroRNA detection by northern blotting using locked nucleic acid probes. Nat Protoc 3:190–196
30. Varallyay E, Burgyan J, Havelda Z (2007) Detection of microRNAs by Northern blot analyses using LNA probes. Methods 43:140–145

HUMAN GENE THERAPY METHODS 25:241–252 (August 2014)

DOI: 10.1089/hgtb.2013.149

Excision Efficiency Is Not Strongly Coupled to Transgenic Rate: Cell Type-Dependent Transposition Efficiency of *Sleeping Beauty* and *piggyBac* DNA Transposons

Orsolya Kolacsek,[1] Zsuzsa Erdei,[1] Ágota Apáti,[1,2] Sára Sándor,[1] Zsuzsanna Izsvák,[3] Zoltán Ivics,[4] Balázs Sarkadi,[1,2] and Tamás I. Orbán[1,5]

Abstract

The *Sleeping Beauty* (SB) and *piggyBac* (PB) DNA transposons represent an emerging new gene delivery technology, potentially suitable for human gene therapy applications. Previous studies pointed to important differences between these transposon systems, depending on the cell types examined and the methodologies applied. However, efficiencies cannot always be compared because of differences in applications. In addition, ''overproduction inhibition,'' a phenomenon believed to be a characteristic of DNA transposons, can remarkably reduce the overall transgenic rate, emphasizing the importance of transposase dose applied. Therefore, because of lack of comprehensive analysis, researchers are forced to optimize the technology for their own ''in-house'' platforms. In this study, we investigated the transposition of several SB (SB11, SB32, SB100X) and PB (mPB and hyPB) variants in various cell types at three levels: comparing the excision efficiency of the reaction by real-time PCR, testing the overall transgenic rate by detecting cells with stable integrations, and determining the average copy number when using different transposon systems and conditions. We concluded that high excision activity is not always followed by a higher transgenic rate, as exemplified by the hyperactive transposases, indicating that the excision and the integration steps of transposition are not strongly coupled as previously thought. In general, all levels of transposition show remarkable differences depending on the transposase used and cell lines examined, being the least efficient in human embryonic stem cells (hESCs). In spite of the comparably low activity in those special cell types, the hyperactive SB100X and hyPB systems could be used in hESCs with similar transgenic efficiency and with reasonably low (2–3) transgene copy numbers, indicating their potential applicability for gene therapy purposes in the future.

Introduction

Transposons were first exclusively considered to be genomic parasites, placing a heavy burden on the host genome. Later it was revealed that these mobile elements have played active roles in shaping the genetic materials of organisms (Kazazian, 2004; Feschotte and Pritham, 2007; Goodier and Kazazian, 2008), and additional studies further supported the idea that transposon activity could have a positive impact during evolution even in higher eukaryotic organisms (Singer *et al.*, 2010; Shen *et al.*, 2011). After cloning and isolating active elements from various sources, transposons have been used as molecular tools for genetic experiments, including mutagenesis screens or transgenesis experiments. DNA transposons are especially favorable for such purposes, mainly because most of them are transposed by a nonreplicative ''cut and paste'' mechanism (Ivics and Izsvák, 2004, 2006). For a long time, however, no DNA transposons were known to be active in vertebrates, which also limited human gene delivery and gene therapy application by these molecular devices. This burden was lifted with the ''resurrection'' of the artificial transposon *Sleeping Beauty* (SB), which can actively transpose in all vertebrate genomes tested, including human cells (Ivics *et al.*, 1997). Since then,

[1]Institute of Enzymology, Research Center for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Hungary.
[2]Max Delbrück Center for Molecular Medicine, 13092 Berlin, Germany.
[3]Divison of Medical Biotechnology, Paul Ehrlich Institute, 63225 Langen, Germany.
[4]Molecular Biophysics Research Group of the Hungarian Academy of Sciences, Semmelweis University and National Blood Service, 1113 Budapest, Hungary.
[5]Chemical Technology Transfer, 1121 Budapest, Hungary.

various other transposons of several species have been shown to be active in vertebrates (Miskey *et al.*, 2003; Ding *et al.*, 2005; Balciunas *et al.*, 2006). In addition, tremendous efforts have been made to develop hyperactive transposase versions (Zayed *et al.*, 2004; Baus *et al.*, 2005; Pledger and Coates, 2005) and with the development of the SB100X system (Mátés *et al.*, 2009), transposon-based gene delivery became a rational alternative to virus-based applications. Indeed, a clinical trial has already been initiated using the SB system (Williams, 2008).

At present, the most efficient and therefore the most widely used transposon-based applications are the SB, *piggyBac* (PB), and *Tol2* systems, the latter two representing naturally active transposons derived from the cabbage looper moth (*Trichoplusia ni*) (Ding *et al.*, 2005) and the medaka fish (Balciunas *et al.*, 2006) genomes, respectively. All have high transgenic potential and can mobilize large cargos (Rostovskaya *et al.*, 2012; Wang *et al.*, 2014), which is a considerable factor as compared with viral packaging systems. For safety considerations in gene therapy, SB seems to be the most ideal system because of its favorable integration profile: there is no obvious preference at the genomic level, and therefore the delivered transgene is integrated randomly, lowering the risk of undesirable insertional mutagenesis (Galvan *et al.*, 2009; Grabundzija *et al.*, 2010; Izsvák *et al.*, 2010; Meir *et al.*, 2011; Burnight *et al.*, 2012; M.A. Li *et al.*, 2013). On the other hand, the clear advantage of the PB system comes when the traceless removal of the delivered transgene is favorable, as exemplified by removal of the reprogramming cassettes after the generation of induced pluripotent stem cells from somatic cells (Kaji *et al.*, 2009; Woltjen *et al.*, 2009; Yusa *et al.*, 2009). In addition, this genetic manipulation is further enhanced with the development of an integration-deficient PB transposase mutant (X. Li *et al.*, 2013). Nevertheless, there are various other aspects (e.g., transgene copy numbers, applications in difficult-to-transfect cell lines) that require careful adjustments and considerations, especially when developing future gene therapy methodologies, and novel DNA transposons with defined integration profiles may also be considered for such purposes (Kojima and Jurka, 2013).

There are only a few studies addressing systematic comparisons of available transposon-based systems, and they often concentrate on a limited number of cell lines, mainly tumor cell lines, that may not be representative for all cell types or for *in vivo* applications (Grabundzija *et al.*, 2010; Huang *et al.*, 2010; Doherty *et al.*, 2012; Sharma *et al.*, 2012). Moreover, specific interactions of the transposase proteins with endogenous cell type-specific factors may also influence the overall outcome of transposition, an important aspect that also requires further investigations. In this study, we systematically analyzed SB and PB transposition from several aspects: examining the excision efficiency of the transposase as compared with the transgenic rate and copy numbers, the overall end products of transposition. Moreover, we examined the phenomenon of overproduction inhibition (OI), a transposase dose-dependent inhibition of the reaction, which is believed to be a general characteristic of DNA transposons (Lohe and Hartl, 1996; Hartl *et al.*, 1997; Lampe *et al.*, 1998; Wilson *et al.*, 2005; Claeys Bouuaert *et al.*, 2013). We found that the choice of transgene, as well as the promoter identity, may also have a strong impact on the outcome of the assays. Besides investigating cultured cell lines, we also addressed all the raised issues in a human embryonic stem cell line, which represents a model system for tissue differentiation studies and for regenerative medicine investigations. The important finding of our investigations is that although the SB100X transposon system seemed to be performing better in several respects in cultured cell lines, embryonic stem cells seem to represent an equally less permissive environment for both SB and PB transposons. Our results underlie the importance of careful optimization of transposon-based gene delivery in the targeted cell type, from the choice of the delivered gene expression cassette, through the careful selection of the transposase system, until the thorough setup of the gene delivery protocol. With all these details cautiously addressed, transposon-based gene methodologies clearly represent promising new technologies for future gene therapy applications.

Materials and Methods

Cell lines and maintenance

Human embryonic kidney cells (HEK-293) and HeLa cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum, 1% L-glutamine, and 1% penicillin–streptomycin (Life Technologies, Carlsbad, CA). The HUES9 embryonic stem cell line (originally provided by D. Melton, Harvard University, Cambridge, MA) was maintained essentially as described previously, using cells from passage 35 (Apáti *et al.*, 2008).

Plasmid constructs, transfection methods, and transposition assays

Structures of the transposon plasmids used are shown in Fig. 1A. During the comparison studies, the exactly same expression cassette sequences were cloned between the corresponding SB or PB inverted repeat sequences. Transposase proteins were expressed from standard expression plasmids under the control of the cytomegalovirus (CMV) promoter; transposase levels were checked by Western blotting. During the experiments, the following transposase protein variants were used: for the SB system, the SB11, SB32, and SB100X variants (Mátés *et al.*, 2009); for the PB system, the mPB (Cadinanos and Bradley, 2007) and hyPB (Yusa *et al.*, 2011) variants.

All transfections were carried out in duplicate. For lipid-based transfections, 2×10^5 HEK-293 cells or 4×10^5 HeLa cells were seeded onto 12-well plates, and transfected with FuGENE 6 or FuGENE HD reagent, respectively, according to the manufacturer's instructions (Roche, Basel, Switzerland). The next day, a maximum of 250 ng of transposon plasmid was transfected with various amounts of transposase plasmid, except when the transposon was rate limiting at 25 ng (see Fig. 4C). All transfections were supplemented with 250 ng of pCMV-EGFP expression plasmid to control transfection efficiencies. In every experiment, transfection cocktails were supplemented with pET-41 plasmid if necessary, to ensure the same amount of total DNA in all reactions. At 48 hr posttransfection, cells were harvested and counted; one-tenth of them were analyzed with a FACSCalibur flow cytometer (BD Biosciences, San Jose, CA) for green fluorescent protein (GFP)-expressing cells to measure transfection efficiency. For

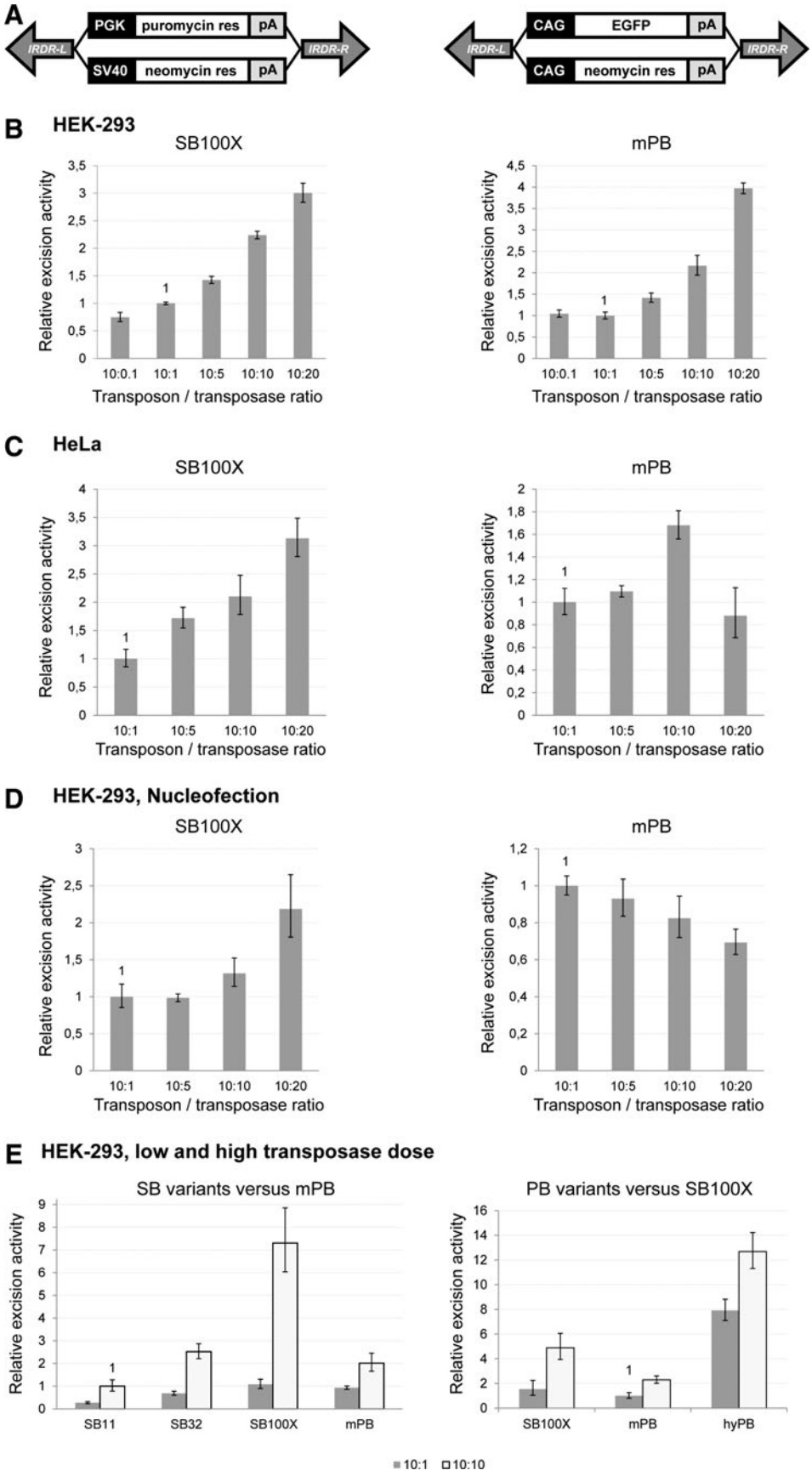


FIG. 1. Transposon constructs and their application in studying overproduction inhibition on the level of excision of DNA transposons. **(A)** Schematic representation of the four DNA transposons used. IRDR-L/-R, inverted repeat-direct repeat sequences of SB and PB transposons (left and right); PGK, phosphoglycerate kinase promoter; SV40, simian virus 40 early promoter; CAG, *C*MV enhancer–chicken *β-a*ctin–rabbit β_1-*g*lobin promoter; pA, polyadenylation signal. **(B–D)** Transposase dose dependence of excision efficiency: when using lipid-based transfection in **(B)** HEK-293 cells or **(C)** HeLa cells, or **(D)** when using nucleofection in HEK-293 cells. **(E)** Effect of the amount of transposase on the excision efficiency in the case of different SB and PB variants in HEK-293 cells.

antibiotic selection, 1% of the transfected cells were seeded onto cell culture Petri dishes, selected for 2 weeks with either puromycin (1 μg/ml; Sigma-Aldrich, St. Louis, MO) or G418 (600 μg/ml; Sigma-Aldrich), and surviving cells were fixed with methanol and stained with Giemsa (Sigma-Aldrich). Colonies were quantified with a universal hood gel imager model 75S, using Quantity One 4.4.0 software (Bio-Rad, Hercules, CA). Transgenic rates were calculated from the colony numbers, normalizing with the number of seeded cells that was also normalized previously with transfection efficiencies; transgenic rates are therefore defined as a percentage of transfected cells.

In nucleofection experiments, 5×10^5 HEK-293 cells were electroporated, using a Nucleofector kit V (Lonza Group, Basel, Switzerland), according to the manufacturer's instructions. One microgram of transposon and various amounts of transposase, supplemented with 1 μg of transfection control plasmid, were transfected and seeded onto 6-well plates. Electroporations were also filled up to equal the amount of DNA. HEK nucleofections were analyzed 48 hr posttransfection for transfection efficiency by fluorescence-activated cell sorting (FACS) and for transposon excision efficiency by qPCR.

HUES9 cells were grown on feeder cells in 6-well plates until confluency; then $\sim2\times10^6$ cells (one entire well per nucleofection) were harvested for nucleofection with a human stem cell Nucleofector kit 1 (Lonza Group). The same amounts of DNA were applied as in HEK-293 cells, but because the transposon in these cases contained the CAG-EGFP-expressing cassette, the transfection control plasmid could be omitted. At 48 hr posttransfection, cells were analyzed by FACS to determine transfection efficiency. At two later time points (1 and 2 weeks later), the ratio of GFP-expressing transgenic cells was also determined by FACS. Normalizing this ratio with the transfection efficiency resulted in the ''GFP index,'' which is proportional to the transgenic rate.

HEK-293 and HeLa transfection efficiencies were in the range of 40–70%, and HUES9 electroporations were within 10–15% efficiencies, but in simultaneous transfection experiments, deviations of transfection efficiencies showed only a small range (3–8%) of alterations.

Quantitative real-time PCR measurements

All quantitative real-time PCR measurements were carried out on a StepOnePlus real-time PCR system platform (Life Technologies); SYBR green and TaqMan reactions were performed according to the standard protocols. Data analysis was done with StepOne software version 2.1 (Life Technologies) and relative quantifications were performed by the $\Delta\Delta C_t$ method; error bars represents confidence intervals of 95%. To quantify SB transposase-interacting factors at the mRNA level by SYBR methodology, the following primer pairs were used: for DNAPKcs (DNA-dependent protein kinase, catalytic subunit): sense primer, 5′-TTGAACACCATGTCCCAAGA; antisense primer, 5′-CTGACATTTTTGTCAGCCAATC; for HMGB1 (high-mobility group protein B1): sense primer, 5′-CATTGAGCTCCATAGAGACAGC; antisense primer, 5′-AGGATCTCCTTTGCCCATGT; for Ku70: sense primer, 5′-AGAGTGAAGATGAGTTGACACCTTT; antisense primer, 5′-CCAAGAGATCTCGATCACTGCT; for Ku80: sense primer, 5′-CATCTGATGCTACCAGATTTTGA; antisense primer, 5′-TCCATGCTCACGATTAGTGC; for HSP90 (heat shock protein 90, as endogenous control): sense primer, 5′-TGGATATCCCATTACTCTTTTTGTG; antisense primer, 5′-TTCTTTTTCTTCTTCTTTGTCTTCCT. Methods for determining transposon excision efficiency and copy numbers are described in the following sections in detail.

Quantifying transposon excision efficiency

After processing cells for transposition assays and for measuring transfection efficiency at 48 hr posttransfection, the remaining cells were used for the excision assay. Direct comparisons of various transposases were carried out in simultaneous experiments with similar transfection efficiency values to exclude minor discrepancies due to high transfection efficiency differences. Duplicated transfections were pooled and plasmids were isolated from the cells, using a modified protocol of the Qiagen plasmid miniprep kit, applying 300 μl of 1.2% sodium dodecyl sulfate (SDS) supplemented with 50 μg of proteinase K for the cell lysis step and protein removal from the DNA. Ten nanograms of each isolated plasmid DNA was used in a two-round nested PCR assay in which primers were specific to the plasmid backbone and resulted in a PCR product only from the excised and repaired plasmid copies due to large transposon cassettes. First-round PCR primers were as follows: 5′-GCGAAAGGGGGATGTGCTGCAAGG, 5′-TCTTTCCTGCGTTATCCCCTGATTC. Fifteen cycles were done in a 20-μl reaction volume with 2×PCR master mix (Promega, Madison, WI) and 200 n*M* primers. The first-round PCR profile was as follows: 94°C for 30 sec, 60°C for 15 sec, 72°C for 1 min. Second-round PCR primers were as follows: 5′-CAGCTGGCACGACAGGTTTCCCG, 5′-CGATTAAGTTGGGTAACGCCAGGG. A five-step fourfold dilution series (1024×) was done from each first-round PCR and 5 μl was introduced to the second-round PCR, which was the real-time PCR step, and it was carried out with 2×*Power* SYBR green PCR master mix (Life Technologies) using 50 n*M* concentrations of primers in triplicate 20-μl reactions. Excision PCR was normalized to the PCR from the ampicillin sequence, which was present exclusively in the transposon donor constructs, thus enabling quantification of the excised ratio of transposon plasmids passed into the cells. In this way we could compare excision efficiencies independent of transfection efficiency and cell type. PCR primers for the ampicillin sequence were as follows: 5′-TTTGCTCACCCAGAAACGC, 5′-AGTTGGCCGCAGTGTTATCAC. Ampicillin-specific PCRs were done similarly to the second-round excision PCR from the first-round excision PCR product, using the same serial dilutions. Efficiencies determined for all assays were in the range of 95–105%, so the $\Delta\Delta C_t$ method proved to be suitable for quantification.

Determining transposon copy numbers

Average copy numbers of transgenic cell populations were measured from transfected and selected cells. Concerning the HEK-293 cell line, puromycin (1 μg/ml; Sigma-Aldrich) selection was carried out for 3 weeks, and then cells were harvested and duplicate transfections were pooled. Genomic DNA was isolated by the standard phenol–chloroform extraction method. For copy number measurements, the amount of input DNA was 90 ng and the

self-designed TaqMan assays were used as previously published (Kolacsek *et al.*, 2011). Briefly, specific assays were applied for the SB inverted repeat-direct repeat sequence-left region (SB-IRDR-L) transposon motif and the GFP-coding sequence; the latter assay was used to determine PB transposon copy numbers. Transposon-specific assays were normalized to RPPH1 endogenous control assay, and genomic DNA of cell clones with one transposon copy were also used as reference samples (Kolacsek *et al.*, 2011). Concerning HUES9 cells, a GFP-expressing transposon cassette was used, and therefore these transfections were sorted at 48 hr for the GFP-expressing cells, and two additional sorts were carried out for the selection of transgenic cells. Average copy numbers of cell populations were measured from these sorted cells approximately 3 weeks after transfection.

Results

Cell type dependence of transposon excision rates: SB100X never exhibits overproduction inhibition at this level

The ''cut and paste'' mechanism of DNA transposons can be divided into at least two well-defined steps: the excision step from the donor locus and the integration step of the transposon into the host genome. We attempted to characterize these steps separately and to compare the PB and SB variants from these aspects. To avoid variability resulting from different DNA sequences, we cloned exactly the same transgene cassette expressing the puromycin resistance gene between the SB or PB inverted repeat (IR) sequences (Fig. 1A). The same amount of transposon-carrying plasmids was transfected with variable amounts of the given transposase-expressing plasmids into HEK-293 or HeLa cells by lipid-based transfection. Excision efficiencies were measured by real-time PCR-based quantification of the repaired donor plasmids isolated from the cells 48 hr posttransfection. Excision PCR was normalized to the PCR of the ampicillin sequence, which was exclusively present in the transposon donor constructs, thus enabling determination of the ratio of plasmids that underwent transposon excision to the total amount of transposon-carrying plasmids passed into the cell. In this way we can compare excision efficiencies in different cell types. In HEK-293 cells, excision efficiencies increased steadily with the amount of transposase for both the SB100X and mPB systems, whereas in HeLa cells, mPB excision activity showed a decline after a certain amount of transposase, resembling the overproduction inhibition phenomenon (Fig. 1B and C). To further increase the transfection efficiency and thereby the amount of expressed transposase in the cell, we used nucleofection technology, which is known to achieve higher transgene delivery efficiency (Aluigi *et al.*, 2006). Using this method, the mPB transposase indeed showed decreasing excision efficiency even in HEK-293 cells with the increasing amount of transposase, whereas the SB100X excision rate still showed a positive correlation (Fig. 1D). In HEK-293 cells we also examined the OI phenomenon for other, less active SB versions (SB32 and SB11), as well as for a hyperactive PB (hyPB) version in several experiments. None of the transposases showed OI when raising the transposase dose from 10:1 to 10:10, and hyPB showed the highest excision activity, exceeding that of SB100X (Fig. 1E). Similarly, no OI was detected for the SB variants in HeLa cells (data not shown).

In comparing excision efficiencies between the two cell lines at fixed transposon/transposase amounts, the HEK-293 cell line seemed the most permissive among all examined systems (Fig. 2A). It is also notable that the hyperactive transposases highly exceeds all other variants in terms of excision efficiency, whereas from this point of view, mPB showed efficiency comparable to different SB variants in different cell lines: to SB32 in HEK-293 cells and to SB11

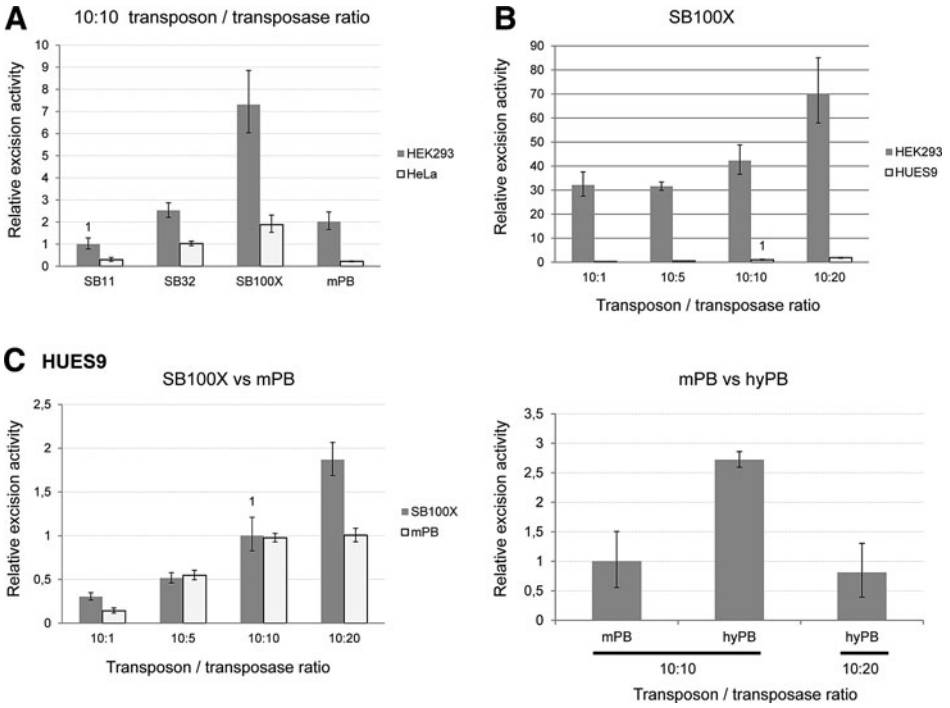


FIG. 2. Comparing the excision efficiencies of SB and PB transposon systems in various cell types. **(A)** Excision efficiencies of various DNA transposons in HEK-293 versus HeLa cells at a fixed transposon-to-transposase ratio. **(B)** Comparison of SB100X excision efficiency in HEK-293 versus HUES9 cells. **(C)** Transposase dose dependence and comparison of excision efficiencies of SB100X, mPB, and hyPB in a human embryonic stem cell line, HUES9.

in HeLa cells. These correlations suggest that the cell type dependence of mPB might not be similar to that of SB. These results were independent of the transposon sequence because similar excision results were obtained with different transcription units such as CAG or CMV promoter-driven EGFP expression cassettes (data not shown).

We also examined excision efficiency in the human embryonic stem cell (hESC) line HUES9, another relevant test system for gene therapy applications and tissue differentiation models. Here, increasing the amount of transposase also resulted in a higher excision rate of SB100X, whereas the efficiency of mPB showed more of a saturation-type reaction kinetic profile (Fig. 2C). However, for various transposase concentration points, mPB and SB100X had similar excision efficiencies, although these values were far below those measured in HEK-293 cells (Fig. 2B), or even in HeLa cells. When comparing mPB with hyPB, the hyperactive version had higher excision efficiency but showed OI at a higher transposase dose (Fig. 2C). To summarize this part of the study, the excision step of DNA transposition showed strong cell type dependence, with the hESCs being the least permissive cell type for this event. In addition, the SB100X reaction could not be saturated at this level, whereas under the conditions applied the PB excision rate (both for mPB and hyPB) was more sensitive to the amount of transposase present in the reaction.

Comparing transgenic rates of various transposon systems in various cell types

To determine the stable integration of transgenes into the host genomes, we performed colony-forming assays after transfection and puromycin selection of cells. Transgenic rates were determined in the same transfections as the excision assays, and were calculated from the number of detected colonies defined as a percentage of seeded and also transfected cells. The most prominent observation was that high excision efficiencies did not correlate with higher transgenic rates, and overproduction inhibition was also detectable at this level. When examining the HEK-293 cell line, overproduction inhibition could not be detected for either the SB100X transposase or the mPB system (Fig. 3A). In the HeLa cell line, transgenic rates of SB100X showed more of a saturation curve when plotted as a function of increasing transposase concentration (Fig. 3B), in contrast to the strong OI results reported previously (Grabundzija *et al.*, 2010). On the other hand, mPB showed a tendency toward slightly higher OI at higher transposase dose (especially at 10:20; see Fig. 3B), in agreement with the excision assay

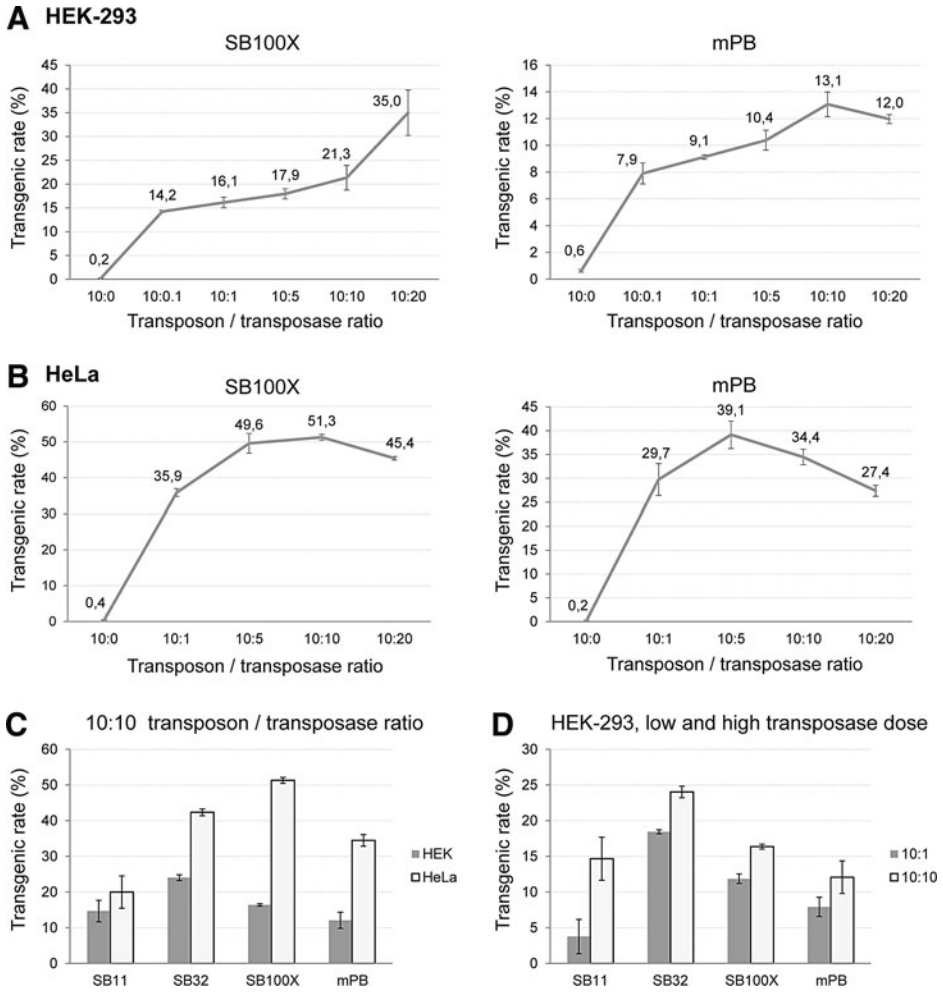


FIG. 3. Transgenic rates of SB and mPB systems and their transposase dose dependence in **(A)** HEK-293 cells and **(B)** HeLa cells. **(C)** Transgenic rates of different transposon systems in the two cell lines at a fixed transposon -to-transposase ratio. **(D)** Effect of low and high transposase doses on transgenic rates in HEK-293 cells. Transgenic rates are defined as the transgene-carrying cells, as a percentage of total transfected cells (see also Materials and Methods). Error bars represent SD values.

and also in line with some previous data (Grabundzija *et al.*, 2010) but not with others (Doherty *et al.*, 2012).

We also compared the two cell types for overall transgenic rate at fixed transposon/transposase amounts (10:10). HeLa cells showed the highest transgenic rates for hyperactive transposases examined, with SB100X having the highest values, and the SB32 variant exceeding the rate of mPB (Fig. 3C). Nevertheless, the HeLa cell line is reported to exhibit overproduction inhibition in terms of transgenic rates for most transposases, although for the SB100X variant our data partially contradict that previously reported, but mPB showed OI at both levels, which was in agreement with previous results (Grabundzija *et al.*, 2010). An interesting feature revealed was that in spite of the high excision efficiencies in HEK-293 cells, transgenic rates for most transposases were found to be significantly lower than in HeLa cells (the only exception was the SB11 variant). Moreover, SB32 seemed to be more efficient in HEK-293 cells than SB100X even at a low transposase dose (Fig. 3D), and a high transposase dose of SB11 resulted in similar transgenic efficiency, like SB100X. In addition, we could not detect inhibition of less active SB variants at the transgenic level (Fig. 3D), and the transgenic rate of SB11 increased to a similar extent as its excision activity when raising the transposase dose (Fig. 1E vs. Fig. 3D). When examining the PB systems, in spite of exhibiting the highest excision activity in HEK-293 cells, the transgenic rate of the hyPB variant was similar to that of mPB even at different transposase doses, and it was lower than that of SB100X (Fig. 4A and B), which is not in agreement with some previous results (Doherty *et al.*, 2012). Taken together, our results pointed out that despite the higher detectable excision rate of DNA transposons, HEK-293 cells seemed to be less permissive for stable genomic integration of transgenes, resulting in lower transgenic rates after transposition.

In previous studies, overproduction inhibition for SB variants was found to be more prominent at low transposon amounts, so we tested these conditions in our assay system with the more susceptible HeLa cell line. However, we could not see a decrease in transgenic rates with increasing amount of transposase and even for less hyperactive variants; rather, a saturation-type curve could be derived when plotting the data (Fig. 4C).

To determine whether the nature of the expression cassette has any effect on the transgenic rate, we tested the neomycin selection system in HeLa cells with different promoters (CAG or SV40) at low transposon dose (Fig. 4D). It turned out that they indeed influenced the kinetics of the

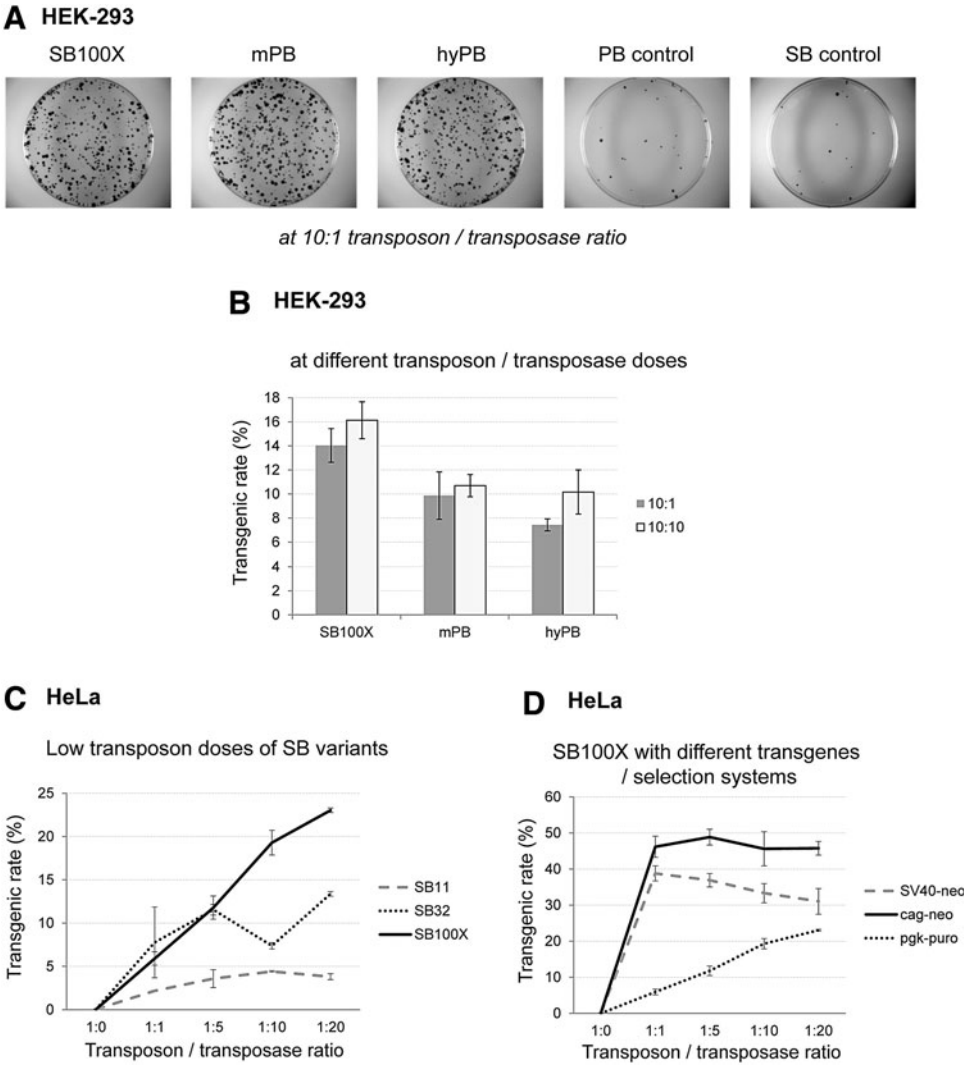


FIG. 4. Comparing the transgenic rates of different transposon systems. **(A)** Representative colony assays and **(B)** transgenic rates of the PB systems versus SB100X. PB control: transfection carried out with the PB transposon cassette but without the transposase; SB control: transfection carried out with the SB transposon cassette and the mutant SB transposase. **(C)** Transposase dose dependence of transgenic rates at low transposon dose in HeLa cells. **(D)** Measured dose dependence of transgenic rates is strongly influenced by the transposon cassette sequences and thereby the selection methodology. Transgenic rates are defined as the transgene-carrying cells, as a percentage of total transfected cells (see also Materials and Methods). Error bars represent SD values.

transgenic rate, and slight inhibition was observed when exactly the same expression cassette (SV40 promoter-driven neomycin resistance gene) was used as in the previously published study (Grabundzija *et al.*, 2010). These results emphasize the importance of the careful design of the expression cassette (e.g., choice of the promoter) carried by the transposon as it can severely influence the outcome and kinetics of the gene delivery and integration processes.

There are a number of proteins described to interact with the SB transposase, most of which are involved in DNA repair processes or in cell cycle regulation (Zayed *et al.*, 2003; Izsvák *et al.*, 2004; Walisko *et al.*, 2006). As these proteins could differ in their cell type-specific distribution and activity, we measured the mRNA expression levels of four SB-interacting factors (DNAPKcs, HMGB1, Ku70, Ku80) in the HeLa and HEK-293 cell lines. Contrary to our expectations, the expression patterns of all examined factors were strikingly similar in both cell lines (data not shown), so at least their expression level differences cannot be attributed to the sharp differences in excision efficiency and in transgenic rates.

Overproduction inhibition at the level of the transgenic rate was also examined in hESCs. However, puromycin selection cannot be reliably applied to these cells as they express the ABCG2 membrane transporter protein, which can extrude the selecting compound from the cells (Sarkadi *et al.*, 2010). Therefore, for these experiments we used a GFP-expressing transposon cassette. We measured the transfected cell ratio 48 hr posttransfection via FACS, and the ratio of transgene-expressing cells 1 and 2 weeks later. Normalizing the latter with the transfection efficiency resulted in the GFP index, which is proportional to the transgenic rate. SB100X showed a decline in GFP index at a 10:20 transposon/transposase dose in these cells (Fig. 5A), which might be attributable to overproduction inhibition detected only at the level of the transgenic rate. Here again, excision efficiencies cannot reliably predict the overall transposition rate, as the excision rate of SB100X showed a continuous increase in this range of transposase concentration (Fig. 2C). The GFP index of mPB showed more of a saturation-type kinetic curve in hESCs (data not shown), similar to what was measured for its excision efficiency (Fig. 2C). Under conditions in which SB100X and PB showed similar excision efficiency (10:10 ratio) we compared the SB and PB variants for the GFP index (Fig. 5B and C). In these experiments lower activity SB variants were notably less efficient for gene delivery in hESCs compared with SB100X, whereas both mPB and hyPB had higher transgenic rates than SB100X by 2 weeks posttransfection (Fig. 5C), which is in line with some previous data (Yusa *et al.*, 2011). These results indicate that there are certain cell types in which the overall PB transposition rate can outweigh that of the SB100X system, and hESCs seem to be one example of these.

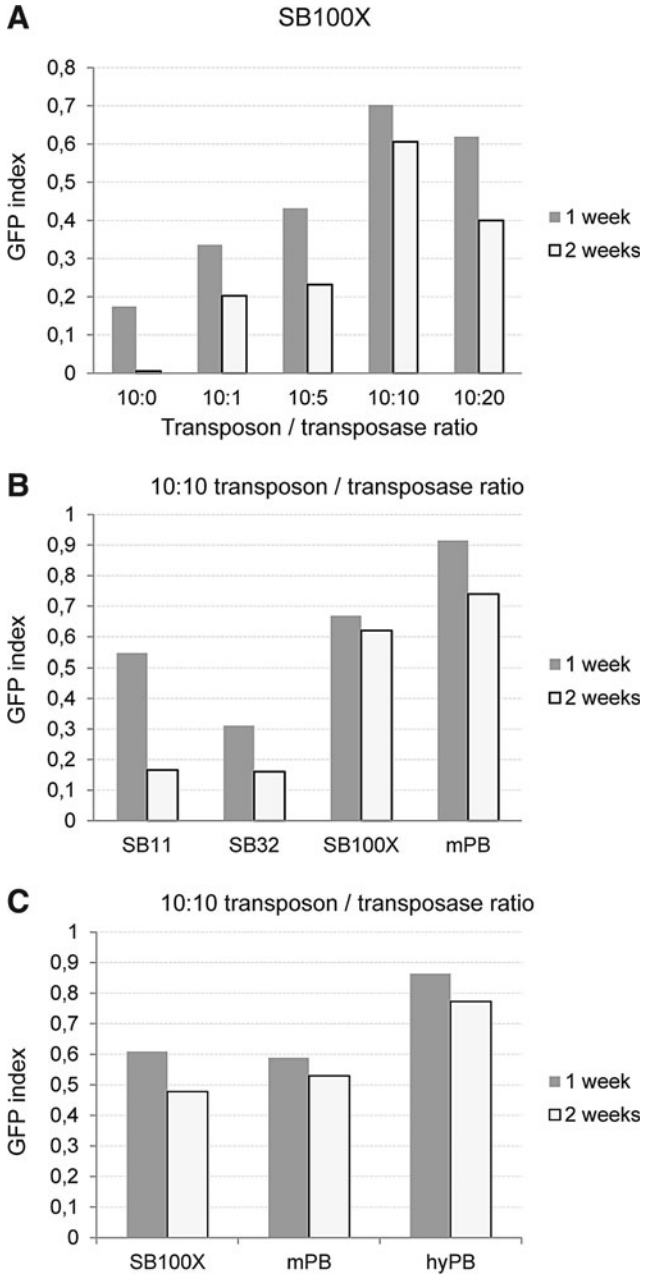


FIG. 5. GFP indexes: the ratio of transfected cells that retain transgene expression in the HUES9 cell line. **(A)** Dose dependency of the GFP index when using SB100X, measured at two time points after transfection. **(B)** GFP indexes at a fixed transposon-to-transposase ratio of different SB variants versus the mPB transposon system. **(C)** GFP indexes at a fixed transposon-to-transposase ratio of different PB variants versus the SB100X transposon system.

Transgenic copy numbers: SB generally exceeds PB

A crucial issue in any gene delivery experiment is to determine the stably integrated transgene copy numbers. In one of the former comparative experiments, in which SB variants and mPB were tested in HEK-293 cells at two different transposase doses (10:1 and 10:10; see Figs. 1E and 3D) we determined the average transposon copy numbers after 3 weeks of selection (Fig. 6A). In general, mPB transposition resulted in the lowest average copy numbers, lower than expected on the basis of excision and the transgenic rate comparisons (Figs. 1E and 3D), a finding that was in agreement with a previous study addressing copy number issues in HeLa cells (Grabundzija *et al.*, 2010). Similar results were obtained with the hyPB version: the copy numbers were comparable to that measured for the mPB version, again showing a clear indication that the high excision activity of a hyperactive transposase may not necessarily

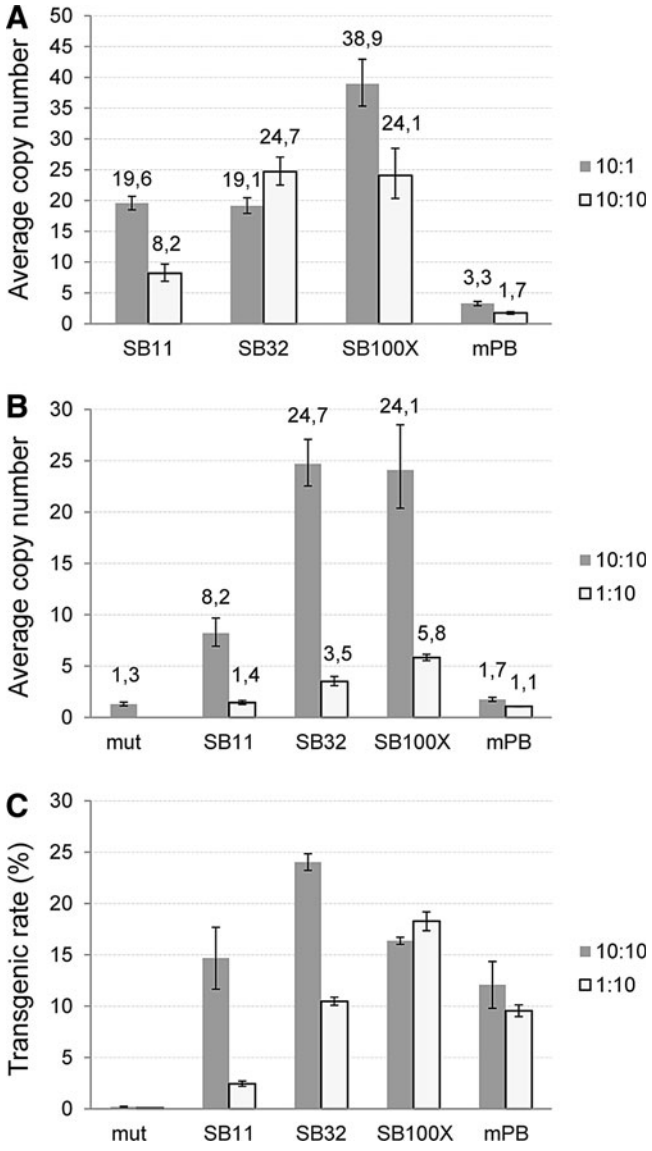


FIG. 6. Copy numbers of DNA transposons in HEK-293 cells. **(A)** Copy numbers resulting from different transposon systems at low versus high transposase dose. Limiting the amount of transposon also influenced **(B)** copy numbers and **(C)** transgenic rates when applying different DNA transposon systems. Mut, control transfection/transposition with a catalytically inactive SB transposase.

indicate higher transgenic rates and/or copy numbers. Using a high amount of transposons with low transposase concentrations (10:1 in Fig. 6A), the SB variants produced high copy numbers (≥10) after 3 weeks of selection, with SB100X resulting in extreme copy numbers of even ≥40 in some of our experiments. In this case, high copy numbers seemed to be associated with the high excision efficiency of SB100X. However, it again does not necessarily mean higher transgenic rates because SB32 could exceed SB100X in this aspect (Fig. 3D).

High transgene copy numbers are clearly unfavorable when considering gene therapy applications because, among other problems, it significantly increases the risk of insertional mutagenesis. It also places a heavy burden on any cellular genome, which is seen in experiments when transgenic cells were kept in culture for 6 weeks and the average copy number dropped significantly. Our experiments with HEK-293 cells indicate that average copy numbers exceeding 20 cannot be stably maintained in long-term culture. The most effective way to avoid such high copy numbers is to substantially lower the amount of transposon in the reaction. This can result in physiologically more acceptable copy numbers of 5–6 when using SB100X, without effectively lowering the transgenic rate (Fig. 6B and C). The same also works for the mPB transposase, lowering its copy number from ~3 to 1.

Considering hESCs, however, the copy number profiles of DNA transposons were found to be different from the other examined cell types. The mPB system produced the characteristically low transgene copy numbers (~2–3) that was observed in HEK-293 cells, whereas the SB100X transposase did not result in such extreme copy numbers as in other cell lines; both transposase systems resulted in an average copy number of ~2 in HUES9 cells. Just like for excision efficiency and transgenic rates, this suggests that PB can work efficiently in hESCs despite the fact that these cell types seem to be less permissive for transgenic manipulations.

Discussion

The rationale behind our study was to carefully examine the two most common and reportedly most efficient vector systems, the *Sleeping Beauty* and *piggyBac* transposon vehicles, in normal (HEK-293) and tumorous (HeLa) model cells, as well as in human embryonic stem cells. We also aimed to characterize aspects of the transposition reaction that have not been investigated in detail so far: we compared the excision efficiency of the transposases with their transgenic rates (the overall outcome of transposition), addressed the presence of overproduction inhibition at both levels, and also shed light on the connection between these properties and the average transgene copy numbers resulting from transposon-based gene delivery.

To examine the excision phase of transposition, we studied this reaction by a real-time PCR method using SYBR green technology. As opposed to a previous study in which TaqMan chemistry was used (Jin *et al.*, 2011), the potential occurrence of imprecise excision by the SB transposase (Liu *et al.*, 2004) prompted us to apply a method that was not biased by the sequence variability of the repaired donor locus after excision, so all possible excision events could be confidently detected. Our studies revealed that transposon excision is strongly cell type dependent, indicating that certain cell types represent a more permissive environment for the reaction. In our experiments, the highest excision rates were detected in HEK-293 cells, in which the level could be a magnitude higher than in hESCs. An important finding was that the selection for hyperactive transposases often results in higher excision efficiency as exemplified by the SB100X and hyPB variants. However, overproduction inhibition (Lohe and Hartl, 1996; Hartl *et al.*, 1997) may still affect the excision reaction in a cell type-dependent manner: hyPB was sensitive to the amount of transposase in hESCs, whereas the excision of SB100X could not be saturated under the conditions tested. The high excision efficiency of a hyperactive transposase, on the other hand, is not necessarily accompanied by

higher transgenic rates or higher copy numbers: in spite of the highest excision values, the transgenic rates measured in HEK-293 cells were lower than in HeLa cells; moreover, SB32 may achieve higher transgenic rates in certain cell types than the hyperactive SB100X. Cell type-specific differences may be due to different cellular defense mechanisms and/or interacting factors. The higher number of transgenic clones in HeLa cells may be related to its cancerous phenotype and its higher ''genomic buffer'' capacity due to the higher number of chromosomes. An important conclusion therefore is that testing the excision efficiency of transposases is a good prescreen for activity; however, it may not be representative of the overall transposition activity.

An important result of our study was that in human embryonic stem cells much lower (one magnitude lower) activity of the examined systems was detected, even at the excision level of transposition. On the other hand, at this lower level of activity, there did not seem to be a significant difference between SB100X and the two PB variants, and both mPB and hyPB seemed to outweigh SB100X in transgenic rates. This was also detected on investigating transgene copy numbers: whereas SB100X delivered copy numbers significantly lower as compared with values in other cultured cell lines, use of the PB system resulted in the ''expected'' number of transgene copies. Nevertheless, we cannot rule out that for the PB system, the more permissive environment in embryonic stem cells is due to specific cell factors; this needs further investigation.

In our experiments, another important result was that the nature of the promoter and the transgene sequence, as well as the nature of the selection method, might influence the phenomenon of overproduction inhibition of the transgenic rate. Changing only the promoter of the neomycin transgene, the transposase dose–response curve showed different, rather saturation-type kinetics in the same cell line. However, when the selection method was changed to puromycin (the resistance gene driven by the housekeeping PGK promoter) the curve changed significantly (Fig. 4D), with lower transgenic rates and showing a constant increase with enzyme activity. These results point to the importance of and need to optimize new transposon constructs, but also indicate that increasing transposition efficiency might be achieved by increasing the transposase concentration when it is not contra-indicated (e.g., minimizing the probability of transposase integration in gene therapy applications).

An important aspect of gene delivery is the transgene copy numbers achieved by a given method. Mutagenesis screens can favor higher copy numbers to maximize the hits, whereas gene therapy applications typically favor low copy numbers to avoid genotoxicity by high gene dose or the unwanted burden of insertional mutagenesis. Our results clearly demonstrate that high excision activity is not necessarily associated with increased copy numbers as hyperactive transposases behave differently: as opposed to hyPB, SB100X can achieve high copy numbers. It can be concluded that excision of DNA transposons might not always be followed by integration, which is more pronounced in the case of the hyPB version. It is possible that for the PB system, the excision and integration phases of transposition are kinetically more distinct than for other DNA transposons. This could explain why the PB transposase could be used to remove the previously integrated transgenes (Kaji *et al.*, 2009; Woltjen *et al.*, 2009; Yusa *et al.*, 2009) and why an integration mutant PB with merely excision activity could be generated (X. Li *et al.*, 2013). For the SB100X variant, the key factor to manipulate copy numbers is to change the amount of the transposon, rather than the transposase (Fig. 6A and B). It is worth noting that the transgenic rate will reflect the transposase *ab ovo* activity only if the transposon amount is rate limiting (Figs. 3D and 6C), and this is exactly the case with gene delivery into difficult-to-transfect cells; that is why hyperactive transposase versions are substantially more efficient in these applications. The sole disadvantage of low average copy number is that it carries the possibility that transgenic rates may decrease even further due to OI because inhibition will raise the number of cells that will not receive even a single copy of transgene. In the case of higher average copy number, inhibition will result only in a decrease in average copy number but not in transgenic cell number. An alternative solution to the problem could be achieved by a method published by Cai and colleagues (2014): they demonstrated that the hyPB transposase embedded in a lentiviral Gag polypeptide could be used efficiently to generate single-copy clones while still keeping a high average transgenic rate.

In this study, we demonstrated strong cell type differences between SB and PB DNA transposons both at the excision level and in transgenic rate of transposition. We could provide additional data that the overproduction inhibition is not such a widespread phenomenon as previously anticipated and that careful experimental design could maximize the efficiency of gene delivery. An important finding of our study is that hyperactive DNA transposase versions seem to be selected for high excision efficiency, which is not necessarily coupled with either higher transgenic rate or increased transgene copy number. Nevertheless, the hyperactive transposases are definitely the method of choice both for insertional mutagenesis and for general gene delivery purposes. Our results also indicate that in terms of efficiency, the PB system seemed to outweigh SB100X in human embryonic stem cells, although its reported nonrandom integration profile (Wilson *et al.*, 2007; Wang *et al.*, 2008; Galvan *et al.*, 2009; Liang *et al.*, 2009) should be considered when such experiments are designed for gene therapy purposes.

Acknowledgments

Research in our laboratory was supported by grants from TransRat (KMR_12-1-2012-0112) and OTKA (NK83533).

Author Disclosure Statement

No competing financial interests exist.

References

Aluigi, M., Fogli, M., Curti, A., *et al.* (2006). Nucleofection is an efficient nonviral transfection technique for human bone marrow-derived mesenchymal stem cells. Stem Cells 24, 454–461.

Apáti, A., Orbán, T.I., Varga, N., *et al.* (2008). High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells. Biochim. Biophys. Acta 1778, 2700–2709.

Balciunas, D., Wangensteen, K.J., Wilber, A., *et al.* (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. PLoS Genet. 2, e169.

Baus, J., Liu, L., Heggestad, A.D., *et al.* (2005). Hyperactive transposase mutants of the *Sleeping Beauty* transposon. Mol. Ther. 12, 1148–1156.

Burnight, E.R., Staber, J.M., Korsakov, P., *et al.* (2012). A hyperactive transposase promotes persistent gene transfer of a *piggyBac* DNA transposon. Mol. Ther. Nucleic Acids 1, e50.

Cadinanos, J., and Bradley, A. (2007). Generation of an inducible and optimized *piggyBac* transposon system. Nucleic Acids Res. 35, e87.

Cai, Y., Bak, R.O., Krogh, L.B., *et al.* (2014). DNA transposition by protein transduction of the *piggyBac* transposase from lentiviral Gag precursors. Nucleic Acids Res. 42, e28.

Claeys Bouuaert, C., Lipkow, K., Andrews, S.S., *et al.* (2013). The autoregulation of a eukaryotic DNA transposon. Elife 2, e00668.

Ding, S., Wu, X., Li, G., *et al.* (2005). Efficient transposition of the *piggyBac* (*PB*) transposon in mammalian cells and mice. Cell 122, 473–483.

Doherty, J.E., Huye, L.E., Yusa, K., *et al.* (2012). Hyperactive *piggyBac* gene transfer in human cells and *in vivo*. Hum. Gene Ther. 23, 311–320.

Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368.

Galvan, D.L., Nakazawa, Y., Kaja, A., *et al.* (2009). Genome-wide mapping of *PiggyBac* transposon integrations in primary human T cells. J. Immunother. 32, 837–844.

Goodier, J.L., and Kazazian, H.H., Jr. (2008). Retrotransposons revisited: The restraint and rehabilitation of parasites. Cell 135, 23–35.

Grabundzija, I., Irgang, M., Mátés, L., *et al.* (2010). Comparative analysis of transposable element vector systems in human cells. Mol. Ther. 18, 1200–1209.

Hartl, D.L., Lozovskaya, E.R., Nurminsky, D.I., *et al.* (1997). What restricts the activity of *mariner*-like transposable elements. Trends Genet. 13, 197–201.

Huang, X., Guo, H., Tammana, S., *et al.* (2010). Gene transfer efficiency and genome-wide integration profiling of *Sleeping Beauty*, *Tol2*, and *piggyBac* transposons in human primary T cells. Mol. Ther. 18, 1803–1813.

Ivics, Z., and Izsvák, Z. (2004). Transposable elements for transgenesis and insertional mutagenesis in vertebrates: A contemporary review of experimental strategies. Methods Mol. Biol. 260, 255–276.

Ivics, Z., and Izsvák, Z. (2006). Transposons for gene therapy! Curr. Gene Ther. 6, 593–607.

Ivics, Z., Hackett, P.B., Plasterk, R.H., *et al.* (1997). Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon from fish, and its transposition in human cells. Cell 91, 501–510.

Izsvák, Z., Stuwe, E.E., Fiedler, D., *et al.* (2004). Healing the wounds inflicted by *Sleeping Beauty* transposition by double-strand break repair in mammalian somatic cells. Mol. Cell 13, 279–290.

Izsvák, Z., Hackett, P.B., Cooper, L.J., *et al.* (2010). Translating *Sleeping Beauty* transposition into cellular therapies: Victories and challenges. Bioessays 32, 756–767.

Jin, Z., Maiti, S., Huls, H., *et al.* (2011). The hyperactive *Sleeping Beauty* transposase SB100X improves the genetic modification of T cells to express a chimeric antigen receptor. Gene Ther. 18, 849–856.

Kaji, K., Norrby, K., Paca, A., *et al.* (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. Nature 458, 771–775.

Kazazian, H.H., Jr. (2004). Mobile elements: Drivers of genome evolution. Science 303, 1626–1632.

Kojima, K.K., and Jurka, J. (2013). A superfamily of DNA transposons targeting multicopy small RNA genes. PloS One 8, e68260.

Kolacsek, O., Krízsik, V., Schamberger, A., *et al.* (2011). Reliable transgene-independent method for determining *Sleeping Beauty* transposon copy numbers. Mob. DNA 2, 5. [Published erratum appears in Mob. DNA 2013;4:11.]

Lampe, D.J., Grant, T.E., and Robertson, H.M. (1998). Factors affecting transposition of the *Himar1 mariner* transposon *in vitro*. Genetics 149, 179–187.

Li, M.A., Pettitt, S.J., Eckert, S., *et al.* (2013). The *piggyBac* transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. Mol. Cell. Biol. 33, 1317–1330.

Li, X., Burnight, E.R., Cooney, A.L., *et al.* (2013). *piggyBac* transposase tools for genome engineering. Proc. Natl. Acad. Sci. U.S.A. 110, E2279–E2287.

Liang, Q., Kong, J., Stalker, J., *et al.* (2009). Chromosomal mobilization and reintegration of *Sleeping Beauty* and *PiggyBac* transposons. Genesis 47, 404–408.

Liu, G., Aronovich, E.L., Cui, Z., *et al.* (2004). Excision of *Sleeping Beauty* transposons: Parameters and applications to gene therapy. J. Gene Med. 6, 574–583.

Lohe, A.R., and Hartl, D.L. (1996). Autoregulation of *mariner* transposase activity by overproduction and dominant-negative complementation. Mol. Biol. Evol. 13, 549–555.

Mátés, L., Chuah, M.K., Belay, E., *et al.* (2009). Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. Nat. Genet. 41, 753–761.

Meir, Y.J., Weirauch, M.T., Yang, H.S., *et al.* (2011). Genome-wide target profiling of *piggyBac* and *Tol2* in HEK 293: Pros and cons for gene discovery and gene therapy. BMC Biotechnol. 11, 28.

Miskey, C., Izsvák, Z., Plasterk, R.H., *et al.* (2003). The *Frog Prince*: A reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. Nucleic Acids Res. 31, 6873–6881.

Pledger, D.W., and Coates, C.J. (2005). Mutant *Mos1 mariner* transposons are hyperactive in *Aedes aegypti*. Insect Biochem. Mol. Biol. 35, 1199–1207.

Rostovskaya, M., Fu, J., Obst, M., *et al.* (2012). Transposon-mediated BAC transgenesis in human ES cells. Nucleic Acids Res. 40, e150.

Sarkadi, B., Orbán, T.I., Szakacs, G., *et al.* (2010). Evaluation of ABCG2 expression in human embryonic stem cells: Crossing the same river twice? Stem Cells 28, 174–176.

Sharma, N., Hollensen, A.K., Bak, R.O., *et al.* (2012). The impact of cHS4 insulators on DNA transposon vector mobilization and silencing in retinal pigment epithelium cells. PLoS One 7, e48421.

Shen, S., Lin, L., Cai, J.J., *et al.* (2011). Widespread establishment and regulatory impact of Alu exons in human genes. Proc. Natl. Acad. Sci. U.S.A. 108, 2837–2842.

Singer, T., McConnell, M.J., Marchetto, M.C., *et al.* (2010). LINE-1 retrotransposons: Mediators of somatic variation in neuronal genomes? Trends Neurosci. 33, 345–354.

Walisko, O., Izsvák, Z., Szabó, K., *et al.* (2006). *Sleeping Beauty* transposase modulates cell-cycle progression through interaction with Miz-1. Proc. Natl. Acad. Sci. U.S.A. 103, 4062–4067.

Wang, W., Lin, C., Lu, D., *et al.* (2008). Chromosomal transposition of *PiggyBac* in mouse embryonic stem cells. Proc. Natl. Acad. Sci. U.S.A. 105, 9290–9295.

Wang, Y., Wang, J., Devaraj, A., *et al.* (2014). Suicidal autointegration of *Sleeping Beauty* and *piggyBac* transposons in eukaryotic cells. PLoS Genet. 10, e1004103.

Williams, D.A. (2008). *Sleeping Beauty* vector system moves toward human trials in the United States. Mol. Ther. 16, 1515–1516.

Wilson, M.H., Kaminski, J.M., and George, A.L., Jr. (2005). Functional zinc finger/*Sleeping Beauty* transposase chimeras exhibit attenuated overproduction inhibition. FEBS Lett. 579, 6205–6209.

Wilson, M.H., Coates, C.J., and George, A.L., Jr. (2007). *PiggyBac* transposon-mediated gene transfer in human cells. Mol. Ther. 15, 139–145.

Woltjen, K., Michael, I.P., Mohseni, P., *et al.* (2009). *piggyBac* transposition reprograms fibroblasts to induced pluripotent stem cells. Nature 458, 766–770.

Yusa, K., Rad, R., Takeda, J., *et al.* (2009). Generation of transgene-free induced pluripotent mouse stem cells by the *piggyBac* transposon. Nat. Methods 6, 363–369.

Yusa, K., Zhou, L., Li, M.A., *et al.* (2011). A hyperactive *piggyBac* transposase for mammalian applications. Proc. Natl. Acad. Sci. U.S.A. 108, 1531–1536.

Zayed, H., Izsvák, Z., Khare, D., *et al.* (2003). The DNA-bending protein HMGB1 is a cellular cofactor of *Sleeping Beauty* transposition. Nucleic Acids Res. 31, 2313–2322.

Zayed, H., Izsvák, Z., Walisko, O., *et al.* (2004). Development of hyperactive *Sleeping Beauty* transposon vectors by mutational analysis. Mol. Ther. 9, 292–304.

Address correspondence to:
Dr. Tamás I. Orbán
Institute of Enzymology
Research Centre for Natural Sciences
Hungarian Academy of Sciences
Budapest, 1117, Magyar Tudósok körútja 2
Hungary

E-mail: orbant@biomembrane.hu; orban.tamas@ttk.mta.hu

Received for publication July 30, 2013;
accepted after revision July 10, 2014.

Published online: July 21, 2014.

Quantitative Analysis of DNA Transposon-mediated Gene Delivery: the *Sleeping Beauty* System as an Example

Orsolya Kolacsek
Institute of Molecular Pharmacology, RCNS
Hungarian Academy of Sciences, Hungary

Zsuzsanna Izsvák
Mobile DNA Group
Max-Delbrück Center for Molecular Medicine, Germany

Zoltán Ivics
Division of Medical Biotechnology
Paul Ehrlich Institute, Germany

Balázs Sarkadi
Membrane Research Group of HAS
Semmelweis University and National Blood Center, Hungary

Tamás I. Orbán
Institute of Molecular Pharmacology, RCNS
Hungarian Academy of Sciences, Hungary

1 Introduction – DNA Transposons as Genetic Tools

Transposons are special genetic elements that are capable of moving from one DNA locus to another. They were discovered in maize by Barbara McClintock in the 1940s, and these ground-breaking experiments later earned her the Nobel Prize in 1983. Transposons can be classified based on their replication intermediate (Class I retrotransposons and Class II DNA transposons), on their replication manner (replicative versus non-replicative transposons), or on their ability to disperse independently (autonomous versus non-autonomous elements) (Burns and Boeke, 2012). Once considered to be selfish units of replication, they are now believed to also represent inevitable driving forces for evolution, proven by their presence in all genomes examined so far (Kazazian, 2004; Hedges & Batzer, 2005; Feschotte & Pritham, 2007). The human genome is not an exception, as approximately 45% of our genetic material is made up of transposons (Biemont & Vieira, 2006; Mills *et al.*, 2006; Wicker *et al.*, 2007; Goodier & Kazazian, 2008; Solyom & Kazazian, 2012). The majority of the human transposons belong to the Class I retrotransposons (or RNA transposons) which move around by the replicative "copy and paste" mechanism. These contain currently active mobile elements which by all means played a significant role in human evolution (Mills *et al.*, 2007; Shen *et al.*, 2011). Moreover, LINE-1 retrotransposons were recently proven to be responsible for certain types of somatic mozaicisms present in vertebrate neurons (Singer *et al.*, 2010). Class II DNA transposons, however, make up a significantly smaller proportion (~3%) of the human genome. This moderate fraction might be attributable to the fact that the majority of them spread by the non-replicative "cut and paste" mechanism. In addition, none of the DNA transposons have been shown to be active in the human genome (Feschotte & Pritham, 2007; Collier & Largaespada, 2007; Izsvak *et al.*, 2010; Solyom & Kazazian, 2012).

Transposons as genetic tools have been widely used in invertebrate model organisms (*Drosophila* species or *Caenorhabditis elegans*), mainly exploiting active DNA transposons of the particular species. The most prominent examples are the P-elements which were used as proof of principle for the two component transposon gene delivery system in *Drosophila melanogaster* (Rubin & Spradling, 1982; Spradling & Rubin, 1982). For vertebrates, however, applications were limited to retrotransposons for a long time, with the obvious disadvantages of higher mutational rate (due to the reverse transcription process) and the long term genetic instability of the modified cells because of the potential remobilization of the integrated transgene (Uren *et al.*, 2005; Ostertag *et al.*, 2007). A significant breakthrough in vertebrate genetics was the creation of an artificial Tc1/Mariner-type transposon, the *Sleeping Beauty* (SB) system, which was the first DNA transposon proven to be active in vertebrates, including human cells (Ivics *et al.*, 1997). Its simple structure (Figure 1) made it easy to modify and to establish a controllable system by separating the transposase from its targets (the originally flanking terminal repeat sequences), thereby allowing the controlled delivery of any gene of interest into the genome (Izsvak *et al.*, 2000). However, the efficiency of the originally resurrected SB variant was still significantly lower than the widely used viral vectors so its potential seemed to be behind those other genetic vehicles, especially in human applications.

The success of "awakening" a new active DNA transposon initiated a wave of research aiming at establishing efficient novel transposon systems applicable in human cells. Apart from other reconstructed species (such as the *Frog Prince* from *Rana pipiens*, Miskey *et al.,* 2003), an active DNA transposon (*Tol2*) was discovered in medaka fish and successfully applied in various vertebrate species (Balciunas *et al.*, 2006). Moreover, a transposon from another insect species (*Trichoplusia ni*) called *piggyBac* (PB) was shown to be highly active in human cells (Ding *et al.*, 2005). In the meantime, hyperactive versions

of the previously used DNA transposons were also established (Zayed *et al.*, 2004; Baus *et al.*, 2005; Pledger & Coates, 2005), opening the possibility of efficient non-viral gene delivery applications. The most promising of all was the 100 times more active form of SB (SB100x), providing a highly effective alternative to the existing viral gene delivery methods (Mates *et al.*, 2009).

In a recent study, the most hyperactive versions of three transposon systems (SB, PB and *Tol2*) were systematically tested and compared in terms of delivery efficiency, copy number and integration profile of the transgene (Grabundzija *et al.*, 2010). It was revealed that SB and PB are the most efficient gene delivery vehicles and, although transposition efficiency is known to decrease with the cargo size (Izsvak *et al.*, 2010), they are able to carry and integrate inserts of ≥10 kb, outweighing the packaging capacity of the most efficient viral vectors. It was also shown that in conditions where the amount of DNA transposon is limiting (modeling most gene therapy applications), SB is superior even to PB. Moreover, the integration profile of SB seems to be the most favorable one: in fact, it is the closest to random on the genomic level among all tested viral- and transposon-based systems so far (Vigdal *et al.*, 2002; Liu *et al.*, 2005; Yant *et al.*, 2005; Grabundzija *et al.*, 2010), providing the lowest risk for insertional mutagenesis and making the SB system particularly suitable for gene therapy applications. In addition, as opposed to PB, no potential endogenous elements resembling SB are present in the human genome (Ivics *et al.*, 1997; Ivics *et al.*, 2004) which is an important safety issue that further supported the initiation of a clinical trial experiment using this transposon system (Williams, 2008).

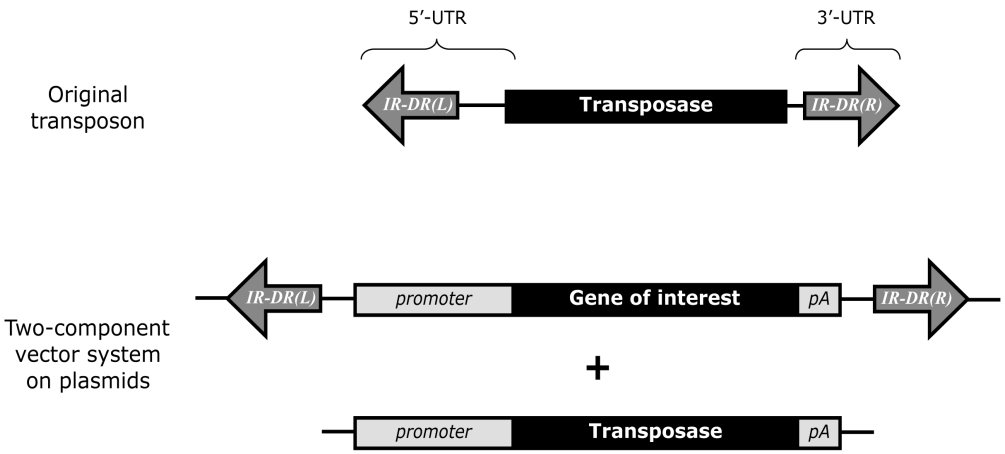


Figure 1: The structure of the *Sleeping Beauty* transposon system. In the natural transposon, the transposase gene is flanked by untranslated regions (UTRs) that include the terminal Inverted Repeat Direct Repeat regions (IRDR-L and IRDR-R), containing binding sites for the transposase. When used as a gene delivery vector system, the transposase coding region is replaced by a gene of interest within the transposable element that is maintained in a plasmid. This non-autonomous transposon can be mobilized if the transposase is supplied in trans by expression from a separate plasmid vector containing a suitable promoter. Co-transfection of the two components into candidate cells provides the platform for transposition from the donor plasmid to the cellular genome.

Considering all aspects, however, other efficient transposon systems with integration preferences into transcriptionally active regions (such as the PB system, Wilson *et al.*, 2007) might be more suitable

when performing "traditional" forward genetic screens (Collier & Largaespada, 2005; Chew *et al.*, 2011; Guo *et al.*, 2011). Also, as the PB transposase is able to "tracelessly" remove the integrated transgene from the genome, it might be a method of choice when such a feature is desirable, exemplified by the removal of the reprogramming cassette after generating induced pluripotent stem cells from fibroblasts (Kaji *et al.*, 2009; Woltjen *et al.*, 2009). Nevertheless, there are still various aspects of the transposon-based technology that should be rigorously tested. Such issues include the potential silencing of the transgene which is often the drawback of viral vectors (Ellis, 2005), especially in the case of embryonic stem cells which are particularly prone to silence viral promoters (Meilinger *et al.*, 2009; Rowe *et al.*, 2010). So far, the already applied DNA transposon sequences did not seem to face this problem as in the case of SB the effect of silencing was shown to depend on the cargo sequence, and not on the transposon vector (Garrison *et al.*, 2007; Zhu *et al.*, 2010).

According to gene therapy guidelines or mutagenesis protocols, one of the most important issues is the exact and fast determination of transgene copy numbers (Bian & Belmont, 2010; Sivalingam *et al.*, 2010; Huang *et al.*, 2010). Various methods are available to perform this, including "traditional" blotting techniques (Southern blotting/dot blotting), or several polymerase chain reaction (PCR)-based techniques (Wicks *et al.*, 2000; Devon *et al.*, 1995). These usually involve the application of radioactively or fluorescently labeled probes, or – depending on the nature of the transgene – utilize the inherent signal originating from the transgene itself (such as quantifying GFP fluorescence, Moeller *et al.*, 2003). Apart from often requiring hazardous chemicals or being laborious, the general problem of these methods is that they are usually set up for a specific transgene, and changing the gene of interest will require optimizing the applied parameters of the method once again. Using the SB system in our laboratory, we aimed to develop an accurate and fast method to quantify transposon copy numbers that is applicable to any SB-based gene delivery experiments without *a priori* optimization of the protocol. We worked out a real-time PCR technique which is independent of the transgene sequence, hence we named it a "transgene independent" quantitative PCR technology (Kolacsek *et al.*, 2011). Apart from being sensitive, accurate and fast, this approach also offers a powerful non-radioactive technique as an alternative against other canonical methodologies.

In this chapter, using the SB transposon system as a prominent example, we address quantitative issues regarding the transposon-based gene delivery methods. We focus in detail on a transgene-independent qPCR method recently developed in our laboratory, providing more information in depth on the theoretical background and the technical aspects of this methodology published earlier (Kolacsek *et al.*, 2011). Additionally, we also cover a technique of measuring transposase activity at the excision phase of the reaction which allows comparative analysis of different transposase variants, different transposon systems or different conditions of application. We believe that these quantitative aspects of transposase activity are of great importance especially in light of the applicability of the DNA transposons for gene therapy purposes.

2 Selecting and Separating Transgenic Clones after Transposition

2.1 Selection Methodology

The basis of generating stable transgenic clones is to apply the most efficient but the least harmful gene delivery into the chosen cell types. The SB transposon is the method of choice in this respect as it offers a stable transgene integration technology with the least mutagenic potential among all available gene de-

livery techniques (see Section 1 above). Nevertheless, the bottleneck of this technology is that it involves the transfection of DNA into the host cells which might have low efficiency rates for certain cell types, such as embryonic stem cells. It is therefore necessary to apply an efficient and preferably non-invasive selection protocol to establish homogenous transgene expressing cells following transfection and transposition. Selection methods may sometimes be carried out utilizing the expression of the transgene but very often it is inescapable to use an additional marker gene even at the expense of having a larger genetic cargo, thereby lower delivery efficiency.

In applications where the cell source is not limiting, various chemical selection methods such as antibiotic selection can be applied to enrich for transgene expressing cells (Figure 2). However, depending on the cell type, this method might significantly disturb cell physiology, therefore other approaches are necessary to be applied. For example, the advantage of using fluorescent markers is that the transfected cells can be separated by Fluorescent Activated Cell Sorting (FACS) analysis (Figure 3), although some cell types may not tolerate such physical stress and this method could also decrease cell viability. The aim is to optimize the transfection/selection procedure for the particular cell type reaching the highest possible gene delivery efficiency with the lowest possible cell mortality rate.

2.2 Separating uniform clones

Some transgenic applications (e.g. transgenic animals) require genetically uniform cell populations, which can be precisely characterized from various aspects, including copy number and integration sites. These features can contribute to the transgenic phenotype to a large extent. To develop a “reliable” designated method for copy number determination, our goal was to detect the lowest (1) stable copy per cell, and to clearly differentiate cells differing in copy numbers by one (e.g. cells carrying 2 copies from those with 1 or 3 copies). In other words, with this method we wanted to detect the copy number as corpuscular units in the cells. This was an important reason why we aimed at working with genetically uniform clones that carry the same transgenic cassette in different copy numbers.

The simplest way for cloning is the threshold limit dilution of heterogeneous transgenic cell population, previously selected by the transgene or the marker gene expression. In this method, a serial dilution of the cells is spread in a 96-well cell culture plate and those wells are considered to represent one clone where only one colony can be seen by microscopy; those cell clones are then further maintained and utilized. However, this method cannot be applied to all cell types, such as human embryonic stem cells which naturally grow only in clumps. In such cases, other manual methods may be applied, including FACS selection or using cloning rings.

2.3 Isolation of a Single Copy Insertion Serving as a Calibration Unit

The major difficulty to start a particular copy number measurement project is the lack of reference samples with known copy numbers, also known as calibrators. If such samples determined by other techniques are not available, the first step is to isolate clones carrying 1 copy of the transgene.

Applying the hyperactive SB transposases often results in high copy numbers and working with such high efficiency transposon delivery may provide only small number of cells carrying 1 copy of the transgene. An obvious 1 copy clone source is the random integration of plasmids that most of the cases results in 1 transgenic copy. These integrations will contain the majority or the whole plasmid sequence, due to a random breakage of the transfected transposon vector. This can be achieved by transfecting the transposon donor plasmid either with the inactive mutant transposase variant or without the transposase expressing helper plasmid. Random integration is very ineffective, occurs usually in less than 1% of the

transfected cells, but they can be selected out and can be cloned as well. This transfection serves as transposition control (Figure 2) resulting in traces of random integrations contrary to the active transposition, which has much higher integration efficiency, and it could provide us an excellent source of one copy clones.

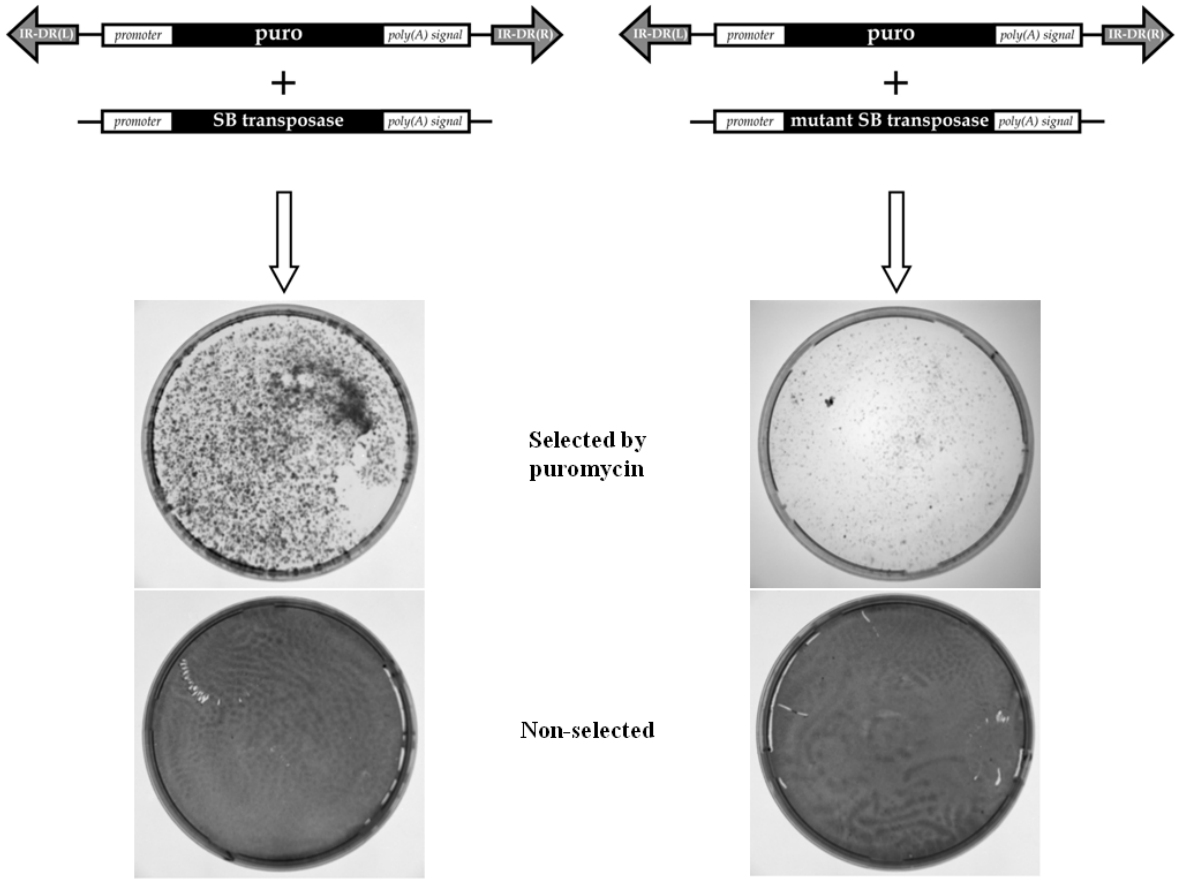


Figure 2: A typical experiments using SB tranposons: establishing transgenic HEK-293 cells expressing a puromycin resistance gene (puro). 2 days posttransfection, cells were passed into puromycin containing medium and selected for 10 days; living cells were visualized by Giemsa-staining following selection. The efficiency of transposition is obvious when comparing the selected cells after co-transfection with the transposase expressing helper plasmid (left) to the control experiment with the mutant transposase (right), the latter one indicating random integration events. As a negative control, non-selected cells are also shown in both experimental setups.

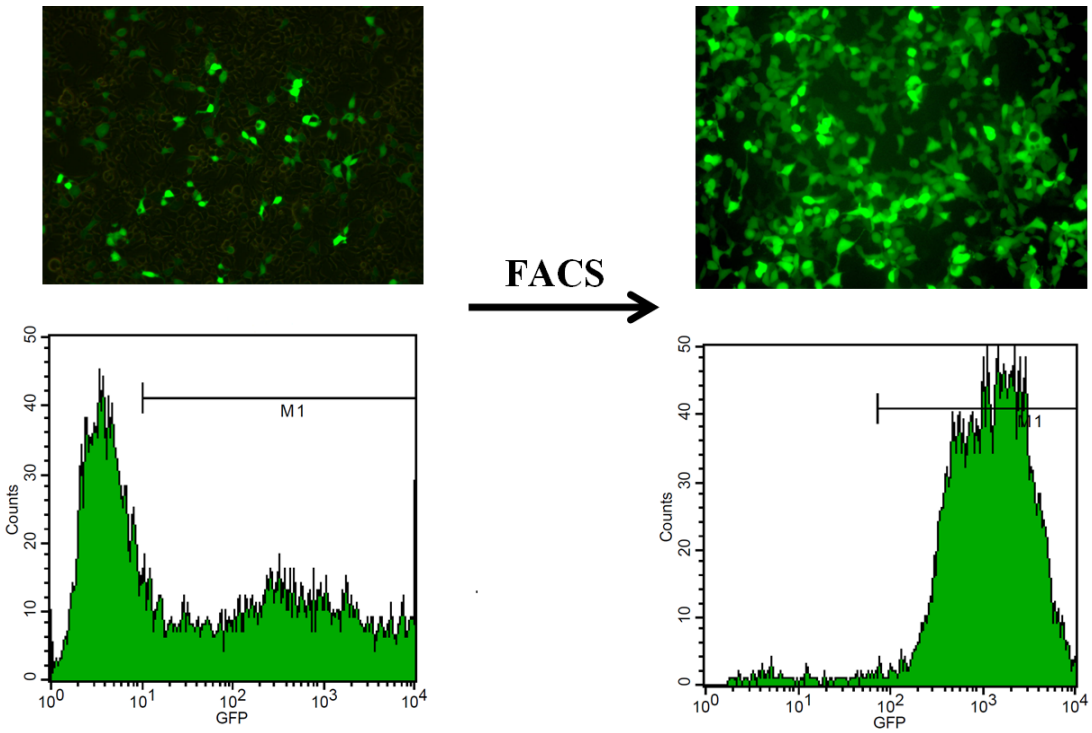


Figure 3: Transgene selection after transposition using GFP expression. The expression cassette is delivered into HEK-293 cells by the SB transposon system resulting in a heterogeneous cell population (left panels). Transgenic cells were selected by FACS experiment (right panels). Fluorescence microscopy images of x200 magnification and FACS histograms of GFP intensities can be seen. Cell numbers (Counts) are shown as a function of fluorescence intensity plotted in a logarithmic scale. M1: marker indicating GFP expressing cells.

3 Real-time PCR for Sequence Quantification

The polymerase chain reaction (PCR) is a technique for the in vitro amplification of specific DNA sequences by the simultaneous primer extension of complementary strands of DNA (Kleppe *et al.*, 1971; Mullis & Falona, 1987). It was a major development in molecular biology because it has simplified existing technologies and enabled a rapid development of new techniques which otherwise would not have been possible. PCR theoretically amplifies DNA exponentially, doubling the number of double stranded sequences present in each amplification cycle. After the logarithmic (log) phase of the reaction, the amount of the PCR product reaches the plateau phase (Figure 4). The amount of the product is proportional to the starting sequence copy number during the log phase, providing the basis for reliable quantitative comparisons.

In traditional (endpoint) PCR, detection and quantification of the amplified sequence are performed at the end of the reaction after the last PCR cycle, and involve post-PCR analysis such as gel electrophoresis, signal detection or image analysis. However, this allows only semi-quantitative analysis

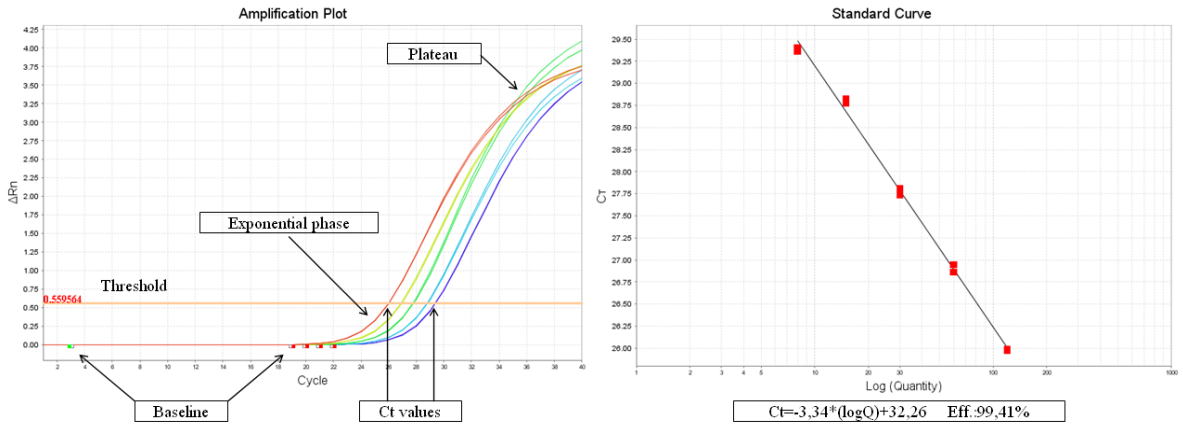


Figure 4: Example of a real-time PCR experiment determining the efficiency of the GFP TaqMan® assay using a standard curve of dilution points. Twofold dilutions were prepared from gDNA of pooled transgenic clones with the same copy number. Measurements were performed on a StepOnePlus™ Real-Time PCR platform (Applied Biosystems, Foster City, CA). Data were analyzed with the StepOne™ Software v2.1. Amplification plot (left panel) shows the increase of fluorescence signal as a function of cycle numbers (baseline fluorescence is subtracted, hence ΔRn; different colors represent different dilution reactions). Typical parameters of the reaction are indicated. The right panel shows the standard curve derived from the experiment, showing Ct values as a function of gDNA input (expressed in logarithm of DNA quantity in nanograms). The line of best fit is calculated by linear regression using the standard points; the equation is also shown from which efficiency of the reaction is calculated.

carried out in samples collected at multiple points throughout the amplification process, thus ensuring the analysis before the plateau is reached. This approach is usually combined with analysis of dilution series of the samples, it also requires known standards and provides a detection range of usually only tenfold difference (Chelly *et al.*, 1988; Wang *et al.*, 1989). In real-time quantitative PCR (qPCR), the amount of PCR product is measured at each cycle by the use of fluorescent dyes (Higuchi *et al.*, 1992; Livak *et al.*, 1995). This ability to monitor the reaction during its exponential phase enables the user to determine the initial amount of target with great precision. Apart from being simple and fast, the powerful benefit of qPCR is the increased dynamic range of comparisons.

The most popular fluorescent detection technologies are double-stranded DNA (dsDNA) binding agents, e.g. SYBR® Green, and fluorescent probes (Livak *et al.*, 1995; Wittwer *et al.*, 1997; Morrison *et al.*, 1998). SYBR® Green signal is measured at the end of each extension step and the intensity depends on the amount of dsDNA that is present. This technology is simple because the dye can be added to any kind of sequence amplification, but lacks specificity because it will also bind to PCR artifacts, e.g. primer-dimers. Good primer design and quality of starting materials are critical to avoid nonspecific products. Specificity of the reaction could be assessed using a melting curve measured at the end of the reaction (Figure 5). Melting curve determines the melting point (Tm) characteristic to the specific PCR product which assures differentiation of valid qPCR reactions from PCR artifacts (Ririe *et al.*, 1997). Those melting curves showing multiple peaks or one peak with rather different Tm than of the main product result from nonspecific PCR products indicating invalid reactions with false Ct values.

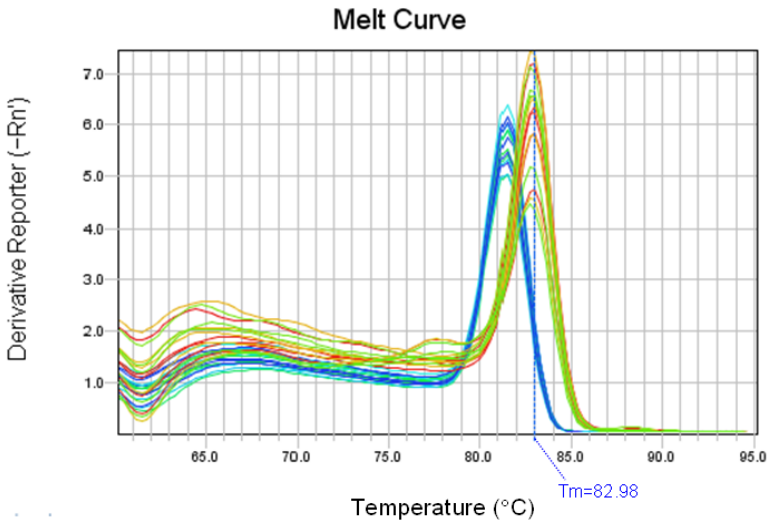


Figure 5: Representative melting curve analysis of a SYBR® Green qPCR experiment. Derivative of the fluorescence signal of the reporter (Rn') is shown as a function of the temperature. Two sequences (a target and an endogenous control) were amplified in various wells of this plate, therefore two melting peaks are visible; the melting temperature of the target is indicated. Different colors represent different reaction wells of the plate.

Fluorescent oligonucleotide probes (e.g. TaqMan® probes) are designed to hybridize to the sequence amplified by the primers. At the annealing step of each cycle, the probe will bind to the target sequence, and will be subsequently cleaved by the 5' nuclease activity of the polymerase during the extension phase. A dual labeled probe with a reporter dye at the 5' end and a quencher dye at the 3' end of the oligonucleotide will generate a fluorescent signal when the probe is degraded which is detected at the end of the extension phase (Livak *et al.*, 1995; Heid *et al.*, 1996). Probe-based systems provide highly specific detection of DNA, however, dual-labeling and complex design make them more expensive.

In qPCR, the cycle number in which the signal appears at the beginning of the log phase is considered to be inversely correlated to the amount of starting template, as a higher amount will result in sooner amplification. Threshold is the level of signal that reflects a statistically significant increase over the baseline fluorescent signal (Figure 4). In most cases, the real-time PCR software automatically sets the threshold at least 10 times the standard deviation of the fluorescence value of the baseline. However, the positioning of the threshold can be set manually at any point in the exponential phase of PCR. Threshold cycle (Ct) is the cycle number determined by the software at which the amplification plot crosses the threshold. Passive reference dyes (usually added to the qPCR master mixes) are frequently used in qPCR to normalize the fluorescent signal of reporter dyes (Rn). This allows the correction of fluctuations in fluorescence that is non-PCR-based, e.g. changes from well to well in reagent concentration or volume, or in instrument scanning (Figure 4).

Validation of the qPCR assay is generally carried out by the analysis of the slopes from standard curves. A standard curve is generated by plotting the results of a dilution series of the template against the Ct for each dilution (Figure 4). In theory, if the reaction is 100% efficient, the PCR duplicates the template in each cycle, and in the log scale of template amount, the slope will be –3.32 (1/[lg(x)-lg(2x)] =

–3.32). The reaction efficiency is related to assay sensitivity, which can be calculated from the slope (Real Time PCR Handbook, http://tools.invitrogen.com/content.cfm?pageid=12257):

$$\text{Efficiency} = 10^{(-1/\text{slope})} - 1.$$

The template used to generate the standard curve should match – as closely as possible – that is being used for the experiment (e.g. the same total RNA or DNA sample). The dilution range or dynamic range should span the concentration range expected for the unknown samples. The simplest way to ensure this is pooling the unknown samples (such as gDNAs from transgenic clones), and using it as a standard. The acceptable range of the efficiency, which most scientists agree on is between 90% – 110%. If efficiency is higher than 100%, it can reflect an inhibitory effect. In this case, scaling down the starting material usually helps by lowering the concentration of the suspected inhibitor. The desirable window of 90 to 110% defines the range of input template quantities that may be measured in a particular qPCR.

In the following three sections, we describe three different approaches which could be applied to determine transgene copy numbers of transgenic clones using a real-time PCR-based strategy.

4 Determining Transgene Copy Numbers of Transgenic Clones

4.1 First Approach: Verifying Presumably One Copy Clones with Absolute Quantification using the Marker Sequence

We generated SB transgenic clones carrying a GFP expressing cassette separating them by FACS analysis. Few clones were prepared also from random integration experiments, being obvious sources of 1 copy integrations as mentioned formerly. Since random integrations can be derived from breakage of the plasmid at any point, the presence of fluorescence in these clones assures the presence of the GFP transcription unit, so the GFP sequence can be reliably used for the copy number analysis. Therefore, specific TaqMan® assays were designed for the two terminal Inverted Repeat Direct Repeat (IRDR) motifs of the SB transposon (left and right, IRDR-L and IRDR-R), as well as for the GFP sequence. Sequences of primers and probes can be found in our previous publication (Kolacsek *et al*., 2011).

In the absence of a reference clone with known copy number, we have to compare the absolute Ct values to known plasmid dilutions containing the transposon sequence. The recommended amount of gDNA input is in the range of 10 to 40 ng (we used 30 ng) but other input size of the starting material can also be accepted if in previous pilot experiments, the efficiency at that point was shown to be in the desired range. For calculations of the required plasmid dilutions, first we need to know how many genome copies are present in the 30ng gDNA input. The average molecular weight of a DNA base pair is 618g/mol, so using the Avogadro's number of 6.02×10^{23} entities/mol, the molecular weight of a single haploid genome is (3×10^9 bp/genome × 618 g/mol) / 6.02×10^{23} = 3.08 pg. Therefore, 30ng gDNA contains 30000 pg / 2 × 3.08 = 4870 copy of diploid cell genome and so a one copy clone must contain 4870 copy of the transgene. (Genome weight of other species can be found in the genome size database at www.genomesize.com.) Since we used a plasmid of 5800 bp in length, for the signal equivalent to the single copy clone, we needed approximately 4870 × (5800 bp × 618 g/mol) / 6.02×10^{23} = 0.029 pg of plasmid input. Considering a plasmid with concentration of 100 ng/µl, at least a 3.45×10^6-fold dilution is required to be in the similar range. The equivalent plasmid amount was put in the middle of our standard curve and two neighboring points of twofold dilutions were taken for the standard curve (Figure 6). With this setup applying the GFP TaqMan® assay, 5317 ± 195 copy was calculated for a randomly inte-

grated GFP expressing clone, slightly differing from the desired 4870, estimating the copy number as (5317 ± 195) / 4870 = 1.09 ± 0.04.

In absolute quantification, each template must have sufficient purity and the input amount needs to be precisely quantified. The accuracy of the assay is directly related to the quality of the standard curve. Several dilution steps preceding each assay have to be performed with special attention, however, no matter how much care is taken, real-time PCR sensitivity amplifies minute human errors. In addition, the plasmid template used to generate the standard curve might not be an ideal specimen as it might not really represent the complex properties of the unknown samples. Due to such difficulties, absolute quantification seemed inconvenient for our routine transposon applications. Nevertheless, once we have successfully selected one copy clones, we could use them as reference samples for comparative analysis.

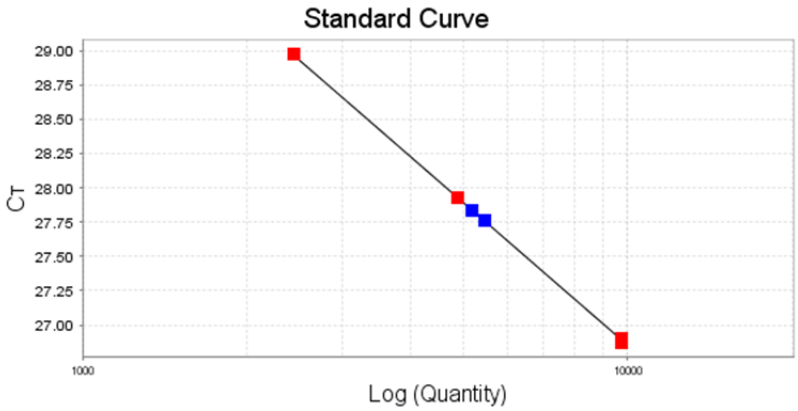


Figure 6: Absolute quantification of a presumably one copy clone using the GFP TaqMan® assay: comparison of the Ct values to known plasmid dilutions. Ct values are shown as a function of the quantity. Standard curve points of plasmid dilutions around the expected quantity are shown in red, whereas Ct values measured for the clone were put on this curve and are shown in blue. Quantity is illustrated as plasmid copy number input. The interpolated copy number for this clone was around 5300. See more informations and calculations in the text.

4.2 Second Approach: Comparative Quantification using One Copy Calibrator Samples

The alternative to identify a potential calibrator clone is to screen for the lowest transgene or marker gene expression. We have analyzed GFP expressing HEK-293 clones by FACS, and measured different green fluorescent intensities (Figure 7). The fluorescence intensities of the selected clones were compared to that of a one copy clone previously identified by absolute quantification. The GFP intensity of this latter one was similar to most of the lowest level expressing clones confirming that these clones are suitable for reference samples with one copy integration for subsequent analysis. Although expression level can be affected by its insertion site, we may say that one copy clones have significantly and uniformly lower GFP intensity than the few copy ones, but other transgene expressions may show wider differences.

Comparative qPCR quantification, while still technically challenging, does not require the same level of stringency. In this approach, the assay for target sequence is compared to a reference sample (a calibrator), and instead of precise copy number determination, it focuses on relative fold changes. In our application the copy number can be calculated based on the relative quantity of a single copy insertion. The method is based on the assumption that the threshold number of the sample and reference molecules

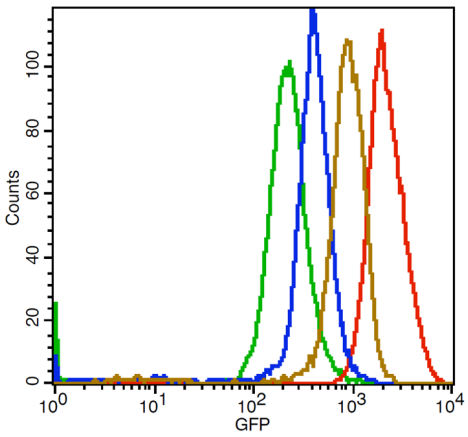


Figure 7: FACS histograms of representative HEK-293 clones expressing GFP as a transgene. Differences in GFP intensities correlate with GFP copy numbers.

is equal (Livak, 1997):

$$R_{Ct\ Sample} = R_{Ct\ Reference}$$

where RCt stands for the number of molecules at the threshold cycle. With the theoretical 100% efficiency, PCR duplicates the target in each cycle:

$$R_{Ct\ Sample} = R_{0\ Sample} \times 2^{Ct\ Sample}$$

$$R_{Ct\ Reference} = R_{0\ Reference} \times 2^{Ct\ Reference}$$

where R0 Sample and R0 Reference are the initial number of molecules of the sample and reference, respectively. The fold difference can therefore be calculated as:

$$R_{0\ Sample} / R_{0\ Reference} = 2^{-\Delta Ct},$$

where:

$$\Delta Ct = Ct_{Sample} - Ct_{Reference}$$

However, when the efficiency is not 100% but it is reproducibly identical, efficiency correction should be incorporated into the ΔCt method.

$$\text{Fold difference: } (1 + E)^{-\Delta Ct}.$$

For each novel assay, it is advisable to determine the efficiency values using the standard curve methodology discussed previously. As formerly mentioned, the most suitable specimen for such standard curve analysis is pooled gDNA of transgenic clones, because it has similar complexity to the unknown samples being analyzed. For comparison, the interpolated values from the standard curves can be used. A concrete example for the copy number calculation is as follows: the derived amount of the unknown sample is 24,99±0,15ng, whereas that of the single copy reference sample is 6.37 ± 0.18ng, so:

Fold difference – unknown/reference: (24.99 ± 0.15ng)/(6.37 ± 0.18ng)

Relative quantity (RQ) = 3.93 RQ Min = 3.80 RQ Max = 4.06

The upper example is still based on comparison of the absolute values of target Ct, however, a single Ct does not always reflect the expected number because of specimen discrepancies. Abandoning absolute quantification did not eliminate the deviations arising from differences in sample quality of the compared clones. Although the input of the template is always standardized, Ct deviations can still be attributable to certain errors, such as DNA concentration measurements. Normalization to an endogenous control sequence can overcome this problem. The control sequence should be similar in abundance to the target sequence and it must be present at a consistent level among all samples being compared. As an endogenous control for the human genome, the RPPH1 gene (the H1 RNA subunit of the RNaseP enzyme complex) was chosen which is a widely-accepted one copy gene of the haploid human genome. As mentioned earlier, without knowing the efficiency values, standard curves of both sequences have to be applied (Figure 8).

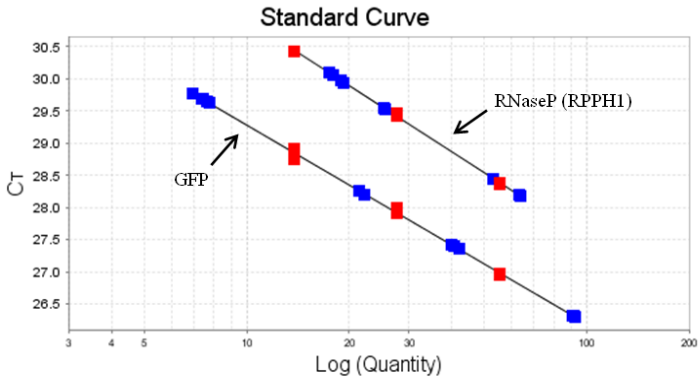


Figure 8: Example of a relative standard curve experiment. Standard curves for GFP and RNaseP (RPPH1) TaqMan® assay can be seen. Dilutions were made from pooled gDNA of unknown clones analyzed in the measurement; quantity is expressed in nanograms of gDNA input. Regression lines calculated from the measurement points of the two assays are parallel to each other, indicating very similar efficiency values. Red points represent points of the standard curves, whereas blue ones represent measurements of examined clones. See more informations and calculations in the text.

A relative standard curve experiment is based on sequential comparisons. First, both the target and the endogenous control are interpolated separately from the respective standard curves. Specimen comparisons showed formerly can be carried out only after normalizations to the endogenous control, for example:

Normalizations – Target/EndCont:

Sample → (54.23 ± 2.60ng) / (42.34 ± 0.82ng)

Reference → (6.13 ± 0.70ng)/(19.43 ± 0.25ng)

Fold difference – Sample/Reference:

Relative quantity (RQ) = 4.07 RQ Min = 3.76 RQ Max = 4.35

Alternatively, comparative analysis could be based on the real efficiency values:

Fold difference – Difference in the target/Difference in the endogenous control:

$$(1 + E_{Target})^{-\Delta Ct\ Target} / (1 + E_{EndCont})^{-\Delta Ct\ EndCont}$$

where:

$$\Delta Ct_{Target} = Ct_{Target\ Sample} - Ct_{Target\ Reference}$$

$$\Delta Ct_{EndCont} = Ct_{EndCont\ Sample} - Ct_{EndCont\ Reference}$$

However, if efficiencies of both the target and the endogenous control are proven to be close to identical, the ΔΔCt method could be chosen:

$$\text{Fold difference: } (1 + E)^{-\Delta\Delta Ct}$$

where:

$$\Delta\Delta Ct = \Delta Ct_{Target} - \Delta Ct_{EndCont}$$

Ideally, efficiencies of both the target and the endogenous control are close to 100%.

$$\text{Fold difference: } 2^{-\Delta\Delta Ct}$$

The main requirement for the ΔΔCt method is that the efficiencies of the assays are identical. The deviation of the two efficiencies can be determined by plotting the ΔCt-s from the standard curves points, and if the slope of this so called relative efficiency plot (Figure 9) is in the range of < 0.1, then it is acceptable to employ the ΔΔCt method. Another way to test the applicability is analyzing the relative standard curve experiment data with the ΔΔCt method. Similar outcome of the two methods will show the acceptability of the comparative Ct method, and in such cases, standard curves can be leaved behind. As the GFP sequence worked well in combination with the RPPH1 endogenous control, we continued to examine if assays designed for SB transposon sequences can also be utilized for copy number determination with similar methodology.

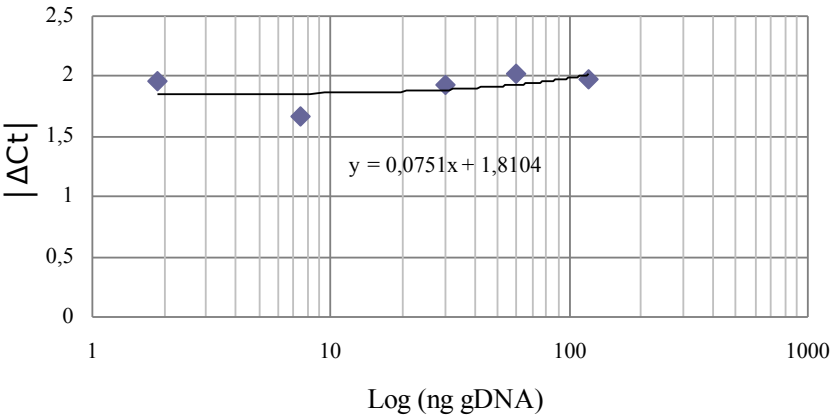


Figure 9: Relative efficiency plot comparing TaqMan® assays of IRDR-L and RNaseP (RPPH1). Dilution series were prepared from pools of gDNA containing one copy of SB transposon. Absolute values of ΔCt (target minus endogenous control) from the standard curve points were plotted against gDNA quantities and the equation of the regression line calculated from the measurement points is shown. A line with a slope of very small value (close to 0) indicates identical efficiencies of the two assays across all input concentrations.

4.3 Third Approach: Comparative Quantification using the SB IRDR Sequence Independently of the Transgene

As the SB transposon system is generally applied in our laboratory, we aimed at developing a real-time PCR-based technique that is transgene-independent, specific for the transposon regions, and therefore widely applicable. As for most transposon flanking sequences, the two IRDR regions are repeat-rich DNA sequences which make PCR primer design relatively difficult. Moreover, the left and the right IR-DRs are very similar to each other which further narrows the possibility to design specific assays for them. Nevertheless, we could still develop specific TaqMan® assays for each: neither of the IRDR-L nor the IRDR-R probe set gives signals in the exclusive presence of the other template.

Next we tested both assays designed for the transposon sequences, whether these fit to the GFP copy numbers. As formerly mentioned, random integrations can be the result from the breakage of the plasmid at any points, therefore randomly integrated one copy clones are not reliable candidates for reference samples when utilizing transposon sequences. Only after validating the one copy candidates cloned from the active transposition experiments with the GFP assay, were these clones used as reference samples with the transposon specific assays. The results based on GFP and the IRDR-L were in agreement with each other (Figure 10) and most standard curve analysis showed similar result with the comparative Ct method indicating that we could directly use the ΔΔCt method. In addition, technical errors could be further decreased using a pool of gDNA samples with known copy number as a reference. Here we show the calculation with the row Ct values by the example of G2C2 clone in Figure 10 with Cts resulted in IRDR-L assay. Ct mean values are calculated as the mean of Cts of 3 simultaneous PCR reactions:

$Ct_{\text{Mean IRDR-L}}$ of G2C2: 31.065

$Ct_{\text{Mean RNaseP}}$ of G2C2: 30.951

$Ct_{\text{Mean IRDR-L}}$ of one copy clone pool: 31.457

$Ct_{\text{Mean RNaseP}}$ of one copy clone pool: 29.773

Calculations of ΔCt – s:

$\Delta Ct_{\text{IRDR-L}} = 31.065\text{-}31.457 = -0.392$

$\Delta Ct_{\text{RNaseP}} = 30.951\text{-}29.773 = 1.178$

Calculation of ΔΔCt: $\Delta\Delta Ct = -0.392 - 1.178 = -1.57$

Relative quantity: $RQ = 2^{-\Delta\Delta Ct} = 2^{1.57} = 2.969$ (3 copies)

These experiments therefore supported the use of the IRDR-L repeat specific assay for transposon copy number determination as it gave the same results as the assay specific for the carried internal transgene (GFP). However similar application of IRDR-R TaqMan® assay for the previously analyzed clones was unreliable to determine the exact copy number and the assay usually showed lower efficiency than the GFP and the IRDR-L. So initially we concluded that we have to leave out the specific but less efficient assay for the IRDR-R region (Kolacsek *et al*., 2011).

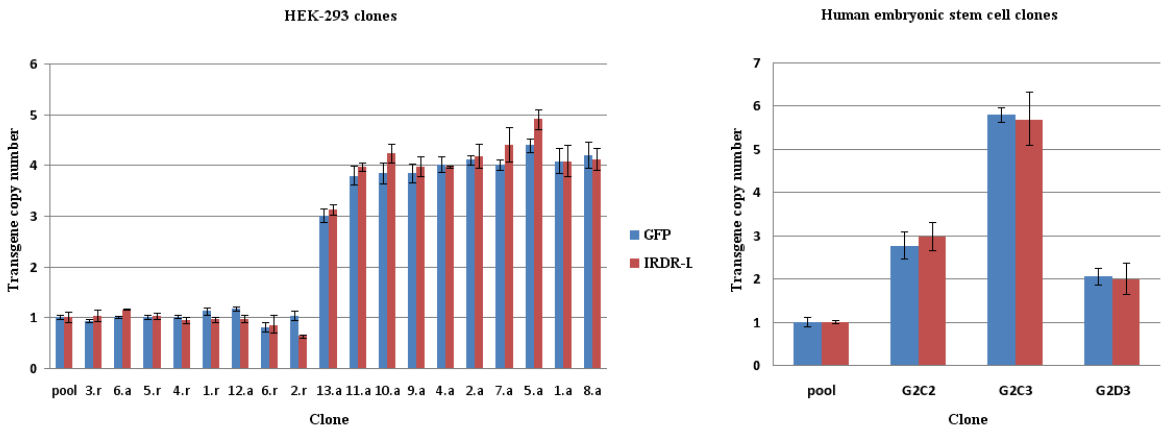


Figure 10: Copy numbers determined by the GFP or the IRDR-L TaqMan® assays in different clones of HEK-293 (left panel) and human embryonic stem cells (right panel). Results of the transgene independent, SB transposon specific IRDR-L assay correlated well with the GFP-based copy numbers. Pool: a mixture of equal amounts of one copy clones. Figure partially adopted from Kolacsek et al., Mobile DNA (2011), 2(1):5, published by BioMed Central.

Concerning SB transposon mutants and variants, "symmetrical" SB transposons with two IRDR-L (but not with two IRDR-R) flanking sequences were shown to be functional (Izsvak *et al.*, 2002). The assay for the left transposon sequence is also applicable for such constructs with a correction factor of 0.5. Searching for such mutant transposon clones, we used the IRDR-R assay to determine the presence of the right terminal repeat sequence. During this presence-absence examination studies, we could achieve higher sensitivity of the IRDR-R TaqMan® with elevating the input range of the gDNA. However, the sensitivity and the reliability of IRDR-R was still below to that of the IRDR-L assay, and only lower copy number clones showed reliable measurements (Figure 11). In addition, we still have to emphasize here that sample quality has a deep impact on reproducibility of all assays.

As a major general recommendation, we routinely analyze more parts of the inserted transgene sequence (e.g. IRDR-L and GFP regions or IRDR-L and IRDR-R regions). It is also advisable using at least two separate single copy clones as reference samples and two few copy (3 – 4) clones as controls in all analyses. The application of more than one reference clone makes it possible to choose the most appropriate one with which low copy control clones give the best precise round copy numbers. In fact, the use of the few (3 – 4) copy clones as references is helpful to approximate extreme copies (> 15) more precisely.

5 Validation of Transgene-independent qPCR Copy Number Quantification

To compare our transgene-independent quantification approach with other techniques, we measured copy numbers of clones that were generated from different cell types by transposons containing various transgene sequences. Such clones were ideal for comparison due to the different transgene sequence and because copy numbers in those cases were also determined either by the Southern/dot blotting techniques,

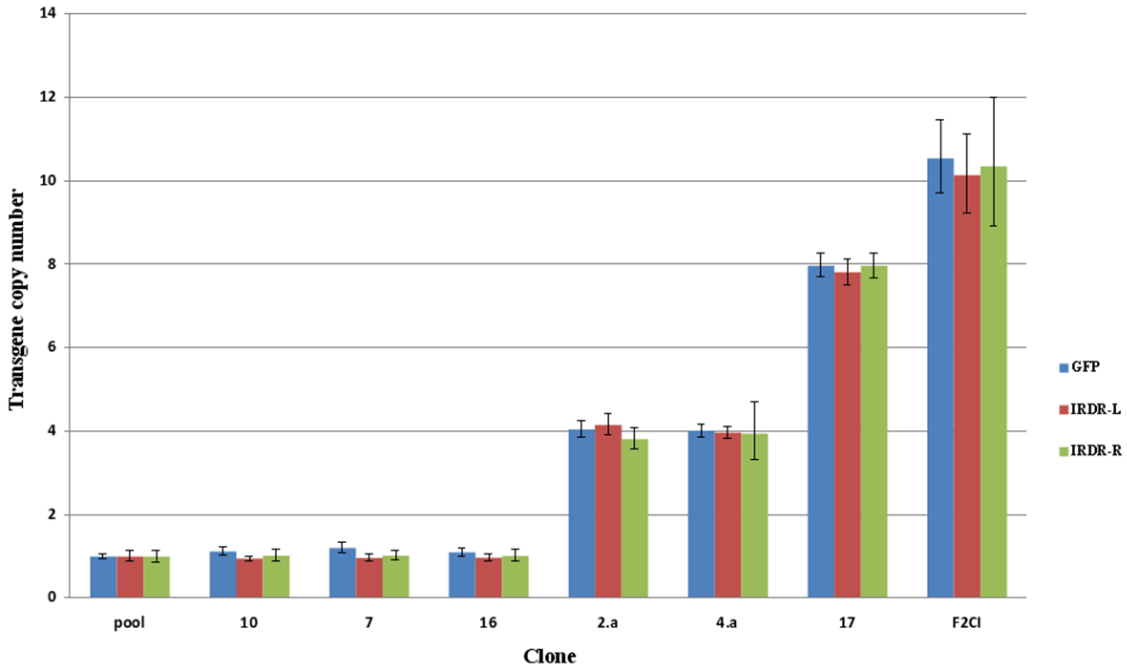


Figure 11: Comparison of transposon copy number determinations based on three different TaqMan® assays. Using higher amount of input gDNA (90ng) resulted in reliable correlation of the IRDR-R assay with the copy numbers determined by the GFP and IRDR-L specific qPCR measurements.

or by the transposon display method, or by estimations from transgene integration assays (Splinkerette PCR/Inverse PCR). Using the IRDR-L assay with the ΔΔCt methodology, copy numbers were estimated ranging from 1 to 50 copies in various clones (Table 1 and 2). The measured transposon copies were almost always the same by the qPCR as by the canonical methods. As canonical methods – with the exception of dot blot – utilize restriction enzyme sites flanking the integration point, which may be sensitive to sequence environment, therefore all integrated copies may not be reliably detected by the canonical methods. Perhaps this is why copy numbers were underestimated in some of the cases comparing to our qPCR method. For higher (> 5) copy-number clones, our method was also accurate, with occasional low relative-error margins (≤ 9%). The slight differences in these cases could be due to the inaccuracy of the standard methods for this range. In addition, it has been suggested that precise values of very high copy numbers are more reliably measured by dot blot rather than transposon display method.

By the above described experiments, the newly developed transgene independent method for determining SB transposon copy numbers could be validated since (i) it provided the same results as the assays specific for the carried transgene sequence and (ii) it could also reliably replace widely used canonical radioactive techniques. The TaqMan® assay designed for the IRDR-L region of the transposon provides the basis for transgene independence as it is present in all SB constructs (Kolacsek *et al.*, 2011).

Clone	Canonical methods	Copy numbers	
		by canonical methods	by qPCR IRDR-L
2/1	Transposon display/Southern blotting	8 – 10	8
2/2	Transposon display/Southern blotting	3	4
2/3	Transposon display/Southern blotting	10 – 12	10
2/9	Transposon display/Southern blotting	1	1
1	Transposon display/Southern blotting	12 – 13	13
4	Dot blot	52	50
5	Transposon display/Southern blotting	15	15
6	Transposon display/Southern blotting	12	11
7	Transposon display/Southern blotting	1	1
8	Transposon display/Southern blotting	2	2
9	Transposon display/Southern blotting	1	1
A3	Splinkerette PCR/Inverse PCR	2	2
A4	Splinkerette PCR/Inverse PCR	4	4
A5	Splinkerette PCR/Inverse PCR	4	4,5
A6	Splinkerette PCR/Inverse PCR	2	2
B1	Splinkerette PCR/Inverse PCR	1	2
B2	Splinkerette PCR/Inverse PCR	2	2
B3	Splinkerette PCR/Inverse PCR	3	3
B5	Splinkerette PCR/Inverse PCR	2	2
C3	Splinkerette PCR/Inverse PCR	1	2
C5	Splinkerette PCR/Inverse PCR	2	2
3	Splinkerette PCR/Inverse PCR	1	1
16	Splinkerette PCR/Inverse PCR	1	1
3.a	Splinkerette PCR/Inverse PCR	1	1
6.a	Splinkerette PCR/Inverse PCR	1	1
12.a	Splinkerette PCR/Inverse PCR	1	1

Table 1: Comparing the IRDR-L qPCR-based method with other canonical techniques for transposon copy number determination. Additional data is added to the table adopted from Kolacsek et al., Mobile DNA (2011), 2(1):5, published by BioMed Central.

Clone	Copy by qPCR	Sequence	Chrom. pos.	Ref.seq.	Nt.pos. of ref.seq.	Region	Seq. orient.	Method
C3	2	TTCCTTGTTACTCCTTCAAAATGCT TAcagttgaagt	**Hs 6p22**	NT_007592.15	51406968	intergenic	**L**eft	splinkerette pGEM
C5	2	CAGTTTTATTATTATTGTCATTGTA TAcagttgaagt	**Hs Xq28**	NT_167198.1	2288320	*GABRA3* intron3	**L**	splinkerette pGEM
		CAGATCCTGATAAATGTTAGTTAC TAcagttgaagt	**Hs 10q21**	NT_030059.13	4645679	*PRKG1* intron7	**L**	inverse PCR pGEM
3	1	TTCCTCCCTGGACTTTAGGACATA TAcagttgaagt	**Hs 5q22**	NT_034772.6	6866220	intergenic	**L**	splinkerette
		AGGAAGTATACACTCACCCTGTGA TAcagttgaagt	Hs 5q22	NT_034772.6	6866220		**R**ight	splinkerette
16	1	GACACAGCTGATTTTGAAGTCAGA TAcagttgaagt	**Hs 4q28**	NT_016354.19	65845752	*SCOC* intron1	**L**	splinkerette
		GAAGACTACCCTATCACTGTATTA TAcagttgaagt	Hs 4q28	NT_016354.19	65845752		**R**	splinkerette
3.a	1	TAGCACAATGAGTACTTTATCACA TAcagttgaagt	**Hs 1p31**	NT_032977.9	82218216	intergenic	**L**	inverse PCR pGEM
6.a	1	AATACATGGAAGGTAGAACAGATC TAcagttgaagt	**Hs 1p31**	NT_032977.9	88638434	intergenic	**L**	splinkerette
		TAAATTATACATGTCTGTTAAAGCA TAcagttgaagt	Hs 1p31	NT_032977.9	88638434		**R**	splinkerette
12.a	1	TCTGGAAAGCCACATTCGGGAACT TAcagttgaagt	**Hs 18q21**	NT_025028.14	21860434	intergenic	**L**	splinkerette

Table 2: Examples of *bona fide* SB transposon integration sites in clones of human cell lines where transposon copy numbers were determined previously. L or R indicates whether the particular transposon integration site was determined from the direction of the left or the right terminal repeat region of SB. The 'TA' sequence (shown in red) next to the transposon terminal repeats provides evidence that these integrations are the results of *bona fide* transposition, as this dinucleotide marks the SB target sequence that is duplicated during the reaction. Hs=Homo sapiens.

6 Quantitative Excision PCR

In our laboratory, we have successfully applied different DNA transposon systems for various in vitro cell culture applications, including studies on directed tissue differentiation from embryonic stem cells and modeling *ex vivo* introduction of therapeutic transgenes into patient cells. Controlling the delivery efficiency is a crucial task for these applications which is usually characterized by testing the transgenic rate of transfectants. However these test methods are time consuming and not applicable in all of the cases, so we have developed a qPCR method to characterize transposition activity. This quantification is extremely useful when optimizing for conditions of a given delivery system, or comparing the efficiency of different transposon systems or variants of a particular transposase enzyme.

There are two major steps involved in transposition, the excision of the transposon from the donor site and the integration of the transposon into the target site. These steps are proven to be coupled since excision frequency of different transposon mutants from a donor plasmid was correlated to overall transposition efficiency (Liu *et al.*, 2004). Excision events can be detected by a PCR reaction with primers flanking the transposon sequence at the donor site (Figure 12). Following excision, the donor plasmid will be circularized by the double-strand DNA break repair mechanism of the cell, and only these excised and repaired plasmids will serve as a template for exponential amplification, because the transposon content is usually large enough not to be amplified from the original uncut sequence. Excision PCR can be carried out on samples taken on the second/third day after transfection, and both isolated plasmids, as well as cell lysates can be used as input material (we generally use isolated plasmids as input.) A semi quantitative version of this excision PCR has been applied by Liu *et al.* (2004) to describe the excision step of SB transposition.

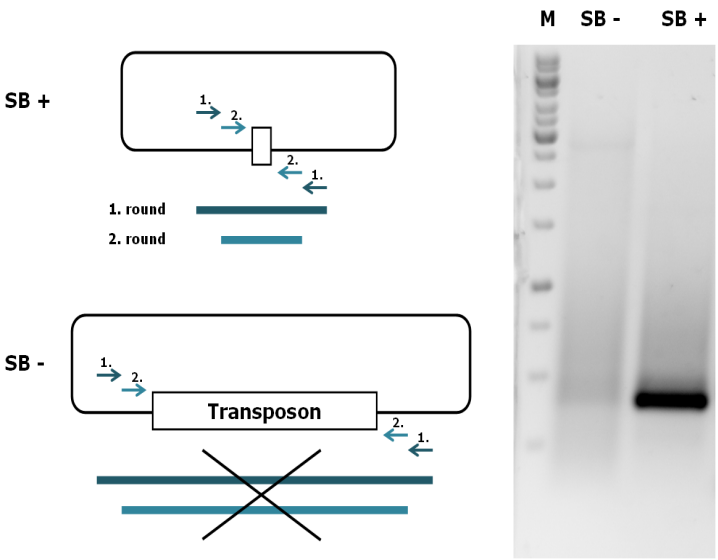


Figure 12: Principle of the excision PCR using transposon containing donor plasmids. This is a nested PCR technique using two sets of primers specific to the transposon flanking sequences, and amplifying the products in two consecutive rounds of PCR. In case of transposition (SB+), there is a distinct product that can be visualized by agarose gel electrophoresis (right panel). In the absence of transposition (SB-), the PCR cannot amplify the target sequence due to its large size. For more details, see the text.

Due to the error-prone double-stranded break repair mechanism, the joining of the donor ends is remarkably imprecise, therefore real-time quantification of excision based on a TaqMan® probe with stringent sequence requirement could not be considered, and hence we adopted the SYBR® Green technology for this application. In some cases, since very small portion of the amount of transposon donor plasmids undergo excision, to increase the traceability, more than one round of nested PCR is necessary to quantify the excision reaction. For this reason, we apply a 10 – 13 cycle pre-amplification (first round of PCR), and after a thousand-fold dilution we measure the real-time round (second round of PCR) with nested primers using SYBR® Green. Amplification from another segment of the plasmid backbone (in our case, the ampicillin resistance gene sequence) could serve as a normalization control. Due to its large excess, ampicillin sequence is not pre-amplified but it is permanently present in our samples during all processes until the second round of excision PCR, so it can be correctly measured in all samples. Although the primers we routinely use are specific to the backbone of our transposon plasmid constructs which can differ among SB users, our primers might be useful for those who apply constructs with the same origin (Table 3). PCR efficiencies for the target and the control sequence were measured by serial dilutions of pooled samples and the ΔΔCt method was proved to be reliable for comparison of excision efficiencies.

	name	sequence 5′ → 3′
1	exc.1 for	GCGAAAGGGGGATGTGCTGCAAGG
2	exc.1 rev	TCTTTCCTGCGTTATCCCCTGATTC
3	exc.2 for	CGATTAAGTTGGGTAACGCCAGGG
4	exc.2 rev	CAGCTGGCACGACAGGTTTCCCG
5	amp. for	TTTGCTCACCCAGAAACGC
6	amp. rev	AGTTGGCCGCAGTGTTATCAC

Table 3: Sequences of primers used for the quantitative excision PCR. Primers with .1 extension are used for the first round, whereas others with .2 extension are used for the second round of the nested PCR; amp primers anneal to the sequence of the ampicillin resistance gene.

An excellent validation of our quantification method developed for measuring the transposition activity was to compare different SB transposase versions resulted from gradual improvement of the activity (Mates *et al.*, 2009). Relative quantification of excision efficiency of SB11x, SB32x and SB100x is correlated to their expected activity in HEK-293 and in HeLa cells as well (Figure 13).

Normalization to the plasmid backbone sequence makes this relative quantification technique independent of the transfection efficiency. This provides the basis for comparing various transposon delivery experiments, including different transposon systems (e.g. SB or PB), different transposase or transposon variants, different host cell types, or different experimental settings. Applying this quantitative technique could therefore be a reliable and fast screening approach for different transposon systems and gene delivery conditions before any applications with the desired transgene.

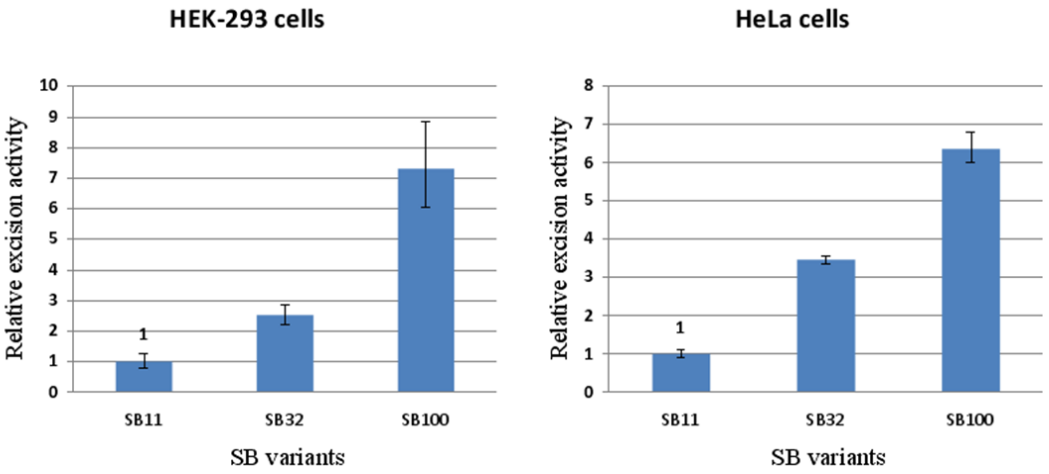


Figure 13: Relative excision efficiencies of SB transposase variants in different cell lines measured by real-time qPCR. Normalizing the excision PCR to the amount of the ampicillin sequence by specific PCR, here we quantify the excision events among all transfected transposon donor plasmids.

7 Related Works

As it was used for quantitative aspects of transposon-based gene delivery experiments, real-time PCR measurements are widely used and have become integral part of the methodology for gene delivery and gene therapy applications. On the other hand, careful design of the experiments or choosing suitable methodology are often missing from such studies, weakening the conclusion of the results. For example, Bian & Belmont (2010) have made absolute quantification of transgenic multi-copy insertions derived from linearized plasmid transfection, although they did not carry out any normalization of the input material. Huang *et al.* (2010) have detected random integration level with absolute quantification normalizing to the RNaseP one copy sequence. They used "empty" gDNA to mix with the target sequence containing plasmid for the standard curve samples; however, normalizing the absolute quantification with external template incorporates additional variability in the measurement. In the meantime, well-designed studies provide excellent new examples of combining existing technologies. Charrier *et al.* (2011) have applied an elegant solution for absolute quantification to determine lentiviral vector copy numbers with normalization to endogenous albumin. Standard curves were made from the same plasmid containing both the target and the endogenous control sequence, resulting in the smallest possible variability in their measurement. Ballester *et al.* (2004) determined copy number of transgenic mice carrying goat lactoglobulin gene using relative quantification with the comparative Ct method. The unique design of the measurement was the application of goat gDNA as a reference sample and the choice of glucagon sequence as an endogenous control, the latter one being strongly conserved between the two species.

Concerning SB transposon applications, a recent paper described a simultaneous analysis of excision activity of *Sleeping Beauty* and the resulting transgene copy number (Jin *et al.*, 2011). Excision was analyzed at days 1 to 3 after electroporation using the ΔΔCt method with a TaqMan® assay where the probe was specific to one of the transposon flanking regions. They also normalized to a plasmid backbone sequence, thus similarly to our qPCR studies, their quantification is independent of the electrotrans-

fer efficiency. However, as the repair of the plasmid after excision is an error-prone process (see Section 7), using a strict TaqMan® probe sequence may not detect all excision products. They also determined the average copy number in transgenic cell population after four weeks of selection. They used an absolute quantification approach using dilution series of a one copy clone gDNA as a standard which is in fact an alternative to relative standard curve method using one copy as a reference.

8 Conclusions

Transposon-based technology is an emerging new method of choice for gene delivery and for gene therapy applications. Compared to viral vectors, transposon systems offer several advantages. Apart from being less expensive in terms of the required safety facilities, for several DNA transposons, the integration profile of the delivered transgene is more close to random, showing no preferences for coding regions, therefore making its application less susceptible for insertional mutagenesis. SB seems to be the safest delivery technique for two reasons: 1) its transgene integration profile is the closest to random among all known gene delivery vehicles; 2) SB was resurrected from an ancient non-functional fish mobile element, therefore no potential transposons are present in vertebrate species, including the human genome, that can remobilize the integrated transgene. Based on these characteristics, the SB system is the most suitable for cell and gene therapy applications, even if compared to other transposon vehicles.

Our work contributed to the quantification of transposon delivery in two important aspects. We have developed sensitive and reliable real-time PCR-based methods to measure (i) the first step of the transposition and (ii) the resulting copy number of the delivery. Quantitative excision PCR is extremely useful to control and to optimize transposon mediated gene delivery, whereas copy number determination is essential to characterize transgenic cells. Comparing our copy number method with widely used canonical methods, it was proved to be just as accurate as those, also offering a faster and non-radioactive approach at the same time. However, the real advantage of this method is the transgene-independence which makes it applicable for any scientists working with SB transposon constructs.

In this chapter, we gave a detailed protocol for designing quantitative measurements of integrated DNA sequences such as transposons, and show examples for developing reliable quantitative assays specific to any sequence. In general, however, this description may also serve as a stepwise guide providing a strategy for similar quantification purposes.

Acknowledgements

Tamás I. Orbán is a recipient of the János Bolyai Scholarship of the Hungarian Academy of Sciences. Research in our laboratory was supported by grants from OTKA (NK83533), STEMKILL (OM00108/2008), and KMOP-1.1.2-07/1-2008-0003.

References

Balciunas, D., Wangensteen, K. J., Wilber, A., Bell, J., Geurts, A., Sivasubbu, S., Wang, X., Hackett, P. B., Largaespada, D. A., McIvor, R. S., & Ekker, S. C. (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. PLoS Genetics, 2(11), e169.

Ballester, M., Castello, A., Ibanez, E., Sanchez, A., & Folch, J. M. (2004). Real-time quantitative PCR-based system for determining transgene copy number in transgenic animals. Biotechniques, 37(4), 610-613.

Baus, J., Liu, L., Heggestad, A. D., Sanz, S., & Fletcher, B. S. (2005). Hyperactive transposase mutants of the Sleeping Beauty transposon. Molecular Therapy, 12(6), 1148-1156.

Bian, Q. & Belmont, A. S. (2010). BAC TG-EMBED: one-step method for high-level, copynumber-dependent, position-independent transgene expression. Nucleic Acids Research, 38(11), e127.

Biemont, C. & Vieira, C. (2006). Genetics: junk DNA as an evolutionary force. Nature, 443(7111), 521-524.

Burns, K. H. & Boeke, J. D. (2012). Human transposon tectonics. Cell, 149(4), 740-752.

Charrier, S., Ferrand, M., Zerbato, M., Pre´cigout, G., Viornery, A., Bucher-Laurent, S., Benkhelifa-Ziyyat, S., Merten, O. W., Perea, J., & Galy, A. (2011). Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. Gene Therapy, 18(5), 479-487.

Chelly, J., Kaplan, J. C., Maire, P., Gautron, S., & Kahn, A. (1988). Transcription of the dystrophin gene in human muscle and non-muscle tissue. Nature, 333(6176), 858-860.

Chew, S. K., Rad, R., Futreal, P. A., Bradley, A., & Liu, P. (2011). Genetic screens using the piggyBac transposon. Methods, 53(4), 366-371.

Collier, L. S. & Largaespada, D. A. (2005). Hopping around the tumor genome: transposons for cancer gene discovery. Cancer Research, 65(21), 9607-9610.

Collier, L. S. & Largaespada, D. A. (2007). Transposable elements and the dynamic somatic genome. Genome Biology, 8 Suppl 1, S5.

Devon, R. S., Porteous, D. J., & Brookes, A. J. (1995). Splinkerettes—improved vectorettes for greater efficiency in PCR walking. Nucleic Acids Research, 23(9), 1644-1645.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., & Xu, T. (2005). Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell, 122(3), 473-483.

Ellis, J. (2005). Silencing and variegation of gammaretrovirus and lentivirus vectors. Human Gene Therapy, 16(11), 1241-1246.

Feschotte, C. & Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annual Reviews of Genetics, 41, 331-368.

Garrison, B. S., Yant, S. R., Mikkelsen, J. G., & Kay, M. A. (2007). Postintegrative gene silencing within the Sleeping Beauty transposition system. Molecular and Cellular Biology, 27(24), 8824-8833.

Goodier, J. L. & Kazazian, H. H., Jr. (2008). Retrotransposons revisited: the restraint and rehabilitation of parasites. Cell, 135(1), 23-35.

Grabundzija, I., Irgang, M., Mates, L., Belay, E., Matrai, J., Gogol-Doring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M. K., VandenDriessche, T., Izsvak, Z., & Ivics, Z. (2010). Comparative analysis of transposable element vector systems in human cells. Molecular Therapy, 18(6), 1200-1209.

Guo, G., Huang, Y., Humphreys, P., Wang, X., & Smith, A. (2011). A PiggyBac-based recessive screening method to identify pluripotency regulators. PLoS One, 6(4), e18189.

Hedges, D. J. & Batzer, M. A. (2005). From the margins of the genome: mobile elements shape primate evolution. Bioessays, 27(8), 785-794.

Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real time quantitative PCR. Genome Research, 6(10), 986-994.

Higuchi, R., Dollinger, G., Walsh, P. S., & Griffith, R. (1992). Simultaneous amplification and detection of specific DNA sequences. Biotechnology, 10(4), 413-417.

Huang, X., Haley, K., Wong, M,, Guo, H., Lu, C., Wilber, A., & Zhou, X. (2010). Unexpectedly high copy number of random integration but low frequency of persistent expression of the Sleeping Beauty transposase after trans delivery in primary human T cells. Human Gene Therapy, 21(11), 1577-1590.

Ivics, Z., Hackett, P. B., Plasterk, R. H., & Izsvak, Z. (1997). Molcular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. Cell, 91(4), 501-510.

Ivics, Z. & Izsvak, Z. (2004). Transposable elements for transgenesis and insertional mutagenesis in vertebrates: a contemporary review of experimental strategies. Methods in Molecular Biology, 260, 255-276.

Izsvak, Z., Ivics, Z., & Plasterk, R. H. (2000). Sleeping Beauty, a wide host-range transposon vector for genetic transformation in vertebrates. Journal of Molecular Biology, 302(1), 93-102.

Izsvak, Z., Khare, D., Behlke, J., Heinemann, U., Plasterk, R. H., & Ivics, Z. (2002). Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in Sleeping Beauty transposition. The Journal of Biological Chemistry, 277(37), 34581-34588.

Izsvak, Z., Hackett, P. B., Cooper, L. J., & Ivics, Z. (2010). Translating Sleeping Beauty transposition into cellular therapies: victories and challenges. Bioessays, 32(9), 756-767.

Jin, Z., Maiti, S., Huls, H., Singh, H., Olivares, S., Mates, L., Izsvak, Z., Ivics, Z., Lee, D. A., Champlin, R. E., & Cooper, L. J. N. (2011). The hyperactive Sleeping Beauty transposase SB100X improves the genetic modification of T cells to express a chimeric antigen receptor. Gene Therapy, 18(9), 849-856.

Kaji, K., Norrby, K., Paca, A., Mileikovsky, M., Mohseni, P., & Woltjen K. (2009). Virus-free induction of pluripotency and subsequent excision of reprogramming factors. Nature, 458(7239), 771-775.

Kazazian, H. H., Jr. (2004). Mobile elements: drivers of genome evolution. Science, 303(5664), 1626-1632.

Kleppe, K., Ohstuka, E., Kleppe, R., Molineux. L., & Khorana, H. G. (1971). Studies on polynucleotides. XCVI. Repair replications of short synthetic DNA's as catalyzed by DNA polymerases. Journal of Molecular Biology, 56(2), 341-361.

Kolacsek, O., Krízsik, V., Schamberger, A., Erdei, Z., Apáti, A., Várady, G., Mátés, L., Izsvák, Z., Ivics, Z., Sarkadi, B., & Orbán, T. I. (2011). Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. Mobile DNA, 2(1), 5.

Liu, G., Aronovich, E. L., Cui, Z., Whitley, C. B., & Hackett, P. B. (2004). Excision of Sleeping Beauty transposons: parameters and applications to gene therapy. The Journal of Gene Medicine, 6(5), 574-583.

Liu, G., Geurts, A. M., Yae, K., Srinivasan, A. R., Fahrenkrug, S. C., Largaespada, D. A., Takeda, J., Horie, K., Olson, W. K., & Hackett, P. B. (2005). Target-site preferences of Sleeping Beauty transposons. Journal of Molecular Biology, 346(1), 161-173.

Livak, K. J, Flood, S. J., Marmaro, J., Giusti, W., & Deetz, K. (1995). Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Research, 4(6), 357-362.

Livak, K. J. (1997). ABI Prism 7700 Sequence Detection System, User Bulletin #2. Applied Biosystems.

Mates, L., Chuah, M. K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., Vandendriessche, T., Ivics, Z., & Izsvak, Z. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nature Genetics, 41(6), 753-761.

Meilinger, D., Fellinger, K., Bultmann, S., Rothbauer, U., Bonapace, I. M., Klinkert, W. E., Spada, F. & Leonhardt, H. (2009). Np95 interacts with de novo DNA methyltransferases, Dnmt3a and Dnmt3b, and mediates epigenetic silencing of the viral CMV promoter in embryonic stem cells. EMBO Reports, 10(11), 1259-1264.

Mills, R. E., Bennett, E. A., Iskow, R. C., Luttig, C. T., Tsui, C., Pittard, W. S., & Devine, S. E. (2006). Recently mobilized transposons in the human and chimpanzee genomes. The American Journal of Human Genetics, 78(4), 671-679.

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? Trends in Genetics, 23(4), 183-191.

Miskey, C., Izsvak, Z., Plasterk, R. H., & Ivics, Z. (2003). The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. Nucleic Acids Research, 31(23), 6873-6881.

Moeller, F., Nielsen, F. C., & Nielsen, L. B. (2003). New tools for quantifying and visualizing adoptively transferred cells in recipient mice. Journal of Immunological Methods, 282(1-2), 73-82.

Morrison, T. B., Weis, J. J., & Wittwer, C. T. (1998) Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques, 24(6), 954-962.

Mullis, K. & Faloona, F. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology, 155, 335-350.

Ostertag, E. M., Madison, B. B., & Kano, H. (2007). Mutagenesis in rodents using the L1 retrotransposon. Genome Biology, 8 Suppl 1, S16.

Pledger, D. W. & Coates, C. J. (2005). Mutant Mos1 mariner transposons are hyperactive in Aedes aegypti. Insect Biochemistry and Molecular Biology, 35(10), 1199-1207.

Ririe, K. M., Rasmussen, R. P., & Wittwer, C. T. (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. Analytical Biochemistry, 245(2), 154-160.

Rowe, H. M., Jakobsson, J., Mesnard, D., Rougemont, J., Reynard, S., Aktas, T., Maillard, P. V., Layard-Liesching, H., Verp, S., Marquis, J., Spitz, F., Constam, D. B., & Trono, D. (2010). KAP1 controls endogenous retroviruses in embryonic stem cells. Nature, 463(7278), 237-240.

Rubin, G. M. & Spradling, A. C. (1982). Genetic transformation of Drosophila with transposable element vectors. Science, 218(4570), 348-53.

Shen, S., Lin, L., Cai, J. J., Jiang, P., Kenkel, E. J., Stroik, M. R., Sato, S., Davidson, B. L., & Xing, Y. (2011). Widespread establishment and regulatory impact of Alu exons in human genes. Proceedings of the National Academy of Sciences of the United States of America, 108(7), 2837-42.

Singer, T., McConnell, M. J., Marchetto, M. C., Coufal, N. G., & Gage, F. H. (2010) LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends in Neurosciences, 33(8), 345-54.

Sivalingam, J., Krishnan, S., Ng, W. H., Lee, S. S., Phan, T. T., & Kon, O. L. (2010). Biosafety assessment of site-directed transgene integration in human umbilical cord-lining cells. Molecular Therapy, 18(7), 1346-1356.

Solyom, S. & Kazazian, H. H. Jr. (2012) Mobile elements in the human genome: implications for disease. Genome Medicine, 4(2), 12.

Spradling, A. C. & Rubin, G. M. (1982). Transposition of cloned P elements into Drosophila germ line chromosomes. Science, 218(4570), 341-7.

Uren, A. G., Kool, J., Berns, A., & van Lohuizen, M. (2005). Retroviral insertional mutagenesis: past, present and future. Oncogene, 24(52), 7656-7672.

Vigdal, T. J., Kaufman, C. D., Izsvak, Z., Voytas, D. F., & Ivics, Z. (2002). Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. Journal of Molecular Biology, 323(3), 441-452.

Wang, A. M., Doyle, M. V., & Mark, D. F. (1989). Quantitation of mRNA by the polymerase chain reaction. Proceedings of the National Academy of Sciences of the United States of America, 86(24), 9717-9721.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12), 973-982.

Wicks, S. R., de Vries, C. J., van Luenen, H. G., & Plasterk, R. H. (2000). CHE-3, a cytosolic dynein heavy chain, is required for sensory cilia structure and function in Caenorhabditis elegans. Developmental Biology, 221(2), 295-307.

Williams, D. A. (2008). Sleeping beauty vector system moves toward human trials in the United States. Molecular Therapy, 16(9), 1515-1516.

Wilson, M. H., Coates, C. J., & George, A. L., Jr. (2007). PiggyBac transposon mediated gene transfer in human cells. Molecular Therapy, 15(1), 139-145.

Wittwer, C. T., Herrmann, M. G., Moss, A. A., & Rasmussen, R. P. (1997). Continuous fluorescence monitoring of rapid cycle DNA amplification. Biotechniques, 22(1), 130-138.

Woltjen, K., Michael, I. P., Mohseni, P., Desai, R., Mileikovsky, M., Hämäläinen, R., Cowling, R., Wang, W., Liu, P., Gertsenstein, M., Kaji, K., Sung, H. K., & Nagy, A. (2009) piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. Nature, 458(7239), 766-70.

Yant, S. R., Wu, X., Huang, Y., Garrison, B., Burgess, S. M., & Kay, M. A. (2005). High-resolution genome-wide mapping of transposon integration in mammals. Molecular and Cellular Biology, 25(6), 2085-2094.

Zayed, H., Izsvak, Z., Walisko, O., & Ivics, Z. (2004). Development of hyperactive sleeping beauty transposon vectors by mutational analysis. Molecular Therapy, 9(2), 292-304.

Zhu, J., Park, C. W., Sjeklocha, L., Kren, B. T., & Steer, C. J. (2010). High-level genomic integration, epigenetic changes, and expression of sleeping beauty transgene. Biochemistry, 49(7), 1507-1521.

SCIENTIFIC REPORTS

OPEN

Received: 01 June 2015
Accepted: 03 July 2015
Published: 03 August 2015

Generation of a Homozygous Transgenic Rat Strain Stably Expressing a Calcium Sensor Protein for Direct Examination of Calcium Signaling

Kornélia Szebényi[1,*], András Füredi[1,*], Orsolya Kolacsek[1,*], Enikő Pergel[1], Zsuzsanna Bősze[2], Balázs Bender[3], Péter Vajdovich[4], József Tóvári[5], László Homolya[1], Gergely Szakács[1], László Héja[6], Ágnes Enyedi[7], Balázs Sarkadi[1,8], Ágota Apáti[1] & Tamás I. Orbán[1]

In drug discovery, prediction of selectivity and toxicity require the evaluation of cellular calcium homeostasis. The rat is a preferred laboratory animal for pharmacology and toxicology studies, while currently no calcium indicator protein expressing rat model is available. We established a transgenic rat strain stably expressing the GCaMP2 fluorescent calcium sensor by a transposon-based methodology. Zygotes were co-injected with mRNA of transposase and a CAG-GCaMP2 expressing construct, and animals with one transgene copy were pre-selected by measuring fluorescence in blood cells. A homozygous rat strain was generated with high sensor protein expression in the heart, kidney, liver, and blood cells. No pathological alterations were found in these animals, and fluorescence measurements in cardiac tissue slices and primary cultures demonstrated the applicability of this system for studying calcium signaling. We show here that the GCaMP2 expressing rat cardiomyocytes allow the prediction of cardiotoxic drug side-effects, and provide evidence for the role of Na^+/Ca^{2+} exchanger and its beneficial pharmacological modulation in cardiac reperfusion. Our data indicate that drug-induced alterations and pathological processes can be followed by using this rat model, suggesting that transgenic rats expressing a calcium-sensitive protein provide a valuable system for pharmacological and toxicological studies.

The importance of proper calcium homeostasis and signaling from the cellular to the complex organ levels is well appreciated: both in physiological and pathological processes cellular free calcium plays a major role[1]. Disruption of the calcium homeostasis by pharmacological agents or pathological conditions correlate with various conditions, including prolonged QT intervals and arrhythmias in the heart[2,3], or

[1]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary. [2]NARIC-ABC, Gödöllő, Hungary. [3]ImmunoGenes Ltd., Budakeszi, Hungary. [4]Department of Clinical Pathology and Oncology, Faculty of Veterinary Science, Szent István University, Budapest, Hungary. [5]Department of Experimental Pharmacology, National Institute of Oncology, Budapest, Hungary. [6]Institute of Organic Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary. [7]2nd Institute of Pathology, Semmelweis University, Budapest, Hungary. [8]Department of Biophysics and Radiation Biology, Semmelweis University, Budapest, Hungary and MTA-SE Molecular Biophysics Research Group, Budapest, Hungary. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to B.S. (email: sarkadi@biomembrane.hu) or A.A. (email: apati@biomembrane.hu) or T.I.O. (email: orbant@biomembrane.hu or orban.tamas@ttk.mta.hu)

ischemic kidney injuries resulting in poor outcome for kidney transplantations[4]. In fact, several drugs with various mechanisms of action had to be withdrawn from the market because of side effects caused by disruption of the calcium homeostasis, including Clobutinol, a cough suppressant[5], Dofetilide, an antiarrhythmic agent[6], Grepafloxacin and Sparfloxacin, antibacterial agents[7], Terfenadine, an antihistamine[8], or Terodiline, a spasmolytic agent[9]. All these findings suggest that in the process of drug discovery an early prediction of toxicity requires the direct examination of the drug effects on cellular calcium homeostasis and signaling in different target tissues, especially in the heart.

Animals stably expressing high-sensitivity cellular calcium indicator proteins are best suitable for direct examination of calcium signaling events in cells, tissues and organs as well. A well-established genetically engineered calcium sensor protein is the GCaMP2, containing a calmodulin-based sensor and a GFP-based fluorescent protein, which can be directly used to determine the changes in cellular calcium concentration[10]. The GCaMP2 protein has already been applied in tissue preparations and in transgenic mice[11–14], as well as in human pluripotent stem cells[15], allowing calcium imaging without additional manipulation. However, a calcium sensor expressing rat model has not been available yet.

Several methods are available for the transgenesis of rats, however, transposase-catalyzed gene delivery provides advantages, such as increased efficiency of chromosomal integration and single-copy insertion, while the system is less prone to genetic mosaicism and gene silencing than lentiviral gene delivery[16]. It has also been documented that the SB100X-mediated transgene integration allows the generation of transgenic lines with tissue-specific expression patterns, specified by selected promoter elements[17].

In the present work we have generated transgenic laboratory rats expressing the fluorescent calcium sensor protein GCaMP2. In order to achieve high-level expression in cardiac tissues, GCaMP2 expression in our model system is driven by a CAG promoter variant proved to be highly active in human embryonic stem cell-derived cardiomyocytes[18]. Additionally to cardiac tissues, characterization of homozygous CAG-GCaMP2 rats demonstrated appreciable GCaMP2 expression in kidney cortex, liver, and blood cells. CAG promoter specific GCaMP2 expression in blood cells allowed the development of a non-invasive, combined approach of genetic and phenotypic selection, yielding rat strains with high sensor protein expression, in spite of a mono-allelic transgene incorporation.

To validate the applicability of this model system in physiological and pharmacological studies, we used *ex vivo* and *in vitro* cardiomyocyte preparations to examine the effects of various ligands and potential drugs, such as the antimalarial agent, mefloquine, reported to disrupt the calcium homeostasis of heart tissue[19]; terodiline, causing prolongation of the QT interval and cardiac arrhythmia[20]; and terfenadine, known to prolong the QT interval through inhibition of the delayed rectifier potassium current of isolated rat ventricular myocytes[21]. Moreover, we examined the function of the Na^+/Ca^{2+} exchanger (NCX) by using an *in vitro* cellular hypoxia-reperfusion model, and found a rapid rise in cellular calcium during reoxygenation, blocked by an NCX inhibitor, KB-R7943. This finding further supports a major role of NCX, working in a reverse mode, in the calcium overload during reperfusion following ischemia[22], and that the inhibition of NCX may decrease calcium overload in ischemia/reperfusion (see[23]).

Results

Generation of a transgenic rat strain by combined genetic and phenotypic selection. To establish a rat strain with a single transgene copy per haploid genome, a combined genotype and phenotype screening procedure was applied. First, microinjected zygotes were implanted into pseudopregnant females to be carried to parturition. In order to promote integration and avoid concatemerization of the transgene, the transposon vector was applied as a circular plasmid, together with the SB100X transposase mRNA; the latter one was necessary to minimize mosaicism by providing a translatable source of the transposase during early cell divisions when the host genome is transcriptionally inactive.

Next, newborn rats were screened for transgene integration based on their tail tissue-derived gDNA, using SB transposon specific PCR primers. Transposase-induced insertion yielded a high percentage, about 21% of transgenic animals (16 carriers among 75 newborns). To avoid high transgene copy number related phenotypic alterations, low copy number carriers were selected at the F0 generation through determining average transgene copy numbers by a methodology developed earlier[24]. SB copy number distributions were more skewed than expected by earlier studies on cell lines[25,26]: apart from 3 extreme cases, the majority of the carriers had copy numbers below 2 (Fig. 1A). The relatively high number of genetically mosaic offspring (copy number <1) was also notable, although two chosen ones showed stable germ line transmission (Table 1).

Founders (F0) with one copy number were crossed with wild type rats to generate low copy number animals with stable transgene integration (F1 generation) (Table 1). Due to its constitutive feature, the CAG promoter driven expression of GCaMP2 was expected to be detectable in leukocytes. In order to accelerate the selection for the desired phenotype, F1 rats with single-copy integration were screened for functional GCaMP2 expression in leukocytes via FACS measurements (Fig. 1B). Single transgene carrier F1 rats with high sensor protein expression were further crossed with wild type counterparts to ensure stable genetic inheritance (F2 generation). GCaMP2 expression levels detected in leukocytes by FACS measurements were later also confirmed to be similar in primary cardiac cells isolated from the same animals (Supplemental Figure 1). Single-copy F2 rats were then inbred to generate homozygous offspring (F3 generation), carrying 2 copy transgenes, one per each haploid genome. Since the F3 generation contains not only homozygous, but also heterozygous and transgene negative animals, our

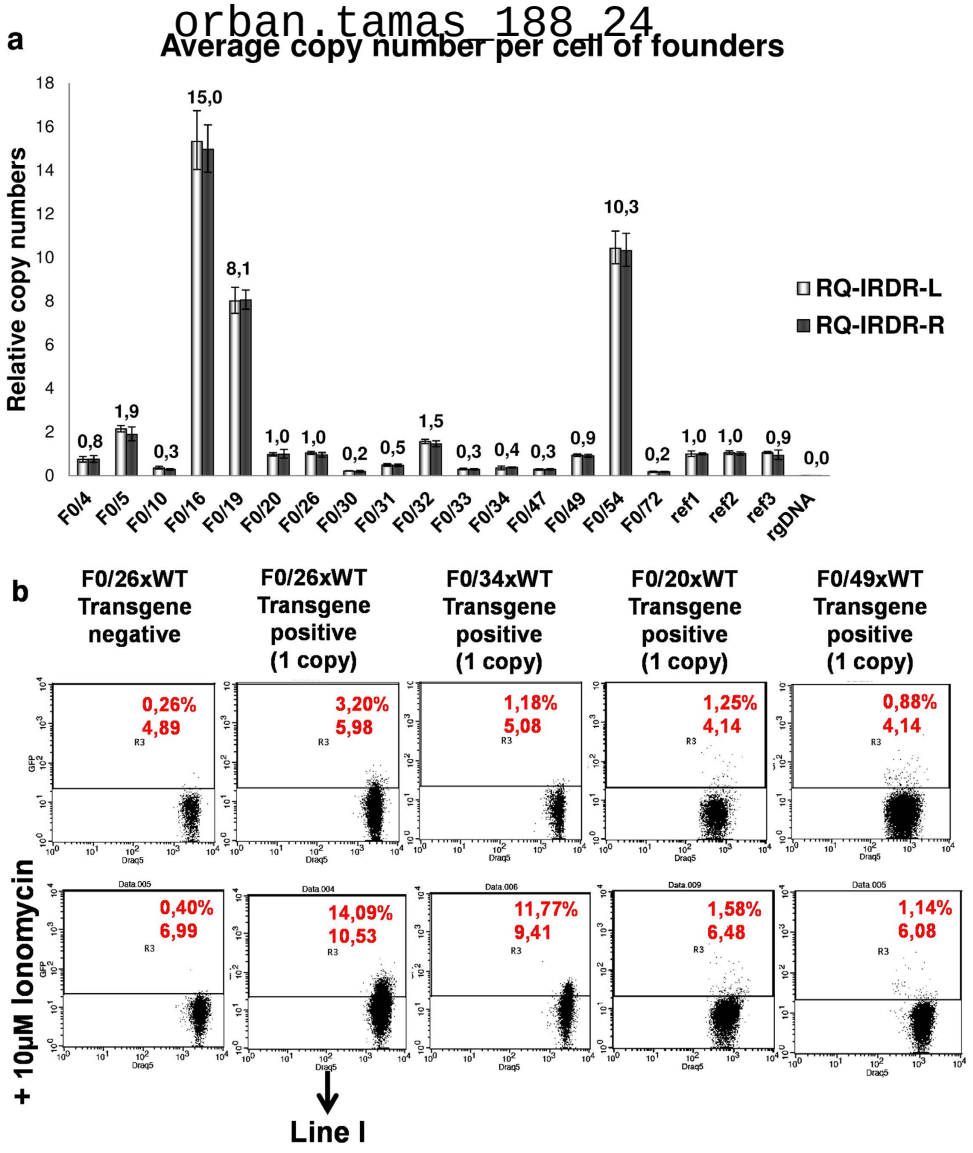


Figure 1. Genotype and phenotype analyses during the establishment of a CAG-GCaMP2 expressing transgenic rat strain. (**A**) Determining the average transgene copy numbers in rat tail tissue samples of the founder (F0) generation using a previously developed, real-time PCR based methodology, applying TaqMan® assays specific for the left and the right inverted repeat-direct repeat transposon sequences (RQ-IRDR-L/R, see Methods). Mean values ± confidence intervals are shown; ref1/2/3 stand for previously genotyped reference clones with one transposon copy, rgDNA is a non-transgenic rat tissue sample (as a negative control). (**B**) Flow cytometry analysis of leukocytes obtained from the blood of transgene positive F1 rats with single-copy integration. A transgene negative F1 rat was used as control. DRAQ5 was used to identify cells with nuclei (leukocytes). 10μM Ionomycin was used to confirm GCaMP2 expression by testing its functionality. Gated % and median values are indicated by red numbers on the histograms. (**C**) The genomic integration locus of the transgene in the homozygous rat strain as determined by splinkerette PCR.

	Wild type crossings of founders							
	(Line II.)		(Line I.)		(Line III.)		(Line IV.)	
Parents	♂	♀	♂	♀	♂	♀	♂	♀
	F0/20	WT	F0/26	WT	F0/34	WT	F0/49	WT
Average copy	**1.0**	–	**1.0**	–	**0.4**	–	**0.9**	–
F1	Progeny	Copy	Progeny	Copy	Progeny	Copy	Progeny	Copy
	F1/36	0	F1/68	1	F1/51	0	F1/17	1
	F1/37	1	F1/69	0	F1/52	0	F1/18	0
	F1/38	1	F1/70	1	F1/53	1	F1/19	0
	F1/39	0	F1/71	1	F1/54	0	F1/20	0
	F1/40	0	F1/72	0	F1/55	1	F1/21	1
	F1/41	0	F1/73	0	F1/56	0	F1/22	0
	F1/42	0	F1/74	1	F1/57	0	F1/23	0
	F1/43	2	F1/75	1	F1/58	1	F1/24	1
	F1/44	1	F1/76	0	F1/59	0	F1/25	1
	F1/45	0	F1/77	1	F1/60	0	F1/26	1
	F1/46	0	F1/78	1	F1/61	1	F1/27	0
	F1/47	0	F1/79	0	F1/62	0	F1/28	1
	F1/48	0	F1/80	0	F1/63	1	F1/29	0
	F1/49	0	F1/81	1	F1/64	0	F1/30	0
	F1/50	1	F1/82	0	F1/65	0	F1/31	0
			F1/83	1	F1/66	0	F1/32	1
			F1/84	0	F1/67	1	F1/33	1
			F1/85	0			F1/34	0
							F1/35	1

Table 1. Genotype analyses of the F1 generation offspring of four chosen F0 individuals crossed with wild type rats. Based on further genetic and phenotypic characterizations, 'Line I.' rat line was used to establish the homozygous rat strain with one copy CAG-GCaMP2 transgene per haploid genome. Of note that although the F0/34 founder of 'Line III.' was a genetic mosaic (average copy number below 1), it stably inherited one transgene copy in its germline (see F1 animals).

approach provided a special advantage at this point: homozygous rats could be selected by copy number determination, instead of crossing transgenic F3 rats with wild type rats for identification of homozygous ancestor.

Variance in the expression profile of the GCaMP2 protein through subsequent generations and among litter mates was monitored via FACS measurements of ionomycin induced changes in intracellular calcium levels of primary cells isolated from the left ventricular wall (Supplemental Figure 2). The established and constantly maintained homozygous rat strain ('Line I.' in Table 1) was proven to be genetically stable: transgene copy numbers and expression profiles were not altered through four subsequent generations. In addition, cell karyotypes were also stable and showed no difference in terms of chromosome numbers and integrity. Transgene integration locus was determined by a splinkerette-PCR method as described earlier[18]: the integration *per se* is not expected to influence gene expression profiles from the neighboring genetic loci (Fig. 1C).

Expression of the calcium sensor protein in homozygous transgenic rats. We examined if the expression of the calcium sensor caused any phenotypic alterations in transgenic (TG) rats. Brain, lung, spleen, kidney, liver and heart sizes were in all cases similar in the wild-type and homozygous TG rats (*data not shown*). Comparison of heart and body weights did not show any significant differences (heart weight/body weight ratios are (in mg/g): 4.808 ± 0.087 for WT and 4.854 ± 0.356 for TG, $n = 8$ animals, see Table 2), only a 15% increase was observed in the apical left ventricle (LV) thickness (close to apex, 2.66 mm $\pm$ 0.42 for WT and 3.1 mm $\pm$ 0.38 for TG, see Table 3), implying a possible LV apical hypertrophy. This result was confirmed by examination of hematoxylin and eosin stained tissue slices prepared from the left ventricle, where only a mild cardiomyocyte hypertrophy was found (Supplemental Figure 3). However, further studies showed no significant pathological alterations in O_2 saturation ($97\% \pm 0.71$ for WT and $95.3\% \pm 0.47$ for TG,), fractional shortening ($23.09\% \pm 6.59$ for WT and 29.74 ± 6.54 for TG, obtained by cardiac ultrasound measurements, see Methods) and LV wall thickness at base (close to

	WT	TG
n	8	8
body weight (g)	198 ± 29.399	213.375 ± 48.382
heart weight (mg)	953.625 ± 152.669	1032.750 ± 226.265
heart weight/body weight ratio (mg/g)	4.808 ± 0.087	4.854 ± 0.356

Table 2. Body and heart weight of wild type (WT) and transgenic (TG) rats. The data represents mean values ± S.E.M (n=8).

	Weight (g, SD)	O2 saturation (%, SD)	Fractional shortening (%, SD)	LV Thickness at base (mm, SD)	LV Thickness at apex (mm, SD)
WT (n = 4)	203.75 ± 14.74	97 ± 0.71	23.09 ± 6.59	3.38 ± 0.6	2.66 ± 0.42
TG (n = 3)	240.67 ± 22.9	95.33 ± 0.47	29.74 ± 6.54	3.57 ± 0.33	3.1 ± 0.38

Table 3. Comparison of body and heart parameters of wild type (WT) and CAG-GCaMP2 (TG) rats.

aorta, 3.38 mm ± 0.6 for WT and 3.57 mm ± 0.33 for TG) in GCaMP2-expressing rats, as compared to wild-type control (Table 3).

Fluorescence images of brain, lung, spleen, kidney, liver, atria and ventricles of TG and WT rats were compared using an *in vivo* imaging system to confirm GCaMP2 expression. Kidney, liver and heart, especially the atria were found to have high GCaMP2 fluorescence intensities (Fig. 2A), while GCaMP2 expression in the lung, brain and spleen failed to reach detectable levels (*data are shown only for lung*). The observed differences among the different tissues could have been caused either by different intracellular calcium levels or different GCaMP2 expression levels. Therefore, we examined GCaMP2 protein expression in samples obtained from different organs of TG rats by immunohistochemistry, and confirmed higher protein levels of GCaMP2 in the kidney, liver and heart samples while in brain, lung and spleen measurable GCaMP2 expression could not be detected (Fig. 2B,C). Functional expression of GCaMP2 could be confirmed in isolated primary cardiac (Supplemental Figure 2), renal[27] and hepatic cells (*unpublished observations by L. Homolya*). Consistent with higher protein levels of GCaMP2, higher mRNA expression could be confirmed in the kidney, liver and heart samples, while brain, lung and spleen were found to express low levels of the transgene mRNA (Supplemental Figure 4A). Low variance in mRNA expression levels among homozygous GCaMP2 rats was confirmed when two rats from the same litter were analyzed (Supplemental Figure 4A). This finding was in good agreement with the low variance of GCaMP2 expression measured by FACS in cardiac primary cells isolated from littermates (Supplemental Figure 2). Differences in protein expression intensities in the kidney between cortex and medulla or in the heart between atrium and ventricles detected by the fluorescent *in vivo* imaging system were also present at the transcriptional level (Supplemental Figure 4B and C).

Direct examination of calcium signaling in cardiomyocytes by using CAG-GCaMP2 transgenic rats. Acute ventricular slices were characterized for ionomycin-induced Ca^{2+} signals to confirm the functional expression of the GCaMP2 protein (Fig. 3A). In the TG tissue slices, intracellular Ca^{2+} overload could be evoked by 10 μM ionomycin, initiating a wave of calcium rise running through the slices, while intracellular calcium could be depleted by 10 mM EGTA (Supplemental movie 1). Ionomycin and EGTA administration also induced small changes in autofluorescence in the WT animal heart slices, however, these changes were two orders of magnitude smaller than the GCaMP2-dependent signals in the TG slices.

Freshly prepared ventricular slices often showed spontaneous contractions, resulting in fluorescence intensity changes in TG slices (Fig. 3B). Spontaneous contractions were localized within the tissue slices and did not generate major artifacts, as confirmed by the absence of changes in the autofluorescence signal in spontaneously contracting WT ventricular slices (Fig. 3B). Thus acute tissue slices can be properly used to follow the calcium transients and contractions in the TG heart tissue samples.

The TG CAG-GCaMP2 rat provides a stable and constant source of primary cardiac cells, suitable for a detailed examination of calcium signals in imaging-based drug screening applications. Therefore we examined the functionality of the GCaMP2 calcium indicator protein in primary cells isolated from the atria or from the ventricles. Cardiomyocyte-identity of the isolated cells was confirmed by immunostaining against the cardiac Troponin T and the cardiac isoform of the NCX protein (Supplemental Figure 5).

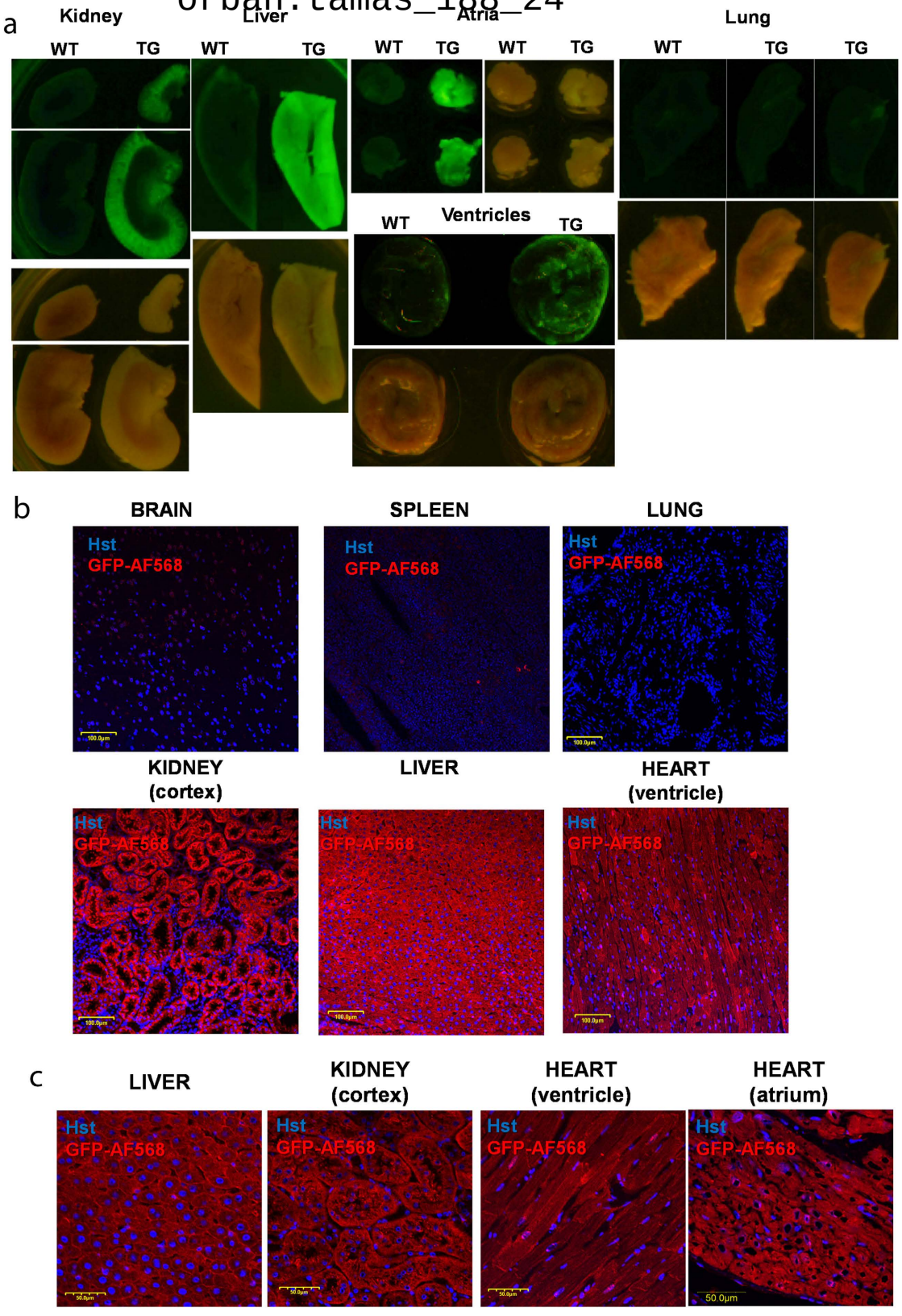


Figure 2. Expression of the GCaMP2 calcium sensor in transgenic rats. (**A**) Fluorescence (upper) and bright field (lower) images of wild type (WT) and transgenic (TG) kidney, liver, heart atria, ventricles and lung. (**B**) Immunohistochemistry of transgenic brain, spleen, lung, renal cortex, liver, and heart ventricle at lower magnification (scale bars represent 100 μm). (**C**) Immunohistochemistry of transgenic liver, renal cortex, heart ventricle and atrium at higher magnification (scale bars represent 50 μm). Immunostaining of the GCaMP2 was performed by a GFP-recognizing antibody. Hoechst (Hst) was used to stain the nuclei.

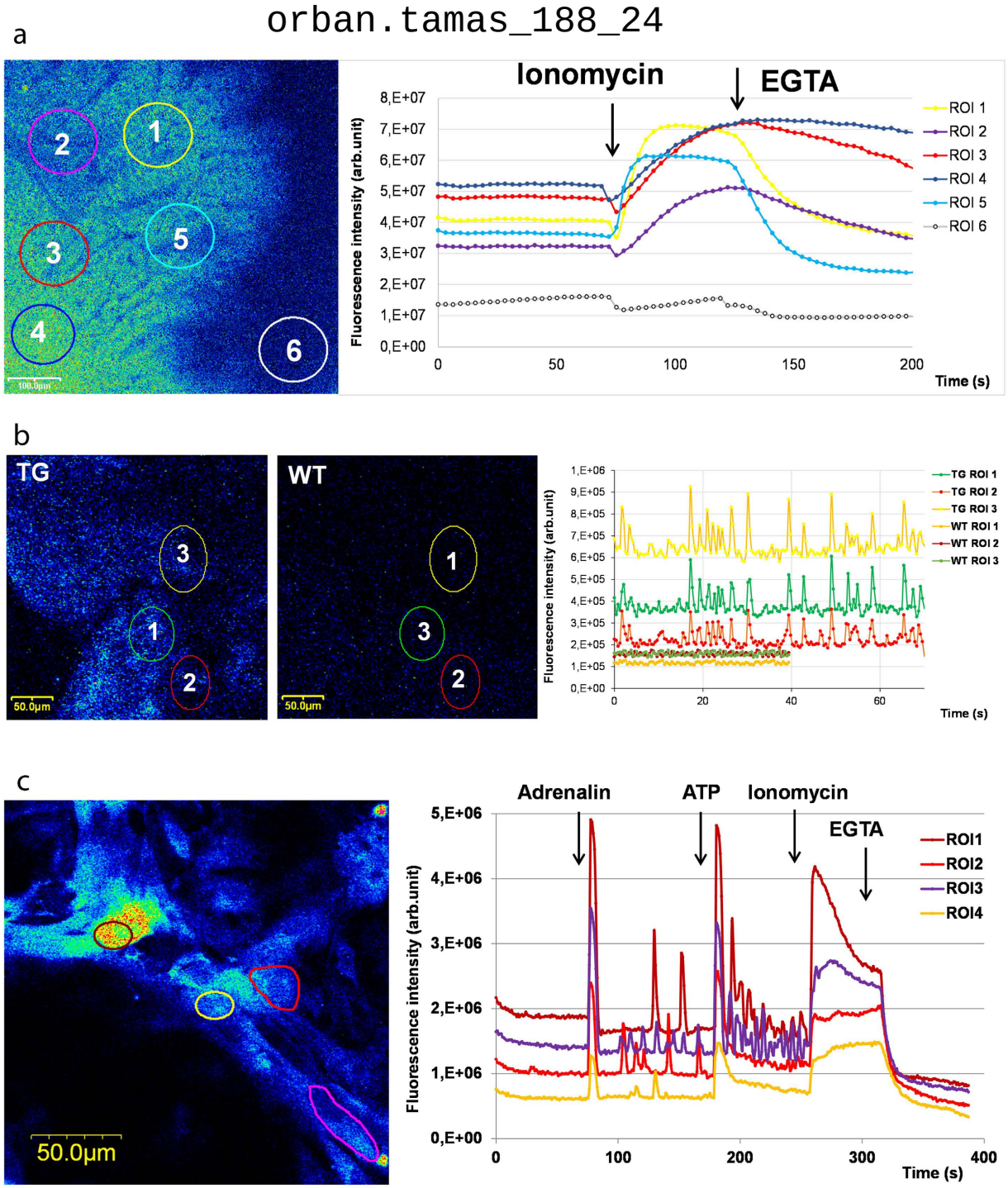


Figure 3. CAG-GCaMP2 expressing ventricular slices and cardiac cultures allow detection of spontaneous and ligand-induced calcium signals. (**A**) The left panel shows the fluorescence image of a ventricular slice obtained from a transgenic rat heart with circled region of interest (ROI). The right panel shows the effect of ionomycin and EGTA administration on the kinetics of calcium signals in the indicated ROIs. (**B**) The left panel shows the fluorescence images of transgenic (TG) and wild type (WT) ventricular slices with circled region of interest (ROI). The right panel shows the fluorescence signals caused by spontaneous calcium waves in TG and WT ventricular slices in the indicated ROIs. (**C**) Spontaneous calcium oscillations were induced by adrenalin and ATP in ventricular cell cultures generated from GCaMP2 expressing rats. Cell cultures obtained from the heart of TG rats were treated with adrenalin, ATP, ionomycin and EGTA. Physiologically relevant ligands such as adrenalin or ATP evoked a strong increase in intracellular Ca^{2+} levels and both compounds induced spontaneous calcium oscillations after addition.

In order to document the applicability and limitations of this system, isolated and cultured rat ventricular cardiomyocytes were characterized for spontaneous and ligand-induced Ca^{2+} signals by confocal microscopy. At high magnification, administration of 10 μg/ml adrenalin caused a well measurable oscillation, while 100 μM ATP further increased the frequency of oscillations in ventricular cells without reaching the limit of detection (Fig. 3C). Detection of calcium signals at lower magnification provided a more high throughput application by calculating the average of Ca^{2+} transients evoked by adrenalin and ATP in atrial and ventricular cultures, while oscillations of single cells was still detectable as shown by the high S.E.M. values (Fig. 4A). For validation of calcium signals evoked by adrenalin and ATP, similar measurements were carried out with Fluo4-AM loaded cardiomyocytes isolated from WT animals (Fig. 4B), resulting in similar calcium signals.

To document the applicability of this system for drug testing, we have examined the effect of 6 μM terfenadine and 10 μM terodiline on the basal calcium levels in ventricular myocytes. We used terfenadine instead of its active metabolite fexofenadine (terfenadine carboxylate), a potent antihistamine agent because it was reported that terfenadine itself is the likely cause of the cardiotoxic effect[8,28]. Treatment with 10 μM terodiline caused elevation in basal calcium levels, moreover, terodiline pre-treatment resulted in a significant increase in the amplitude of adrenalin-induced calcium transients (Fig. 4C). The ATP-evoked calcium influx was also elevated when ventricular cultures were pre-treated with terodiline or terfenadine, while terfenadine had only moderate effect on the adrenalin-induced calcium influx (Supplemental Figure 6).

We used the GCaMP2 rat system to examine the effect of an antimalarial, and potentially cardiotoxic agent, mefloquine, on the calcium signal responses in the primary cardiomyocyte cultures. As shown (Fig. 4D), higher mefloquine concentrations induced an increase in cellular calcium levels both in the atrial and ventricular CMs, but mefloquine had different effects on the ligand-induced signals. Atrial CMs showed a decrease in ligand-induced Ca^{2+} influx, however, the upstroke kinetics of the calcium transients were not significantly altered (Supplemental Figure 7A and B). Ligand-induced calcium signals in ventricular CMs were affected in their upstroke kinetics by mefloquine, as shown in Supplemental Figure 7A and B, that is mefloquine-treated ventricular CMs showed higher 'time to peak' values as compared to control ventricular CMs. It also appeared that adrenalin induced calcium signals were more affected by mefloquine than the ATP-induced signals.

We found that a 24 hours pre-treatment with near therapeutic (2.6 μM) concentrations of mefloquine did not increase the basal calcium levels in cardiomyocytes, however, calcium signals induced by adrenalin or ATP were similarly altered as after the immediate addition of 37.5 μM mefloquine (Supplemental Figure 7A).

After the occurrence of a cardiac hypoxia, (e.g.: by acute myocardial infarction), reperfusion causes major cell death in patients mostly through an altered calcium handling[29]. To mimic ischemia (hypoxia) and reperfusion (reoxygenation) at an *in vitro* cellular level, we pre-treated rat primary ventricular cells with 400 μM $CoCl_2$, and after 24 hours we rapidly exchanged the medium for an oxygenated one, containing no $CoCl_2$. Reoxygenation caused a significant calcium increase, while the medium change without $CoCl_2$ pre-treatment did not alter the intracellular calcium levels (Fig. 5A,B).

At the time of reperfusion/reoxygenation, following the ATP depletion and intracellular Na^+ increase produced by hypoxia, the reverse-mode activation of the NCX has been suggested to be a key factor in cellular calcium loading[22]. Thus the inhibition of NCX and the related calcium influx during initial reperfusion may reduce myocardial injury[23]. In order to explore this effect, different concentrations of the NCX inhibitor, KB-R7943, were added to the primary ventricular cells at the time of reoxygenation. This treatment, in a concentration dependent manner, decreased or eliminated the rise in cellular calcium due to reoxygenation (Fig. 5C,D), although the amplitude of ATP-driven Ca^{2+} influx was also lowered at the higher concentration.

Discussion

In the present study we describe the generation and validation of a new GCaMP2 expressing rat strain with a single transgene copy per haploid genome. This transgene model was generated by a *Sleeping Beauty* transposon-based method, and the transgene expression is driven by a CAG promoter variant[18]. Utilizing the constitutive nature of this promoter, a combined genotype/phenotype screening approach was applied to establish a homozygous rat strain in a relatively short period of time. In accordance with our earlier findings with embryonic stem cells[30], the highest GCaMP2 expression was found in the cardiac tissues, especially in the atria, while appreciable transgene expression was found in the kidney and in the liver. Anatomical analyses found no differences in the heart size and weight between WT and TG animals, while slight left ventricular wall thickening was observed at the apex of the heart in transgenic rats. However, this did not cause any alterations in the heart function and other circulatory parameters, since physiological parameters of the CAG-GCaMP2 expressing rats were similar to the WT rats.

As opposed to the traditional detection methods of filling up the cells with a calcium-sensitive, often toxic fluorescent dyes, the genetically encoded and stably expressed calcium indicators have a clear advantage of providing a non-invasive platform for calcium measurements. Although new generation calcium sensors (e.g. GCaMP3–8) are already available[31–34], most of them have higher rates of induced cytomorbidity which limits their applications[35]. In addition, they were developed to follow action potentials of neurons more reliably, while GCaMP2 offers an effective system of Ca^{2+} signal measurements

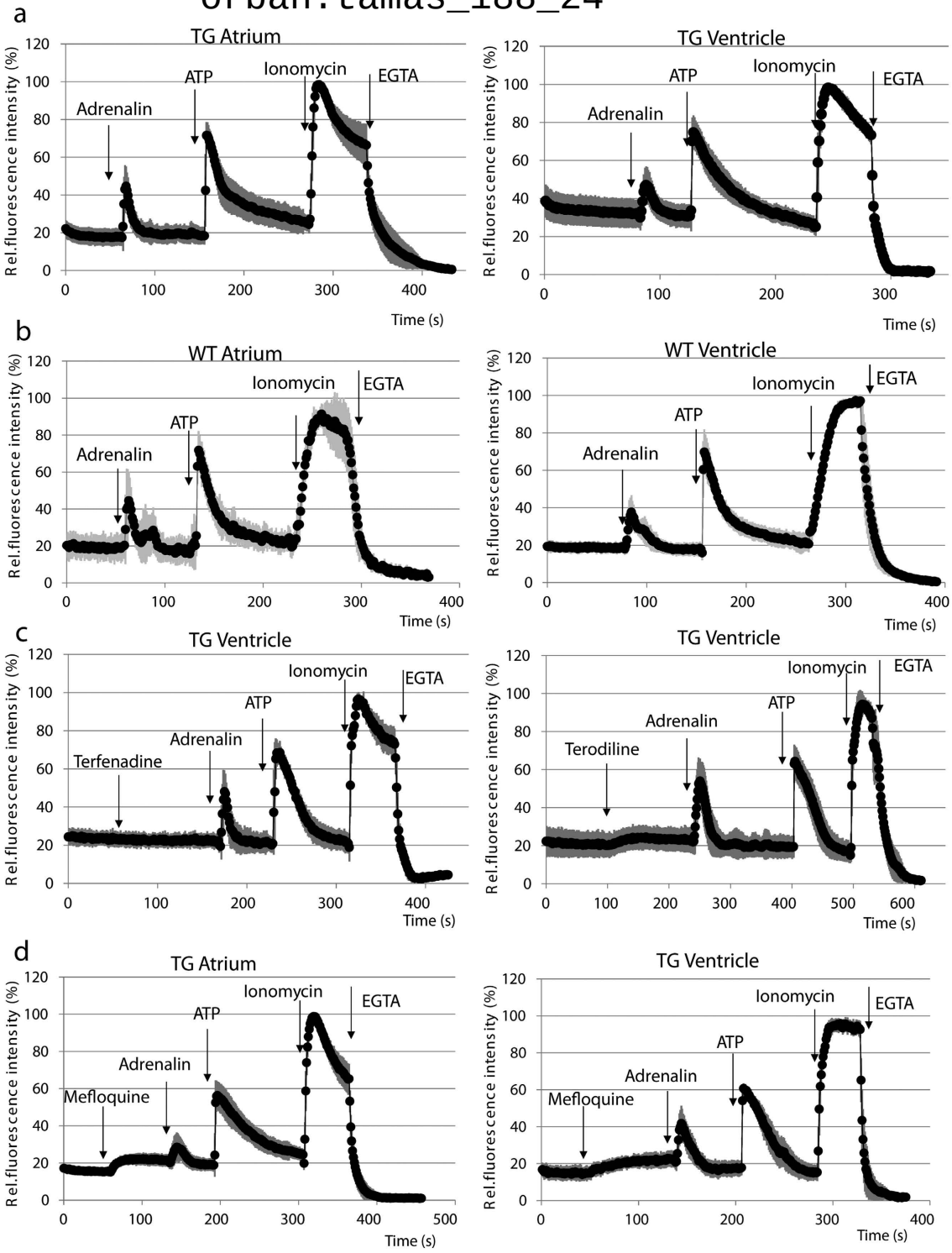


Figure 4. CAG-GCaMP2 expressing cardiomyocyte cultures allow detection of ligand-induced calcium signals and drug-induced alteration of physiologically relevant ligand-induced calcium signals. (**A**) The effects of adrenalin, ATP, ionomycin and EGTA administration are shown on the kinetics of calcium signals in atrial (left panel) and in ventricular (right panel) cardiomyocyte cultures. (**B**) Ligand-induced calcium signals in wild type (WT) primary cardiac cultures loaded with the calcium-sensitive Fluo-4 AM dye. (**C**) The effects of 6.0 μM terfenadine (left panel) or 10.0 μM terodiline (right panel) followed by adrenalin, ATP, ionomycin and EGTA administration are shown on the kinetics of calcium signals in ventricular cell cultures. (**D**) The effects of 37.5 μM mefloquine, as well as adrenalin, ATP, ionomycin and EGTA administration are shown on the kinetics of calcium signals in atrial (left panel) and in ventricular (right panel) cell cultures. For each measurement, a total number of at least 25 cells was examined; light gray lines represent S.E.M. values.

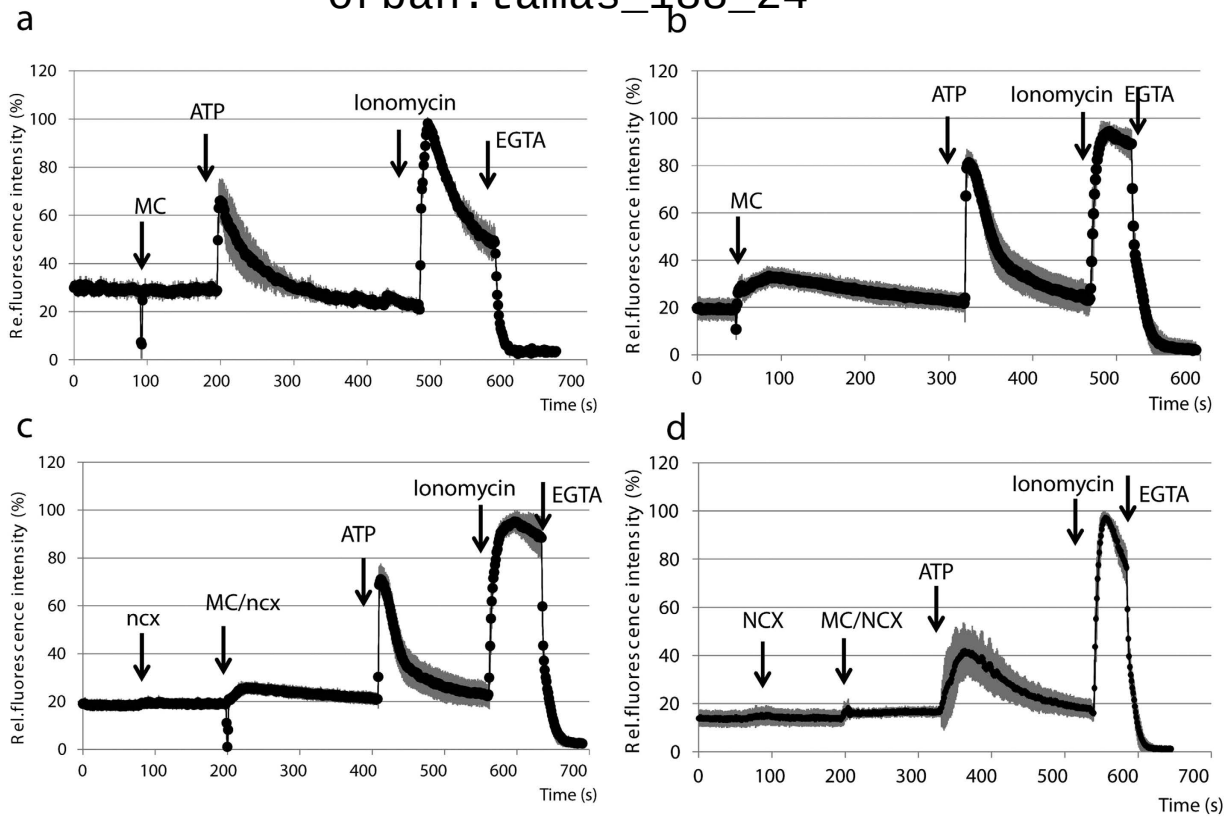


Figure 5. Modeling hypoxia/reoxygenation in CAG-GCaMP2 expressing ventricular cultures allows detection of the protecting effect of KB-R7943. (**A**) Control measurement in ventricular cultures without $CoCl_2$ pre-treatment, showing no changes in fluorescence by medium change (MC). (**B**) In $CoCl_2$-pretreated ventricular cultures, reoxygenation through medium change (MC) causes a significant calcium increase in hypoxic ventricular cells. (**C**) In $CoCl_2$-pretreated ventricular cultures, administration of 0.5 μM KB-R7943 (ncx) reduced the calcium influx caused by reoxygenation. (**D**) In $CoCl_2$-pretreated ventricular cultures, administration of 1 μM KB-R7943 (NCX) eliminates the calcium influx caused by reoxygenation.

in all other tissues with potentially less toxicity and side effects. The physiological relevance of the calcium sensor expressing rat model was verified by the characterization of ligand-induced Ca^{2+} signals both *ex vivo* and *in vitro*. In acute ventricular slices, when spontaneous contractions were observed, the GCaMP2 fluorescence intensity increased parallel with the contractions. Experiments in primary cell cultures isolated from TG rats were performed using both atrial and ventricular cells. As documented in Supplemental 5, these cultured cells were mainly troponin T (cTnT) positive cardiomyocytes, and also showed high expression of the calcium sensor protein GCaMP2.

In this report we have examined the effect of an antimalarial agent, mefloquine, on the calcium signal responses in isolated cardiomyocytes. Mefloquine, a widely used agent to prevent and treat *Plasmodium falciparum* malaria, has a suspected cardiotoxicity. Mefloquine alone did not cause significant alterations in a healthy heart muscle[36], while this compound induced changes in the heart calcium levels through the modulation of the L-type Ca^{2+} current[19]. In addition, pre-treatment with mefloquine potentiated the effects of halofantrine on the prolongation of QT intervals in anaesthetized rabbits[37].

These studies raised warnings against the use of mefloquine in the case of cardiac problems. Our present experiments reported here indicate that in the rat cardiomyocyte model a long (24 hours) pre-treatment with low (near therapeutic) concentrations of mefloquine did not increase the basal calcium level, but considerably decreased the calcium signals induced by adrenaline or ATP, especially in the atrial cells (Supplemental Figure 7A and B). Higher mefloquine concentrations induced a significant increase in intracellular calcium levels (Fig. 4D). Moreover, mefloquine treatment also altered the properties of Ca^{2+} transients both in atrial and in ventricular cells (Supplemental Figure 7A and B). These findings raise the possibility for a direct effect of mefloquine on cardiomyocyte calcium homeostasis. In addition, two drugs, terodiline and terfenadine withdrawn from the market due to QT interval prolonging effect, were used to validate the CAG-GCaMP2 rat system as a potent tool for predicting cardiotoxic side effects. Both terodiline[38] and terfenadine inhibits the rapid component of the delayed-rectifier K^+

current, therefore the increase in calcium influx detected may be a consequence of the reduction of active K^+ channels, normally compensating calcium influx with K^+ efflux in ventricular cardiomyocytes.

Our results, obtained in an *in vitro* hypoxia-reoxygenation model, were in agreement with previous studies showing that cellular Ca^{2+} levels increase at the time of reoxygenation (Fig. 5B)[22,29], and this elevation can be eliminated by the NCX inhibitor KB-R7943 (Fig. 5D)[23]. Moreover, we found a reduced calcium signal induced by ATP after the addition of 1 μM KB-R7943, showing the potential role of NCX in the calcium signaling of ventricular cells.

This study is the first publication dealing with the generation of a TG rat strain, stably expressing a calcium sensor fluorescent protein. While mouse models expressing genetically engineered calcium indicators have already been described[14,39] the present rat model with calcium sensor expression may significantly improve the experimental background examining these phenomena in a widely used system for pharmacological and toxicological studies. The transgenic rats expressing a calcium-sensitive protein combined with recently developed, genetically engineered disease-specific rat strains (see SAGE, KO Rat Consortium, http://www.sageresearchlabs.com/research-models), could provide a valuable system for further elucidation of the mechanisms and potential treatments of heart diseases.

Methods

All animal protocols were approved by the Hungarian Animal Health and Animal Welfare Directorate according to the European Union's most recent directives. All surgical procedures were performed according to the Committee on the Care and Use of Laboratory Animals of the Council on Animal Care at the National Institute of Oncology in Budapest, Hungary (22.1/722/3/2010).

Rat transgenesis. To establish a rat strain with a single transgene copy per haploid genome (two copies per animal), a combined genotype and phenotype screening procedure was applied (see Results). Transgenic rats were derived by microinjection of zygotes from superovulated donor females of an outbred Sprague-Dawley strain, and transgenic embryos were implanted into pseudopregnant Wistar female rat recipients (experiment license number in Hungary: PEI/001/2197-2/2013). Circular transposon plasmid vector[15] was applied together with the SB100X transposase mRNA. The latter translation competent mRNA was prepared by *in vitro* transcription using the mMessage mMachine® T7 kit (Life Technologies). For genotyping, genomic DNA (gDNA) was extracted from rat tails as described earlier[17], and the following PCR primers were used to detect the transposon: 5′- AATTCCCTGTCTTAGGTCAGTTAGGA and 5′- TTCAGGTTCAGGGGGAGGTGTGGG. As a rat gDNA control, primers 5′- AAGATTGAATGTCT GTAAGTTCGAG and 5′- TGTAATTGGTTTGGGGTTAT were used to amplify a sequence from intron 1 of the RN-LOC500546 gene. SB transposon copy numbers were determined as described earlier[24].

Organ processing. Animals were sacrificed by overdosing a 4-component anesthetic mixture (20 mg/kg zolazepam, 12.5 mg/kg xylazine, 3 mg/kg butorphanol, 20 mg/kg tiletamine). The organs were removed, washed twice in PBS and were either archived for later use by snap freezing with Isopentane or fluorescent images were made by placing the organs into an *in vivo* imaging system (LT-9MACIMSYSPLUSC, Lightools Research). For experiments with ventricular slices the left ventricle was cut into 300 μm thin section with a McIlwain tissue chopper (Mickle Laboratory Engineering Comp.).

Structural and functional characterization of cardiac muscle in transgenic and wild type animals. Fractional shortening, basal and apical thickness of the left ventricle (LV) and oxygen saturation were measured in both transgenic and wild type rats. The end-diastolic and end-systolic dimensions (EDD and ESD) of the LV were measured in M-mode by a GE LOGIQ 5 Pro ultrasound system (General Electric Healthcare) under 4% isoflurane anaesthesia, which was introduced directly without previous prenarcotic pretreatment, and fractional shortening (FS) was calculated using the following equation:

$$\frac{EDD - ESD}{EDD} \times 100\%$$

All ultrasound examinations were performed in a 5 minutes interval after anaesthesia to avoid possible pathological alterations due to anesthetic inhalation, and O_2 saturation was monitored in the hind paw with a pulse oximeter (Pulse-BLT M700 Vet). After echocardiography the animals were euthanized with a single intracardial injection of barbiturate (Euthasol® 40% inj. A.U.V., Produlab Pharma b.v.), the heart was removed, and the basal (aprox. 3 mm from the aortic valve) and apical (3 mm from apex of the heart) LV wall thickness was measured by a caliper at least twice. The hearts were fixed in formalin, embedded in paraffin and haematoxylin-eosin stained for micro-structural analysis.

Real-time PCR analysis. Frozen rat samples were pulverized under liquid nitrogen and were homogenized in TRIzol™ Reagent (Life Technologies). Total RNA was isolated from tissue samples and RNA degradation was monitored by gel electrophoresis. cDNA samples were prepared from 1 μg

total RNA using the Promega Reverse Transcription System Kit. The Pre-Developed TaqMan® assay for beta-2-microglobulin (B2M) (Life Technologies) was used as endogenous control in real-time qPCR experiments; for quantifying EFGP mRNA, specific TaqMan® assay was designed for the cDNA[18]. Real time PCR analyses were carried out using the StepOne™ Real-Time PCR System (Life Technologies); mRNA fold changes were determined using the $2^{-\Delta\Delta Ct}$ method. Relative mRNA levels were presented as mean values ± S.E.M. of 3 independent experiments.

Generation of primary cell cultures from cardiac atria and ventricles. Primary cell cultures were established from the atria or the ventricles of 10–30 days old male TG and wild type rats. Briefly, the left and right atria were separated from the ventricles and were cut into ~1 mm^3 pieces, similarly to the walls of the right and left ventricles. The pieces of the atria or the ventricles were transferred to a 50 ml conical tube containing 20 ml enzyme mixture of 200 U/ml type IV collagenase and 0.6 U/ml dispase (Life Technologies), vortexed for 1 minute then digested for 15 minutes at 37 °C. After 15 minutes the larger pieces were settled, the supernatant was transferred to a sterile conical tube, centrifuged at 300 × g for 5 minutes and the pellet was dissolved in pure fetal bovine serum (FBS). The remaining tissue pieces were re-digested with fresh enzyme mix. This procedure was repeated until no macroscopic pieces were observed. The cell suspension was filtered through a 70 micron nylon filter insert (BD Biosciences), centrifuged for 5 minutes at 300 × g, the supernatant was removed and the pellet was dissolved in complete DMEM (2% penicillin/streptomycin + 1% glutamine + 10% FBS, 25 mM HEPES). Cells in 300 μL of complete DMEM were plated onto 8-well Nunc Lab-Tek II Chambered Cover glass (Nalgen Nunc International) covered with 0.1% gelatin. After 90 minutes the supernatant from all chambers were transferred to a clean, gelatin coated chamber to enrich cardiomyocytes and decrease the number of fibroblasts. For experiments only secondary chambers were used and culture media were changed on every second day.

Immunohistochemistry. Snap-frozen tissues were cut with a cryostat into 5 μm thin sections. The tissue slices were transferred onto microscope slides and were incubated with 10 mM EGTA for 5 minutes before fixation with methanol. Tissue sections were then washed in PBS, blocked with 2% bovine serum in PBS (1 hr) and incubated with a rabbit polyclonal antibody (1:500) targeting GFP (Abcam, ab290) and with a mouse monoclonal antibody (1:100) against cardiac Troponin T (Abcam, ab8295) in 0.5% bovine serum in PBS overnight at 4 °C. After a 90 minutes of PBS wash, tissue slices were incubated (1 hr) with Alexa Fluor 488 or Alexa Fluor 568-conjugated secondary antibody against rabbit IgG and with Alexa Fluor 568-conjugated secondary antibody against mouse IgG (all by 1:200; Life Technologies), washed in PBS and the nuclei were stained with Hoechst 33342 dye. Stained sections were mounted with ProLong Gold (Life Technologies). Wild type tissues were used as negative controls. Immunohistochemistry images were examined by an Olympus FV500-IX confocal laser scanning microscope.

Immunocytochemistry. Seven day old cultures of primary cells isolated from rat atrium or ventricle were incubated with 10 mM EGTA for 5 minutes before methanol fixation and washed with PBS. Blocking and immunostaining were performed as described above, nuclei were stained with DAPI. For staining of the Na^+/Ca^{2+} exchanger (NCX) the R3F1 antibody was used, kindly provided by Michela Ottolia from the Cedars-Sinai Heart Institute. Isotype and secondary antibody controls were used, and image detection was performed as described above.

Flow cytometry measurements (FACS). Leukocytes were obtained through red blood cell lysis, while primary ventricular cells were obtained as described earlier. The isolation of primary rat hepatocytes were performed by a three-step collagenase perfusion method[40]. Renal cells were isolated from the cortex after decapsulation and separation from the renal medulla, followed by enzymatic digestion steps as described previously[27]. Measurements were carried out in HBSS (Life Technologies). In the case of blood samples DRAQ5 was used to identify cells with nuclei (leukocytes) and 10 μM Ionomycin was used to confirm GCaMP2 expression by testing its functionality. For primary ventricular, renal cortical cells, and hepatocytes 5 μM Ionomycin was used to confirm GCaMP2 functionality. Forward scattered light (FSC channel), GCaMP2 expression (FL1 channel) and DRAQ5 staining (FL4 channel) were detected by a BD FACSCalibur flow cytometer.

Calcium signal measurements. GCaMP2 expressing primary cardiomyocytes were seeded onto eight-well Nunc Lab-Tek II Chambered Coverglass (Nalge Nunc International) covered by 0.1% gelatine. Culture media was removed by changing the medium to HBSS. In the case of *ex vivo* measurements 300 μm thin ventricular sections were placed into HBSS containing MatTek glass-bottom dishes. Ligand concentrations were chosen according to literature: ATP (100 μM), adrenalin (10 μg/ml) or ionomycin (5 μM for *in vitro* cultures and 10 μM for tissue slices). For calibration, EGTA was used in 5x excess of calcium concentration in the medium[41]. Calcium signal measurements were carried out by following time lapse sequences of cellular fluorescence recordings, and images were analyzed with the FluoView Tiempo (v4.3, Olympus) software. For GCaMP2 imaging, the 488 nm laser line was used for excitation and emission was measured between 505 and 535 nm. In confocal images artificial coloring was used for better visualization. Fluo-4 imaging was carried out as described previously[41].

Induction of hypoxia-reoxygenation in primary cells. To mimic hypoxia in cultured cardiomyocytes, 400 μM of cobalt chloride ($CoCl_2$) was added in 500 μL phenol-free completed media and incubated in normoxia (21% O_2). 24 hours later the cells were transferred to the confocal microscope stage and the baseline cellular calcium levels were measured. Thereafter the culture medium was rapidly changed to phenol-free HBSS without $CoCl_2$, and changes in fluorescent intensities were detected continuously.

References

1. Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodelling. *Nat Rev Mol Cell Biol* **4,** 517–529 (2003).
2. Hummel, Y. M. *et al.* Ventricular dysfunction in a family with long QT syndrome type 3. *Europace* **15,** 1516–1521 (2013).
3. Liu, N. & Priori, S. G. Disruption of calcium homeostasis and arrhythmogenesis induced by mutations in the cardiac ryanodine receptor and calsequestrin. *Cardiovasc Res* **77,** 293–301 (2008).
4. Edelstein, C. L. Calcium-mediated proximal tubular injury-what is the role of cysteine proteases? *Nephrol Dial Transplant* **15,** 141–144 (2000).
5. Bellocq, C. *et al.* A common antitussive drug, clobutinol, precipitates the long QT syndrome 2. *Mol Pharmacol* **66,** 1093–1102 (2004).
6. Srivastava, S. *et al.* Paradoxical effect of dofetilide on action potential duration and calcium transient amplitude in newborn rabbit ventricular myocytes. *J Cardiovasc Pharmacol* **45,** 165–174 (2005).
7. Patmore, L., Fraser, S., Mair, D. & Templeton, A. Effects of sparfloxacin, grepafloxacin, moxifloxacin, and ciprofloxacin on cardiac action potential duration. *Eur J Pharmacol* **406,** 449–452 (2000).
8. Roy, M., Dumaine, R. & Brown, A. M. HERG, a primary human ventricular target of the nonsedating antihistamine terfenadine. *Circulation* **94,** 817–823 (1996).
9. Ogura, T., Jones, S., Shuba, L. M., McCullough, J. R. & McDonald, T. F. Block and modified gating of cardiac calcium channel currents by terodiline. *Br J Pharmacol* **127,** 1837–1845 (1999).
10. Zhao, Y. *et al.* An expanded palette of genetically encoded Ca(2)(+) indicators. *Science* **333,** 1888–1891 (2011).
11. Bagher, P. & Segal, S. S. The mouse cremaster muscle preparation for intravital imaging of the microcirculation. *J Vis Exp* (2011).
12. Diez-Garcia, J., Akemann, W. & Knopfel, T. *In vivo* calcium imaging from genetically specified target cells in mouse cerebellum. *Neuroimage* **34,** 859–869 (2007).
13. Peppiatt-Wildman, C. M., Crawford, C. & Hall, A. M. Fluorescence imaging of intracellular calcium signals in intact kidney tissue. *Nephron Exp Nephrol* **121,** e49–58 (2012).
14. Tallini, Y. N. *et al.* Imaging cellular signals in the heart *in vivo*: Cardiac expression of the high-signal Ca2+ indicator GCaMP2. *Proc Natl Acad Sci USA* **103,** 4753–4758 (2006).
15. Apati, A. *et al.* Characterization of calcium signals in human embryonic stem cells and in their differentiated offspring by a stably integrated calcium indicator protein. *Cell Signal* **25,** 752–759 (2013).
16. Mates, L. Rodent transgenesis mediated by a novel hyperactive Sleeping Beauty transposon system. *Methods Mol Biol* **738,** 87–99 (2011).
17. Katter, K. *et al.* Transposon-mediated transgenesis, transgenic rescue, and tissue-specific gene expression in rodents and rabbits. *Faseb J* **27,** 930–941 (2013).
18. Orban, T. I. *et al.* Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery. *Stem Cells* **27,** 1077–1087 (2009).
19. Coker, S. J., Batey, A. J., Lightbown, I. D., Diaz, M. E. & Eisner, D. A. Effects of mefloquine on cardiac contractility and electrical activity *in vivo*, in isolated cardiac preparations, and in single ventricular myocytes. *Br J Pharmacol* **129,** 323–330 (2000).
20. Stewart, D. A. *et al.* Terodiline causes polymorphic ventricular tachycardia due to reduced heart rate and prolongation of QT interval. *Eur J Clin Pharmacol* **42,** 577–580 (1992).
21. Ohtani, H. *et al.* Inhibitory effects of the antihistamines epinastine, terfenadine, and ebastine on potassium currents in rat ventricular myocytes. *J Pharm Pharmacol* **51,** 1059–1063 (1999).
22. Schafer, C. *et al.* Role of the reverse mode of the Na+/Ca2+ exchanger in reoxygenation-induced cardiomyocyte injury. *Cardiovasc Res* **51,** 241–250 (2001).
23. Inserte, J. *et al.* Effect of inhibition of Na(+)/Ca(2+) exchanger at the time of myocardial reperfusion on hypercontracture and cell death. *Cardiovasc Res* **55,** 739–748 (2002).
24. Kolacsek, O. *et al.* Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. *Mob DNA* **2,** 5 (2011).
25. Grabundzija, I. *et al.* Comparative analysis of transposable element vector systems in human cells. *Mol Ther* **18,** 1200–1209 (2010).
26. Kolacsek, O. *et al.* Excision Efficiency Is Not Strongly Coupled to Transgenic Rate: Cell Type-Dependent Transposition Efficiency of Sleeping Beauty and piggyBac DNA Transposons. *Hum Gene Ther Methods* **25,** 241–252 (2014).
27. Szebenyi, K. *et al.* Visualization of Calcium Dynamics in Kidney Proximal Tubules. *J Am Soc Nephrol* (2015).
28. Gibson, J. K., Yue, Y., Bronson, J., Palmer, C. & Numann, R. Human stem cell-derived cardiomyocytes detect drug-mediated changes in action potentials and ion currents. *J Pharmacol Toxicol Methods* **70,** 255–267 (2014).
29. Garcia-Dorado, D., Ruiz-Meana, M., Inserte, J., Rodriguez-Sinovas, A. & Piper, H. M. Calcium-mediated cell death during myocardial reperfusion. *Cardiovasc Res* **94,** 168–180 (2012).
30. Szebenyi, K. *et al.* Efficient generation of human embryonic stem cell-derived cardiac progenitors based on tissue-specific EGFP expression. *Tissue Eng Part C Methods* **21,** 35–45 (2014).
31. Akerboom, J. *et al.* Optimization of a GCaMP calcium indicator for neural activity imaging. *J Neurosci* **32,** 13819–13840 (2012).
32. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
33. Dana, H. *et al.* Thy1-GCaMP6 transgenic mice for neuronal population imaging *in vivo*. *PLoS One* **9,** e108697 (2014).
34. Ohkura, M. *et al.* Genetically encoded green fluorescent Ca2+ indicators with improved detectability for neuronal Ca2+ signals. *PLoS One* **7,** e51286 (2012).
35. Weitz, A. C. *et al.* Imaging the response of the retina to electrical stimulation with genetically encoded calcium indicators. *J Neurophysiol* **109,** 1979–1988 (2013).
36. Kinoshita, A., Yamada, H., Kotaki, H. & Kimura, M. Effects of anti-malarial drugs on the electrocardiographic QT interval modelled in the isolated perfused guinea pig heart system. *Malar J* **9,** 318 (2010).
37. Lightbown, I. D., Lambert, J. P., Edwards, G. & Coker, S. J. Potentiation of halofantrine-induced QTc prolongation by mefloquine: correlation with blood concentrations of halofantrine. *Br J Pharmacol* **132,** 197–204 (2001).
38. Jones, S. E., Ogura, T., Shuba, L. M. & McDonald, T. F. Inhibition of the rapid component of the delayed-rectifier K+ current by therapeutic concentrations of the antispasmodic agent terodiline. *Br J Pharmacol* **125,** 1138–1143 (1998).

39. Direnberger, S. *et al.* Biocompatibility of a genetically encoded calcium indicator in a transgenic mouse model. *Nat Commun* **3,** 1031 (2012).
40. Lengyel, G., Veres, Z., Szabo, P., Vereczkey, L. & Jemnitz, K. Canalicular and sinusoidal disposition of bilirubin mono- and diglucuronides in sandwich-cultured human and rat primary hepatocytes. *Drug Metab Dispos* **33,** 1355–1360 (2005).
41. Pentek, A., Paszty, K. & Apati, A. Analysis of Intracellular Calcium Signaling in Human Embryonic Stem Cells. *Methods Mol Biol* (2014).

Acknowledgements

The authors are grateful to Irén Bodrogi-Mayer, Anita Hídvégi, Gabriella Szoták and Katalin Parragné Derecskei for the excellent technical assistance and to Ildikó Pál for preparing the ventricular slices. We are grateful to János Matkó for imaging facility support. We appreciate the gift of R3F1 antibody against the Na^+/Ca^{2+} exchanger (NCX), kindly provided by Michela Ottolia from the Cedars-Sinai Heart Institute. This work was supported by the Hungarian Scientific Research Fund [grant number K112112, NK83533, K84173, PD101733]; by the National Development Agency [grant numbers KTIA_AIK_12-1-2012-0025, KMR_12-1-2012-0112].

Author Contributions

K.Sz. and A.F. designed and performed experiments, analyzed the data and wrote the manuscript; O.K., E.P. and T.I.O. generated the plasmids and genotyped the TG rat; Zs.B. and B.B. generated the TG rat strain and edited the manuscript; P.V. and J.T. collected and analyzed the data and edited the manuscript; L.Ho., G.Sz., L.Hé. and E.Á. analyzed the data and discussed the interpretation of the results; T.I.O., Á. A. and B.S. designed the concept of the study, analyzed the data and wrote the paper. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: A patent application (number P1400500) has been filed on the generation of a transgenic laboratory rat strain which partially relates to the data described in the paper.

How to cite this article: Szebényi, K. *et al.* Generation of a Homozygous Transgenic Rat Strain Stably Expressing a Calcium Sensor Protein for Direct Examination of Calcium Signaling. *Sci. Rep.* **5**, 12645; doi: 10.1038/srep12645 (2015).

Gene 595 (2017) 43–49

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Ct shift: A novel and accurate real-time PCR quantification model for direct comparison of different nucleic acid sequences and its application for transposon quantifications

Orsolya Kolacsek, Enikő Pergel, Nóra Varga, Ágota Apáti, Tamás I. Orbán *

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

ARTICLE INFO

Article history:
Received 6 September 2016
Received in revised form 20 October 2016
Accepted 24 October 2016
Available online 26 October 2016

Keywords:
Copy number
PGBD3
Sleeping Beauty
Promoter activity
Quantitative PCR
Splice variant

ABSTRACT

There are numerous applications of quantitative PCR for both diagnostic and basic research. As in many other techniques the basis of quantification is that comparisons are made between different (unknown and known or reference) specimens of the same entity. When the aim is to compare real quantities of different species in samples, one cannot escape their separate precise absolute quantification. We have established a simple and reliable method for this purpose (Ct shift method) which combines the absolute and the relative approach. It requires a plasmid standard containing both sequences of amplicons to be compared (e.g. the target of interest and the endogenous control). It can serve as a reference sample with equal copies of templates for both targets. Using the ΔΔCt formula we can quantify the exact ratio of the two templates in each unknown sample. The Ct shift method has been successfully applied for transposon gene copy measurements, as well as for comparison of different mRNAs in cDNA samples. This study provides the proof of concept and introduces some potential applications of the method; the absolute nature of results even without the need for real reference samples can contribute to the universality of the method and comparability of different studies.

1. Introduction

Use of real-time quantitative polymerase chain reaction (qPCR) amplifying DNA molecules has become a routine tool in molecular biology to study the abundance of nucleic acid sequences in different samples. While monitoring reaction products in each cycle the basis of quantification is that during the log phase of the reaction, the signal of the product is proportional to the starting material (Livak et al., 1995; Morrison et al., 1998; Freeman et al., 1999). However, in spite of few trials the exact quantities cannot be defined well enough empirically (Liu and Saint, 2002; Ramakers et al., 2003; Schefe et al., 2006); the only thing that can be supposed is that the proportion is probably constant for the same PCR target between different samples (Livak and Schmittgen, 2001). This is why we can make comparisons only between same PCR targets measured from different samples. On this basis two types of quantification approaches are possible: relative (Livak and Schmittgen, 2001; Pfaffl, 2001) and absolute (Morrison et al., 1998; Bustin, 2000) quantifications. Relative quantification is doubly relative because it measures relative content of a target sequence comparing to that in a reference sample, while the ratio of the target is normalized to the ratio of the endogenous control (comparative Ct or ΔΔCt formula). Since the abundance of latter sequence is constant among samples, relative quantification does not require precise quantity input measurements before setting up the qPCR. To investigate changes in gene expression or relative abundance of an mRNA between different types of tissues or cells, relative quantification is easy and adequate for most purposes. However, using relative approach for gene copy quantifications, a reference sample with known copy number is essential. Absolute quantification is based on a calibration curve of an artificial sequence (most often a plasmid) that allows to calculate the exact copy of the target in the starting quantities of dilutions. Ct value of the unknown is compared to the standard curve and copy number is calculated by its Ct value from the curve. Contrary to relative approach, absolute quantification requires precise quantity input measurements not only for the standard but for all of the samples to be examined.

If one would like to match copy numbers of different sequences in the same specimen – e.g. expression levels of RNA products belonging to the same cellular or genetic pathway – the adequate method is

Abbreviations: qPCR, quantitative polymerase chain reaction; Ct, threshold cycle; ΔΔCt, comparative Ct; RT-PCR, reverse transcription PCR; *PGBD3*, PiggyBac derived transposon family member 3; CS, Cockayne syndrome; *CSB*, Cockayne syndrome B; SB, Sleeping Beauty; *Rcor1*, REST corepressor 1; EGFP, enhanced green fluorescent protein; IRDR-L, inverted repeat direct repeat – left; IRDR-R, inverted repeat direct repeat – right; plst, plasmid standard; *RPPH1*, Ribonuclease P RNA component H1; Hsp90, heat shock protein 90; cds, coding sequence; RQ, relative quantity.

* Corresponding author: Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Magyar Tudósok krt. 2, Hungary.
E-mail address: orban.tamas@ttk.mta.hu (T.I. Orbán).

http://dx.doi.org/10.1016/j.gene.2016.10.035
0378-1119/

absolute quantification for all sequences. The most frequent and simple standard is a plasmid containing the qPCR target sequence. For the construction of this plasmid, it is sufficient to clone the regular PCR amplicon into any plasmid. Absolute quantification has to be made separately for all sequences that we want to compare. In this case all the errors are added up in the resulting comparison. The best way to minimize the variability between two particular absolute quantifications is to measure them simultaneously using a single plasmid that contains both target sequences to be compared, thus errors of quantity inputs and dilution steps will not vary between them. In spite of this clear advantage, doubly targeted plasmid solution is rarely chosen, e.g. in gene copy number quantification projects, and if so it is used for determining the dilution curves from the same plasmid aliquot (Charrier et al., 2011). The requirement of this unique plasmid standard provided the idea of Ct shift method: since the application of such plasmid *ab ovo* guarantees the determination of correct relation of sequences, we presumed it can be used as a reference for quantifications by the ΔΔCt method.

We applied the Ct shift method successfully for gene copy number measurement purposes, as well as for comparing the abundance of different mRNA splicing variants in cDNA samples. In these scenarios, the reference plasmid contains the target of interest, as well as the sequence of a reference gene (often called the endogenous control), the abundance of which is presumably constant in all samples. This plasmid can be applied either for absolute quantification in parallel standard curve experiments, or as a single calibrator sample containing one copy of the target amplicon and one copy of the endogenous control amplicon. In this way the plasmid standard can serve as a reference sample with equal copy number of templates for the two sequences, and the Ct difference of the amplicons is denoted as the Ct shift parameter. With the use of the ΔΔCt formula referring to the plasmid standard, we can quantify the exact ratio of the two sequences in each sample. An important advantage of the Ct shift method is that similarly to relative quantification, it does not necessarily require precise quantity inputs when setting up qPCR.

As a proof of concept, we started using the Ct shift method to measure *Sleeping Beauty* (SB) transposon copy numbers in transgenic cell cultures and in transgenic animals. Few years ago we developed diagnostic copy number assays for such purposes which required tedious preliminary work of establishing reference samples with known transposon copy numbers (Kolacsek et al., 2011). Applying the Ct shift method we provided evidence that this approach is applicable not only for the routine copy number determinations but can also be easily and reliably extended to new copy number projects, even without the laborious pilot experiments to generate reference samples with known copies.

We also show the advantage of the Ct shift method in gene expression studies. Cryptic transcription start sites and alternative splicing often results in several mRNA species produced from a single gene and the ratio among the mRNA variants is certainly crucial for the function of the gene. The Ct shift method offers a good solution for monitoring the levels of these gene products, and this could be applied for splice variants of certain domesticated transposase genes whose cellular functions have only been partially revealed (Sinzelle et al., 2009) or have not yet been elucidated (Alzohairy et al., 2013; Vogt et al., 2013). Some members of the PiggyBac transposase family express themselves by a gene trapping strategy, and in the absence of an own promoter they can only be transcribed as an alternative mRNA fusion product from the host (Newman et al., 2008). *PGBD3*, one of the PiggyBac derived transposons in the human genome is abundantly expressed in such a fusion form with Cockayne syndrome B (CSB) protein (http://biogps.org/). CSB protein has functions in transcription-coupled nucleotide excision repair (Sarker et al., 2005; Laine and Egly, 2006) and chromatin remodeling (Newman et al., 2006), and mutations of the *CSB* gene are responsible for 70% of CS disease (Laugel et al., 2010). *PGBD3* has been conserved in the 5th intron of *CSB* gene in Primates, and alternative splicing produces a protein fusion product with the N-terminal part of CSB protein (Newman et al., 2008). In addition, using a semi-quantitative method, the authors showed that a cryptic transcript is also expressed starting at the end of exon 5 (Newman et al., 2008). The role of the PGBD3 transposon as part of these transcripts is not known but it is probably beneficial for normal cell physiology. Our first aim in this topic was to reveal the ratio of these mRNA variants and whether this ratio is under cell type specific regulations. In this study while interpreting our results by the Ct shift model we have also verified the reliability of the method also in splicing pattern determinations.

2. Materials and methods

2.1. Cell lines and culturing

Human embryonic kidney cells (HEK-293) and HeLa cells were maintained as described previously (Kolacsek et al., 2014a); the HUES9 embryonic stem cell line was cultured as described earlier (Apati et al., 2008). The establishment and maintenance of the MSCL-2 mesenchymal-like cell line was described in detail previously (Varga et al., 2011).

2.2. Calculation with Ct shift

Ct deviation of two different PCRs – e.g. the target and the endogenous control – with equal input of template we have named as Ct_{shift} and determined by the plasmid standard:

$$Ct_{shift} = \Delta Ct_P = Ct_{TP} - Ct_{EP} \quad (1)$$

where P stands for plasmid standard, T for target and E for endogenous control.

In a particular measurement Ct shift is likely a uniform characteristic for all specimens as reflected by all standard points of the plasmid standards (Fig. 1A). Thus the value of Ct shift is constant and independent of the plasmid input quantity but only if efficiencies of the target and the endogenous control PCRs are approximately identical which can be tested empirically with the help of the slopes of dilution plots (Kolacsek et al., 2014b). This constant is probably characteristic to an unknown sample containing equal copy of templates.

$$Ct_{shift} = Ct_{TS} - Ct_{ES} \quad (2)$$

where S stands for sample. So with the help of Ct shift we can estimate a theoretical Ct for the target of an unknown sample when it would be present in equal copy number to endogenous control in the specimen.

$$Ct_{T(Theor)S} = Ct_{ES} + Ct_{shift} \quad (3)$$

Now the difference between the real Ct of target and this theoretical Ct can be calculated to determine the fold difference between them, which in fact gives the ratio of target and endogenous control of the unknown.

$$\Delta Ct_{T(Theor)S} = Ct_{TS} - Ct_{T(Theor)S} = Ct_{TS} - Ct_{ES} - Ct_{shift} \quad (4)$$

Fold difference of initial number of the same nucleic acid molecule between reactions can be expressed as an exponential function of the difference of Cts (Livak and Schmittgen, 2001) and we can calculate our theoretical difference with $\Delta Ct_{T(Theor)S}$.

$$\text{Fold difference} = [1 + E]^{-\Delta Ct\ T(Theor)S} \quad (5)$$

where E is PCR-efficiency of the two PCR. When E = 1, then:

$$\text{Fold difference} = 2^{-(\Delta Ct\ S\ -\ Ct\ shift)} \quad (6)$$

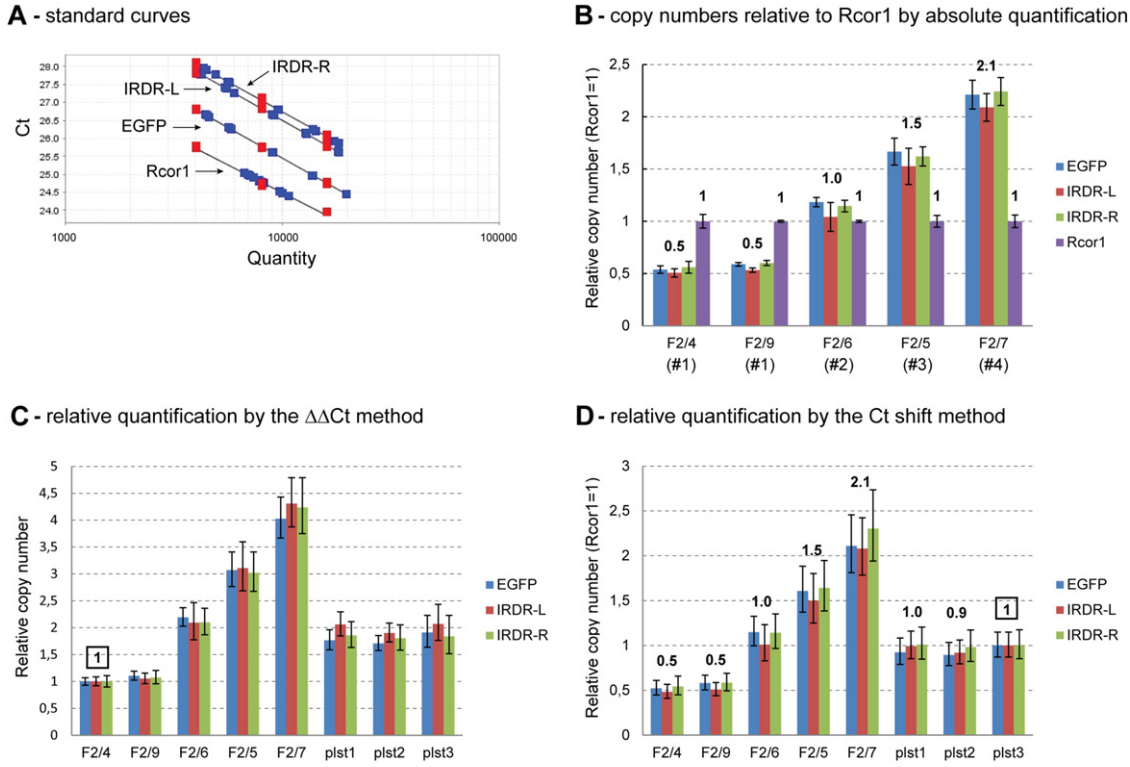


Fig. 1. Validation of the Ct shift method with transgenic copy number quantification of progeny of founder rats established by the Sleeping Beauty transposon based gene delivery. (A) Standard curves of absolute quantification. Red points are plasmid standard dilutions, blue ones are transgenic rat measurements. Copy inputs estimated from the plasmid quantities are on the X axis, resulting Cts are on the Y axis. Efficiencies calculated from the slopes are: Rcor1 (endogenous control) – 95%, EGFP – 96%, IRDR-L – 96%, IRDR-R – 98%. (B) Relative copy numbers by absolute quantification were calculated from the Ct values of targets and *Rcor1*. *Rcor1* represents 1 copy per haploid genome so 2 copies per cell, 0.5 copy of a target means 1 copy per cell (hemizygous for the locus). Previously determined copy numbers are shown in brackets. (C) Relative quantification of the same dataset with comparative Ct method (ΔΔCt). The one copy F2/4 rat sample was used as a reference sample, the three plasmid standard points (plst) were measured as a 2-copy-gDNA sample. (D) Relative quantification of the same dataset with the Ct shift method using one of the plasmid standard points as a reference sample. Means of at least 3 independent measurements are shown, error bars represent 95% confidence intervals.

where:

$$\Delta Ct_S = Ct_{TS} - Ct_{ES} \quad (7)$$

2.3. Copy number quantifications of SB transposons

For endogenous controls, genes present in one haploid copy were chosen: the *RPPH1* gene encoding for the H1 RNA subunit of the RNaseP complex for human samples, whereas the *Rcor1* (REST Corepressor 1) gene for rat samples. For construction of plasmid standards, *RPPH1* and *Rcor1* PCR amplicons were gel purified and first cloned into a pGEM®-T vector system (Promega). After verification by sequencing, amplicons were subcloned into a SB-CMV-EGFP transposon containing plasmid separately for *RPPH1* and *Rcor1* (pSB-EGFP-RPPH1, pSB-EGFP-Rcor1). These plasmids can be used as quantification standards since the targeted amplicons, the EGFP transgene, the SB transposon inverted repeat-direct repeat (IRDR) regions, and the control gene are all present in exactly one copy on them. Plasmid standards were linearized with restriction digestion leaving intact amplicon sequences and purified from gel electrophoresis. For absolute quantification measurements input quantity of plasmid was calculated precisely (Kolacsek et al., 2014b), but for the Ct shift method it is sufficient to input them into reactions in a quantity where the endogenous control amplification from the plasmid approximates the amplifications of samples.

Human transgenic cell clones (Kolacsek et al., 2011) and the transgenic rat line (Szebenyi et al., 2015b) were established and genomic DNA was isolated as described earlier. TaqMan® primer pairs with FAM-MGB probes were also detailed previously (Kolacsek et al., 2011; Kolacsek et al., 2013) with the exception of *Rcor1* for which an Applied Biosystems pre-developed gene expression assay (Rn03810960_s1) was used. Quantitative PCRs were performed on a real-time PCR platform (StepOne™ or StepOnePlus™; Applied Biosystems) with the following thermal profile: 95 °C 10 min, 40 cycles of 95 °C 15 s and 60 °C 1 min. 100 ng gDNA was inputted into a 20 μl reaction volume measured in triplicate for each target, final concentrations of primers and probes were 900 and 250 nM, respectively.

2.4. RT-qPCR for gene expression analyses of CSB gene products

RNA variants from the CSB locus were detected by five different amplicons as shown on Fig. 3A. For construction of plasmid standards, first the endogenous control PCR amplicon of Hsp90 and the five target PCR amplicons were gel purified, cloned into separate pGEM®-T vectors, and after verification by sequencing, the target amplicons were subcloned into the Hsp90 containing pGHsp90 pGEM® plasmid separately (pGHsp-tr1svPG3, pGHsp-tr1 + 2svPG3, pGHsp-tr1svCSB, pGHsp-tr1 + 2svCSB, pGHsp-PG3cds). Hsp90 containing pGHsp90 were digested with Spe I and Hinc II. Spe I and Zra I sites were utilized to cut out target amplicons. Plasmid standards were processed into qPCRs as described in the former section.

Total RNA was isolated from cells in near confluent 6-well plates using RNeasy Plus Mini kit (QIAGEN). 1 μg of DNase I (NEB) treated total RNA was reverse transcribed using random oligonucleotides with the High-Capacity cDNA Reverse Transcription Kit (Life Technologies). Ten-fold diluted cDNAs were inputted into qPCRs in 20 μl reaction volume containing 50 nM of primers and Power SYBR® Green PCR Master Mix (Life Technologies). Reactions were performed in triplicates for

each target in the same platform with same heat profile as for copy number measurements except that melting curve profile was added. For gDNA content control, equivalent DNase I digested RNA was tested in PG3cds PCR and no remarkable amplification was detected. All other PCRs were intron spanning. Primers were designed with Primer-BLAST software at NCBI website; the sequences are as follows (see Fig. 3A for CSB locations):

Hsp90 forward 5′-TGGATATCCCATTACTCTTTTTGTG, reverse 5′-TTCTTTTTCTTCTTCTTTGTCTTCCT; tr1svPG3 forward 5′-TGAAGAG TCTGAGTATTTCCCCAC, reverse 5′-TGCTTCTATGCTGTCATCTGTCTC; tr1 + 2svPG3 forward 5′-AGGTCGGAAAGTGGGAAGATAC, reverse primer is the same as for tr1svPG3; tr1svCSB forward primer is the same as for tr1svPG3, reverse 5′-TTCCTCAGAATCGTCCTCCAG; tr1 + 2svCSB forward primer is the same as for tr1 + 2svPG3, reverse primer is the same as for tr1svCSB; PG3cds forward 5′-TCATGGGAGGCGTAGACAGA, reverse 5′-CACGTCGACGAAACTCCAGA.

3. Results

3.1. Validation of the Ct shift method: SB transposon copy numbers in transgenic rats and human cell clones

We provide evidence for the reliability of the Ct shift method first by applying it for copy number determination. Gene copies are considered as corpuscular units, which uniformly present in individual cells in case of a cell clone or in the germline of an animal, offering a good candidate to prove that a particular qPCR quantification method is reliable. We have established a transgenic rat line expressing fluorescent calcium sensor protein which contains an SB-CAG-GCaMP2 transposon cassette in a homozygous form (Szebenyi et al., 2015a; Szebenyi et al., 2015b). Phenotypic analysis combined with transgene copy number determination of founders and progenies of subsequent generations enabled an effective and rapid selection of animals, resulting in a stable homozygous line in a short time schedule of crossings. One form of validation of the Ct shift method is whether we can gain the same copy number results arising from different sequences of the transgene. Analyses were done for left- and right transposon repeat sequences (IRDR-L and IRDR-R) with TaqMan® assays developed earlier. Since our EGFP TaqMan® assay can detect the *GCaMP2* fluorescent calcium sensor sequence, this was also analyzed. Plasmid standard was constructed from a transposon containing plasmid, cloning the *Rcor1* TaqMan® amplicon also into this plasmid, serving as the endogenous control as it is present in one copy per haploid genome.

Evidence for the existence of the Ct shift can be seen in Fig. 1A. Measurement of serial dilution of the plasmid standard produced good efficiencies for all sequences. However, applying the same volume of an individual plasmid standard for the reactions, it produced different separate Ct values for all of the analyzed sequences, and these Ct differences were consequent for all of the standard points originated from any of the dilutions; in this way, each standard point represents the same copy input of different sequences. These Ct delays occurring from the same copy input of targets and endogenous control were denominated as Ct shifts and were determined by multiple targeted plasmid standards in this quantification model. Ct shift is unique for a particular target and to some extent, for the actual run. Here we applied the measurement on the progeny derived from the crossing of F1 generation transgenic animals having two hemizygous copies of transgene on separate loci. As they descended from different founders, they contained different integration sites, thus the possible copy numbers of their offspring varies between 0 and 4. We have chosen one progeny for all representing copies and used the plasmid standard for parallel absolute quantification to determine the relative copy numbers – target/Rcor1 – precisely. After counting the total copy inputs of samples from the plasmid standard curves for all targets and the endogenous control (Fig. 1A), relative copies were calculated for each target (Fig. 1B). As clearly seen, the three different targets present on the same transgene produced very similar results and the multiple targeted plasmid standards provided credible relative copies (*Rcor1* control has 2 endogenous haploid copies, so the relative copy of 0.5 means 1 hemizygous copy per cell). While absolute quantification based on only one sequence and DNA quantities of inputs are often error prone, this standard seems to guarantee correct relations of sequences. Fig. 1C represents our formerly used routine copy number quantifications: it shows relative quantification of the same samples normalized to the endogenous control plus to a reference rat gDNA with one transgene copy. In addition, plasmid dilution points (plst) – that were previously used for absolute quantification to determine standard curves – were quantified as unknown samples: these plasmid inputs were determined as they were gDNA samples with 2 copies, proving that plasmid standards can serve as an excellent reference sample mimicking the reference gDNAs with known transgene copies. Therefore, plasmid standards could be used as references for the Ct shift method (Fig. 1D), providing exactly the same results as previous quantification models (Fig. 1B and C).

We have also tested the Ct shift quantification model in previously established human cell clones (Kolacsek et al., 2011). We have constructed a similar plasmid standard that contained the *RPPH1* human endogenous control amplicon. Using the ΔΔCt method with a one copy reference clone, relative quantification of this plasmid standard again mimicked a 2 copy clone (Fig. 2A). When we applied the Ct shift method for the same dataset, it resulted in similar copy numbers for all clones (Fig. 2B). Moreover, it produced more balanced results for all targets, especially smoothening the discrepancies in the case of the higher copy number clones (such as G2C3).

3.2. Application of the Ct shift method for RT-qPCR: comparison of the complex human CSB-PGBD3 mRNA variant pool

In RT-qPCR application of the Ct shift method, intra-sample relative gene expression measurements of different targets can be achieved using a single reference gene with unique calibrator plasmids for quantifying each target. All calibrator plasmids should contain the particular target-, as well as the common endogenous control amplicon, thus each target is quantified with separate calculations relative to the same endogenous control. Since all targets will be compared precisely to the same control gene, relations between them will also be just as accurate as the relative quantification measurements referring to even different individual plasmid standards.

Here we investigated the relative abundance of diverse mRNA products expressed from the *CSB* gene in four different human cell lines – HEK293, HeLa, HUES9 and MSCL-2. The latter is a mesenchymal stem cell like cell line which was differentiated in vitro in our laboratory from the HUES9 embryonic stem cell line (Varga et al., 2011). Fig. 3A shows the exon-intron structure of the CSB locus containing the domesticated *PGBD3* transposase gene, and also depicts primer pairs used to differentiate various transcripts and splice variants. We have chosen Hsp90 for endogenous control as it is proven that even the confluence of cells does not alter expression of this gene (Greer et al., 2010). PCR products of the different variants have been cloned into calibrator vectors also containing the *Hsp90* amplicon, to produce individual plasmid standards for each target. As shown previously for copy number quantifications, one proof for reliability is if different assays for the same target produce similar results. With the use of the Ct shift method, this was exactly the case in our RT-qPCR application shown in Fig. 3B. cDNAs were tested with two different primer pairings detecting all mRNAs containing *PGBD3* coding sequence. The assays produced very similar results for all cell types though they are targeting different parts of the same mRNA, providing further evidence for the reliability of the Ct shift method (Fig. 3B).

The Ct shift method is directly suitable for comparison of normal- and fusion splice variants (svCSB and svPG3) derived from the canonical (tr1) transcription start site (tr1svPG3 vs tr1svCSB, Fig. 3C), because these primer pairs detect solely the canonical transcripts. Quantifying

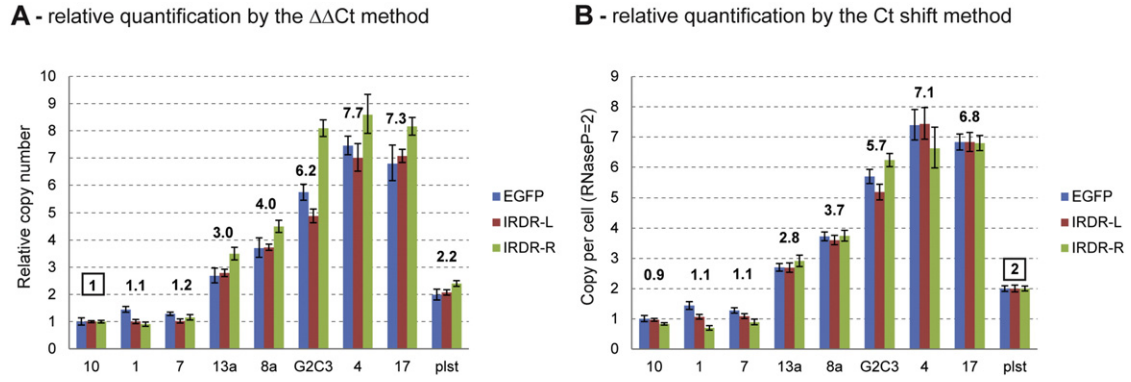


Fig. 2. The Ct shift method is applied for copy number quantification of different human cell clones, formerly created by the Sleeping Beauty transposon system. (A) Relative quantification with the ΔΔCt method using a previously established one copy reference clone, and the human *RPPH1* as the endogenous control. Values represent the average results of the three different assays, plasmid standard was measured as a clone with 2 copies. (B) Relative quantification with the Ct shift method using a single plasmid standard (plst) as a reference. A correction factor of 2 was applied to gain copies per cell results. Means of at least 3 independent measurements are shown, error bars represent 95% confidence intervals.

their levels revealed that they are expressed at similar levels only in two cell lines, Hela and MSCL-2. However, when the sum of the canonical and the cryptic promoter activities are measured, higher splice variant expression levels suggest significant cryptic promoter activity (tr1svPG3 vs tr1 + 2svPG3 and tr1svCSB vs tr1 + 2svCSB, Fig. 3C). These variants (tr1 + 2) showed similar pattern in cell types such as canonical splice variants but higher levels in MSCL-2 cell line (compare tr1 + 2svPG3 and tr1 + 2svCSB, Fig. 3C). As signal levels of tr1 + 2 splice products showed more diversity we have tested the reproducibility of the method using different cDNA inputs in separate measurements (Fig. 3D). The data clearly show that regardless of the amount of template applied, the absolute values of relative quantities are reproducible. The measurements also indicate that there is a substantial variation among the expression levels of the PG3 splice variants in the four cell lines examined (Fig. 3C and D).

Since in our gene expression analyses with the Ct shift method the relative quantities (RQs) are measured in the endogenous control expression unit of *Hsp90*, different RQ signals quantified in the same sample might be subtracted or added to extract more information from the result. Because of the "absolute" nature of resulting signal we can recover the products from the cryptic promoter by simply subtracting the canonical transcript splice variant (tr1 values) levels from the entire transcription activities (tr1 + 2 values) from Fig. 3C (Fig. 4A). Utilizing this feature we can reveal real differences between cell types when referring to one of the samples: when normalizing the values to the ones in HEK293 (Fig. 4B), the results show that the highest variability

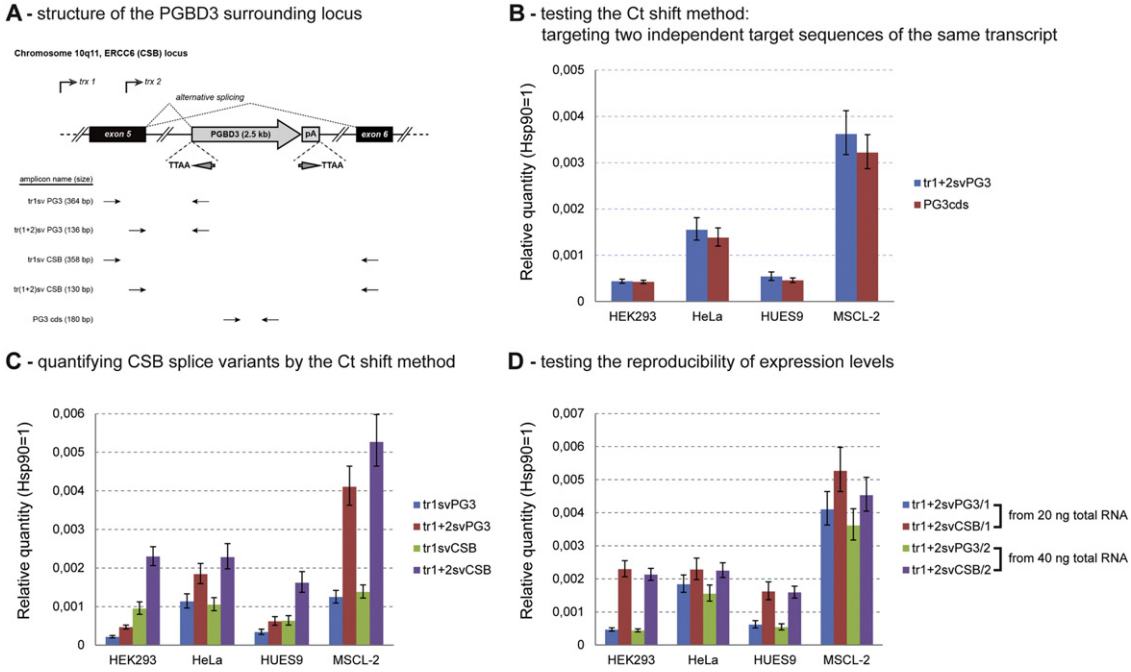


Fig. 3. Applying the Ct shift method for RT-qPCR. (A) Structure of the human *CSB* gene locus containing the domesticated *PGBD3* transposase, and primer pairs used to differentiate splice variants and promoter activities. Arrows depicted as "trx 1" indicates transcription from the canonical gene promoter, whereas "trx 2" shows transcription starting from a cryptic promoter located in exon 5. pA: polyA signal; "TTAA" sequences are the original target site duplications of the domesticated *PGBD3* transposase; Sv stands for splice variant. (B) Validation of the Ct shift method for gene expression quantification in four human cell lines. Results of two different primer pairs detecting separate parts of PGBD3 mRNA. Signal levels of plasmid standards were set as 1. (C) Quantification measurements differentiating all variants expressed from the CSB gene. Tr1 stands for the canonical, whereas tr1 + 2 for the sum of canonical and cryptic promoter activity. (D) Testing the reproducibility of the Ct shift method. RQs from two separate measurements (/1 and /2) with different cDNA inputs derived from 20 ng or 40 ng total RNA, respectively. Representative experiments are shown, error bars indicate mean values ± 95% confidence intervals.

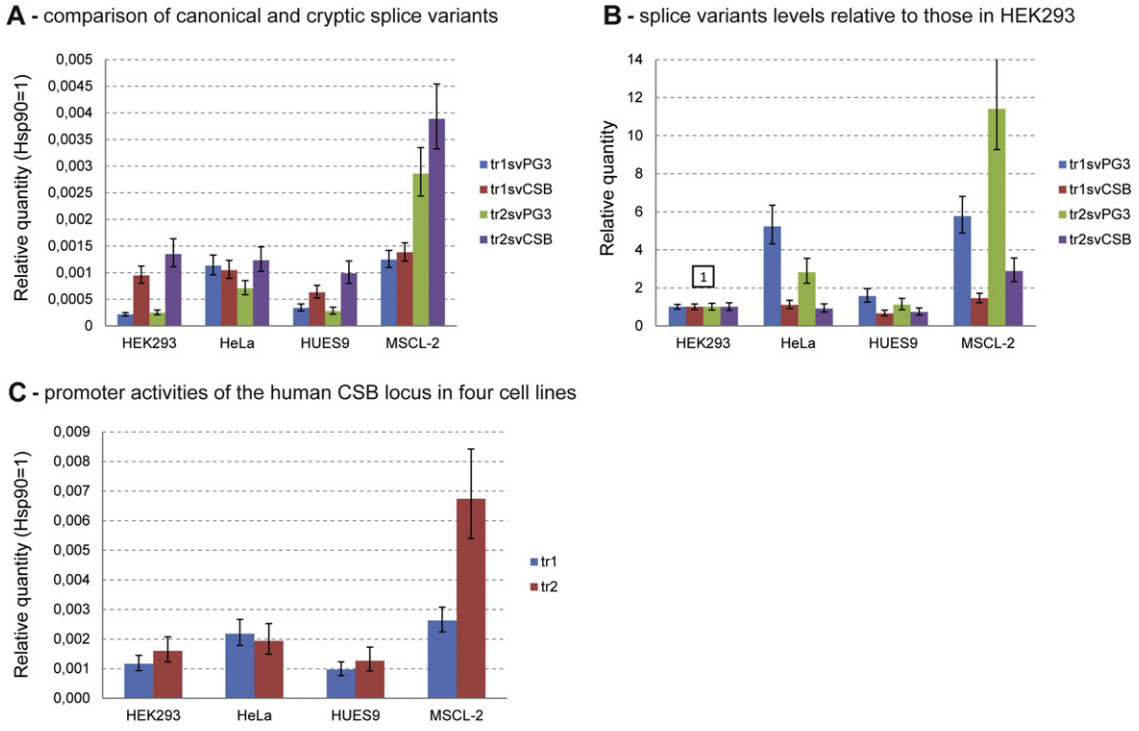


Fig. 4. Quantification of canonical and cryptic promoter products of *CSB* gene in four cell lines. (A) Splice variants produced from the canonical and cryptic promoter. RQs of splice variants from the cryptic promoter were calculated by subtracting splice variant levels accounting for solely the canonical promoter activity (tr1) from measurements accounting for the sum of the appropriate splice variants produced from the two promoters (tr1 + 2) in Fig. 3C. (B) Splice variant RQ values normalized to the levels measured in HEK293 cell line (C) Comparison of individual products of the two promoters in *CSB* gene. It was calculated by the addition of RQs of two possible splice variants produced from the appropriate promoters separately from Fig. 4A. Representative experiments are shown, error bars indicate mean values ±95% confidence intervals.

is detected in the MSCL-2 cell line. Different calculations also allow us to reflect the activities of the two promoters separately: with the addition of RQs of appropriate splice variants deriving from the same transcript from Fig. 4A, total promoter activities can be expressed (Fig. 4C). It suggests that in some of the cell types, transcription from the cryptic promoter can be elevated by at least two fold as compared to the canonical promoter activity.

4. Discussion

In this study we provide evidence that the existence of the "Ct shift", featured as the real-time PCR based Ct delay produced from different PCRs of same copy number, can be utilized for reliable quantification measurements. If the molar ratio of two sequences in PCR inputs of the plasmid standard is 1:1, one would expect that Ct values will be the same for both assays, especially when PCR efficiencies are similar and close to 100% for all of them – as proved by the slope of standard curves fitted onto plasmid dilution points. However, Ct values in qPCR measurements will differ in most of sequence cases (see Fig. 1A), even with application of the same threshold levels and detection chemistry. Based on our investigations, this deviation shows a partially stochastic nature: Ct shifts represent characteristic values for the particular measurements – this is the basis of our method, but it could show remarkable inter-assay variability. The reason behind the phenomenon is not entirely clear, but the most likely explanation is that the few initial cycles of the latent phase of PCR do not follow the kinetics measured by the reaction efficiency. This kinetic change during cycling may occur due to template change from the initial longer DNA template to the smaller PCR amplicon present in excess in later cycles. Because of these Ct delays, Ct values of different targets cannot directly be compared, and it is incorrect to represent ΔCt values as a fold difference by these Cts of different PCR target reactions, e.g. to compare the expression levels of different genes even if they were similarly compared to the same control gene.

Using the Ct shift method for three different targets and two different endogenous controls, we have provided evidence that precise gene copy numbers can be accurately determined similarly to traditional quantification models. The calibrator plasmids (containing amplicons of both the target gene and the endogenous control) can serve as excellent references, mimicking the gDNA of real reference samples containing single copies of target and control per haploid genome, and ensuring exact results by the ΔΔCt formula. Using such easy-to-make calibrators, copy number quantification projects do not require the tedious establishment of real reference samples, the copy numbers of which also need to be confirmed with other unrelated techniques. Here we provide evidence that Ct shift can be well characterized by plasmid standards and is also uniformly characteristic for the real samples. It definitely means that it is solely attributable to the particular amplicons independently of sequential or sample environment. So the Ct shift method is perfectly suitable for other qPCR studies regardless of targets and specimens, providing good reaction efficiencies and sample qualities.

Application of the Ct shift method for RT-qPCR studies is exemplified by our measurements of the splice variants from the human CSB-PGBD3 locus in different cell lines. The complexity of the mRNA pool is provided by the alternative splicing of the domesticated *PGBD3* transposase as part of a fusion CSB mRNA product, as well as by a cryptic promoter activity in addition to the canonical transcriptional start site of the *CSB* gene. The reliability of our method was also shown when the same RNA sequence region (PGBD3) was targeted with different PCRs, and the intra- and intersample splice variant patterns could also be reproducibly determined. Because of the "absolute" nature of Ct shift resulting signal, the method also allows to calculate indirectly determined transcript levels by subtraction or addition of different target RQs, in this case differentiating between promoter activities and alternative splicing. Using this approach, we could show remarkable variability of the cryptic promoter usage among cell types (Fig. 4C), and the comparable expression of the "host" (CSB) mRNA and the CSB-PGBD3 fusion product (Fig. 4A), further supporting the importance of

the domesticated PGBD3 expression, despite its as yet undetermined function (Newman et al., 2008). It is clearly seen that higher variance is observed among cell types for the PGBD3 splice variants than normal CSB splice variants; moreover, we can suspect important function of the PGBD3 splice product from the cryptic promoter in the MSCL-2 fibroblast cell line, as its expression is significantly higher than in the other cell lines (Fig. 4B). In fact, an autonomous PGBD3 protein product corresponding to this splice variant has been already detected by western blots (Newman et al., 2008), underlining its potential function. Although we also detected a significant level of the "normal" CSB splice product from this cryptic promoter, it is currently unknown if it is translated. The final conclusion of this part of our study is that the newly developed Ct shift method is readily applicable to decipher the fine-tuned complex regulation of mRNA patterns, such as the human CSB-PGBD3 locus, and could help to point toward potential functional studies, in this case examining the regulation of these mRNA variants in fibroblast like cell lines, such as the examined MSCL-2 cell line.

Up to now, several principally different technologies have been developed for nucleic acid quantifications. Some new methodologies, such as next generation sequencing, are powerful tools to analyze large datasets, but despite their applicability for quantitative comparisons, they are not optimal for quantifying small sample sizes, such as determining single transgene copy numbers or splice variant variability of a given locus in a small scale. Another powerful technology is droplet digital PCR which performs absolute quantitation without the assumption of equal primer efficiencies, and in principle it offers high precision in a relatively large dynamic range. Although it seems ideal for the quantitative measurement of nucleic acids in very low concentrations, it is often difficult to optimize the reaction conditions without a priori knowledge of the approximate concentration range, not to mention the demand for a relatively expensive machinery (Weaver et al., 2010; Majumdar et al., 2015). Thus, traditional qPCR remains substantial to gain precise results in cases of low throughput analyses, and our newly developed Ct shift method can become a useful contributor in the field of nucleic acid quantification providing a more absolute nature of expression measurements. This model allows direct comparison of target and endogenous control expressing the target level in quantity units of the control, so different measurements or studies become more comparable even for independent targets when same endogenous control is used in these assays. This model combines absolute and relative quantification approaches: similarly to absolute quantification, it uses artificially made standards with known copy number content. However, since these plasmid standards contain both the target and the control amplicons with known copies, they can be directly applied as a reference sample regardless of precise quantity input. With the application of this modified ΔΔCt method we can avoid the traditional absolute quantification measurements, and avoid the tedious establishment of known copy controls, such as for transgene copy number determinations. The Ct shift method might not be limited to the two most frequent quantification forms demonstrated here but may also find potential use in quality control areas or in healthcare, e.g. determination of microbiological or foreign load of any biological material.

Acknowledgement

We would like to thank Kornélia Némethy for excellent technical assistance and Dr. Douglas Melton for originally providing the HUES9 cell line. Tamás I. Orbán is a recipient of the János Bolyai Scholarship of the Hungarian Academy of Sciences. This work was supported by the Hungarian Scientific Research Fund (OTKA) K112112 grant.

References

Alzohairy, A.M., Gyulai, G., Jansen, R.K., Bahieldin, A., 2013. Transposable elements domesticated and neofunctionalized by eukaryotic genomes. Plasmid 69, 1–15.

Apati, A., Orban, T.I., Varga, N., Nemeth, A., Schamberger, A., Krizsik, V., Erdelyi-Belle, B., Homolya, L., Varady, G., Padanyi, R., Karaszi, E., Kemna, E.W., Nemet, K., Sarkadi, B., 2008. High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells. Biochim. Biophys. Acta 1778, 2700–2709.

Bustin, S.A., 2000. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. J. Mol. Endocrinol. 25, 169–193.

Charrier, S., Ferrand, M., Zerbato, M., Precigout, G., Viornery, A., Bucher-Laurent, S., Benkhelifa-Ziyyat, S., Merten, O.W., Perea, J., Galy, A., 2011. Quantification of lentiviral vector copy numbers in individual hematopoietic colony-forming cells shows vector dose-dependent effects on the frequency and level of transduction. Gene Ther. 18, 479–487.

Freeman, W.M., Walker, S.J., Vrana, K.E., 1999. Quantitative RT-PCR: pitfalls and potential. Biotechniques 26 (112–22), 124–125.

Greer, S., Honeywell, R., Geletu, M., Arulanandam, R., Raptis, L., 2010. Housekeeping genes; expression levels may change with density of cultured cells. J. Immunol. Methods 355, 76–79.

Kolacsek, O., Krizsik, V., Schamberger, A., Erdei, Z., Apati, A., Varady, G., Mates, L., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2011. Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. Mob. DNA 2, 5.

Kolacsek, O., Krizsik, V., Schamberger, A., Erdei, Z., Apati, A., Varady, G., Mates, L., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2013. Correction: reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. Mob. DNA 4, 11.

Kolacsek, O., Erdei, Z., Apati, A., Sandor, S., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2014a. Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. Hum. Gene Ther. Methods 25, 241–252.

Kolacsek, O., Izsvák, Z., Ivics, Z., Sarkadi, B., Orbán, T.I., 2014b. Quantitative analysis of DNA transposon-mediated gene delivery: the Sleeping Beauty system as an example. Genomics III - Methods, Techniques and Applications. iConcept Press Ltd., pp. 97–123.

Laine, J.P., Egly, J.M., 2006. Initiation of DNA repair mediated by a stalled RNA polymerase IIO. EMBO J. 25, 387–397.

Laugel, V., Dalloz, C., Durand, M., Sauvanaud, F., Kristensen, U., Vincent, M.C., Pasquier, L., Odent, S., Cormier-Daire, V., Gener, B., Tobias, E.S., Tolmie, J.L., Martin-Coignard, D., Drouin-Garraud, V., Heron, D., Journel, H., Raffo, E., Vigneron, J., Lyonnet, S., Murday, V., Gubser-Mercati, D., Funalot, B., Brueton, L., Sanchez Del Pozo, J., Munoz, E., Gennery, A.R., Salih, M., Noruzinia, M., Prescott, K., Ramos, L., Stark, Z., Fieggen, K., Chabrol, B., Sarda, P., Edery, P., Bloch-Zupan, A., Fawcett, H., Pham, D., Egly, J.M., Lehmann, A.R., Sarasin, A., Dollfus, H., 2010. Mutation update for the CSB/ERCC6 and CSA/ERCC8 genes involved in Cockayne syndrome. Hum. Mutat. 31, 113–126.

Liu, W., Saint, D.A., 2002. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal. Biochem. 302, 52–59.

Livak, K.J., Schmittgen, T.D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25, 402–408.

Livak, K.J., Flood, S.J., Marmaro, J., Giusti, W., Deetz, K., 1995. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Res. 4, 357–362.

Majumdar, N., Wessel, T., Marks, J., 2015. Digital PCR modeling for maximal sensitivity, dynamic range and measurement precision. PLoS One 10, e0118833.

Morrison, T.B., Weis, J.J., Wittwer, C.T., 1998. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques 24, 954–958 (960, 962).

Newman, J.C., Bailey, A.D., Weiner, A.M., 2006. Cockayne syndrome group B protein (CSB) plays a general role in chromatin maintenance and remodeling. Proc. Natl. Acad. Sci. U. S. A. 103, 9613–9618.

Newman, J.C., Bailey, A.D., Fan, H.Y., Pavelitz, T., Weiner, A.M., 2008. An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. PLoS Genet. 4, e1000031.

Pfaffl, M.W., 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. 29, e45.

Ramakers, C., Ruijter, J.M., Deprez, R.H., Moorman, A.F., 2003. Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci. Lett. 339, 62–66.

Sarker, A.H., Tsutakawa, S.E., Kostek, S., Ng, C., Shin, D.S., Peris, M., Campeau, E., Tainer, J.A., Nogales, E., Cooper, P.K., 2005. Recognition of RNA polymerase II and transcription bubbles by XPG, CSB, and TFIIH: insights for transcription-coupled repair and Cockayne syndrome. Mol. Cell 20, 187–198.

Schefe, J.H., Lehmann, K.E., Buschmann, I.R., Unger, T., Funke-Kaiser, H., 2006. Quantitative real-time RT-PCR data analysis: current concepts and the novel "gene expression's CT difference" formula. J. Mol. Med. (Berl) 84, 901–910.

Sinzelle, L., Izsvak, Z., Ivics, Z., 2009. Molecular domestication of transposable elements: from detrimental parasites to useful host genes. Cell. Mol. Life Sci.: CMLS 66, 1073–1093.

Szebenyi, K., Furedi, A., Kolacsek, O., Csohany, R., Prokai, A., Kis-Petik, K., Szabo, A., Bosze, Z., Bender, B., Tovari, J., Enyedi, A., Orban, T.I., Apati, A., Sarkadi, B., 2015a. Visualization of calcium dynamics in kidney proximal tubules. J. Am. Soc. Nephrol. 26, 2731–2740.

Szebenyi, K., Furedi, A., Kolacsek, O., Pergel, E., Bosze, Z., Bender, B., Vajdovich, P., Tovari, J., Homolya, L., Szakacs, G., Heja, L., Enyedi, A., Sarkadi, B., Apati, A., Orban, T.I., 2015b. Generation of a homozygous transgenic rat strain stably expressing a calcium sensor protein for direct examination of calcium signaling. Sci. Rep. 5, 12645.

Varga, N., Vereb, Z., Rajnavolgyi, E., Nemet, K., Uher, F., Sarkadi, B., Apati, A., 2011. Mesenchymal stem cell like (MSCl) cells generated from human embryonic stem cells support pluripotent cell growth. Biochem. Biophys. Res. Commun. 414, 474–480.

Vogt, A., Goldman, A.D., Mochizuki, K., Landweber, L.F., 2013. Transposon domestication versus mutualism in ciliate genome rearrangements. PLoS Genet. 9, e1003659.

Weaver, S., Dube, S., Mir, A., Qin, J., Sun, G., Ramakrishnan, R., Jones, R.C., Livak, K.J., 2010. Taking qPCR to a higher level: analysis of CNV reveals the power of high throughput qPCR to enhance quantitative resolution. Methods 50, 271–276.

Gene 676 (2019) 184–188

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Transcription activity of transposon sequence limits *Sleeping Beauty* transposition

Orsolya Kolacsek, Tamás I. Orbán*

Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

ARTICLE INFO

Keywords:
Sleeping Beauty
piggyBac
Promoter
Copy number

ABSTRACT

Sleeping Beauty (SB) transposon based technology has been extensively applied in basic research and biotechnology for routine cell culture gene delivery and vertebrate transgenesis, and it is also investigated in various gene therapy applications. Cell tolerance for the transgene is a key factor during transgenesis and is modulated not only through the type but by the dose of expression. Our experimental results exemplify that transgenes regulated with high activity promoters can reduce the overall success of gene delivery. Observations connected to transposon donors regulated by different promoters have also revealed inverse correlation between transcription activity and the hyperactive variant SB100X excision efficiency. This competition between transcription and transposition was independent of the transgene coding sequence and did not alter the transgenic efficiency in general. However, promoters applied in the transgene cassette can produce different average copy numbers depending on the transcriptional activity of the transposon. Unlike the *piggyBac* (PB) transposon system, this phenomenon allows a fine balance of expression using the high copy potential SB system that adjusts the copy number of lower activity promoter driven transgenes to a higher expression level. All this contributes to a well-tolerated and satisfactory transgenesis, and would be important to consider in gene therapy applications.

1. Introduction

Mobility of transposons can be utilized for gene transfer in both experimental and therapeutic applications. Among the DNA transposon systems, the *Sleeping Beauty* (SB) transposon is a reconstructed element from fish genomes and it is one of the best-characterized transposons in terms of activity in vertebrate cells (Kawakami et al., 2017; Narayanavari et al., 2017). The most active SB100X variant supports gene integration at a level approaching the efficiency of integrating viral vector systems (Mates et al., 2009). SB has a close-to-random integration profile, which contributes to its enhanced safety in therapeutic applications (Gogol-Doring et al., 2016; Kebriaei et al., 2017). In addition, its relative independence on species or cell type specific cellular factors allows its application in a wide range of cell types, including various stem cells.

Efficient genomic integration of SB enables stable expression of large variety of transgenes as exemplified by several preclinical animal models and cell culture studies (Izsvak et al., 2010). It is generally accepted that the size of the transgene lowers transposition efficiency, although BAC clones over the size of 150 kb can be stably integrated by the SB system into vertebrate genomes (Rostovskaya et al., 2012; Wang et al., 2014). However, cellular tolerance to transgenesis can vary depending on the type and expression level of the transgene which might also modulate the overall efficiency or outcome of transgenesis. This is to say that the nature of the promoter and the transgene are not expected to influence transposition rates, rather their effects are assumed to be manifested by the interference with the cellular expression machinery and by the genomic defense activity of the particular cell type. Different transgenes can be tolerated differently in a cell type specific manner, so the expression dose is recommended to be carefully adjusted to avoid this selection effect. This is often achieved by the appropriate choice of the promoter, and some of our previous experiments supported this idea: certain viral promoters (e.g. CMV, cytomegalovirus promoter) are strongly silenced in human embryonic stem cells as compared to other constitutive promoters (Orban et al., 2009), suggesting that lower activity promoters, especially those ones derived from endogenous sequences, could lead to a more stable long term transgene expression.

Contrary to that, when optimizing the parameters of the SB and the *piggyBac* (PB) transposon systems, some experiments indicated that

Abbreviations: SB, *Sleeping Beauty*; PB, *piggyBac*

* Corresponding author at: Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Magyar Tudósok krt. 2, Hungary.

E-mail address: orban.tamas@ttk.mta.hu (T.I. Orbán).

https://doi.org/10.1016/j.gene.2018.07.045
Received 26 April 2018; Received in revised form 11 July 2018; Accepted 13 July 2018

transcription activity of different promoters may additionally influence transposition efficiencies (Kolacsek et al., 2014). Follow up studies of long term cell culture experiments revealed that, particularly for the hyperactive SB100X transposase, transposon copy numbers are influenced by the choice of promoters, as if there was a balancing effect on the overall transgene expression. Since there was some indications that a strong promoter activity could interfere with SB transposition efficiency, this important aspect of transposon-based gene delivery awaited further careful examinations. Therefore, in this study we have experimentally tested how transgene expression provided by different activity promoters in the delivered transposon cassette could affect transposition efficiency. We have also addressed the question of which step of the transposition reaction is affected by the level of transcription, and what would the overall outcome of this interference have on the resulting transgenic cell number.

2. Materials and methods

2.1. Plasmid constructs

SB100X transposase variant was expressed from a standard expression plasmid regulated by a CMV promoter (pCMV-SB100X) (Mates et al., 2009). SB transposon donor plasmids were created from the pT2 variant of the SB inverted repeat-direct repeats (IRDRs) (Geurts et al., 2003). Puromycin antibiotic resistance gene was cloned into the initial transposon plasmid with PCR technique; CAG, CMV, PGK, and SV40 promoters were cloned with standard restriction cassette replacements. For GFP expression plasmids, the puromycin coding sequences in these transposon donor plasmids were changed to GFP using restriction enzymes. For transfection control, the pCMV-EGFP (Clontech) expression plasmid was used.

2.2. Cell maintenance and transfection method

Human embryonic kidney cells (HEK-293) were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal calf serum, 1% L-glutamine, and 1% penicillin–streptomycin. 2×10^5 HEK-293 cells were seeded onto 12-well plates, and transfected with FuGENE 6 reagent (Promega) in duplicates according to the manufacturer's instructions. For colony assays, 10:1 or 1:10 donor – helper dose combinations were applied, e.g. 400 ng puromycin containing transposon donor plasmids were co-transfected with 40 ng pCMV-SB100X helper vector or vice versa. All transfections were supplemented with 400 ng pCMV-EGFP transfection control. Two days posttransfection cells were harvested and one tenth of the transfections were analyzed with Attune Acoustic Focusing Cytometer (Thermo Fisher Scientific) for GFP expressing cells. Transfection efficiencies were comparable within experiments, ranging from 72% to 84%.

2.3. Transposition assay

For colony forming assay, 48 h posttransfection 1% of the transfected cells were seeded onto cell culture Petri dishes and selected for 3 weeks with 1 μg/ml puromycin. Surviving cells were fixed with methanol and stained with 0.05% crystal violet in 25% methanol. Colonies were quantified with a universal hood gel imager model 75S, using Quantity One 4.4.0 software (Bio-Rad, Hercules, CA). Transgenic rates were calculated from the colony numbers, normalizing with the number of seeded cells and also for transfection efficiencies; transgenic rates are therefore defined as a percentage of transfected cells.

2.4. Transposon copy number measurements

For average copy number quantification, cells were harvested two days posttransfection and one tenths of the cells in parallels were selected in puromycin for 3 weeks. Cells from parallels were pooled and genomic DNAs were isolated by the standard phenol-chloroform extraction method. Average copy numbers were quantified with a TaqMan® assay specific for the IRDR-L (left transposon sequence), normalizing to the level of the RPPH1 endogenous control gene as described previously (Kolacsek et al., 2011). Briefly, all samples were run in triplicate singleplex reactions, using 100 ng of gDNA template, in a final volume of 20 μl. Real-time PCR measurements were performed on a StepOnePlus™ platform, and data were analyzed by the StepOne software version 2.1, according to the manufacturer's instructions (Thermo Fisher Scientific).

2.5. Expressivities of the promoters

HEK-293 cells were transfected as described above, except that 800 ng GFP transposon donor plasmids were solely applied. Two days posttransfection cells were harvested and parallels were pooled. GFP expressing cells were separated with FACSAria flow cytometer according to the manufacturer's instructions (Becton, Dickinson and Company) and seeded onto a 12-well-plate. Next day they were lysed with 500 μl peqGOLD TriFast (VWR Peqlab) and total RNAs were extracted by standard protocol of the manufacturer. RNAs were treated with DNase I to remove plasmid DNA contamination. Reverse transcriptions were carried out with High-Capacity cDNA Reverse Transcription kit (Thermo Fischer Scientific) according to the manufacturer's instructions, including negative controls without adding reverse transcriptase. For qPCR reactions, 5 μl of tenfold diluted cDNAs were applied in 20 μl reaction volume in triplicates. Reactions were carried out with TaqMan® Gene Expression Master Mix (Thermo Fisher Scientific), and were run on StepOnePlus™ real-time PCR system platform using standard protocol; results were analyzed by StepOne software version 2.1. TaqMan® primers and probe for the GFP were: forward primer, 5′-GAGCGCACCATCTTCTTCAAG; reverse primer, 5′-TGTCGCCCTCGAACTTCAC; TaqMan® labeled probe, 5′-ACGACGGC AACTACA. Final concentrations of primers and probes were 900 nM and 250 nM, respectively. Ribosomal protein P0 (RPLP0, TaqMan® assay Hs9999902_m1) was used as endogenous control and the ΔΔCt method was applied for relative quantification.

2.6. Excision activities

To quantitate the excision efficiency of the hyperactive SB100X transposase on different transposon donors, HEK-293 cells were co-transfected with 400 ng of transposon donor and 400 ng of SB100X expressing helper plasmids. Two days posttransfection cells were harvested and parallels were pooled. Isolation of plasmid DNA and real-time qPCRs were carried out essentially as described previously (Kolacsek et al., 2014). Briefly, the Qiagen plasmid miniprep kit was used for DNA isolation with a modified protocol: the cell lysis/protein removal step solution was supplemented with 300 μl of 1.2% sodium dodecyl sulfate solution and with 50 μg of proteinase K. 10 ng of the purified DNA samples were used as templates in a two-round nested PCR where primers specific for the transposon donor plasmid backbone were used, amplifying sequences from post-excision plasmids that underwent DNA repair. Plasmids without excision do not yield PCR amplicons due to the large size of the transposon cassettes. The second round of the nested PCR is carried out on a StepOnePlus™ real-time PCR platform, using SYBR Green chemistry (Thermo Fisher Scientific). For normalization, PCR primers specific for the ampicillin sequence were used which is exclusively present on the transposon donor plasmid; excision efficiencies were then calculated by the ΔΔCt method.

3. Results and discussion

3.1. Choice of promoters influence SB transgene copy numbers

Our previous works with the most frequently and routinely applied

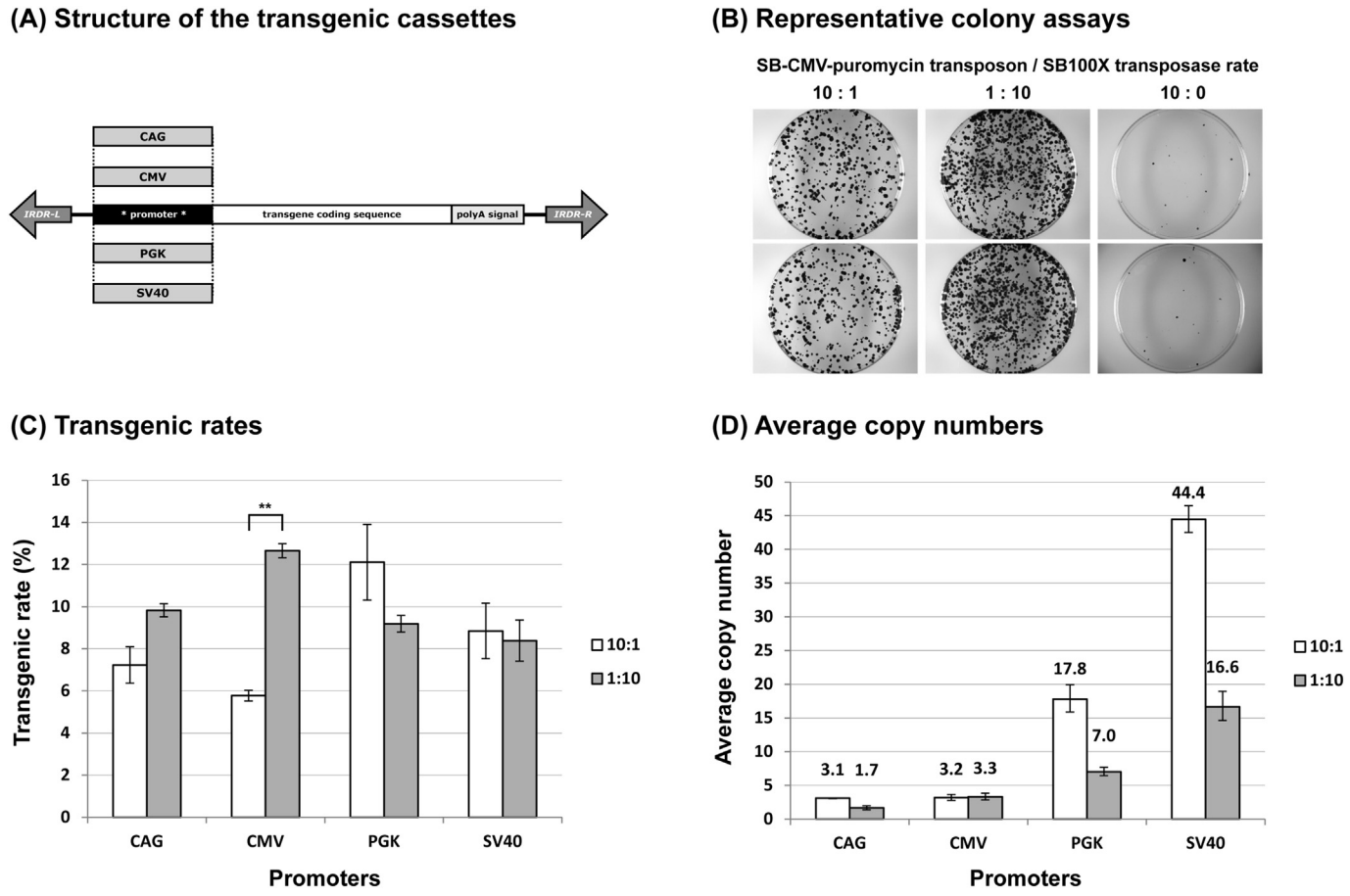


Fig. 1. Transgenic efficiencies of SB100X in HEK-293 cells depending on promoters of the transgene, as well as on transposon/transposase dose combinations. Comparison of four frequently used promoters in puromycin resistance gene containing transposon deliveries, using high and low transposon doses. (**A**) The structure of the used expression cassettes differ only by the indicated promoter sequences. (**B–C**) Transgenic rates determined by colony assays, and photos of CMV driven transgene colonies (two parallels shown for each dose). Error bars represent standard deviations of parallel transfections. For statistical analysis, two-sided Student's *t*-tests were applied, **: $p < 0.01$. (**D**) Average copy number quantification of the same deliveries by qPCR specific for the SB transposon IRDR-L sequence. Error bars represent 95% confidence intervals of parallel measurements.

two-component SB transfection system have showed that dose combinations of transposon donor and transposase expressing helper plasmids define different average transgene copy numbers after co-transfection and selection of transgenic cells (Kolacsek et al., 2014). Ten to one (10:1) donor – helper ratio resulted in the highest average copy number, while one to ten (1:10) dose combination supported transgenic cell population with low average copy number and comparably high transgenic efficiency when hyperactive SB100X was applied. Therefore these two dose combinations providing high and low copy numbers are well suitable to monitor the cell tolerance toward different promoters in the transgenesis. Four frequently used promoters were cloned into SB transgene cassettes expressing puromycin antibiotic resistance gene (Fig. 1A). These were the CAG composite promoter with CMV (cytomegalovirus) enhancer and two different chicken- and rabbit endogenous promoter elements, the CMV promoter, PGK- (phosphoglycerate kinase), and SV40 (simian virus 40) promoter. Evaluating the transgene delivery experiments in HEK-293 cells with colony assays (Fig. 1B), transgenic cell numbers were not considerably different between promoters when low transposon donor dose was applied (Fig. 1C, 1:10 donor-helper ratio). It suggests that low transgene load is well tolerated independent of the type of the promoter regulating the transgene expression. More fluctuations were seen among promoters at high transposon donor dose (Fig. 1C, 10:1 donor-helper ratio), and CMV-puromycin transgene cassette resulted the lowest number of transgenic colonies suggesting lower tolerance of cells for CMV driven transgene expression. Moreover, when we compare the two dose combinations (10:1 vs. 1:10), in the case of CMV two fold difference is observed in transgenic efficiency, while the tolerance of the other types of promoters did not seem to depend on transposon load (Fig. 1B,C).

We also examined the average copy numbers of these selected transgenic cell populations which revealed unexpected differences between the examined promoters (Fig. 1D). Copy numbers produced by the high activity CAG and CMV promoter driven transposons were significantly lower at both dose combinations than in the cases of PGK and SV40 promoters. Next we have verified the expression activity of these promoters by measuring transcription with qRT-PCR. Using GFP as the transgene with different promoters, cells transfected with the SB constructs were sorted for GFP expressing cells by FACS two days following transfection, and then GFP mRNA levels were measured in those cells. CAG and CMV promoters showed much higher transcription activities compared to PGK and SV40 (Fig. 2).

3.2. Promoter activity influences SB excision rate

At this point one could expect that differences in copy numbers of selected cells merely represent lower tolerance of higher transgene expression levels. This might be true in the case of CAG and CMV promoter. However, the much lower activity PGK and SV40 promoters also showed at least two fold difference in copy numbers at both dose combinations, and expression levels produced by them differentiated to a similar extent (Figs. 1D and 2). This raises the possibility that transposase function can be directly influenced by the expression activity of

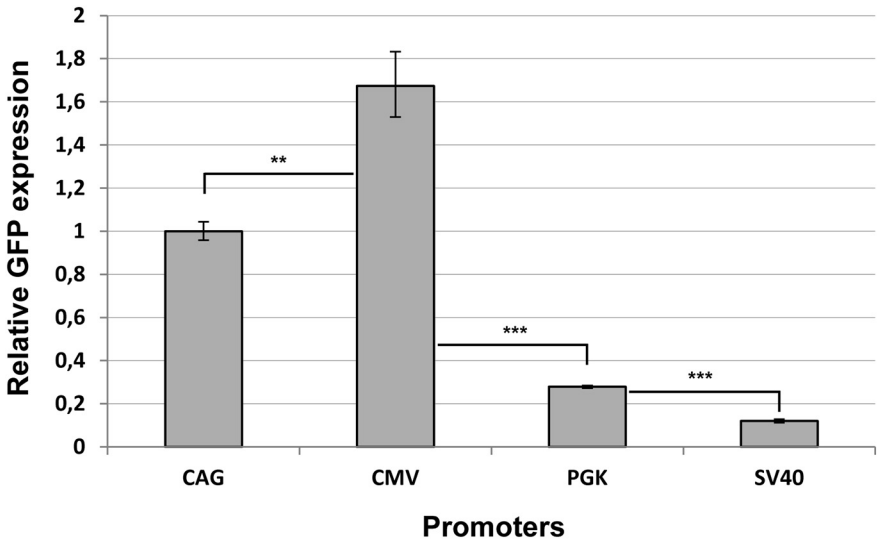


Fig. 2. Comparison of promoter activities in HEK-293 cells by qRT-PCR.
Four different promoters driven GFP containing SB transposon donor plasmids were transfected and sorted by FACS two days following transfection. Messenger RNA levels were measured by qPCR assay specific for GFP. No RT controls were performed in each case, and no amplifications were detected for either the endogenous control (P0) or GFP. Error bars represent 95% confidence intervals of parallel measurements. Two-sided Student's *t*-tests were performed for statistical analysis, **: $p < 0.01$, ***: $p < 0.001$.

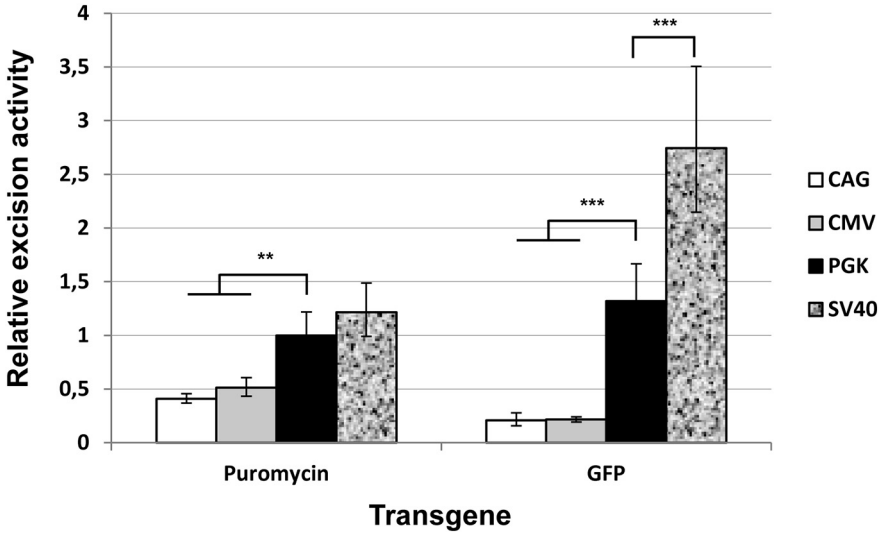


Fig. 3. Excision activities of SB100X in HEK-293 cells depending on promoters of the transgene.
Comparison of four different promoters driven transposon donor excision with relative quantification by qPCR. Error bars represent 95% confidence intervals of parallel measurements. For statistical analysis, two-sided Student's t-tests were applied; **: $p < 0.01$, ***: $p < 0.001$.

a particular promoter in the transposon. Cell tolerance level for an exogenous protein expression is difficult to determine, but definitely detectable over a longer period of time. Immediate transposase activity in co-transfected cells is certainly not dependent on cell tolerance that develops later during delivery and selection of transgenic cells. Transposase activity is often characterized by measuring the transposon excision product left behind after excision and reparation of double stranded breaks on the donor plasmid (Mates et al., 2009; Yusa et al., 2011; Claeys Bouuaert et al., 2013; Kolacsek et al., 2014). Quantification of excised and re-circularized donor plasmids by real-time PCR proved the transcription dependence of transposase activity (Fig. 3). Transfecting the two component SB100X transposon system into HEK-293 cells, the amount of excision products inversely correlated with the transcription activity of promoters applied in the transposon cassette (compare Figs. 2 and 3). Transposase activity correlates more obviously with the resulting copy number (Figs. 3 and 1D). This phenomenon was independent of the type of the transgene, as both puromycin and GFP expressing transposon showed similar excision tendency between the examined promoters (Fig. 3).

These promoters are intended to express the transgene from a genomic locus after delivery; however, they also constantly regulate transcription frequency ectopically from the plasmid before chromosomal integration. We can draw the conclusion that this episomal expression could limit transposition efficiency. This idea is supported by earlier observations when donor plasmids methylated in vitro allows more efficient delivery in mice transgenesis, and the transgene will be released from repression thereafter during epigenetic reprogramming of embryonic stem cells (Carlson et al., 2011).

Our experimental data led to the assumption that continuous presence of the transcriptional machinery owing to higher activity promoters may develop steric hindrance for transposition of an actively transcribed DNA substrate. We expected this competition between transcription and transposition to be a ubiquitous function of DNA transposons. However, in the case of similarly widely used *piggyBac* (PB) transposon system, representing a different superfamily of DNA transposases, we could not reproduce this phenomenon (data not shown), suggesting that PB transposition may also be sensitive to other factors binding to different promoters. Nevertheless we may conclude that both transposon systems are similarly efficient in establishing transgenic cells independently of promoter sequence, but SB tends to make different copy numbers depending on the transcription activity of the transgene and on the transposon donor load in delivery (Fig. 1D). In connection with the latter aspect of SB100X, promoters with extremely high activity (like the viral-derived CMV promoter) are not recommended to be applied in high transposon donor dose because the expected high copy numbers can reduce transgenic cell number owing to bad tolerance for high level of transgene expression (Fig. 1B,C). Furthermore, a remarkable selective disadvantage appears for CMV

driven highly expressing transgenic cells when the population is not under selective pressure (unpublished data).

Competition between transcription and SB transposition allows balancing the overall expression level of transgene, as lower activity promoters can in fact produce higher transgene expression owing to higher copy number; nevertheless, this also depends on the dose of transposon donor and is also balanced by the tolerability of the protein. While PB generally produce much lower copy numbers (Grabundzija et al., 2010; Kolacsek et al., 2014), the high copy "potential" of SB gives the opportunity for fine adjustment of copy number in delivery of transgene, and in addition, it makes SB quite suitable for multiple gene delivery (Kacherovsky et al., 2015). Based on our studies, for increased efficiency of multiple transgene delivery applications, using lower activity promoters is worth to consider.

3.3. Conclusions

Using several widely used eukaryotic promoters in SB transposon cassettes, we concluded that the transcription activity of the promoters applied in the transposon cassette interfere with SB100X transposition, and this interference is manifested at the excision step of transposition. The competition between transcription and transposition elevates the transgene copy numbers of lower activity promoters during transgenesis and as a result, the overall transgene expression level will be higher from weaker promoters. This "fine-tuning" of the gene expression produces similar expression levels from lower copy number of strong promoter-driven transgenes that would otherwise be less tolerated for the cells. Based on these new observations, we could provide an "updated user guide" for SB transposon delivery applications, in order to achieve the most optimal long term transgene expression levels. As a first issue, transfection efficiency should be taken into considerations: at low efficiency rate (e.g. for hard-to-transfect cell lines), high transposon donor dose will result in sufficient number of transgenic cells, and can produce low average transgene copies per cell. On the other hand, in cases of higher transfection efficiency, the application of strong promoters (like CMV), especially at high donor doses, can be unfavorable due to the potential of the transgene silencing, as well as to the bad tolerance of cells toward highly expressing copies, reducing transgenic efficiency. In such cases, promoters with moderate or low transcription activity (typically from endogenous promoters) can be beneficial, as they are mobilized more effectively due to the less interference with transposition, and the level of expression can be optimal with higher number of integrated transgene copies. When transgene expression level needs more adjustments, transposon donor dosage can be used to further manipulate the resulting copy number. At high transfection rates, increasing the dose of transposase might help in increasing transposition efficiency, however, although not observed in our experiments, this could have an adverse effect due to the overproduction inhibition phenomenon (Bire et al., 2013).

Declarations of interest

None.

Acknowledgments

The authors are grateful to Kornélia Némethy for excellent technical assistance.

Funding

Tamás I. Orbán is a recipient of the János Bolyai Scholarship of the Hungarian Academy of Sciences, and a supported fellow of the "Szinergia VIII. – MedInProt" program. This work was supported by the Hungarian Scientific Research Fund (NKFIH-OTKA, grant number K112112).

Authors' contributions

OK: designed the study, performed the experiments, analyzed the data, wrote the manuscript.

TIO: designed the study, analyzed the data, wrote and finalized the manuscript.

References

Bire, S., Casteret, S., Arnaoty, A., Piegu, B., Lecomte, T., Bigot, Y., 2013. Transposase concentration controls transposition activity: myth or reality? Gene 530, 165–171.
Carlson, D.F., Geurts, A.M., Garbe, J.R., Park, C.W., Rangel-Filho, A., O'Grady, S.M., Jacob, H.J., Steer, C.J., Largaespada, D.A., Fahrenkrug, S.C., 2011. Efficient mammalian germline transgenesis by cis-enhanced Sleeping Beauty transposition. Transgenic Res. 20, 29–45.
Claeys Bouuaert, C., Lipkow, K., Andrews, S.S., Liu, D., Chalmers, R., 2013. The autoregulation of a eukaryotic DNA transposon. eLife 2, e00668.
Geurts, A.M., Yang, Y., Clark, K.J., Liu, G., Cui, Z., Dupuy, A.J., Bell, J.B., Largaespada, D.A., Hackett, P.B., 2003. Gene transfer into genomes of human cells by the sleeping beauty transposon system. Mol. Ther. 8, 108–117.
Gogol-Doring, A., Ammar, I., Gupta, S., Bunse, M., Miskey, C., Chen, W., Uckert, W., Schulz, T.F., Izsvak, Z., Ivics, Z., 2016. Genome-wide profiling reveals remarkable parallels between insertion site selection properties of the MLV retrovirus and the piggyBac transposon in primary human CD4(+) T cells. Mol. Ther. 24, 592–606.
Grabundzija, I., Irgang, M., Mates, L., Belay, E., Matrai, J., Gogol-Doring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M.K., VandenDriessche, T., Izsvak, Z., Ivics, Z., 2010. Comparative analysis of transposable element vector systems in human cells. Mol. Ther. 18, 1200–1209.
Izsvak, Z., Hackett, P.B., Cooper, L.J., Ivics, Z., 2010. Translating Sleeping Beauty transposition into cellular therapies: victories and challenges. Bioessays 32, 756–767.
Kacherovsky, N., Liu, G.W., Jensen, M.C., Pun, S.H., 2015. Multiplexed gene transfer to a human T-cell line by combining Sleeping Beauty transposon system with methotrexate selection. Biotechnol. Bioeng. 112, 1429–1436.
Kawakami, K., Largaespada, D.A., Ivics, Z., 2017. Transposons as tools for functional genomics in vertebrate models. Trends Genet. 33, 784–801.
Kebriaei, P., Izsvak, Z., Narayanavari, S.A., Singh, H., Ivics, Z., 2017. Gene therapy with the Sleeping Beauty transposon system. Trends Genet. 33, 852–870.
Kolacsek, O., Krizsik, V., Schamberger, A., Erdei, Z., Apati, A., Varady, G., Mates, L., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2011. Reliable transgene-independent method for determining Sleeping Beauty transposon copy numbers. Mob. DNA 2, 5.
Kolacsek, O., Erdei, Z., Apati, A., Sandor, S., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2014. Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. Hum. Gene Ther. Methods 25, 241–252.
Mates, L., Chuah, M.K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D.P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., Vandendriessche, T., Ivics, Z., Izsvak, Z., 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat. Genet. 41, 753–761.
Narayanavari, S.A., Chilkunda, S.S., Ivics, Z., Izsvak, Z., 2017. Sleeping Beauty transposition: from biology to applications. Crit. Rev. Biochem. Mol. Biol. 52, 18–44.
Orban, T.I., Apati, A., Nemeth, A., Varga, N., Krizsik, V., Schamberger, A., Szebenyi, K., Erdei, Z., Varady, G., Karaszi, E., Homolya, L., Nemet, K., Gocza, E., Miskey, C., Mates, L., Ivics, Z., Izsvak, Z., Sarkadi, B., 2009. Applying a "double-feature" promoter to identify cardiomyocytes differentiated from human embryonic stem cells following transposon-based gene delivery. Stem Cells 27, 1077–1087.
Rostovskaya, M., Fu, J., Obst, M., Baer, I., Weidlich, S., Wang, H., Smith, A.J., Anastassiadis, K., Stewart, A.F., 2012. Transposon-mediated BAC transgenesis in human ES cells. Nucleic Acids Res. 40, e150.
Wang, Y., Wang, J., Devaraj, A., Singh, M., Jimenez Orgaz, A., Chen, J.X., Selbach, M., Ivics, Z., Izsvak, Z., 2014. Suicidal autointegration of Sleeping Beauty and piggyBac transposons in eukaryotic cells. PLoS Genet. 10, e1004103.
Yusa, K., Zhou, L., Li, M.A., Bradley, A., Craig, N.L., 2011. A hyperactive piggyBac transposase for mammalian applications. Proc. Natl. Acad. Sci. U. S. A. 108, 1531–1536.

RNA BIOLOGY
2021, VOL. 18, NO. 8, 1170–1180
https://doi.org/10.1080/15476286.2020.1836457

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

Tissue-specific and transcription-dependent mechanisms regulate primary microRNA processing efficiency of the human chromosome 19 MicroRNA cluster

Ábel Fóthi[a], Orsolya Biró[b], Zsuzsa Erdei[a], Ágota Apáti[a], and Tamás I. Orbán [a]

[a]Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary; [b]Department of Obstetrics and Gynaecology, Semmelweis University, Budapest, Hungary

ABSTRACT

One of the longest human microRNA (miRNA) clusters is located on chromosome 19 (C19MC), containing 46 miRNA genes, which were considered to be expressed simultaneously and at similar levels from a common long noncoding transcript. Investigating the two tissue types where C19MC is exclusively expressed, we could show that there is a tissue-specific and chromosomal position-dependent decrease in mature miRNA levels towards the 3′ end of the cluster in embryonic stem cells but not in placenta. Although C19MC transcription level is significantly lower in stem cells, this gradual decrease is not present at the primary miRNA levels, indicating that a difference in posttranscriptional processing could explain this observation. By depleting Drosha, the nuclease component of the Microprocessor complex, we could further enhance the positional decrease in stem cells, demonstrating that a tissue-specific, local availability of the Microprocessor complex could lie behind the phenomenon. Moreover, we could describe a tissue-specific promoter being exclusively active in placenta, and the epigenetic mark analysis suggested the presence of several putative enhancer sequences in this region. Performing specific chromatin immunoprecipitation followed by quantitative real-time PCR experiments we could show a strong association of Drosha with selected enhancer regions in placenta, but not in embryonic stem cells. These enhancers could provide explanation for a more efficient co-transcriptional recruitment of the Microprocessor, and therefore a more efficient processing of pri-miRNAs throughout the cluster in placenta. Our results point towards a new model where tissue-specific, posttranscriptional 'fine-tuning' can differentiate among miRNAs that are expressed simultaneously from a common precursor.

ARTICLE HISTORY
Received 3 September 2020
Revised 4 October 2020
Accepted 9 October 2020

KEYWORDS
C19MC; miRNA; Drosha; DGCR8; enhancer; ChIP-qPCR

Introduction

MicroRNAs (miRNAs) are short non-coding RNAs that form ribonucleoprotein complexes with Argonaute (AGO) proteins to fine-tune the expression of their target mRNA molecules. These approximately 22-nucleotide-long single-stranded nucleic acids are formed via consecutive cleavage and maturation steps from long primary transcripts (pri-miRNAs): during the canonical pathway in animal cells, the imperfect secondary structured hairpins are cleaved from the transcript by the Drosha/Dgcr8 Microprocessor complex, and the so-formed precursor miRNAs (pre-miRNAs) are transported out from the nucleus by the Exportin-5 system. In the cytoplasm, another RNaseIII type enzyme, Dicer removes the apical loop of the hairpin, forming short, double-stranded RNA molecules with 3′ overhangs of two nucleotides. Several subsequent maturation events result in a formation of an RNA-induced silencing complex (RISC) where one strand of the Dicer cleavage product becomes associated with an AGO protein and matures to an effector complex. The so-formed RISC starts scanning the mRNA population in the cell to find its target, typically a short sequence in the 3′ untranslated region complementary to the 'seed sequence' (two to eight nucleotides at the 5′ part) of the miRNA; the target mRNA is then either degraded or its translation is inhibited by various mechanisms [1,2].

miRNA genes can be located in introns (or even in exons) of protein-coding and non-coding transcripts (even in the coding region of DGCR8 itself, see [3]), where their expression is linked to the host gene transcription, but not necessarily coupled to splicing and maturation of the host transcript [4]. On the other hand, miRNA genes can be regulated by their own promoters, even if positioned in an intron [5]. Previous studies revealed that miRNAs are often clustered in the human genome [6], and such clustered miRNAs are functionally linked [7]. During evolution, several long miRNA clusters have been formed, and in primates, they are typically involved in stem cell regulation and placenta physiology [8,9]. Two exceptionally long miRNA clusters (MC) in the human genome are located on chromosome 14 (the C14MC, with 52 miRNA genes) and on chromosome 19 (the C19MC, with 46 miRNA genes), and there is emerging evidence that at least C19MC controls migration and invasion of human trophoblasts and take part in cell-to-cell communication during pregnancy [8–10]. The genomic loci of both clusters are complex: the C14MC is interrupted by a C/D

CONTACT Tamás I. Orbán orbant@biomembrane.hu Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

Supplemental data for this article can be accessed here.

snoRNA cluster and it is currently unknown if they are regulated independently [8,9], whereas the C19MC is located nearby the short miR-371-3 cluster which seems to have a distinct regulation [11]. Another common aspect of these long clusters is that their expression is regulated by imprinting: in the placenta, the C14MC is expressed from the maternally inherited allele, whereas the C19MC is expressed exclusively from the paternally inherited chromosome [8,9,11].

Although previous reports indicated that all miRNAs in a cluster share a common regulation, several recent studies found individual miRNAs showing distinct expression patterns, indicating more elaborate regulatory mechanisms (for a recent review, see [12]). As a prominent example, components of the let-7 cluster mature differently from the same polycistronic transcript, and due to the high variability of individual processing, pri-miRNA levels were found to be a better indicator for mature miRNA expression than primary transcription [13]. Moreover, RNA binding proteins such hnRNPA1 can bind to the stem-loop structure of distinct pre-miRNAs in a cluster, providing separate regulation for individual miRNAs, as shown for the hsa-miR-18a in the miR-17-92 (oncomiR-1) cluster [14–16]. Clearly, common transcriptional regulation is a necessary but may not be a sufficient condition for similar expression levels of miRNAs within the same cluster.

The C19MC has an additional unique feature: there are several Alu elements dispersed throughout the cluster, located typically flanking the pre-miRNA sequences [17–19]. This genomic arrangement and the fact that many of the miRNAs in the cluster share similar seed sequences indicate that the C19MC were formed by gene duplication events mediated by the Alu elements during primate evolution, and the common ancestor sequence was most likely an ancient miRNA similar to the miR-371 species [17,20]. The cluster is expressed predominantly in embryonic stem cells (ESCs) and in the reproductive system including the placenta [21,22], but some individual miRNAs from the cluster (such as hsa-miR-498) were found to be expressed in the foetal brain [23]. Moreover, previous studies of C19MC regulation remained controversial: earlier investigators claimed that due to the presence of Alu sequences, the cluster is transcribed by RNA polymerase III [18]. In a later study, however, it was revealed that this is not the case but the C19MC is transcribed by RNA polymerase II as a primate-specific, long non-protein-coding transcript with a complex splicing pattern, and the miRNA genes are intron-encoded [19]. Following that, Bellemer *et al.* provided evidence that during miRNA maturation, the Microprocessor complex recognizes the intron-coding long transcript near the site of transcription, and after Drosha cleavage, the DGCR8 remains attached to the precursors for longer [24]. However, the functional role of this latter association remains unclear, as well as the elaborate, potentially distinct regulatory processes of individual miRNAs from this large transcript.

In this study, we aimed to understand the tissue-specific transcriptional and posttranscriptional regulation of miRNAs expressed from the C19MC cluster. By examining and comparing human ESC lines with placenta-derived cells, we provided evidence that there is a strong correlation between the position and the expression level of a given miRNA from the cluster: significantly lower miRNA steady-state levels could be detected towards the 3′ region of the cluster in ESCs. We could reveal that the major source of this difference is originated at the posttranscriptional level of miRNA maturation, showing a decreasing pri-miRNA processing efficiency towards the 3′ region of C19MC, mediated by the lowered local availability of Drosha. Moreover, we could show that there is a placenta-specific promoter that contributes to tissue specificity, and by performing chromatin immunoprecipitation with antibody against Drosha, followed by quantitative real-time PCR (ChIP-qPCR), we could provide evidence that in placenta, Drosha is associated with certain enhancer regions located upstream of C19MC. These data point towards a potential enhancer-mediated, cotranscriptional recruitment of the Microprocessor to the transcription complex, providing the basis for the more efficient processing of miRNAs throughout the entire cluster in placenta. Our results can contribute to a new model of miRNA expression from long clusters, where tissue-specific promoters or enhancers could influence the processing of miRNAs from a common long non-coding transcript.

Results

C19MC miRNA expression levels show a position-dependent profile

Previous studies indicated that C19MC is expressed predominantly in embryonic stem cells and placenta [21,22] so we started analysing small RNA sequencing data from these tissue types (see Methods). Examining the abundance of miRNAs expressed from this cluster, we found that in placenta samples, these miRNAs contribute to a very high fraction, up to 40% of the total cellular miRNA population (Fig. 1A). This high expression level was in contrast to that in ESCs, where all C19MC species were detected but their overall expression was around 1% of the total miRNA pool (Fig. 1A and 1B). Individual miRNAs of the cluster showed heterogeneous expression levels in both tissue types, being more abundant in placenta samples (Fig. 1B). When the miRNA expression levels were compared with their genomic positions two different trends were revealed: in placenta, the expression of miRNAs slightly increases towards the 3′ end of the cluster, while in ESCs, the expression levels show a decreasing tendency (Fig. 2A). Moreover, if data were normalized to the corresponding placenta expression level, this position-dependent relative decrease in miRNA levels was further enhanced (Fig. 2B). Thorough analysis of the data reveals that the expression difference between the two tissue sample sets doubles at approximately every 12th miRNA, so the ~10-fold difference at the beginning of the cluster emerges to a ~ 135-fold difference at the 3′ end. Similar tendency could also be observed when independent datasets from Okae et al. [25] or from Mong et al. [10] were analysed (Suppl. Fig. 1). The data represents steady-state expression levels which are the combined results of transcription efficiency and RNA decay processes. Therefore, we aimed to investigate which

orban.tamas_188_24

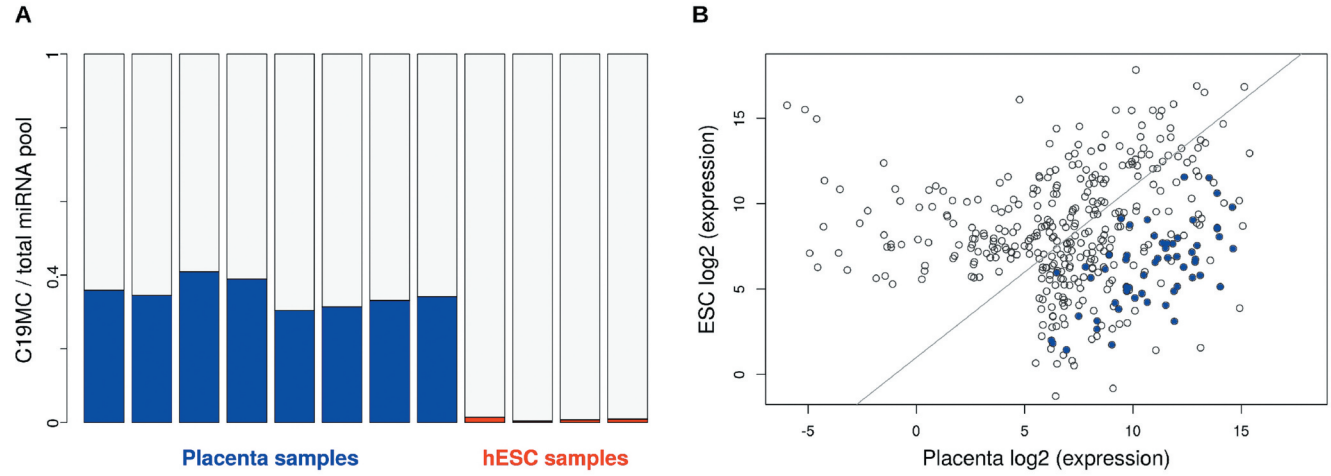


Figure 1. Investigation of mature C19MC miRNA expression. (A) Abundance of C19MC miRNAs in the total cellular miRNA population of placenta (n = 8; blue bars) and ESC (n = 4; red bars) samples (see Materials and Methods for dataset identifiers). (B) Expression levels of individual C19MC miRNAs. Averages of log_2-transformed CPMs are shown in placenta (x-axis) and in ESC (y-axis). C19MC miRNAs are marked by blue dots.

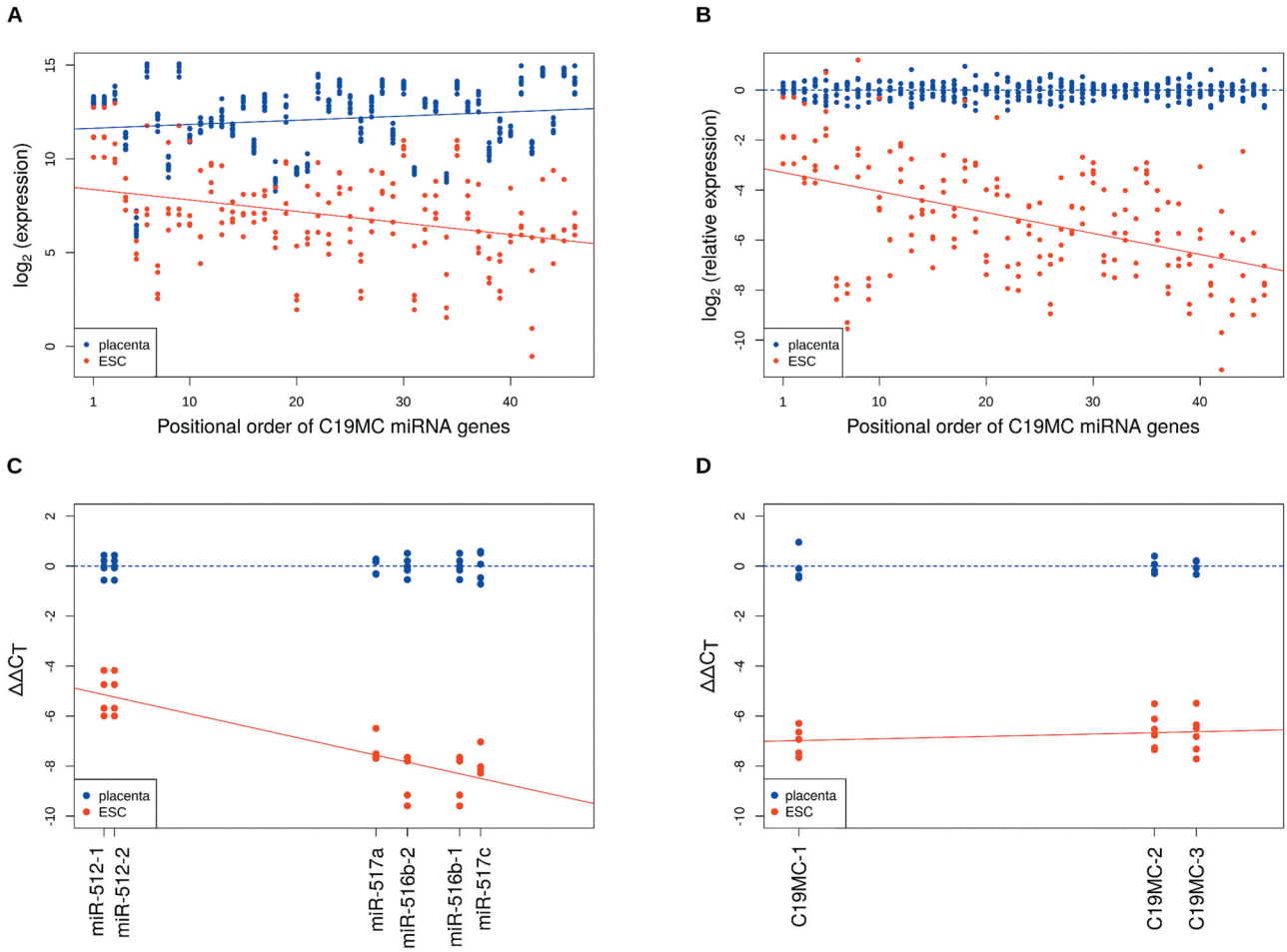


Figure 2. C19MC miRNA levels are genomic position-dependent. MiRNAs encoded by the same pre-miRNA (5p and 3p arms) were summed and ordered by their genomic position from first to 46th (x-axis). Y-axis shows log_2-transformed (A) CPMs and (B) CPMs normalized to average placenta levels for each sample (placenta (n = 8; blue dots); ESC (n = 4; red dots)). Blue and red lines illustrate fitted linear models for placenta and ESC samples, respectively. For panel **(B)**, the β-coefficient of regression is −0.08401, with *p* value of 3.5e-12. (C) Small RNA-seq results were reproduced by qRT-PCR on selected miRNA targets (β-coefficient: −0.08134, *p* value: 2.53e-06; blue dots are placenta samples (n = 5), red dots are ESC samples (n = 4)). (D) Pri-miRNA expression levels were measured by qRT-PCR and were found to be position-independent (β-coefficient: 0.00928, *p* value: 0.321). Primers for detecting the indicated PCR amplicons are listed in Suppl. Table 1.

level of regulation is responsible for this tissue-specific, position-dependent miRNA expression profile.

The position effect is attributed to tissue-specific pri-miRNA processing efficiency

To validate the expression profiles in the two tissues, we measured the steady-state levels of selected miRNA species by quantitative real-time PCR in the placenta-derived JAR and in the human embryonic stem cell originated HuES9 cell lines. The position-dependent profile was indeed confirmed (Fig. 2C): expression levels in hESCs were lower in general but when compared to their corresponding levels in placenta samples, miRNAs located at the 5′ regions of the cluster had higher abundance than those from the 3′ regions (e.g. hsa-miR-512 versus hsa-miR-517c). This tendency of gradual decrease towards the 3′ end in stem cells was statistically significant when linear regression models were fitted on the data (Fig. 2B and 2C). However, when pri-miRNA transcript levels were examined, this position-dependent decrease in abundance was not detected, although the steady-state expression level difference between the two cell types could still be revealed (Fig. 2D).

In order to further characterize this phenomenon, we quantified the processing efficiency by measuring the ratios of the unprocessed and the total pri-miRNAs in the two tissue types (Fig. 3A, and see also Materials and Methods). Interestingly, the processing was more efficient in hESCs but showed a position-dependent decrease towards the end of the cluster when compared to placenta cells (Fig. 3B).

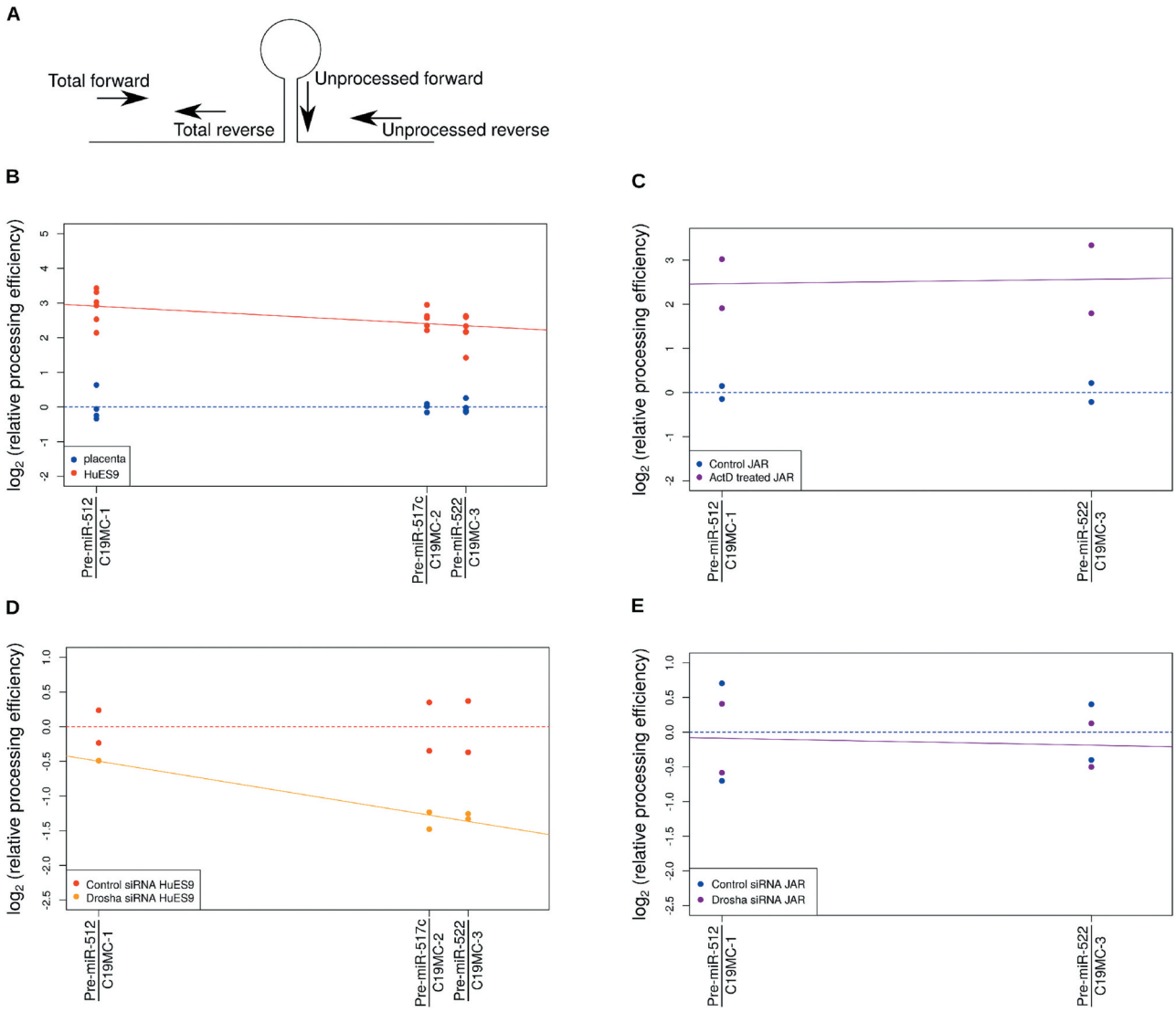


Figure 3. Position dependence of pri-miRNA processing efficiency. (A) Processing efficiency was measured at multiple positions of the cluster. At each position, two pairs of PCR primers were used: for the total pri-miRNA level measurements, both primers are located outside the pre-miRNA stem-loop; for the unprocessed pri-miRNA levels, a forward primer on the pre-miRNA stem-loop and a reverse primer downstream from the pre-miRNA were used. (B) Processing efficiency shows a position-dependent difference between placenta (n = 4; blue dots) and ESC samples (n = 6, red dots) (β-coefficient: −0.014825, *p* value: 0.0184). (C) Processing efficiency difference between control (n = 2; blue dots) and ActD treated JAR cells (n = 2; purple dots) was not affected by chromosomal position (β-coefficient: 0.002638, *p* value: 0.9256). (D) Drosha depletion by siRNA (n = 2; orange dots) reduced processing efficiency in a position-dependent manner in HuES9 cells (β-coefficient: −0.0228201, *p* value: 0.00131). (E) In JAR cells, Drosha depletion (n = 2; purple dots) had no effect compared to the control siRNA treatment (n = 2, blue dots) (β-coefficient: −0.002625, *p* value: 0.881).

These results indicated that there is a tissue-specific, post-transcriptional process acting on the pri-miRNA species.

Position-dependent processing is influenced by Microprocessor recruitment

We first hypothesized that pri-miRNA processing could be related to the transcript levels: in placenta, the much higher transcript level might saturate the miRNA processing machinery, causing a lower processing efficiency. To test this theory in the JAR cell line, we performed actinomycin D treatment to inhibit transcription and quantified the processing efficiency of pri-miRNAs in the C19M cluster. The transcripts in JAR cells had half-lives of less than 1 hour, and after 8 hours of treatment, the levels decreased to what was comparable to that in hESCs (Suppl. Fig. 2). At that time point, when measured at the two ends of the cluster, the processing efficiency did not show the position-dependent difference detectable in hESCs, although the processing efficiency increased because of the lower pri-miRNA input due to transcription inhibition (Fig. 3C). These results demonstrated that transcription level *per se* could not explain the position-dependent decrease in processing efficiency.

There are several RNA binding proteins involved in miRNA stability and processing, and a significant portion of them have been shown to bind the stem-loop structure of various pri-miRNAs, influencing their further maturation steps [12]. We searched for such factors that could potentially regulate miRNAs of the C19MC and found that hnRNPA1 and SRSF1 could be predicted to bind to several sequences in the cluster. However, when the scores of these binding sites were examined, they showed no correlation with the miRNA expression levels in the cluster (Suppl. Fig. 3). This indicated that although these (and potentially other) factors could regulate miRNA members of the cluster, their activity could not explain the tissue-specific, position-dependent processing difference.

To investigate whether the Microprocessor level or its recruitment is connected to processing efficiency, we examined the levels of Drosha in the cells. The mRNA levels were comparable in the two tissues so we knocked down Drosha by siRNA treatment (Suppl. Fig. 4) to see if this influences the processing. It was intriguing that in contrast to JAR cells (Fig. 3E), Drosha depletion in ESCs caused a strong decrease in processing efficiency towards the 3′ end of the cluster and it further enhanced the already present position-dependent processing decline within the cluster (Fig. 3D). When performing Drosha overexpression in JAR cells, no changes in the overall processing could be detected. On the other hand, when Drosha was expressed at higher levels in ESCs, the processing efficiency showed an increase towards the 3′ end of C19MC (Suppl. Fig. 5). These results supported the idea that during miRNA maturation from this cluster, tissue-specific processes regulate the local concentration of Drosha and thereby the Microprocessor, significantly influencing its access to the RNA template and therefore its processivity.

A placenta-specific promoter could cause more efficient co-transcriptional Microprocessor recruitment

To further study the tissue-specific aspect of the position-dependent processing regulation, we analysed epigenetic datasets from the NIH Roadmap Epigenomics Project. By examining data from human embryonic stem cell and placenta samples, we could reveal that there are distinct promoter activities in the two cell types. CAGE-seq data analysis [26] points towards 5 different transcription start sites, 4 of which are placenta-specific (TSS1-4), whereas TSS5 is active in ESCs (Fig. 4A). Moreover, a previously identified CpG island overlaps TSS1, and its methylated status in several tissues but placenta [11] indicates that this promoter region is exclusively active in the placenta, initiating the transcription of a longer RNA species. In contrast to that, in ESCs transcription is initiated 3′ downstream of this region which is also supported by the validated binding site of the stem cell-specific Nanog transcription factor, as well as the TATA-Box-Binding protein (TBP) site at TSS5 (Fig. 4A, see also the H3K4me3 marks of active transcription start sites [27]). By investigating our placenta and ESCs samples, we could indeed show that a longer transcript is present only in placenta samples (Fig. 4B). The two putative promoters show distinct epigenetic marks (different patterns of modified histones and DNase hypersensitive sites, Fig. 4A) which could well explain the tissue-specific differences in transcription intensity. In addition, the extensive regions with H3K27 acetylation and H3K4 trimethylation in placenta samples not only indicate active enhancers and strong transcription start sites, but their presence in broad regions could also boost Microprocessor recruitment [28]. In order to investigate this potential aspect of regulation, we carried out ChIP-qPCR experiments using antibody against Drosha to analyse its binding to selected regions upstream of C19MC. We used the DNase I hypersensitive (DHS) sites to predict putative enhancers and designed real-time PCR primers to amplify seven selected DNA segments (DHS 1–7, Fig. 4A). For background normalization, we used two irrelevant promoter sequences, the elongation factor 1-alpha (EF1α) promoter being constitutively active, and the cardiac-specific alpha-actin (ACTC1) promoter being inactive in both tissues, to exclude potential binding signals related to Drosha's non-canonical functions (e.g. in DNA repair, see [29]). By performing the experiments we could reveal Drosha association with selected sequences in placenta, showing a highest peak on the DHS 6 putative enhancer region; in ESCs, however, only very weak Drosha signals could be obtained (Fig. 4C and Suppl. Fig. 6). These findings were in line with the previous results on the distinct, tissue-specific promoters. In addition, an independent support for the tissue-specific promoter effect was revealed when we analysed the datasets from Mong et al.: the authors used the CRISPR/dCas9 Synergistic Activation Mediator system to activate the stem cell-specific C19MC promoter in HEK293 cells where it is normally transcriptionally silent. [10] By analysing their miRNA expression data, we could indeed provide evidence that the miRNA expression profiles show the similar, genomic position-dependent gradual decrease that

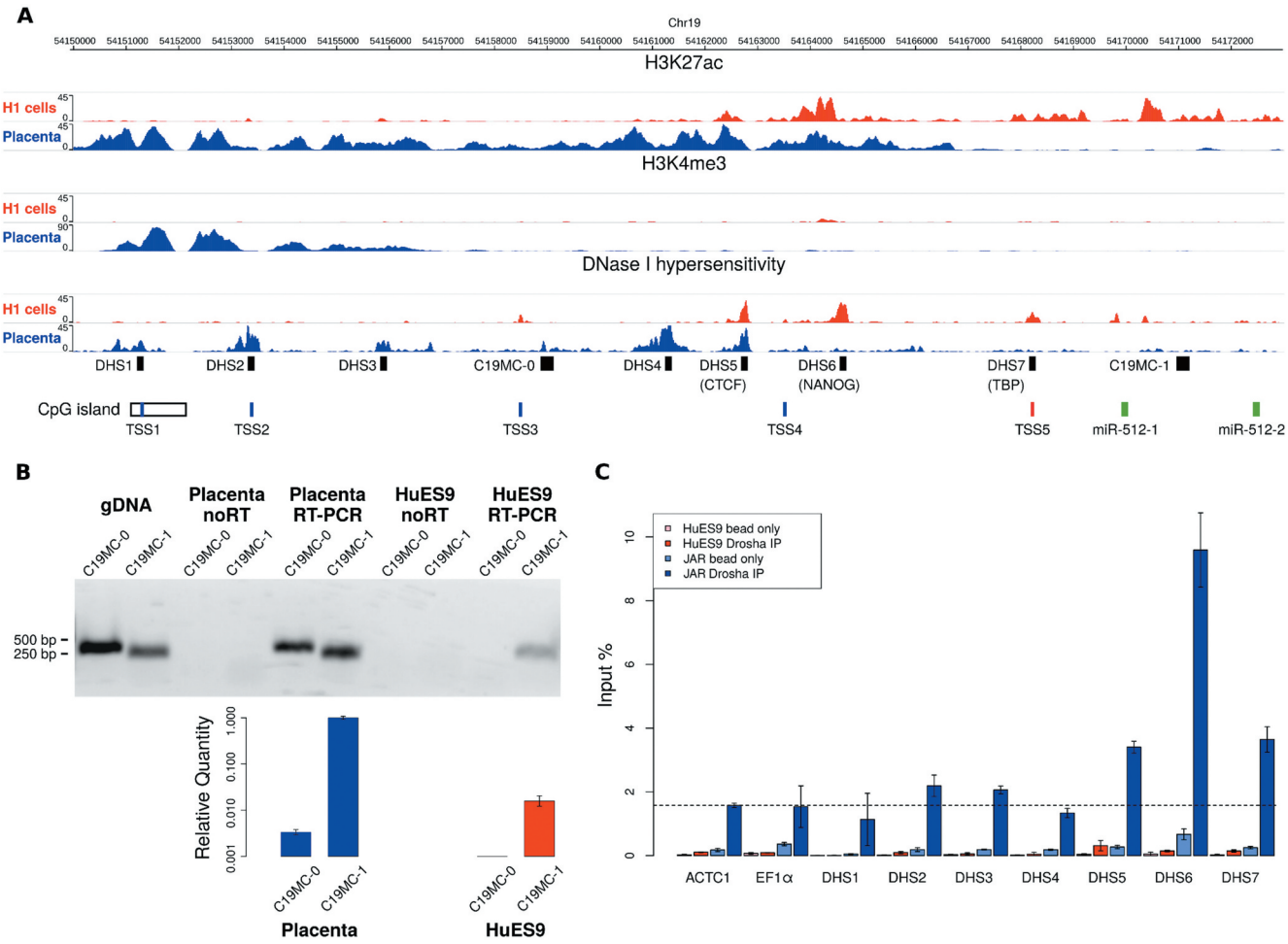


Figure 4. Promoter usage analysis in placenta tissues and in embryonic stem cells (H1 cell line as example) by investigating the epigenetic marks on a 22 kb DNA segment upstream of C19MC. (A) H3K4me3 modifications indicate two putative tissue-specific promoters with distinct active enhancer regions (H3K27ac modifications and DNase hypersensitive regions). The upper promoter is active in placenta, with four identified transcription start sites (TSS1-4 by CAGE-Seq data [26], shown in blue); while a more downstream alternative promoter initiates transcription in ESC (TSS5, shown in red). The TSS1 is located in a longer CpG island shown to be methylated in non-placental tissues [11]. PCR primer targets are marked by black boxes: the seven putative enhancers (DNase I hypersensitive sites, DHS1-7) and the two pri-miRNA transcripts (C19MC-0 for the longer, and C19MC-1 for the shorter form). The first two miRNAs of the cluster (hsa-miR-512-1 and –2) are marked by green boxes (B) The pri-miRNA with a longer 5′ end (C19MC-0) was detected only in the placenta sample (RT-PCR end point detection, gel image on upper panel). JAR genomic DNA (gDNA) was used for primer testing; the 'noRT' samples serve as negative controls. For the placenta and the HuES9 samples, C19MC-0 and C19MC-1 transcripts were also quantified by real-time PCR, using PolR2A as endogenous control; error bars show S.E.M. values. (C) ChIP-qPCR results analysing Drosha binding to the DHS sites and two unrelated promoter regions (ACTC1 and EF1α), the latter two being used for background control (indicated with a dashed line). Significant enrichment for DHS5-7 regions were detected in JAR cells (blue bars), but only weak binding in HuES9 cells (red bars, see separately also in Suppl. Fig. 6). Beads without antibody were used as ChIP negative controls (light blue and light red bars for JAR and HuES9, respectively), error bars indicate standard deviations.

were detected in other ESCs (Suppl. Fig. 1B). As a conclusion of our various analyses, the strong association of Drosha with the promoter/enhancer regions in placenta cells could explain the more efficient co-transcriptional recruitment of Drosha/DGCR8 in these cell types which ensures that the Microprocessor could efficiently process the entire region of the long non-coding transcripts from the miRNA cluster on the human chromosome 19.

Discussion

In this study, we investigated the regulation of the miRNA cluster on human chromosome 19, especially focusing on the tissue-specific differences between placenta and human embryonic stem cells. The C19MC is one of the longest human miRNA clusters and the presence of scattered Alu sequences among the pre-miRNAs indicates that retrotransposon elements-mediated gene duplication events could have been responsible for the formation of the cluster [17–19]. For this reason, originally it had been proposed that the cluster is transcribed entirely by RNA polymerase III [18], however, it was later proved that RNA polymerase II transcribes this heavily spliced long RNA molecule [19]. Those studies all suggested that the numerous encoded miRNA species are processed simultaneously and at similar levels after transcription, however, we could prove that this is in fact not the case: apart from differences in individual miRNA levels, in human ESCs, there is a tissue-specific, chromosomal position-dependent miRNA expression profile which is different from what can be detected in placenta. We could provide evidence that there is a gradual decrease in the steady-state levels of miRNAs when moving towards the 3′ end of the cluster in

ESCs, indicating tissue-specific regulatory differences (Fig. 2A-C). Investigating the potential mechanism(s) behind this observation we could show that such positional difference could not be detected on the pri-miRNA levels, demonstrating that at least on the transcript level, the 5′ and 3′ regions of this long noncoding RNA are present at similar levels in the examined tissues (Fig. 2D). Moreover, when transcription was hindered in a placenta cell line and when it was comparable to the lower level found in hESCs, the pri-miRNA processing from the 5′ and the 3′ end of C19MC did not show the gradual decrease present in stem cells (Fig. 3B-C). These results clearly proved that transcription level *per se* cannot explain the observed phenomenon and posttranscriptional mechanism(s) are responsible for the tissue-specific positional effect.

There are examples of other mammalian miRNA clusters where individual miRNAs are regulated distinctly from their neighbours: the hsa-miR-18a, for instance, shows different expression in its cluster, and posttranscriptional regulation by the hnRNPA1 protein has been found to be responsible for that [14–16]. We have tested some RNA regulatory proteins for the C19MC and although pri-miRNA binding was clearly predicted throughout the cluster, an overall effect on the miRNAs expression could not be detected (Suppl. Fig. 3). When testing the pri-miRNA processing efficiency, a clear difference was revealed between the two examined tissues: the efficiency showed a position-dependent gradual decrease in hESCs as compared to placenta (Fig. 3B). A previous study revealed that the DGCR8 protein component of the Microprocessor seems to stay longer on the pre-miRNA after the Drosha cleavage but the authors could not explain the significance or the consequence of this observation [24]. Based on this we hypothesized that the Microprocessor availability or the difference in its local activity might be responsible for the positional decrease in miRNA processing in hESCs. To investigate the mechanism, we tested the miRNA maturation by knocking down or overexpressing Drosha in the examined tissues. It was intriguing to see that changes in the level of Drosha could strongly influence the positional effect in hESCs but did not affect the miRNA processing efficiency in placental cells (Fig. 3 and Suppl. Fig. 5). The results indicated that in stem cells, the availability of the active Microprocessor complex could be limited which would explain the lower releasing efficiency of pre-miRNAs towards the 3′ end of this particularly long primary noncoding transcript.

In a recent study [30], Donayo and colleagues revealed that apart from previous findings of the selective regulation by hnRNPA1 [14], an additional hierarchical processing of pre-miRNAs in the miR-17-92 cluster results in different expression levels of the mature miRNA species. Moreover, they showed that the oncogenic amplification of this cluster leads to the sequestration of Microprocessor complexes in the cells, lowering the processing efficiency of other miRNA clusters. Such a mechanism could be in line with our findings but would not explain the tissue-specific differences in processing of C19MC. So, what could cause the decrease in availability of the Drosha/DGCR8 complex in embryonic stem cells as compared to placenta cells? One explanation could lie in the transcriptional regulation of the cluster. Here we described a placenta-specific promoter that overlaps with the previously described promoter of the cluster [11] and it is very active in placenta but silent in hESCs (Fig. 4A and 4B). This clearly explains the much stronger transcription activity observed in placenta cells but the detected epigenetic histone marks also point to the idea that this region can work as a potential enhancer: as described earlier for certain 'super-enhancer' regions [31,32], it can more efficiently recruit proteins and regulators to the RNA polymerase II apparatus, carrying out a more robust co-transcriptional processing of the transcript. Such a mechanism was shown to exist for long noncoding RNAs [33], including co-transcriptional recruitment of the Drosha/DGCR8 complex [13,28,34]. As a similar regulation could well explain our results showing tissue-specific differences in formation of miRNAs from C19MC, we performed ChIP-qPCR experiments using an antibody against Drosha, and could provide evidence that Drosha is indeed associated with several potential enhancer sequences located in the region upstream of the C19MC locus (Fig. 4C). This result fits well into the model of Microprocessor recruitment to the transcription apparatus already at the initiation site of transcription [28,34], and the level and the local availability of Drosha could also contribute to explaining the efficient processing of miRNA precursors throughout the C19MC region in placenta (Fig. 5). By investigating similar, super-enhancer mediated recruitment mechanisms it seems that it is not limited to the more efficient co- or posttranscriptional processing of the transcripts, such as splicing or pre-miRNA cleavage, but could also contribute to the more efficient recruitment of RNA export factors [35]. In addition, such tissue-specific sequestration of the miRNA processing apparatus could help in understanding why the misregulation of C19MC could be connected to tumour formation, as elevated level of C19MC transcription was found to be an associated marker of triple-negative breast cancers [36], or when a genomic rearrangement mediated fusion of C19MC with the TTYH1 gene resulted in the formation of embryonal brain tumours [37]. On the other hand, it still needs to be addressed how the previously described imprinting regulation of

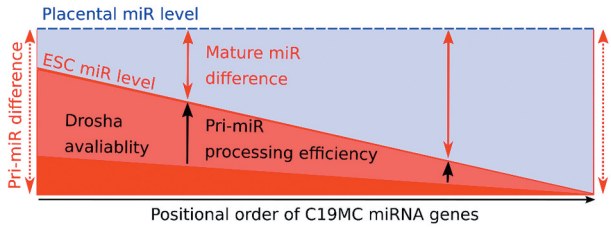


Figure 5. Cumulative model of the position-dependent miRNA maturation at C19MC. Normalized mature miRNA levels of C19MC show a position-dependent difference between placenta and ESCs (indicated with a blue and a red line, respectively). The constant difference of primary transcript levels (pri-miRNAs levels, represented by vertical dotted red arrows) could not explain this phenomenon, however, pri-miRNA processing efficiencies (represented by vertical black arrows) significantly contribute to this positionally increasing difference. Based on the results of the present study, tissue-specific difference in processing efficiency of C19MC is regulated by Drosha/Microprocessor level and local availability.

C19MC [11] is connected to the identified enhancer region and Microprocessor recruitment.

The Drosha recruitment model supported by our ChIP-qPCR data could explain the local Drosha availability and the efficient processing of C19MC but fails to explain why the ESCs are more sensitive to changes in Drosha levels whereas placenta cells are not (Fig. 3 and Suppl. Fig. 5). When considering expression levels, the transcription of C19MC in placenta is very high, and miRNAs from the cluster make up a huge amount (40%) of the total miRNA pool in the cell (Fig. 1A). Together with the efficient recruitment of Drosha to the promoter in placenta, this massive amount of miRNA could 'titrate out' available Drosha (e.g. Microprocessor complexes) from other, less transcriptionally active miRNA loci, similarly to what was shown for the highly expressed miR-17 ~ 92 cluster [30]. In stem cells, however, the C19MC is expressed in a much lower level (approx. 1% of the total cellular miRNA pool, see Fig. 1A), and other strongly expressed miRNA loci could 'titrate out' Drosha locally, making the Microprocessor complexes less available for the C19MC in stem cells, therefore being more sensitive to the overall Drosha level. Our data analyses indicate that the mentioned miR-17 ~ 92 cluster (making up 10–30% of the total miRNA pool) or the miR-302 ~ 367 cluster (also making up a significant, 10–20% of the total miRNA pool) could act such local "recruiters" in stem cells. Nevertheless, to gain a more detailed picture on the molecular mechanisms, further studies should also address the level and potential recruitment of DGCR8, or other associated factors of the Microprocessor complex in relation to C19MC processing. Activating the placenta-specific promoter by the already described CRISPR/dCas9 Synergistic Activation Mediator system [10] could also provide further details about the connection among transcription-recruitment-processing; however, these experiments are beyond the scope of the current investigation.

In conclusion, our results could reveal a new transcription-coupled, tissue-specific regulatory mechanism of miRNA maturation from long clusters. Although we could provide evidence that the rate-limiting step in the processing of the 3′ region of C19MC is the local availability of the effector Drosha/DGCR8, further studies are required to elucidate the exact details of this mechanism. Nevertheless, together with the described regulation of other oncogenic clusters, our results can well contribute to a novel model of a tissue-specific regulation of miRNA maturation from long genomic clusters.

Materials and methods

Cell culture maintenance and treatments

The JAR placental choriocarcinoma cell line was maintained under standard conditions in 5% CO2 incubator at 37°C, in RPMI-1640 medium supplemented with 10% FBS and 1% penicillin-streptomycin (Gibco). Transcription inhibition was achieved by 5 μg/mL of actinomycin D (Sigma-Aldrich) treatment, and samples were taken at 1, 2, 4 and 8 h after treatment.

The HuES9 embryonic stem cell line was originally provided by Dr. Douglas Melton (HHMI). The cells were cultured on Matrigel (Corning) coated six-well plates in mTeSR medium (Stemcell Technologies) and were grown to 70% density before transfection. For nucleofection, cells were treated overnight with the rho-associated protein kinase (ROCK) inhibitor Y-27,632 (Selleckchem), then detached with Accutase (Thermo Fisher Scientific) and washed with 1× PBS at 37°C.

10^6 cells per reaction were used for electroporation using the A-023 program with the Amaxa Human Stem Cell Nucleofector Kit 1 (cat. #: VPH-5012, Lonza) for HuES9 cells, or the X-005 program with the Amaxa Cell Line Nucleofector Kit V (cat. #: VVCA-1003, Lonza) for JAR cells, according to the manufacturer's protocol; the cells were seeded on six-well plates and harvested 24 hrs after transfection. For knock-down experiments, 25 nM of siRNA targeting Drosha (catalogue #4,390824) and a negative control (cat. #4,390843) were used as recommended by the manufacturer (Thermo Fisher Scientific). For Drosha overexpression experiments, GFP-tagged Drosha expressing plasmid was used (Addgene #62520 plasmid, [38]) and transfection efficiency was verified using fluorescence microscopy. In case of HuES9 cells, the similar nucleofection method was applied as described above. For efficient plasmid transfection into JAR cells, the FuGENE® HD reagent (Promega) was used according to the manufacturer's instruction.

Study participants, placenta sample collection and handling

Study participants had been recruited during routine prenatal care or following hospital admission during the third trimester of pregnancy at 1st Department of Obstetrics and Gynaecology, Semmelweis University, Budapest, Hungary. In this study, five placenta samples were collected right after C-sections at term pregnancies without any indications of gestational complications. The study protocol was approved by the Scientific and Research Ethics Committee of the Medical Research Council (ETT TUKEB) [No: 24387–2/2016] and written informed consent was obtained from each patient. The research was conducted in accordance with the Declaration of Helsinki.

Placenta samples were collected according to the protocol described by Pasupathy *et al.* [39]. Four areas suitable for sampling were located on the maternal surface; damaged areas (calcification, haematoma, etc.) were excluded. About 1–2 mm from the basal membrane was removed and pea-sized tissue samples were taken from the placental cotyledons. The samples were washed twice in 1× PBS solution at 4°C and placed in RNAlater™ stabilizing solution (Thermo Fisher Scientific) to avoid RNA degradation.

RNA isolation

Total RNA isolation was done by using TRIzol™ Reagent (Thermo Fisher Scientific) as described in the user guide. RNA integrity was analysed by agarose gel electrophoresis,

sample purity and concentration were measured by a Nanodrop spectrophotometer (Thermo Fisher Scientific).

cDNA preparations and real-time PCR quantifications

For mRNA or pri-miRNA analysis, 1 µg total RNA was reverse transcribed by random oligomers using the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific); cDNA samples were diluted 1:10 before subsequent amplifications. Drosha mRNA level was measured by using TaqMan® Gene Expression Master Mix (Thermo Fisher Scientific) and pre-designed Drosha TaqMan® assay (cat. #4331182). In the case of C19MC pri-miRNA, RT-PCR was done by using SYBR Green PCR Master Mix with custom-made PCR primers (Suppl. Table 1). Real-time PCR measurements were done on a StepOnePlus™ platform (Thermo Fisher Scientific) according to the manufacturer's instructions. The ΔΔCt method was applied for relative quantifications, using a set of endogenous control mRNAs for normalization: for TaqMan® analyses, the PolR2A (assay Hs00172187_m1), and the RPLP0 (assay Hs9999902_m1); for SYBR® Green assays, custom-made primers for PolR2A and RPLP0 (for details, see Suppl. Table 1).

For mature miRNA quantification, the expression analysis was performed using the miRCURY LNA™ Universal RT miRNA PCR Assay (Qiagen), according to the manufacturer's instructions. Briefly, RNA samples (5 ng/µl) were reverse-transcribed and the UniSp6 RNA spike-in template was added to each reaction for controlling the quality of cDNA synthesis. cDNA samples were diluted 1:80 before subsequent amplifications. RT-PCR was done by using miRCURY SYBR® Green master mix (Qiagen) and real-time PCR reactions were run on a StepOnePlus™ platform (Thermo Fisher Scientific) according to the manufacturer's protocol. Pre-designed assays were used to measure the levels of hsa-miR-512, hsa-miR-517a, hsa-miR-516b and hsa-miR-517 c. In these cases, the hsa-miR-103a internal control miRNA was used for normalization during the relative quantifications by the ΔΔCt method.

Pri-miRNA processing efficiency was calculated with normalization of uncleaved pri-miRNA level to the total amount of pri-miRNA (see also Fig. 3A); primers are listed in Suppl. Table 1.

Chromatin immunoprecipitation (ChIP) qPCR

ChIP-qPCR measurements were done using the ChIP Kit (cat. #ab500, Abcam) according to the manufacturer's instructions. Briefly, 10^6 cells were collected per ChIP, chromatin was cross-linked by formaldehyde (Sigma), and cells were lysed and sonicated for 10 minutes. Sheared DNA fragment length was analysed by gel electrophoresis. Immunoprecipitations were done by anti-Drosha antibody (cat. #ab12286, Abcam), and for positive control, the anti-H3 antibody (cat. #ab1791, Abcam); as a negative control, only Protein A beads were used. For normalization purposes, input chromatin DNA was used. qPCR was done by using SYBR Green PCR Master Mix with custom-made PCR primers (Suppl. Table 1) on a StepOnePlus™ platform (Thermo Fisher Scientific) according to the manufacturer's instructions.

Western blot

Cell lysates were collected 24 h after transfection with siRNAs. Samples were briefly sonicated and protein concentration was measured by Lowry method. About 30 µg of protein samples per lane were run on 8% acrylamide gels and electroblotted onto PVDF membranes (BioRad). After washing with TBS-Tween, membranes were blocked by 5% milk/TBS-Tween, and subsequently incubated with Anti-Drosha antibody (cat. #ab12286, Abcam) on 4°C overnight. Membranes were washed three times with TBS-Tween and then incubated in Anti-Rabbit IgG secondary antibody (Jackson ImmunoResearch) solution for 1 h at room temperature. Membranes were washed twice with TBS-Tween, and for signal detection ECL reagent (Thermo Fisher Scientific) was used, and the membranes were exposed to Agfa films. Anti-beta Actin antibody (cat. #ab20272, Abcam) was used as a loading control and to normalize Drosha expression. Expression levels were determined by densitometry of the scanned images using the ImageJ software [40].

Next-generation sequencing data analyses, visualization and statistics

Publicly available sequencing data were reanalysed from the PRJNA187509 NCBI BioProject (placenta samples) [41] and from SRA data with accession name: SRR026761, SRR1616134, SRR1616135, SRR1203788 (hESC samples) [42–44]. Histone modification data (H3K27ac and H3K4me3) from hESC and placenta samples (E003 H1, E008 H9, E014 HuES48, E015 HuES6, E016 HuES64 and E091 Placenta) were downloaded from NIH Roadmap Epigenomics [45].

Raw data were trimmed by CutAdapt (1.10) [46] and reads were mapped to hg38 assembly of the human genome by BWA aln (0.7.12) algorithm [47]. Read counting was carried out by FeatureCounts from the Rsubread package (1.26.1) [48] using miRBase (v.21) [49] annotations. Normalization was done by edgeR package (3.18.1) [50] and 'count per million' (CPM) values were used. For quantification of miRNA genes, the two arms of each miRNA were summed. Logarithmic transformed relative expression levels were calculated similarly to the ΔΔCt method: logarithmic transformed values of the target samples were normalized to the mean of logarithmic transformed values of the control samples.

Statistics and visualization were done in R(3.4.4) [51] software environment. Simple linear regression models were fitted by the lm function [52] and associated p-values were used to decide whether the explanatory variable (e.g. genomic position) has a significant influence on the response variable (e.g. relative miRNA expression). The relationship between the response and the explanatory variable was described by the slope (β-coefficient) of the fitted linear model.

Epigenomics data was displayed with the WashU Epigenome Browser [53].

Acknowledgments

The authors are grateful to Kornélia Némethy and Gerda Wachtl for excellent technical help.

Disclosure of potential conflicts of interest

No potential conflict of interest was reported by the authors.

Funding

This study was supported by the grants VEKOP-2.1.1-15-2016-00156 and VEKOP-2.3.3-15-2017-00014, and project no. 2018-1.2.1-NKP-2018-00005 from the National Research, Development and Innovation Fund of Hungary.

ORCID

Tamás I. Orbán http://orcid.org/0000-0002-3424-3428

References

[1] Bartel DP, Metazoan MicroRNAs. Cell. 2018;173:20–51.
[2] Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. Nat Rev Mol Cell Biol. 2019;20:21–37.
[3] Guo WT, Wang Y. Dgcr8 knockout approaches to understand microRNA functions in vitro and in vivo. Cell Mol Life Sci. 2019;76:1697–1711.
[4] Kim YK, Kim VN. Processing of intronic microRNAs. Embo J. 2007;26:775–783.
[5] Ramalingam P, Palanichamy JK, Singh A, et al. Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. RNA. 2014;20(1):76–87.
[6] Altuvia Y, Landgraf P, Lithwick G, et al. Clustering and conservation patterns of human microRNAs. Nucleic Acids Res. 2005;33:2697–2706.
[7] Kim YK, Yu J, Han TS, et al. Functional links between clustered microRNAs: suppression of cell-cycle inhibitors by microRNA clusters in gastric cancer. Nucleic Acids Res. 2009;37:1672–1681.
[8] Morales-Prieto DM, Ospina-Prieto S, Chaiwangyen W, et al. Pregnancy-associated miRNA-clusters. J Reprod Immunol. 2013;97(1):51–61.
[9] Malnou EC, Umlauf D, Mouysset M, et al. Imprinted MicroRNA gene clusters in the evolution, development, and functions of mammalian placenta. Front Genet. 2019;9:706.
[10] Mong EF, Yang Y, Akat KM, et al. Chromosome 19 microRNA cluster enhances cell reprogramming by inhibiting epithelial-to-mesenchymal transition. Sci Rep. 2020;10(1):3029.
[11] Noguer-Dance M, Abu-Amero S, Al-Khtib M, et al. The primate-specific microRNA gene cluster (C19MC) is imprinted in the placenta. Hum Mol Genet. 2010;19(18):3566–3582.
[12] Michlewski G, Caceres JF. Post-transcriptional control of miRNA biogenesis. RNA. 2019;25(1):1–16.
[13] Conrad T, Marsico A, Gehre M, et al. Microprocessor activity controls differential miRNA biogenesis in vivo. Cell Rep. 2014;9(2):542–554.
[14] Concepcion CP, Bonetti C, Ventura A. The microRNA-17-92 family of microRNA clusters in development and disease. Cancer J. 2012;18(3):262–267.
[15] Michlewski G, Guil S, Semple CA, et al. Posttranscriptional regulation of miRNAs harboring conserved terminal loops. Mol Cell. 2008;32(3):383–393.
[16] Michlewski G, Caceres JF. Antagonistic role of hnRNP A1 and KSRP in the regulation of let-7a biogenesis. Nat Struct Mol Biol. 2010;17(8):1011–1018.
[17] Lehnert S, Van Loo P, Thilakarathne PJ, et al. Evidence for co-evolution between human microRNAs and Alu-repeats. PLoS One. 2009;4(2):e4456.
[18] Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol. 2006;13(12):1097–1101.
[19] Bortolin-Cavaille M-L, Dance M, Weber M, et al. C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts. Nucleic Acids Res. 2009;37(10):3464–3473.
[20] Zhang R, Wang Y-Q, Su B. Molecular evolution of a primate-specific microRNA family. Mol Biol Evol. 2008;25(7):1493–1502.
[21] Bar M, Wyman SK, Fritz BR, et al. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells. 2008;26(10):2496–2505.
[22] Ren J, Jin P, Wang E, et al. MicroRNA and gene expression patterns in the differentiation of human embryonic stem cells. J Transl Med. 2009;7(1):20.
[23] Flor I, Bullerdiek J. The dark side of a success story: microRNAs of the C19MC cluster in human tumours. J Pathol. 2012;227(3):270–274.
[24] Bellemer C, Bortolin-Cavaille ML, Schmidt U, et al. Microprocessor dynamics and interactions at endogenous imprinted C19MC microRNA genes. J Cell Sci. 2012;125:2709–2720.
[25] Okae H, Toh H, Sato T, et al. Derivation of Human Trophoblast Stem Cells. Cell Stem Cell. 2018;22(1):50–63 e6.
[26] Noguchi S, Arakawa T, Fukuda S, et al. FANTOM5 CAGE profiles of human and mouse samples. Sci Data. 2017;4(1):170112.
[27] Liang G, Lin JC, Wei V, et al. Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A. 2004;101(19):7357–7362.
[28] Suzuki HI, Young RA, Sharp PA. Super-enhancer-mediated RNA processing revealed by integrative microrna network analysis. Cell. 2017;168(6):1000–14 e15.
[29] Lu WT, Hawley BR, Skalka GL, et al. Drosha drives the formation of DNA:RNA hybrids around DNA break sites to facilitate DNA repair. Nat Commun. 2018;9:532.
[30] Donayo AO, Johnson RM, Tseng HW, et al. Oncogenic biogenesis of pri-miR-17 approximately 92 reveals hierarchy and competition among polycistronic MicroRNAs. Mol Cell. 2019;75:340–56 e10.
[31] Pott S, Lieb JD. What are super-enhancers? Nat Genet. 2015;47(1):8–12.
[32] Sengupta S, George RE. Super-enhancer-driven transcriptional dependencies in cancer. Trends Cancer. 2017;3(4):269–281.
[33] Schlackow M, Nojima T, Gomes T, et al. Distinctive patterns of transcription and RNA processing for human lincRNAs. Mol Cell. 2017;65(1):25–38.
[34] Church VA, Pressman S, Isaji M, et al. Microprocessor recruitment to elongating RNA polymerase II is required for differential expression of MicroRNAs. Cell Rep. 2017;20(13):3123–3134.
[35] Viphakone N, Sudbery I, Griffith L, et al. Co-transcriptional loading of RNA export factors shapes the human transcriptome. Mol Cell. 2019;75(2):310–23 e8.
[36] Jinesh GG, Flores ER, Brohl AS. Chromosome 19 miRNA cluster and CEBPB expression specifically mark and potentially drive triple negative breast cancers. PLoS One. 2018;13(10):e0206008.
[37] Kleinman CL, Gerges N, Papillon-Cavanagh S, et al. Fusion of TTYH1 with the C19MC microRNA cluster drives expression of a brain-specific DNMT3B isoform in the embryonal brain tumor ETMR. Nat Genet. 2014;46(1):39–44.
[38] Tang X, Zhang Y, Tucker L, et al. Phosphorylation of the RNase III enzyme Drosha at Serine300 or Serine302 is required for its nuclear localization. Nucleic Acids Res. 2010;38(19):6610–6619.
[39] Pasupathy D, Dacey A, Cook E, et al. Study protocolA prospective cohort study of unselected primiparous women: the pregnancy outcome prediction study. BMC Preg Childbirth. 2008;8:51.
[40] Schneider CA, Rasband WS, Eliceiri KW. NIH Image to ImageJ: 25 years of image analysis. Nat Methods. 2012;9(7):671–675.
[41] Williams Z, Ben-Dov IZ, Elias R, et al. Comprehensive profiling of circulating microRNA via small RNA sequencing of cDNA libraries reveals biomarker potential and limitations. Proc Natl Acad Sci U S A. 2013;110(11):4255–4260.

[42] Morin RD, O'Connor MD, Griffith M, et al. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res. 2008;18:610–621.
[43] Asikainen S, Heikkinen L, Juhila J, et al. Selective microRNA-Offset RNA expression in human embryonic stem cells. PLoS One. 2015;10(3):e0116668.
[44] Chen T, Xiang JF, Zhu S, et al. ADAR1 is required for differentiation and neural induction by regulating microRNA processing in a catalytically independent manner. Cell Res. 2015;25:459–476.
[45] Roadmap Epigenomics C, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–330.
[46] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–12.
[47] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25 (14):1754–1760.
[48] Liao Y, Smyth GK, Shi W. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–930.
[49] Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47 (D1):D155–D62.
[50] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140.
[51] Ihaka R, Gentleman R. R: a language for data analysis and graphics. J Comput Graph Stat. 1996;5:299–314.
[52] Chambers JM, Hastie TJ. Chapter 4:Linear models. Chambers JM, Hastie TJ, eds. Statistical models in S.California: Wadsworth & Brooks/Cole; 1992.
[53] Zhou X, Maricque B, Xie M, et al. The human epigenome browser at Washington University. Nat Methods. 2011;8 (12):989–990.

Gene 824 (2022) 146609

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Functional indications for transposase domestications – Characterization of the human *piggyBac* transposase derived (PGBD) activities

Orsolya Kolacsek [a], Gerda Wachtl [a,b], Ábel Fóthi [a], Anita Schamberger [a], Sára Sándor [a], Enikő Pergel [a], Nóra Varga [a], Tamás Raskó [c], Zsuzsanna Izsvák [c], Ágota Apáti [a], Tamás I. Orbán [a,*]

[a] Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary
[b] Doctoral School of Biology, Institute of Biology, ELTE Eötvös Loránd University, Budapest, Hungary
[c] Max Delbrück Center for Molecular Medicine in the Helmholtz Society, Berlin, Germany

ARTICLE INFO

Edited by Andre van Wijnen

Keywords:
PGBD
PGBD5
MER75
MER85
Transposon
Cellular stress

ABSTRACT

Transposable elements are widespread in all living organisms. In addition to self-reproduction, they are a major source of genetic variation that drives genome evolution but our knowledge of the functions of human genes derived from transposases is limited. There are examples of transposon-derived, domesticated human genes that lost (SETMAR) or retained (THAP9) their transposase activity, however, several remnants in the human genome have not been thoroughly investigated yet. These include the five human *piggyBac*-derived sequences (PGBD1-5) which share ancestry with the *Trichoplusia ni* originated *piggyBac* (PB) transposase. Since PB is widely used in gene delivery applications, the potential activities of endogenous PGBDs are important to address. However, previous data is controversial, especially with the claimed transposition activity of PGBD5, it awaits further investigations. Here, we aimed to systematically analyze all five human PGBD proteins from several aspects, including phylogenetic conservation, potential transposase activity, expression pattern and their regulation in different stress conditions. Among PGBDs, PGBD5 is under the highest purifying selection, and exhibits the most cell type specific expression pattern. In a two-component vector system, none of the human PGBDs could mobilize either the insect PB transposon or the endogenous human PB-like MER75 and MER85 elements with intact terminal sequences. When cells were exposed to various stress conditions, including hypoxia, oxidative or UV stress, the expression profiles of all PGBDs showed different, often cell type specific responses; however, the pattern of PGBD5 in most cases had the opposite tendency than that of the other *piggyBac*-derived elements. Taken together, our results indicate that human PGBD elements did not retain their mobilizing activity, but their cell type specific, and cellular stress related expression profiles point toward distinct domesticated functions that require further characterization.

1. Introduction

Transposable elements (TEs) are discrete DNA segments that can move into new chromosomal locations either by a "cut and paste" or by a "copy and paste" mechanism. They fall into two classes according to whether their transposition intermediate is RNA (Class I or retrotransposons) or DNA (Class II or DNA transposons). TEs are ubiquitous in living organisms and several remnants of them occupy large portions of genomes (Kidwell, 2002). While they are generally considered as selfish sequences being able to cause harmful mutations, they are also important evolutionary factors (Cosby et al., 2019). Active LINE-1 retrotransposon copies continue to produce genetic diversity in human populations (Kazazian, 1999), and have a potential to contribute to individual somatic mosaicism (Coufal et al., 2009; Singer et al., 2010). One of the most direct contributions of TEs to host genome evolution is the emergence of new genes, also known as "molecular domestication" or "exaptation," where TE sequences acquire new functions. This process can be illustrated by the RAG1 recombinase (Kapitonov and Jurka,

Abbreviations: TE, Transposable Element; TIR, Terminal Inverted Repeat; TSD, Target Site Duplication; PB, *PiggyBac*; PGBD, *PiggyBac*-Derived; SB, *Sleeping Beauty*; MITE, Miniature inverted repeat transposable elements; NPC, neural progenitor cell.

* Corresponding author at: Institute of Enzymology, Research Centre for Natural Sciences, Eötvös Loránd Research Network, 1117 Budapest, Magyar Tudósok körútja 2, Hungary.
E-mail addresses: orban.tamas@ttk.hu, orbant@biomembrane.hu (T.I. Orbán).

https://doi.org/10.1016/j.gene.2022.146609
Received 15 October 2021; Received in revised form 17 May 2022; Accepted 18 May 2022
Available online 21 May 2022

2005) or the SET histone methyltransferase (Cordaux et al., 2006), showing domestication examples of a Transib or a Mariner transposon, respectively. Though domestication is usually associated with the loss of their mobilizing capacities, there are also examples of genes retaining their transposase activity, such as the human THAP9, which is able to transpose the distantly related *Drosophila P-element* (Majumdar et al., 2013),

PiggyBac (PB) DNA transposon elements, first characterized in the cabbage looper moth *Trichoplusia ni* (Fraser et al., 1983), have then been identified in a variety of eukaryotes from protozoa to primates (Lander et al., 2001; Pritham et al., 2005; Bouallegue et al., 2017). A PB element generally contains a 1.8 kb open reading frame (ORF) encoding a 68 kDa transposase; the ends of the transposon sequence consist of 13–15 nt terminal inverted repeats (TIRs), which are flanked by the duplication of the target site (TSD) sequence 'TTAA'. Among DNA transposons, the PB superfamily is peculiar because precise excision of the element restores the pre-integration site (Mitra et al., 2008), making transposition mechanism less genotoxic than in case of other families, where restoration of the excision site is an error prone process (Chen et al., 2020). There are active PB copies even in higher phylogenetic orders like mammals (Mitra et al., 2013), and domestication can also be observed for functions such as DNA binding and excision, or recombination of host DNA in ciliates (Baudry et al., 2009; Cheng et al., 2010) and in *Xenopus* (Hikosaka et al., 2007). The PB of the *T. ni* has been developed as an effective nonviral gene transfer tool (Ding et al., 2005; Yusa et al., 2011), with the potential for human gene therapy applications (Manuri et al., 2010). However, these applications raise concerns about the presence of ~2000 PB-like elements in the human genome (Mandal and Kazazian, 2008), especially for PGBD5 (*piggyBac* derived 5) which was suggested to have transposase activity for *piggyBac* TIR sequences (Henssen et al., 2015) and recently for distantly related *piggyBac*-like elements (Helou et al., 2021b), although these findings have later been challenged (Beckermann et al., 2021). Cross-reactivities of these distant PB-related elements enhance the possibility of genomic instability during transgenesis performed by the *T. ni* PB transposon system (Ivics, 2016).

The five PB derived elements in the human genome (PGBD1-5) are variously conserved among vertebrates; PGBD5 dates back to the cephalochordate/vertebrate split, PGBD1 and PGBD2 are found in mammals, whereas PGBD3 and PGBD4 are present predominantly in primates (Lander et al., 2001; Sarkar et al., 2003; Pavelitz et al., 2013; Bouallegue et al., 2017). Miniature inverted repeat transposable elements (MITEs) are internally deleted versions of invading transposon copies which could be spread effectively as nonautonomous variants throughout the genome (Hartl et al., 1992). They were transposed *in trans* by master copies of the complete autonomous transposon being able to express a transposase. MITEs can be illustrated by their consensus sequence: MER85 and MER75 have been identified belonging to PGBD3 and to PGBD4 respectively, because they TIR sequences have been preserved and is recognizably identical (Lander et al., 2001). In total, there are 1587 copies of them in the human genome, one-fifth of which have intact TIR ends and TSDs (Henssen et al., 2015). PGBD3, PGBD4 and their MITEs are the most recently amplified DNA transposon families in the human genome as they invaded the primate lineage. For the other PGBDs, the corresponding recognition TIR sequences can no longer be identified (Sarkar et al., 2003).

Genetic structures of PGBDs are very different. PGBD1, PGBD2 and PGBD3 are transcribed as extra exons within host genes. The transposase-derived sequences usually present as one single exon except for *PGBD5* which has six introns in vertebrates. PGBD1 members are the result of an ancestral fusion between exons containing SCAN domains (leucine-rich regions) in the N terminal regions and the last exon encoding for the transposase (Sarkar et al., 2003; Raskó et al., 2021); the human PGBD1 fusion product was also called as HUCEP-4 earlier. PGBD2 mRNA is transcribed with three exons, and its translation starts form the last one. PGBD3 is inserted into the fifth intron of the *Cockayne Syndrome group B* gene (CSB), which is a chromatin remodeling factor (Citterio et al., 2000; Newman et al., 2006) and also has a role in transcription coupled nucleotide excision repair (Fousteri and Mullenders, 2008; Gray and Weiner, 2010). Unlike other PGBDs, the PGBD3 transposon potentially functions as a natural exon trap and contains a splicing acceptor site in the 5′ region and a polyadenylation signal in the 3′ region. Thus, an alternative splicing of this region leads to a regular CSB product or to the CSB-PGBD3 fusion protein, but also to an extra transcript starting from the fifth exon and producing a single PGBD3 protein (Newman et al., 2008). Indeed, closest relatives of PGBD3 from *Acyrthosiphum pisum* and *Stegodyphus mimosarum* also share exon trap features (Bouallegue et al., 2017). Moreover, the authors suggest that these RNA processing signals were present in the common ancestor of PGBD1, PGBD2 and PGBD3, and were then lost along the branches leading to PGBD1 and PGBD2. This might be the reason for utilizing host promoter for the expression of all of these elements, which is also a characteristic of PGBD4 but in a different manner. PGBD4 has been inserted near the ORF of *Endoplasmic reticulum Membrane Complex 7* (EMC7), and is expressed via the bidirectional activity of the shared promoter (https://www.ncbi.nlm.nih.gov/gene/161779).

Recently, the potential activity of PGBD5, the most ancestral member of the human PGBD genes has become a very controversial issue. Originally, its transposase activity was considered to be very unlikely (Pavelitz et al., 2013; Saha et al., 2015), however, later papers claimed a transposase (Henssen et al., 2015; Helou et al., 2021b), or rather moderately, a genome "slicing" activity of PGBD5 (Henssen et al., 2016; Henssen et al., 2017). On the other hand, a recent systematic study provided evidence that PGBD5 cannot perform canonical DNA transposition (Beckermann et al., 2021). These results clearly awaited further systematic examinations of all human PGBD variants, and the importance to investigate their potential reactivity either with the MITEs of MER75/MER85, or with the insect derived PB vector system. In the current study, we have characterized all the five PGBDs present in the human genome from various aspects. We have analyzed phylogenies and conservation levels among the primates and investigated their expression profiles in different cell lines, together with some co-regulated transcripts produced from the same loci. Potential transposase activities have been tested with the two-component transposon system, providing evidence for the lack of *bona fide* transposition ability. In addition, we have examined the cross-reactivities between the insect PB transposase and MER75/MER85 elements of the human genome, but these were found to be non-functional. Lastly, to get closer to exploring the functions of *PGBD1-5*, we have tested the regulation of expression levels under different stress conditions. Our results support the hypothesis of human PGBDs being domesticated with a simultaneous loss of their transposase activity, and the use of the insect-derived PB system for potential gene therapy applications.

2. Material and methods

2.1. Phylogeny and conservation analysis by Ka/Ks

Human PGBD1-*5* protein sequences (NCBI Ref Seqs: NP_001171672.1, NP_733843.1, NP_736609.2, NP_689808.2, NP_001245240.1) were analyzed in Pfam 33.1 protein families database (https://pfam.xfam.org/) (El-Gebali et al., 2019) to search for the transposase domain DDE_Tnp_1_7 (PF13843, transposase IS4). The defined sequences and the narrow primate homologous flanking region (+0/1 amino acids (aa) at start and +4 aa at end) were used as basis (nucleotide sequences are shown in Suppl. Table 1). Amino acid sequences were determined by TranslatorX (Abascal et al., 2010), protein alignments were done by Clustal Omega program (Sievers et al., 2011). Phylogenetic tree was visualized by Archaeopteryx (Han and Zmasek, 2009). Nonsynonymous and synonymous substitution rates (denoted as *Ka* and *Ks*, respectively) were estimated using the software KaKs_Calculator (Zhang et al., 2006), with the method of model averaging. *Ka/Ks*

calculations are detailed in Suppl. Table 2. Significance levels (p-values) of differences from the *Ka/Ks* statistics were calculated by the two tailed *t*-test.

2.2. Cell lines, culturing and toxicity assays

Human embryonic kidney cells (HEK-293) and HeLa cells were maintained as described previously (Kolacsek et al., 2014); the HUES9 embryonic stem cell line was cultured as described earlier (Apáti et al., 2008). The establishment and maintenance of the MSCL-2 mesenchymal-like cell line was described in detail previously (Varga et al., 2011). Maintenance of 62F iPSC cells and differentiation into hippocampal neural progenitors (NPCs) were performed as described earlier (Vofely et al., 2018).

For toxicity assays, 6×10^4 of HeLa or 8×10^4 of HEK-293 cells were seeded onto 24-well plates and transfected with the FuGENE® 6 reagent, according to the manufacturer's instructions (Roche Applied Science). For the assays, 150 ng of pPB-CAG-GFP transposon donor plasmid (Kolacsek et al., 2014) and 150 ng of transposase (mPB or PGBD1-5) expression plasmid were co-transfected into the cells. For controls, mock transfections (transfection reagents without DNA) or using the pET41 control plasmid (Merck) instead of the transposase expression plasmid were applied in the assays. At 48 h posttransfection, cells were collected by trypsinization and after the addition of propidium iodide (PI, in the final concentration of 2 μg/ml), cell viability was measured by the percentage of PI-negative cells using the Attune NxT flow cytometer (Thermo Fisher Scientific). The use of a GFP-expressing donor plasmid allowed to determine not only the transfection efficiency, but also to determine and compare the percentage of viability/toxicity among the transfected and the non-transfected cells. All transfections were carried out in duplicates.

2.3. RT-qPCR for gene expression analyses

Total RNA was isolated from cells in near confluent 6-well plates using RNeasy Plus Mini kit (QIAGEN). Isolated RNA was digested with DNaseI (New England BioLabs) according to manufacturer instructions, and 1 μg of DNA-free total RNA was reverse transcribed in the volume of 10 μl using random oligonucleotides with the High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific). 5 μl ten-fold diluted cDNAs were used in qPCRs in 20 μl reaction volume, containing 50 nM of primers and Power SYBR® Green PCR Master Mix (Thermo Fisher Scientific). Reactions were performed in triplicates on StepOnePlus™ real-time PCR platform (Thermo Fisher Scientific) with the following thermal profile: 95 °C 10 min, 40 cycles of 95 °C 15 s and 60 °C 1 min, and melting curve profile was added. With the exception of PGBD4 all the qPCR primer pairs were designed for intron spanning with the use of Universal Probe Library Assay Design Center (Roche Applied Science) and listed in the Suppl. Table 3.

Expression levels in different tissue cell lines were determined with the Ct shift method described previously (Kolacsek et al., 2017). Briefly, we have constructed plasmid standards for each target as described, containing PCR amplicon sequences of the target and the Hsp90 endogenous control. Plasmid standards were used as reference samples containing equal copies of the target and the endogenous control. Plasmid standard and cDNA samples were measured in qPCRs using the ΔΔCt formula, determining the exact ratio of the target and the endogenous control. Different targets are comparable by this method as they are measured in the same endogenous control expression units plotted on the y-axis.

RNA variants from the CSB locus were detected by three amplicons (Fig. 1D, left panel). One amplifies the whole CSB-PGBD3 fusion mRNA splice variant expressed from the canonical promoter (tr1svPG3), another amplifies the whole CSB mRNA splice variant also expressed from the canonical promoter (tr1svCSB). The third amplicon detects the PGBD3 exon splice variant, but measures the sum of the fusion and the sole PGBD3 mRNA (tr1 + 2svPG3), the latter being expressed from a cryptic promoter located in exon 5 (tr2svPG3). The single PGBD3 can be calculated by subtracting the appropriate PGBD3 splice variant (tr1svPG3) from the sum of the transcripts (tr1 + 2svPG3). This mRNA gene product is denoted as sPGBD3 and produces a single PGBD3 protein.

Gene expression quantification after stress treatments (see Section 2.6.) were carried out by the standard ΔΔCt formula using the untreated control cells as reference sample and $^{-}$ΔΔCt was plotted. Quantification of p21 and VEGFA were used as positive controls for controlling the efficiency of treatments. In hypoxia and UV stress, analysis were normalized to Hsp90 and HMGB1 as multiple endogenous controls. For oxidative stress condition, HMGB1 and B2M endogenous controls were used for normalization. These endogenous controls were selected from a panel based on their stable expression levels (Hsp90, HMGB1, B2M, PolR2A, and RPLP0).

2.4. Plasmid constructs

All PCR primer pairs used for construction of plasmids are listed in Suppl. Table 3. Structures of the transposon donor plasmids are shown in Fig. 2A. Donor plasmid sequences are exactly the same except TIR sequences and some untranslated regions of the original transposon. Transposon donors contain a phosphoglycerate kinase (PGK) promoter regulated puromycin antibiotic resistance expression cassette. Candidates of MER75B and MER85 sequences with intact TIRs and TSDs were PCR amplified from gDNA with a short flanking region from 4 and 21 chromosomes, respectively. PCR products were inserted into pGEM®-T vector (Promega) and verified by Sanger sequencing. MER candidates were ligated into the excision backbone identical with pPB-puro donor plasmid. HindIII restriction site naturally present in the MER85 candidate sequence was used to insert the PGK-puromycin cassette. For this purpose, HindIII site was inserted into the middle region of MER75B by site directed PCR mutagenesis. As the CSB-PGBD3 protein was shown to bind the palindromic inner sequence previously (Gray et al., 2012), this palindromic sequence was reconstituted with PCR mutagenesis in the cloned MER85 candidate. Maps and sequences of the puromycin resistance cassette expressing plasmids are shown in Suppl. Fig. 1.

For the construction of transposase expressing helper plasmids, the coding sequence of PGBD1, PGBD2, PGBD3, CSB-PGBD3, PGBD4, and PGBD5 proteins were PCR amplified with Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher Scientific) from cDNA and inserted into pGEM®-T vector (Promega). Sequences were verified by Sanger sequencing. Coding sequences were cloned into CMV promoter driven pEGFP-C1 or pEGFP-N1 expression vectors from Clontech by replacing the GFP (pPGBD1, pPGBD2, pPGBD3, pCSB-PGBD3, pPGBD4, pPGBD5). In pPGBD3, the predicted catalytic amino acid triad DND (Keith et al., 2008) was reconstituted to DDD by point mutation of aa 352 from N to D (pPGBD3N/D) by site-directed PCR mutagenesis.

For PGBD5, apart from the amplified, full length protein encoding cDNA (identical to the reference sequence of Gene ID #79605), we also used two other protein species for the transposition experiments. At the beginning of our studies, expression databases for various Primate species listed a PGBD5 protein variant lacking or different in the region encoded by exon 5. Although some of these entries have recently been curated in NCBI we still used this exon 5 deletion protein named PGBD5del as a control. In the pPGBD5del helper plasmid, exon 5 was deleted from pPGBD5 by the FastCloning method (Li et al., 2011) with the amplification of the plasmid by primers with overlapping complementary 5′ ends. As a third PGBD5 protein, we requested the plasmid pRecLV103-GFP-PGBD5 from Addgene which was deposited from the work of Henssen et al. (2015). The coding sequence of this expression plasmid differs from that we have cloned (identical to the reference sequence of Gene ID #79605), as it contains an alternative exon 1 (most likely an earlier cloning artefact which has been permanently suppressed in the NCBI database, see https://www.ncbi.nlm.nih.gov/nucco

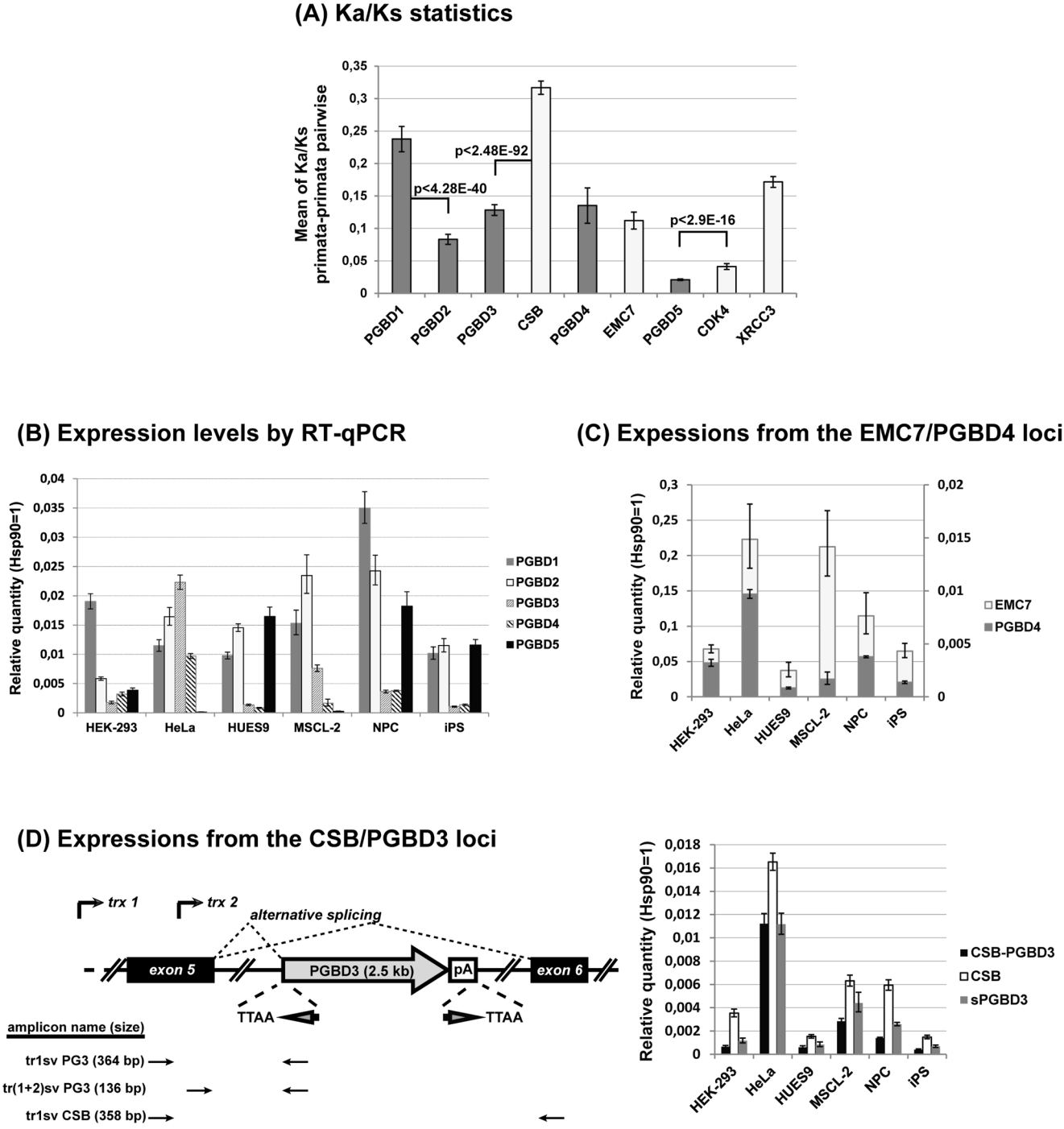


Fig. 1. Evolutionary analyses and the expression profiles of human PGBDs (A) *Ka/Ks* statistics of PGBDs based on comparative primate-primate pairwise analyses. Dark columns represent PGBD1-5, white ones are other endogenous genes; p values are indicated for certain selected cases (see text for explanation). **(B)** Gene expression analysis of PGBDs in different human cell lines by RT-qPCR. **(C)** Comparison of expression levels from the EMC7/PGBD4 loci. Left y-axis scale is for EMC7, right scale is for PGBD4. **(D)** Comparison of expression levels from the CSB/PGBD3 locus. (Left panel) Structure of the locus and primer pairs used for the analysis. (Right panel) Expression analysis of three different mRNA variants. Arrow depicted as "trx 1" indicates transcription from the canonical gene promoter, whereas "trx 2" shows transcription starting from a cryptic promoter located in exon 5; pA: polyA signal; "TTAA" sequences are the original target site duplications of the domesticated PGBD3 transposase; Sv stands for splice variant. Tr1svPG3 primer pair detects mRNA producing CSB-PGBD3; tr1svCSB pair detects mRNA producing CSB; tr(1 + 2)svPG3 detects all PGBD3 splice variants including the one from the cryptic transcript. The latter produces a solo PGBD3 protein (sPGBD3) encoding transcript, which can be calculated from the measurements of tr(1 + 2)svPG3 and tr1svPG3. For the quantification of PGBD3 in Fig. 1B, tr(1 + 2)svPG3 primer pair was used detecting both PGBD3 splice variants. In all graphs, mean values of at least 3 independent measurements are given, and error bars represent 95% confidence intervals.

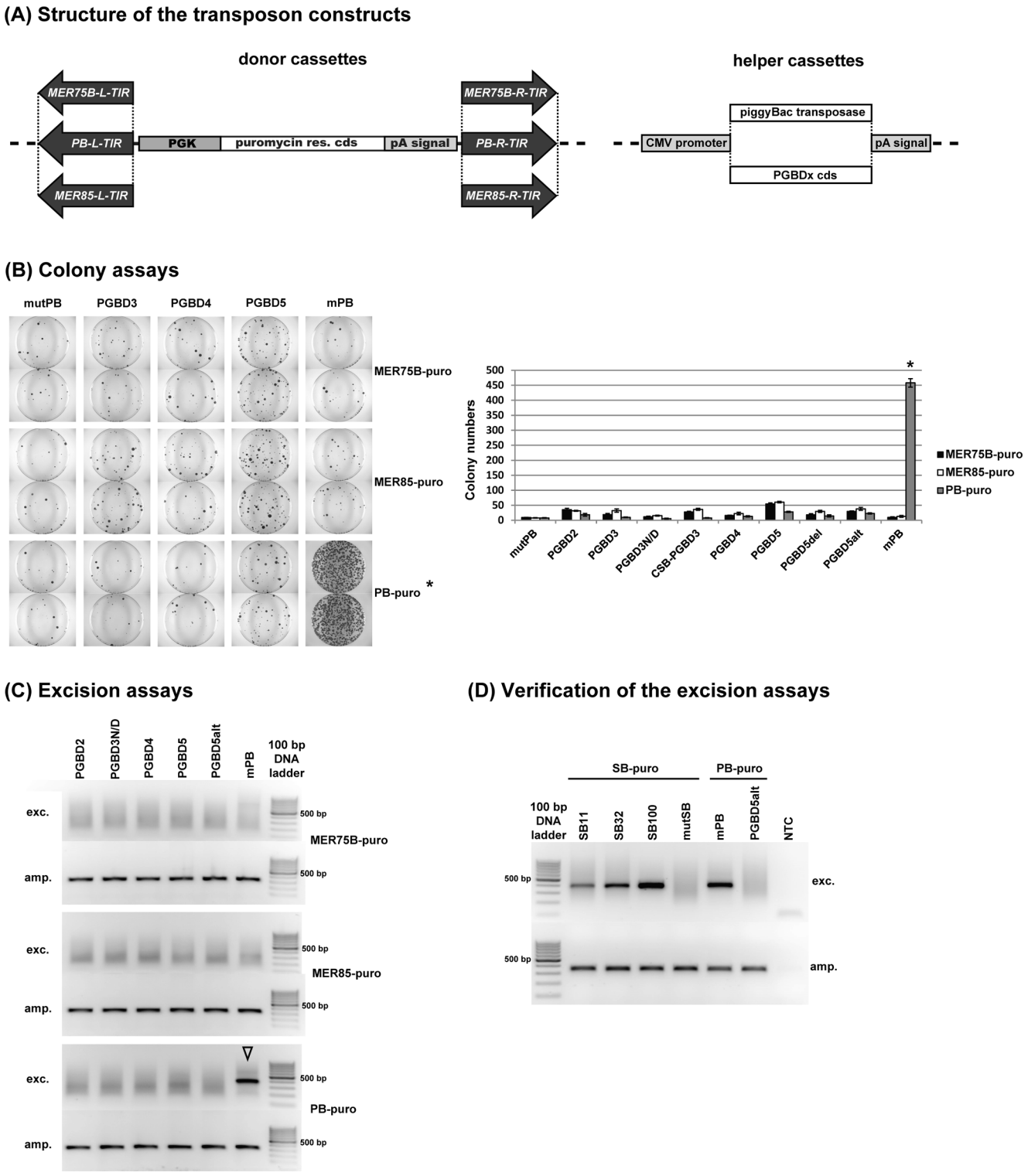


Fig. 2. Transposition assays of PGBDs in HEK-293 cell. (A) Three different transposon donors were tested: MER75B, MER85, and PB. They carry an antibiotic selection cassette, and their structures differ only in the specific transposon TIRs. The helper constructs express either the *piggyBac* transposase or the different PGBDs to be tested (x = 1–5). L-TIR: left-TIR; R-TIR: right-TIR; PGK: phosphoglycerate kinase promoter; CMV: cytomegalovirus promoter; cds: coding sequence; pA signal: polyadenylation signal. **(B)** Cells transfected with the two components (the selected donor and transposase expressing helper plasmids) were then selected by puromycin, and the surviving transgenic cells were stained and counted by the colony assay. (Left panel) Selected representative photos of the colony assays with the appropriate controls, showing two technical replicates for each assay. (Right panel) Quantifications of colony assays represented by the means of at least three independent experiments, error bars indicating standard deviations. ***:** The shown positive control mPB/PB-puro co-transfections were selected from fivefold dilutions. **(C)** Transfected cells were tested for excision events by PCR, where the product (381 bp, 'exc.') indicates the excised and repaired donor plasmid copies. Assay control was the ampicillin sequence, present on the donor constructs, with a 340 bp PCR product. **(D)** Sensitivity of the excision assay was tested by three SB variants with different activities, and compared to the PGBD5alt variant (a GFP-tagged construct used by Henssen et al., 2015) in parallel transfections. A representative result of several independent experiments is shown. NTC: non-template (water) control.

re/NM_024554) and N-terminally fused GFP; hence we named this variant as PGBD5alt (pPGBD5alt). The plasmid also contains a puromycin selection cassette but since our transposon donor system is based on puromycin selection, we partially deleted the puromycin cassette in this plasmid by SacII digestion and re-ligation. This was a necessary modification, otherwise random integration of the pPGBD5alt helper plasmid would increase colony numbers during the transposition assay.

The following transposase protein variants were used as positive and negative controls: for the PB system, the original *piggyBac* in a mammalian codon-optimized form, denoted as mPB (Cadiñanos and Bradley, 2007); for the *Sleeping Beauty* (SB) system, the SB11, SB32, SB100X and mutSB variants (Mates et al., 2009). Transposon donor for SB was identical to pPB-puro except TIR sequences (pSB-puro) (Kolacsek and Orban, 2018). To construct a negative control for the PB system, we generated the Asp268Gly mutant at the DDD catalytic site of mPB by PCR mutagenesis (mutPB).

2.5. Transposition assay and Western blotting

Transfections were carried out in duplicates. 2×10^5 of HEK-293 or HeLa cells were seeded onto 12-well plates, and transfected with FuGENE® 6 according to the manufacturer's instructions (Roche Applied Science). Next day, 200 ng transposon donor and 400 ng transposase expressing helper plasmid and 200 ng GFP expressing pEGFP-C1 transfection control plasmid from Clontech were transfected (the optimal ratios were chosen based on our previous study (Kolacsek et al., 2014)). In the case of pPGBD5alt helper plasmid that expresses GFP as an N-terminal fusion, the transfection control plasmid was replaced with pET-41 (Novagen). At 48 h post-transfection, cells were harvested and one tenth of them were analyzed by Attune NxT flow cytometer (Thermo Fisher Scientific) for GFP expressing cells to control transfection efficiency. Transfection efficiencies were between 53 and 59%. For antibiotic selection, 1% of the transfected cells were seeded onto cell culture Petri dishes, selected for two weeks in 1 μg/ml puromycin (Sigma-Aldrich), then surviving cells were fixed in ice cold methanol and stained with 0,04% Crystal Violet (Sigma-Aldrich) in 25% methanol. Colonies were quantified with the Universal Hood Gel Imager Model # 75S, using the Quantity One 4.4.0 software (BioRad).

After processing the transfected cells for colony assays and for flow cytometry measurements at 48 h post-transfection, the remaining cells were used for excision assay. Duplicated transfections were pooled and plasmids were isolated from the cells using a modified protocol of the QIAGEN plasmid Miniprep Kit, applying 250 μl of 1.2% SDS supplemented with 50 μg of Proteinase K for cell lysis step and protein removal from the DNA. 10 ng of isolated plasmid DNA was used in a two-round nested PCR assay where primers were specific to the plasmid backbone and resulted in a PCR product only from the excised and repaired plasmid copies due to large transposon cassettes. PCRs were carried out in 50 μl reaction volumes using 200 nM primers with 2xPCR master mix (Promega) and with the following thermal profile: 95 °C 5 min, 35 cycles of 95 °C 30 s 60 °C 30 s and 72 °C 1 min, 72 °C 5 min. 5 μl of 200-fold diluted first round product was used for the second round. Ampicillin sequence presenting in all donor construct was used as assay control, PCR conditions were the same as for the first round PCR of excision. Products were separated in a 2% agarose gel and visualized by ethidium bromide staining. PCR primers for excision analysis are listed in Suppl. Table 3.

Separate wells were transfected for Western blotting to detect transient PGBD protein expression. Cells were lysed at 48 h post-transfection by the addition of TE sample buffer (0.1 M TRIS-PO_4, 4% SDS, 40 mM Na-EDTA, 40% glycerol, 0.04% bromophenol blue, and 4% β-mercaptoethanol; materials from Sigma-Aldrich). Samples were sonicated to shear DNA and equal amounts were loaded on a 7.5% acrylamide gel (SDS-PAGE) (Sigma-Aldrich). Proteins were transferred to a 0.2 μm PVDF membrane (BioRad). Blots were probed with the following primary antibodies: anti-PGBD1 (Abgent, cat. AP17268C), anti-PGBD2 (Biorbyt, cat. orb519701), anti-PGBD3 (Sigma-Aldrich, cat. HPA025825), anti-PGBD4 (Sigma-Aldrich, cat. HPA040896), anti-PGBD5 (Biorbyt, cat. orb13159), and anti-β-actin (Sigma, cat. A3854) HRP conjugate. Donkey anti-rabbit IgG (H + L) HRP conjugate secondary antibody (Jackson ImmunoResearch, cat. 711–035-152) were used to visualize the results. Detection was performed with Pierce™ ECL Western Blotting Substrate (BioRad, cat. 32106).

2.6. Stress conditions

1.5×10^5 HEK-293 and 1×10^5 HeLa cells were seeded onto 24-well plates. Treatments were carried out in 6 parallels. Next day medium was replaced with 50 μM deferasirox (DFX) (MedChem Express) or 4 mM $KBrO_3$ (Sigma-Aldrich) for inducing hypoxic or oxidative stress, respectively. Concerning $KBrO_3$, we chose the higher possible concentration without severe cell loss (Luan et al., 2007). After 24 h cells were harvested, and parallels were pooled for total RNA purification, reverse transcription and RT-qPCR (see 2.3). For UV light exposure, the medium was removed and cells were rinsed with 40 μl PBS. UV irradiation was performed using the Stratalinker UV Crosslinker Model 1800, delivering 20 J/m^2 dose of UV light. Cells were recovered by adding back the culture medium, and samples were harvested after 6 h or 18 h posttreatment.

3. Results

3.1. Evolution of human piggyBac-derived sequences

Active TEs spread throughout the genome but protection of genome stability acts as a selective pressure and mutations can eventually inactivate the TE copies. However, several examples reveal that by gaining new functions, domesticated copies can also be preserved in the evolution by purifying selection. While classical TEs are present in multi-copies within and between species, domesticated elements are generally found in single orthologous copies. The PB DNA transposon family is unique in that domestication for useful function exemplifies the retention of some transposase-like aspects such as the targeted removal of endogenous DNA fragments by the cut and paste mechanism (Hikosaka et al., 2007; Baudry et al., 2009; Cheng et al., 2010).

The first part of our study focuses on the evolution of the human PGBD elements. As described earlier (Sarkar et al., 2003), the N-terminal region (positions 1–130, using *T. ni* transposase as reference) is not well conserved, therefore as opposed to other phylogenetic analyses, we have restricted our conservation analysis to primate sequences and to the IS4/5 transposase-like domain (Suppl. Table 1). The phylogenetic tree inferred from 24 species and 111 PGBD sequences (Suppl. Fig. 2) revealed that PGBD1 and PGBD2 are more closely related, and they are found mostly in mammals (Sarkar et al., 2003; Bouallegue et al., 2017; Raskó et al., 2021). PGBD3 is clustered closest to them (Suppl. Fig. 2), invaded into the primate lineage, but relatives of PGBD3 can also be found in the aphid *Acyrthosiphum pisum* and in the spider *Stegodyphus mimosarum,* although they do not appear to be domesticated copies (Bouallegue et al., 2017). PGBD4 is also widespread among primates but clustered separately from PGBD3, PGBD1 and PGBD2 (Suppl. Fig. 2). The closest relatives of PGBD4 are in bat flying fox *Pteropus vampyrus*, in aphid *Acyrthosiphon pisum*, and in moth *Spodoptera frugiperda* (Bouallegue et al., 2017). Inconsistencies between the transposase tree and the species phylogeny suggest that the PB family might be frequently and successfully horizontally transferred. In addition, PGBD4 is more related to domesticated elements in lower phyla: PGM and TPB2 found in ciliates, and KOBUTA in *Xenopus* (Bouallegue et al., 2017). PGBD5 is the most ubiquitous member present in a large spectrum of species and is clustered quite distinctly from the other PGBDs (Suppl. Fig. 2). As previously described (Pavelitz et al., 2013), this reflects an early domestication event of PGBD5 in the cephalochordate lineage. Moreover, no sequence similarity of the PGBD5 flanking regions can be found in

hemichordates, echinoderms, and urochordates, supporting the conservation of the entire region, including the flanking sequences (Bouallegue et al., 2017).

During evolution if a new function has been acquired, low ratio (*Ka/Ks* < 1) of nonsynonymous (*Ka*) to synonymous (*Ks*) nucleotide substitution rates suggests that the sequence is under strong purifying selection, while a high ratio (*Ka/Ks* > 1) indicates that adaptive peak is not yet reached (Hurst, 2002). Earlier calculations of the *Ka/Ks* ratio of PGBDs provide arguments in favor of domestication of these sequences (Newman et al., 2008; Bouallegue et al., 2017). Here we analyzed the level of conservation of the primate PGBD elements, comparing them to some endogenous genes, including those expressed from the same loci (sequences are in Suppl. Table 1, *Ka/Ks* calculations are in Suppl. Table 2). Comparison of *Ka/Ks* statistics revealed that PGBD5 is the most conserved among PGBDs, it is subjected to a higher selective pressure than the conserved cell cycle regulator gene CDK4 (Fig. 1A). Values of other PGBDs were found to be comparable to endogenous genes like the DNA repair factor XRCC3, falling into the category of protein coding genes under strong purifying selection (Hurst, 2002). PGBD4 and EMC7, regulated by the same promoter, are similarly conserved, however, the sequence of PGBD3 is under much higher selective pressure than the coding region of the *CSB* "host" gene. The latter one has been described to have DNA repair activity (Sarker et al., 2005), yet it seems less conserved than the other analyzed endogenous repair factor, XRCC3 (Fig. 1A). Another notable result is that there is a three-fold difference between the conservation of the separately clustered and closely related PGBD1 and PGBD2 (Fig. 1A), suggesting that PGBD2 is also a strongly conserved, domesticated gene with a currently unknown function. Our analysis may differ from the previous ones (Sarkar et al., 2003; Bouallegue et al., 2017; Helou et al., 2021b), because it is restricted to primate lineage and is limited strictly to the DDD transposase-derived domain omitting the N-terminal regions.

3.2. Cell type specific expression profiles of human PGBDs

When obtaining information on the expression profiles of PGBDs in online databases, such as BioGPS (Wu et al., 2013) or the Human Protein Atlas (https://www.proteinatlas.org/), PGBD5 shows strong neural-specific expression, which was confirmed by Pavelitz et al. (2013) by *in situ* hybridization. Concerning the other PGBD genes, however, they show more diverse expression patterns. We have developed a qPCR-based quantification method (the Ct shift method) to accurately compare the levels of different nucleic acids (Kolacsek et al., 2017) (see also section 2.3.). Using this method, we measured expression levels of all PGBDs in six cell lines: HEK-293, HeLa, a human embryonic stem cell line (HUES9), a mesenchymal stem cell-like cell line (MSCL-2, differentiated *in vitro* from HUES9 (Varga et al., 2011)), an induced pluripotent stem cell line (62F), and hippocampal neural progenitor cells (differentiated *in vitro* from the 62F iPSC line (Vofely et al., 2018)). Comparing to the Hsp90 as a reference gene, the PGBD expression levels were consistently found to be two orders of magnitude lower (Fig. 1B). This magnitude may indicate a general expression level that is well tolerated by the cell in regard of this type of external sequences. While PGBD1 and PGBD2 are similarly expressed in all cell lines examined, PGBD3, PGBD4 and PGBD5 show more cell line specific differences (Fig. 1B). Interestingly, the expression profiles of PGBDs in NPCs were similar to those in HUES9 and iPS cell lines: PGBD3 and PGBD4 are expressed at a relatively low level, whereas PGBD5 is similarly abundant as PGBD1 and PGBD2. Our result supports the Primary Cell Atlas dataset of the BioGPS database in relation of PGBD5 (Mabbott et al., 2013), where in addition to neural expression, embryonic stem cells and some iPSs also show elevated expression levels. Our data suggests that PGBD5 might function in pluripotent cells, and also in differentiated cells of the neural lineage.

We compared mRNA levels expressed from shared loci in cases of PGBD4 and PGBD3 (Fig. 1C and D). We have measured EMC7 mRNA as one order of magnitude higher than PGBD4 (Fig. 1C), which confirmed that the promoter shared with PGBD4 induces transcription originally in the direction of EMC7. Expression differences between EMC7 and PGBD4 were ranging from ~ 20- to ~ 50 fold, with the exception of cell line MSCL-2 where difference was ~ 130 fold. These results indicate the regulated promoter directionality but post-transcriptional regulation (e.g. mRNA stability) may also contribute to this high difference between EMC7 and PGBD4 mRNA steady state levels.

Human PGBD3 is of special interest, because its host gene, CSB is mutated in the Cockayne Syndrome (CS) that is a rare genetic disease characterized by neurological problems, growth failure and premature ageing. The study of Newman et al. (2008) showed that the mutation spectrum of CSB gene in CS in general prevents the production of the full length CSB but allows to express the fusion CSB-PGBD3 gene product. However, the complete absence of the CSB locus products does not cause severe progeria of CS, only mild UV sensitivity (Horibata et al., 2004). Thus, the exclusive presence of CSB-PGBD3 may play a causative role in CS (Weiner and Gray, 2013). Nevertheless, it is possible that CSB-PGBD3 fusion protein is important in both health and disease, because it may confer a metabolic advantage presumably in DNA damaging stress, but exerts this effect only if CSB-PGBD3 is co-expressed with CSB (Bailey et al., 2012). The complex nature of the locus results in a complex pattern of mRNA and protein isoforms, with the solo PGBD3 expressed from a cryptic promoter located in exon 5 (Fig. 1D, left panel) (Newman et al., 2008; Kolacsek et al., 2017). Examining all the gene products separately by the Ct shift method, we could reveal that both PGBD3 splice variants are expressed in a significant amount as compared to the canonical CSB mRNA (Fig. 1D, right panel). These PGBD3 splice variants are present in lower amounts than the CSB transcript in all cell types, but the extent is different, suggesting tissue specific differences in splicing regulation. Post-transcriptional regulation cannot be ruled out, but correlation between the fusion and solo PGBD3 mRNA product rather indicates the regulation of splicing (Fig. 1D, right panel). Overall, when compared to the CSB transcript, the dose of the PGBD3 containing mRNAs is the highest in HeLa and MSCL-2 cell lines which may confer the metabolic advantage described earlier for the co-presence of CSB and CSB-PGBD3 (Bailey et al., 2012).

Taken together, our data indicate that all human PGBDs are regulated distinctly from each other and in a cell type specific manner. These results are in favor of the hypothesis that these transposase-originated genes represent individual domestication events. Nevertheless, it is a crucial question whether they retained their transposase activity, therefore we continued to systematically investigate this issue for all PGBDs.

3.3. Lack of transposase activity of all PGBDs

Concerning the human of the PB superfamily, the predicted DDD catalytic aspartate amino acids at positions of 268, 346, and 447 of the *T. ni* PB transposase are not conserved among the domesticated PGBD sequences (Sarkar et al., 2003; Newman et al., 2008; Pavelitz et al., 2013). However, mutation analysis revealed a fourth aspartate amino acid (D450) that is also essential for transposition of *T. ni* PB, but it has a tolerance for a glutamate substitution (Keith et al., 2008). Intriguingly, we have found the fourth D to be conserved among all the primate PGBD orthologs and paralogs (data not shown), so the position is likely to play an essential role in the function of these PB family members. PGBD3 and PGBD4 are the youngest domesticated elements, and presumably have acquired only a few inactivating mutations. Nevertheless, only PGBD4 retained all the essential catalytic amino acids DDDD, so it cannot be excluded to have transposase activity (Mitra et al., 2008). PGBD3 has gained a D to N point mutation at position 352, so it has a DNDD motif (Sarkar et al., 2003; Newman et al., 2008). On the other hand, Gray et al. observed that CSB-PGBD3 is capable of binding to the TIR of PGBD3 transposon and of its residual MER85 elements *in vitro* and *in vivo*, providing a hint that MER85 could have been the integral part of an

ancient PGBD3-like functional transposon (Gray et al., 2012). This prompted us to test the presumably active, "reconstituted" PGBD3 mutant having the expected DDDD motif. Apart from testing the reactivity with the insect-derived PB system, we also tested the MER75 and the MER85 elements of recognizable PB origin present in the human genome, the potential substrates of ancient PGBD4 and PGBD3 transposases, respectively. We have cloned representative copies of these MER sequences from the human genome, and constructed transposon donor plasmids by inserting a selection marker of puromycin resistance into the internal region (Fig. 2A). The two component transfection system was used in HEK-293 cells, where helper plasmids expressed the appropriate human PGBD variant proteins. After two weeks of antibiotic selection, transgenic efficiencies were tested by colony counting. As a result, neither PGBD3 nor PGBD4 showed any indication of transposition when compared to the background control (the insect PB catalytic mutant, Fig. 2B).

To continue with the systematic evaluation of all *piggyBac*-related transposition activity in the human genome, we tested all human PGBD proteins for their potential interaction with either the insect PB vector system or the two MER elements. This was also urged by a somewhat stunning previous study showing that PGBD5 has significant classical transposase activity for the TIR of *T. ni* PB transposon, in spite of this ortholog being the oldest among human PGBDs, clearly lacking recognizable flanking TIR sequences. The authors also claimed that it is not the components of the originally predicted 'DDD' motifs but rather amino acids in different positions were the ones responsible for this unexpected activity (Henssen et al., 2015). However, due to the ambiguity of the encoded protein prediction this group used, we tested not only their GFP-tagged version (PGBD5alt), but also the currently available updated sequence variant (PGBD5, Gene ID: 79605), as well as an alternative variant lacking exon 5 (PGBD5del) which also appeared earlier in expression databases (see section 2.4.). Successful transient exogenous expressions of all human PGBD proteins were confirmed by Western blot (Suppl. Fig. 3). As a result of the colony assays, none of the PGBDs showed reactivity with either the insect PB vector system or the MER elements: the colony counts produced by all PGBDs and their variants were far below the mPB activity (Fig. 2B). On the other hand, we detected somewhat elevated colony numbers in case of PGBD1 (see details in our concurrent publication, (Raskó et al., 2021)), and to a lower extent, for PGBD5. This prompted us to test whether the first step of transposition occurs in case of any combinations, therefore we performed excision assays detecting the capability of the human PGBDs to cut out the transposon cassettes. However, with the exception of the positive control (the insect mPB), no excision events were detectable by a two round nested PCR (Fig. 2C), indicating the lack of transposition mediated by the human PGBDs or their variants. To test the sensitivity of our excision PCR, we have made parallel transfections by the *Sleeping Beauty* transposon variants with different activity (Kolacsek et al., 2014) and the PGBD5alt used by Henssen et al. (2015), for which authors detected transposase activity. Based on this, it can be concluded that even the low excision activity of the SB11 variant is still well-detectable by this PCR assay but we were unable to reproduce the result of Henssen et al. (2015) for their used PGBD5alt variant (Fig. 2D).

Another potential explanation for not detecting transposition activity would be the toxicity caused by the expression of PGBDs: in that case, the death of the transfected cells would prevent transposition to occur. To exclude that, we performed a cell viability assay comparing the overexpression of all PGBDs to the effect of mock transfection and to the expression of other unrelated expression vectors in both HeLa and HEK-293 cells. As a result, no elevated toxicity could be detected for overexpressing any of the human PGBD proteins as compared to the controls or to the insect mPB transposase (Suppl. Fig. 4). Taken together, all these results indicate the lack of *bona fide* transposition activity for all human PGBDs, including all the tested PGBD5 variants.

3.4. Expression changes of PGBDs in stress conditions

It is likely that *PGBD* transposases have been conserved due to their interactions with various cellular factors, thus we can assume that they also retained their expression pattern regulation. Several transposases interact with DNA repair factors to increase their efficiency, or to modulate the cell cycle, often depending on the environmental conditions (Zayed et al., 2003; Izsvak et al., 2004; Walisko et al., 2006). We hypothesized that PGBDs may still have these cellular connections, therefore we investigated the main cell cycle regulating circumstances, the DNA damaging stress conditions, whether they can influence the expression patterns of the domesticated primate PGBD proteins. What is common in hypoxia, oxidative stress and UV stress is that they cause DNA damage, and by activating p53 responsive genes like *p21*, they regulate cell cycle progression (Bunz et al., 1998; Pan et al., 2004; Pires et al., 2010). Expression studies were carried out in two cell lines, HEK-293 and HeLa, and hypoxic conditions resulted in similar induction of all domesticated PGBD elements in both cell lines (Fig. 3A, left panel). Expressions of mRNAs from the shared loci of *PGBD3* and *PGBD4* were concordantly induced, but to different extents (Fig. 3A, right panel). Induction of CSB was two times higher than PGBD3, which again confirms strong regulation difference between the two splicing products expressed from the same promoter. Induction of EMC7 also differed from PGBD4 in HEK-293 cells, but not in HeLa cells, suggesting that posttranscriptional regulation may also contribute to their expression profiles.

In oxidative stress conditions, PGBD3 and PGBD4 expressions responded concordantly in the two cell lines, both being reduced (Fig. 3B, left panel), but the effect on PGBD3 in HeLa cells was more pronounced, producing an eight-fold reduction ($-\Delta\Delta Ct \sim -3$). Measuring the effect on CSB revealed that the entire locus seems to be downregulated in HeLa (Fig. 3B, right panel), but not in HEK-293 cell line where only the level of the PGBD3 splice variant is reduced. Concerning the other locus with a shared promoter, the expression level of EMC7 in oxidative stress conditions was not significantly changed in either cell line, in contrast to PGBD4 (Fig. 3B, right panel). However, in UV stress conditions, only PGBD5 showed a significant response of 4–5-fold downregulation (Fig. 3C), that was an immediate effect after 6 h of treatment (data not shown). The distinct expression patterns indicate that these domesticated genes may be involved in completely different cellular processes.

4. Discussion

There are several *piggyBac*-related elements present in the human genome and both the transposase-derived protein coding genes and the repetitive TIR sequences raised concerns about the safe applications of the widely used insect *piggyBac*-based genetic tools. An intense debate was therefore initiated when the PGBD5 protein was recently claimed to have transposase activity (Henssen et al., 2015; Ivics, 2016). Based on this unexpected result, we have decided to systematically investigate this issue by examining whether these *piggyBac*-related elements still have transposase activity or to the contrary, whether we can provide indications for them being domesticated endogenous genes.

Focusing on the PGBD copies present in primates, our phylogenetic analysis supports the hypothesis about the frequent horizontal transfer of PB elements also in this lineage (Pagan et al., 2010), with PGBD5 being the most ancient homolog without obvious flanking TIR sequences, pointing towards a domesticated function, rather than towards a conserved transposase activity. Our results are in line with the study of Pavelitz *et al.*, concluding that PGBD5 protein is a neuron specific endogenous element, and also claiming that it is unlikely that it can bind DNA in brain nuclei, which would be necessary for transposition (Pavelitz et al., 2013). Analyzing the gene structures of the other homologs, the lack of TIR sequences and the examples of fusion transcripts with endogenous protein coding genes are also in favor of the

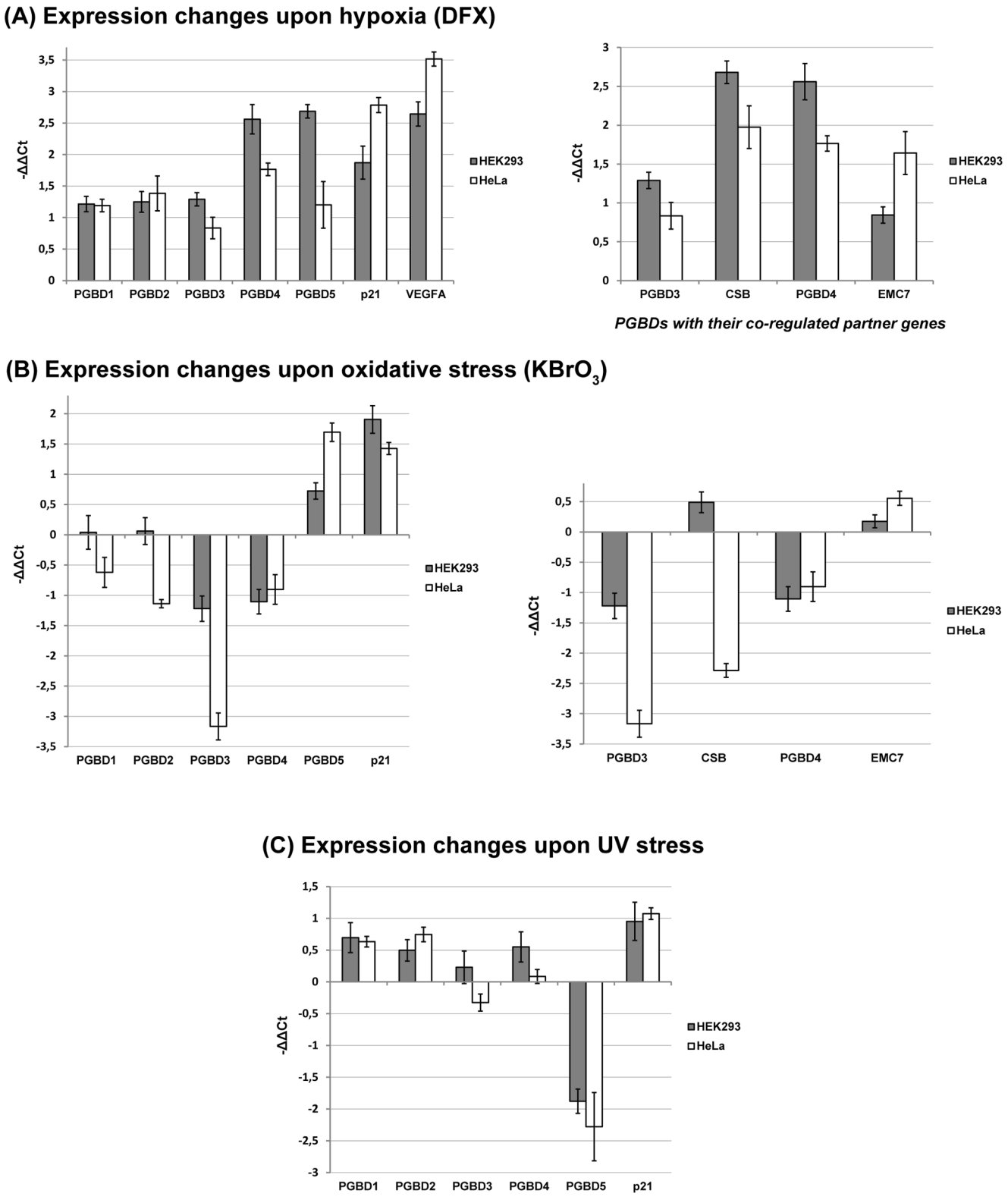


Fig. 3. Expression changes of PGBDs upon different stress conditions. Stress treatments were carried out in HEK-293 and HeLa cells, mRNA levels were measured by RT-qPCR. **(A)** Cells were treated with deferasirox (DFX) to induce hypoxic response. (Left panel) PGBD1-5 expression changes were measured, p21 and Vascular Endothelial Growth Factor A (VEGFA) were used as stress condition controls. (Right panel) Expression changes shown for the shared loci of CSB/PGBD3 and EMC7/PGBD4. **(B)** Cells were treated with KBrO3 to induce oxidative stress. (Left panel) PGBD1-5 expression changes were measured, p21 was used as stress condition control. (Right panel) Expression changes shown for the shared loci of CSB/PGBD3 and EMC7/PGBD4. **(C)** Cells were radiated with UV light, and PGBD1-5 expression changes were measured 18 h after treatment; p21 was used as stress condition control. For the analysis of PGBD3, the tr(1 + 2)svPG3 primer pair was used to detect both PGBD3 splice variants (see Fig. 1D, left panel). Mean values of at least three independent measurements are given, error bars represent 95% confidence intervals.

domestication hypothesis. Only PGBD3 seems to have preserved the structure of an active ancient transposon with MER85 TIR sequences present on both flanking regions; however, the mutation present in the DDD motif indicates the inability to perform the "canonical" cut-and-paste transposition reaction. When examining the expression profiles of human PGBDs in several types of cell lines, the selective and carefully regulated expression patterns also argue for specific endogenous functions. Our conclusion was that *PGBDs* have most likely been domesticated in primates but their functions are unlikely to be associated with canonical transposase activity.

To directly examine the ability of human PGBDs to perform transposition, we tested each transposase-like proteins for mobilization of the insect transposon sequence, as well as for the MER85 and the MER75 sequences. The most striking result was that we could show the lack of transposase activity for all five human PGBD genes. Even the best candidate, PGBD3 and CSB-PGBD3 fusion product gave negative results: although both were shown to bind to MER85 sequences (Bailey et al., 2012; Gray et al., 2012), our "restored" catalytic domain variant (PGBD3N/D) was not able to perform transposition either. This result shows that even the "youngest" PGBD homolog has lost the canonical DNA transposase activity. In addition and of most importance, we could not detect transposase activity for any of the tested PGBD5 variants, even in various cell lines (e.g. in HeLa, data not shown). Although in a later publication, Henssen et al. presents a more careful interpretation of PGBD5 having "only" some genome-cutting ability, possibly causing oncogenic mutations (Henssen et al., 2017), our results are in contrast with two recent papers claiming a lower but detectable transposase activity of PGBD5 (Helou et al., 2021a; Helou et al., 2021b). On the other hand, a more recent study is in agreement with our results, presenting the lack of cross-reactivity among the insect *piggyBac*, the bat *piggyBat* transposase and human PGBD5 (Beckermann et al., 2021). They found no evidence of PGBD5 transposase activity, moreover, they showed that the two other transposases have restricted activity toward their own cognate sequences, showing no cross-mobilization of each other's transposon elements (Beckermann et al., 2021). Similarly to our approach, this group also failed to reproduce the results of Henssen et al., either using the original GFP-tagged PGBD5 or the corrected PGBD5 transposase sequence variant. Our and their results are all in favor of PGBD5 being a domesticated endogenous gene, having lost the ancient transposase activity.

In an earlier study, Saha *et al.* have found a few solo 14–16 bp sequences in the human genome that are identical with the end of the TIR of *T. ni* PB transposon sequence (Saha et al., 2015). Authors examined whether the insect PB transposase used in gene delivery is able to mobilize elements with these genomic sequences. Their result showed that the PB vector system is safe in this aspect. Here we examined the MER75 and the MER85 elements in a similar manner and our MER donor constructs do not show reactivity with the mPB transposase. Our results confirm the safety of the application of PB system for gene transfer in all aspects: neither endogenous PGBDs can remobilize PB vectors integrated in the genome, nor the PB transposase can react with endogenous residual PB-like sequences of the human genome. Both scenarios would pose a risk of genetic instability, however, increasing number of observations suggests that it should not be feared of.

In spite of their inability to perform a cut-and-paste reaction, we still detected somewhat elevated colony numbers resulting from the co-transfection of PGBDs and donors, as compared to the background values obtained with the inactive variant, mutPB (Fig. 2B, right panel). We hypothesize that high exogenous PGBD expression may aid the random integration of donor sequences, especially given the number of MER sequences in the human genome. This view is supported by the fact that PGBDs with the most elevated background (PGBD1 and PGBD5) lack the C-terminal cysteine rich domain (CRD) responsible for sequence specific DNA binding of the *T. ni* PB transposase (Keith et al., 2008; Morellet et al., 2018; Chen et al., 2020), making it unlikely that these would mobilize any transposon in a sequence-specific manner (concerning PGBD1, see our concurrent paper (Raskó et al., 2021)). The two recently published works on PGBD5 showed similar results (Helou et al., 2021a; Helou et al., 2021b): cotransfecting the human PGBD5 and PB transposon substrates, they obtained a few fold elevation in the background level of colony numbers; however, the authors evaluate these results as transposase activity of the domesticated PGBD5. On the other hand, these may be explained alternatively: as a general behavior of transposases, sequence specific or nonspecific DNA binding, together with oligomerization may facilitate the formation of transposase-like synaptic complexes, resulting in enhanced repair of randomly damaged regions by integrating unrelated DNA sequences, including transfected plasmid templates.

In the last part of our study, we aimed to investigate the potential cellular functions of the human PGBDs. In general, transposon activity and expression are sensitive to and regulated by stress conditions, therefore we tested if this response have been retained for the domesticated PGBDs. There were some indications for that: in the case of the co-regulated CSB and PGBD3, the former one plays a central role in the cellular response to stress as a chromatin remodeling factor through modulating transcriptional changes, e.g. response to hypoxia, DNA damage, and also response to insulin-like growth factor-1 (Filippi et al., 2008; Velez-Cruz and Egly, 2013). Also, in the promoter of the CSB gene there are two hypoxia response elements where the HIF-1α transcription factor can bind (Filippi et al., 2008). In addition, the elevated random integration of donor constructs with co-transfection of certain PGBDs raises the possibility that the transposase originated proteins could interact with DNA repair factors during processing of randomly damaged plasmid DNA. We tested stress conditions causing potential DNA damage and in response to hypoxia, we detected elevated expression levels for all PGBDs, although to a different extent. On the other hand, no common expression changes were found under oxidative stress and UV stress conditions: as opposed to the other homologs, PGBD3 and PGBD4 responded with a decrease to oxidative stress in all cell types examined, while PGBD5 was upregulated in oxidative stress but downregulated in UV stress conditions. These opposite behaviors indicate that these domesticated genes may be involved in different cellular processes, however, the distinct and often cell type specific expression responses support the hypothesis that all PGBDs are domesticated genes with precise endogenous regulations.

In conclusion, our systematic investigations revealed that none of the human PGBD elements retained their original ability for transposition. On the other hand, several pieces of evidence support the hypothesis that these elements have been "co-opted" during evolution and serve distinct domesticated functions. The low Ka/Ks values for many of them indicate that they are under a strong purifying selection, and in some cases, they seem more conserved that the co-regulated transcripts of the same locus (e.g., PGBD3 vs CSB). In addition, for such co-regulated transcripts for PGBD3 and PGBD4, the splicing and the transcription pattern is markedly different and actively regulated in a cell type specific manner, indicating endogenous functions. Finally, the expression profile changes of all human PGBDs in various stress conditions also point to the fact that the roles of these genes can likely be placed in different cellular pathways. Taken together, although the exact and distinct domesticated functions of human PGBDs still require further investigations, the lack of their transposition ability provides a necessary safety background for using the *piggyBac* gene delivery system for potential gene therapy applications in the future.

CRediT authorship contribution statement

Orsolya Kolacsek: Investigation, Methodology, Data curation, Validation, Writing – original draft, Writing – review & editing. **Gerda Wachtl:** Investigation, Methodology. **Ábel Fóthi:** Data curation, Visualization. **Anita Schamberger:** Investigation. **Sára Sándor:** Investigation. **Enikő Pergel:** Investigation. **Nóra Varga:** Investigation. **Tamás Raskó:** Validation, Formal analysis, Writing – review & editing.

Zsuzsanna Izsvák: Validation, Formal analysis, Writing – review & editing. **Ágota Apáti:** Investigation, Validation, Funding acquisition. **Tamás I. Orbán:** Conceptualization, Supervision, Data curation, Validation, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are indebted to Kornélia Némethy and Beáta Haraszti for their excellent technical help. The human embryonic stem cell line (HUES9) and iPSC line (62F) were kindly provided by Douglas Melton (HHMI) and Fred H. Gage (Salk Institute), respectively. This study was supported by grants from the National Research, Development and Innovation Fund of Hungary: VEKOP-2.1.1-15-2016-00156, VEKOP-2.3.3-15-2017-00014, 2018-1.2.1-NKP-2018-00005 grants to T.I.O., NKFIH-OTKA grants K112112 to T.I.O., PD121287 and FK124661 to A. S., and by the National Brain Research Program (NAP) of Hungary (grant number: 2017-1.2.1-NKP-2017-00002, to Á.A.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2022.146609.

References

Abascal, F., Zardoya, R., Telford, M.J., 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucl. Acids Res. 38, W7–W13.

Apáti, Á., Orbán, T.I., Varga, N., Németh, A., Schamberger, A., Krizsik, V., Erdélyi-Belle, B., Homolya, L., Várady, G., Padányi, R., Karászi, É., Kemna, E.W.M., Német, K., Sarkadi, B., 2008. High level functional expression of the ABCG2 multidrug transporter in undifferentiated human embryonic stem cells. Biochim. Biophys. Acta 1778 (12), 2700–2709.

Bailey, A.D., Gray, L.T., Pavelitz, T., Newman, J.C., Horibata, K., Tanaka, K., Weiner, A. M., 2012. The conserved Cockayne syndrome B-piggyBac fusion protein (CSB-PGBD3) affects DNA repair and induces both interferon-like and innate antiviral responses in CSB-null cells. DNA Repair (Amst) 11 (5), 488–501.

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., Bétermier, M., 2009. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev. 23 (21), 2478–2483.

Beckermann, T.M., Luo, W., Wilson, C.M., Veach, R.A., Wilson, M.H., 2021. Cognate restriction of transposition by piggyBac-like proteins. Nucl. Acids Res. 49, 8135–8144.

Bouallegue, M., Rouault, J.D., Hua-Van, A., Makni, M., Capy, P., 2017. Molecular Evolution of piggyBac Superfamily: From Selfishness to Domestication. Genome Biol. Evol. 9, 323–339.

Bunz, F., Dutriaux, A., Lengauer, C., Waldman, T., Zhou, S., Brown, J.P., Sedivy, J.M., Kinzler, K.W., Vogelstein, B., 1998. Requirement for p53 and p21 to sustain G2 arrest after DNA damage. Science 282 (5393), 1497–1501.

Cadinanos, J., Bradley, A., 2007. Generation of an inducible and optimized piggyBac transposon system. Nucl. Acids Res 35, e87.

Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M.H., Dyda, F., 2020. Structural basis of seamless excision and specific targeting by piggyBac transposase. Nat. Commun. 11, 3446.

Cheng, C.-Y., Vogt, A., Mochizuki, K., Yao, M.-C., Bloom, K.S., 2010. A domesticated piggyBac transposase plays key roles in heterochromatin dynamics and DNA cleavage during programmed DNA deletion in Tetrahymena thermophila. Mol. Biol. Cell 21 (10), 1753–1762.

Citterio, E., Van Den Boom, V., Schnitzler, G., Kanaar, R., Bonte, E., Kingston, R.E., Hoeijmakers, J.H.J., Vermeulen, W., 2000. ATP-dependent chromatin remodeling by the Cockayne syndrome B DNA repair-transcription-coupling factor. Mol. Cell. Biol. 20 (20), 7643–7653.

Cordaux, R., Udit, S., Batzer, M.A., Feschotte, C., 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc. Natl. Acad. Sci. USA 103 (21), 8101–8106.

Cosby, R.L., Chang, N.-C., Feschotte, C., 2019. Host-transposon interactions: conflict, cooperation, and cooption. Genes Dev. 33 (17-18), 1098–1116.

Coufal, N.G., Garcia-Perez, J.L., Peng, G.E., Yeo, G.W., Mu, Y., Lovci, M.T., Morell, M., O'Shea, K.S., Moran, J.V., Gage, F.H., 2009. L1 retrotransposition in human neural progenitor cells. Nature 460 (7259), 1127–1131.

Ding, S., Wu, X., Li, G., Han, M., Zhuang, Y., Xu, T., 2005. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. Cell 122 (3), 473–483.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D., 2019. The Pfam protein families database in 2019. Nucl. Acids Res. 47, D427–D432.

Filippi, S., Latini, P., Frontini, M., Palitti, F., Egly, J.-M., Proietti-De-Santis, L., 2008. CSB protein is (a direct target of HIF-1 and) a critical mediator of the hypoxic response. EMBO J. 27 (19), 2545–2556.

Fousteri, M., Mullenders, L.HF., 2008. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. Cell. Res. 18 (1), 73–84.

Fraser, M.J., Smith, G.E., Summers, M.D., 1983. Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses. J. Virol. 47 (2), 287–300.

Gray, L.T., Fong, K.K., Pavelitz, T. and Weiner, A.M., 2012. Tethering of the conserved piggyBac transposase fusion protein CSB-PGBD3 to chromosomal AP-1 proteins regulates expression of nearby genes in humans. PLoS Genet, 8, e1002972.

Gray, L.T., Weiner, A.M., 2010. Ubiquitin recognition by the Cockayne syndrome group B protein: binding will set you free. Mol. Cell. 38 (5), 621–622.

Han, M.V., Zmasek, C.M., 2009. phyloXML: XML for evolutionary biology and comparative genomics. BMC Bioinf. 10, 356.

Hartl, D.L., Lozovskaya, E.R., Lawrence, J.G., 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica 86 (1-3), 47–53.

Helou, L., Beauclair, L., Dardente, H., Arensburger, P., Buisine, N., Jaszczyszyn, Y., Guillou, F., Lecomte, T., Kentsis, A. and Bigot, Y., 2021a. The C-terminal Domain of piggyBac Transposase Is Not Required for DNA Transposition. J. Mol. Biol., 433, 166805.

Helou, L., Beauclair, L., Dardente, H., Piegu, B., Tsakou-Ngouafo, L., Lecomte, T., Kentsis, A., Pontarotti, P., Bigot, Y., 2021b. The piggyBac-derived protein 5 (PGBD5) transposes both the closely and the distantly related piggyBac-like elements Tcr-pble and Ifp2. J. Mol. Biol., 433, 166839.

Henssen, A.G., Henaff, E., Jiang, E., Eisenberg, A.R., Carson, J.R., Villasante, C.M., Ray, M., Still, E., Burns, M., Gandara, J., Feschotte, C., Mason, C.E., Kentsis, A., 2015. Genomic DNA transposition induced by human PGBD5. Elife 4.

Henssen, A.G., Jiang, E., Zhuang, J., Pinello, L., Socci, N.D., Koche, R., Gonen, M., Villasante, C.M., Armstrong, S.A., Bauer, D.E., Weng, Z., Kentsis, A., 2016. Forward genetic screen of human transposase genomic rearrangements. BMC Genomics 17, 548.

Henssen, A.G., Koche, R., Zhuang, J., Jiang, E., Reed, C., Eisenberg, A., Still, E., MacArthur, I.C., Rodríguez-Fos, E., Gonzalez, S., Puiggròs, M., Blackford, A.N., Mason, C.E., de Stanchina, E., Gönen, M., Emde, A.-K., Shah, M., Arora, K., Reeves, C., Socci, N.D., Perlman, E., Antonescu, C.R., Roberts, C.W.M., Steen, H., Mullen, E., Jackson, S.P., Torrents, D., Weng, Z., Armstrong, S.A., Kentsis, A., 2017. PGBD5 promotes site-specific oncogenic mutations in human tumors. Nat Genet 49 (7), 1005–1014.

Hikosaka, A., Kobayashi, T., Saito, Y., Kawahara, A., 2007. Evolution of the Xenopus piggyBac transposon family TxpB: domesticated and untamed strategies of transposon subfamilies. Mol. Biol. Evol. 24, 2648–2656.

Horibata, K., Iwamoto, Y., Kuraoka, I., Jaspers, N.G.J., Kurimasa, A., Oshimura, M., Ichihashi, M., Tanaka, K., 2004. Complete absence of Cockayne syndrome group B gene product gives rise to UV-sensitive syndrome but not Cockayne syndrome. Proc. Natl. Acad. Sci. USA 101 (43), 15410–15415.

Hurst, L.D., 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet. 18, 486.

Ivics, Z., 2016. Endogenous Transposase Source in Human Cells Mobilizes piggyBac Transposons. Mol. Ther. 24, 851–854.

Izsvak, Z., Stuwe, E.E., Fiedler, D., Katzer, A., Jeggo, P.A., Ivics, Z., 2004. Healing the wounds inflicted by sleeping beauty transposition by double-strand break repair in mammalian somatic cells. Mol. Cell. 13, 279–290.

Kapitonov, V.V., Jurka, J., 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biol. 3, e181.

Kazazian Jr., H.H., 1999. An estimated frequency of endogenous insertional mutations in humans. Nat. Genet. 22, 130.

Keith, J.H., Schaeper, C.A., Fraser, T.S., Fraser Jr., M.J., 2008. Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the piggyBac transposase. BMC Mol. Biol. 9, 73.

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115, 49–63.

Kolacsek, O., Erdei, Z., Apati, A., Sandor, S., Izsvak, Z., Ivics, Z., Sarkadi, B., Orban, T.I., 2014. Excision efficiency is not strongly coupled to transgenic rate: cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. Hum. Gene .Ther. Methods 25, 241–252.

Kolacsek, O., Orban, T.I., 2018. Transcription activity of transposon sequence limits Sleeping Beauty transposition. Gene 676, 184–188.

Kolacsek, O., Pergel, E., Varga, N., Apati, A., Orban, T.I., 2017. Ct shift: A novel and accurate real-time PCR quantification model for direct comparison of different nucleic acid sequences and its application for transposon quantifications. Gene 598, 43–49.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K.,

Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A. T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al., 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921.

Li, C., Wen, A., Shen, B., Lu, J., Huang, Y., Chang, Y., 2011. FastCloning: a highly simplified, purification-free, sequence- and ligation-independent PCR cloning method. BMC Biotechnol. 11, 92.

Luan, Y., Suzuki, T., Palanisamy, R., Takashima, Y., Sakamoto, H., Sakuraba, M., Koizumi, T., Saito, M., Matsufuji, H., Yamagata, K., Yamaguchi, T., Hayashi, M., Honma, M., 2007. Potassium bromate treatment predominantly causes large deletions, but not GC>TA transversion in human cells. Mutat. Res. 619, 113–123.

Mabbott, N.A., Baillie, J.K., Brown, H., Freeman, T.C., Hume, D.A., 2013. An expression atlas of human primary cells: inference of gene function from coexpression networks. BMC Genomics 14, 632.

Majumdar, S., Singh, A., Rio, D.C., 2013. The human THAP9 gene encodes an active P-element DNA transposase. Science 339, 446–448.

Mandal, P.K., Kazazian Jr., H.H., 2008. SnapShot: Vertebrate transposons. Cell 135 (192–192), e1.

Manuri, P.V., Wilson, M.H., Maiti, S.N., Mi, T., Singh, H., Olivares, S., Dawson, M.J., Huls, H., Lee, D.A., Rao, P.H., Kaminski, J.M., Nakazawa, Y., Gottschalk, S., Kebriaei, P., Shpall, E.J., Champlin, R.E., Cooper, L.J., 2010. piggyBac transposon/transposase system to generate CD19-specific T cells for the treatment of B-lineage malignancies. Hum. Gene. Ther. 21, 427–437.

Mates, L., Chuah, M.K., Belay, E., Jerchow, B., Manoj, N., Acosta-Sanchez, A., Grzela, D. P., Schmitt, A., Becker, K., Matrai, J., Ma, L., Samara-Kuko, E., Gysemans, C., Pryputniewicz, D., Miskey, C., Fletcher, B., Vandendriessche, T., Ivics, Z., Izsvak, Z., 2009. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. Nat. Genet. 41, 753–761.

Mitra, R., Fain-Thornton, J., Craig, N.L., 2008. piggyBac can bypass DNA synthesis during cut and paste transposition. EMBO J. 27, 1097–1109.

Mitra, R., Li, X., Kapusta, A., Mayhew, D., Mitra, R.D., Feschotte, C., Craig, N.L., 2013. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. Proc. Natl. Acad. Sci. USA 110, 234–239.

Morellet, N., Li, X., Wieninger, S.A., Taylor, J.L., Bischerour, J., Moriau, S., Lescop, E., Bardiaux, B., Mathy, N., Assrir, N., Betermier, M., Nilges, M., Hickman, A.B., Dyda, F., Craig, N.L., Guittet, E., 2018. Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. Nucl. Acids Res. 46, 2660–2677.

Newman, J.C., Bailey, A.D., Fan, H.Y., Pavelitz, T., Weiner, A.M., 2008. An abundant evolutionarily conserved CSB-PiggyBac fusion protein expressed in Cockayne syndrome. PLoS Genet 4, e1000031.

Newman, J.C., Bailey, A.D., Weiner, A.M., 2006. Cockayne syndrome group B protein (CSB) plays a general role in chromatin maintenance and remodeling. Proc. Natl. Acad. Sci. USA 103, 9613–9618.

Pagan, H.J., Smith, J.D., Hubley, R.M., Ray, D.A., 2010. PiggyBac-ing on a primate genome: novel elements, recent activity and horizontal transfer. Genome Biol. Evol. 2, 293–303.

Pan, Y., Oprysko, P.R., Asham, A.M., Koch, C.J., Simon, M.C., 2004. p53 cannot be induced by hypoxia alone but responds to the hypoxic microenvironment. Oncogene 23, 4975–4983.

Pavelitz, T., Gray, L.T., Padilla, S.L., Bailey, A.D., Weiner, A.M., 2013. PGBD5: a neural-specific intron-containing piggyBac transposase domesticated over 500 million years ago and conserved from cephalochordates to humans. Mobile DNA 4, 23.

Pires, I.M., Bencokova, Z., Milani, M., Folkes, L.K., Li, J.L., Stratford, M.R., Harris, A.L., Hammond, E.M., 2010. Effects of acute versus chronic hypoxia on DNA damage responses and genomic instability. Cancer Res. 70, 925–935.

Pritham, E.J., Feschotte, C., Wessler, S.R., 2005. Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol. Biol. Evol. 22, 1751–1763.

Raskó, T., Pande, A., Radscheit, K., Zink, A., Singh, M., Sommer, C., Wachtl, G., Kolacsek, O., Inak, G., Szvetnik, A., Petrakis, S., Bunse, M., Bansal, V., Selbach, M., Orbán, T.I., Prigione, A., Hurst, L.D., Izsvák, Z., 2021. PiggyBac Transposable Element-derived 1 controls Neuronal Progenitor Identity, Stress Sensing and mammal-specific paraspeckles. bioRxiv. https://doi.org/10.1101/2021.05.19.444448v1.

Saha, S., Woodard, L.E., Charron, E.M., Welch, R.C., Rooney, C.M., Wilson, M.H., 2015. Evaluating the potential for undesired genomic effects of the piggyBac transposon system in human cells. Nucl. Acids Res. 43, 1770–1782.

Sarkar, A., Sim, C., Hong, Y.S., Hogan, J.R., Fraser, M.J., Robertson, H.M., Collins, F.H., 2003. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. Mol. Genet. Genomics 270, 173–180.

Sarker, A.H., Tsutakawa, S.E., Kostek, S., Ng, C., Shin, D.S., Peris, M., Campeau, E., Tainer, J.A., Nogales, E., Cooper, P.K., 2005. Recognition of RNA polymerase II and transcription bubbles by XPG, CSB, and TFIIH: insights for transcription-coupled repair and Cockayne Syndrome. Mol. Cell. 20, 187–198.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539.

Singer, T., McConnell, M.J., Marchetto, M.C., Coufal, N.G., Gage, F.H., 2010. LINE-1 retrotransposons: mediators of somatic variation in neuronal genomes? Trends Neurosci. 33, 345–354.

Varga, N., Vereb, Z., Rajnavolgyi, E., Nemet, K., Uher, F., Sarkadi, B., Apati, A., 2011. Mesenchymal stem cell like (MSCl) cells generated from human embryonic stem cells support pluripotent cell growth. Biochem. Biophys. Res. Commun. 414, 474–480.

Velez-Cruz, R., Egly, J.M., 2013. Cockayne syndrome group B (CSB) protein: at the crossroads of transcriptional networks. Mech. Ageing Dev. 134, 234–242.

Vofely, G., Berecz, T., Szabo, E., Szebenyi, K., Hathy, E., Orban, T.I., Sarkadi, B., Homolya, L., Marchetto, M.C., Rethelyi, J.M., Apati, A., 2018. Characterization of calcium signals in human induced pluripotent stem cell-derived dentate gyrus neuronal progenitors and mature neurons, stably expressing an advanced calcium indicator protein. Mol. Cell. Neurosci. 88, 222–230.

Walisko, O., Izsvak, Z., Szabo, K., Kaufman, C.D., Herold, S., Ivics, Z., 2006. Sleeping Beauty transposase modulates cell-cycle progression through interaction with Miz-1. Proc. Natl. Acad. Sci. USA 103, 4062–4067.

Weiner, A.M., Gray, L.T., 2013. What role (if any) does the highly conserved CSB-PGBD3 fusion protein play in Cockayne syndrome? Mech. Ageing Dev. 134, 225–233.

Wu, C., Macleod, I., Su, A.I., 2013. BioGPS and MyGene.info: organizing online, gene-centric information. Nucl. Acids Res. 41, D561–D565.

Yusa, K., Zhou, L., Li, M.A., Bradley, A., Craig, N.L., 2011. A hyperactive piggyBac transposase for mammalian applications. Proc. Natl. Acad. Sci. USA 108, 1531–1536.

Zayed, H., Izsvak, Z., Khare, D., Heinemann, U., Ivics, Z., 2003. The DNA-bending protein HMGB1 is a cellular cofactor of Sleeping Beauty transposition. Nucl. Acids Res. 31, 2313–2322.

Zhang, Z., Li, J., Zhao, X.Q., Wang, J., Wong, G.K., Yu, J., 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genom. Proteom. Bioinform. 4, 259–263.

International Journal of
Molecular Sciences

MDPI

Article

Functional Characterization of the N-Terminal Disordered Region of the *piggyBac* Transposase

Gerda Wachtl [1,2], Éva Schád [1], Krisztina Huszár [1], Antonio Palazzo [3], Zoltán Ivics [4], Ágnes Tantos [1] and Tamás I. Orbán [1,*]

[1] Institute of Enzymology, Research Centre for Natural Sciences, Eötvös Loránd Research Network, 1117 Budapest, Hungary
[2] Doctoral School of Biology, Institute of Biology, ELTE Eötvös Loránd University, 1117 Budapest, Hungary
[3] Department of Biology, University of Bari "Aldo Moro", 70125 Bari, Italy
[4] Transposition and Genome Engineering, Division of Medical Biotechnology, Paul Ehrlich Institute, 63225 Langen, Germany
* Correspondence: orban.tamas@ttk.hu; Tel.: +36-1-382-6638

Abstract: The *piggyBac* DNA transposon is an active element initially isolated from the cabbage looper moth, but members of this superfamily are also present in most eukaryotic evolutionary lineages. The functionally important regions of the transposase are well described. There is an RNase H-like fold containing the DDD motif responsible for the catalytic DNA cleavage and joining reactions and a C-terminal cysteine-rich domain important for interaction with the transposon DNA. However, the protein also contains a ~100 amino acid long N-terminal disordered region (NTDR) whose function is currently unknown. Here we show that deletion of the NTDR significantly impairs *piggyBac* transposition, although the extent of decrease is strongly cell-type specific. Moreover, replacing the NTDR with scrambled but similarly disordered sequences did not rescue transposase activity, indicating the importance of sequence conservation. Cell-based transposon excision and integration assays reveal that the excision step is more severely affected by NTDR deletion. Finally, bioinformatic analyses indicated that the NTDR is specific for the *piggyBac* superfamily and is also present in domesticated, transposase-derived proteins incapable of catalyzing transposition. Our results indicate an essential role of the NTDR in the "fine-tuning" of transposition and its significance in the functions of *piggyBac*-originated co-opted genes.

Keywords: DNA transposon; *piggyBac*; PGBD; intrinsically disordered protein

Citation: Wachtl, G.; Schád, É.; Huszár, K.; Palazzo, A.; Ivics, Z.; Tantos, Á.; Orbán, T.I. Functional Characterization of the N-Terminal Disordered Region of the *piggyBac* Transposase. *Int. J. Mol. Sci.* **2022**, *23*, 10317. https://doi.org/10.3390/ijms231810317

Academic Editor: Benoît Chénais

Received: 22 July 2022
Accepted: 3 September 2022
Published: 7 September 2022

1. Introduction

Transposons are mobile genetic elements capable of relocating from one genomic locus to another, and they often make up significant portions of the genomes of organisms, including humans [1]. There are two classes of transposons: members belonging to the RNA transposons (or retrotransposons) use an RNA intermediate for their replication cycle and use a 'copy-and-paste' mechanism for transposition. In contrast, the class of DNA transposons contains elements that rely on DNA intermediates for their propagation, and the reaction mainly occurs via 'cut-and-paste' mechanisms [2,3]. Transposons were initially considered mutagenic, "selfish" genomic parasites. However, they could also contribute to the adaptive evolution of organisms by either carrying advantageous genes for survival (such as antibiotic resistance genes for bacteria) or by becoming harmless "domesticated" genes that serve a novel endogenous function of the host [4–10].

Due to their simple structure and the cut-and-paste mechanism of non-replicative nature, some DNA transposons became favorable genetic tools for mutagenesis or gene delivery experiments [11]. For mammalian applications, the insect-derived *piggyBac* (PB) [12,13], the resurrected artificial *Sleeping Beauty* (SB) [14], and the medaka fish-originated *Tol2* [15,16] transposons are all widely used, the two former ones being the most favorable systems,

partly because of the development of their high efficiency ("hyperactive") variants [17,18]. The SB system has a clear advantage for human gene therapy applications due to its safest integration profile. Among all integrating gene delivery vehicles, SB has the least tendency to integrate into endogenous genes, providing the lowest likelihood for a genotoxic event. In addition, the lack of potentially cross-reacting endogenous elements makes it an attractive choice for human applications [19]. Considering the PB system, there are PB-like elements in the human genome, and their cross-reactivity with the insect transposon is still a matter of debate [20–22]. On the other hand, the PB transposase performs a seamless excision of the transposon unit from a genomic location [23–25]. This ability to remove the DNA substrate without leaving any footprints is convenient for specific applications: a prominent example is genetic reprogramming, when the transgenic cassette can be safely eliminated after the desired cellular phenotype has been achieved [26,27]. Apart from exploiting the beneficial characteristics of a given transposon system, there are continuous efforts to further optimize the transposase enzymes for more efficient and controllable reactions [28]. To achieve that, there is also a constant need to understand the structure-function correlations of the transposases and, thereby, understand the transposition reactions in more detail.

Our studies focus on the functional characterization of the PB transposase and its relatives. They represent a widespread superfamily of DNA transposons [29–31] and contain active mobile elements in various genomes, including bats [32]. The first discovery of a PB transposase was made in the cabbage looper moth (*Trichoplusia ni*) cells [33,34], and this insect-derived autonomous element was proven to be functioning and applicable in various other organisms, including human cells [12]. When characterizing its structure, this transposase was shown to have a domain with an RNase H-like fold containing a DDD motif [35]: this triad of aspartates is responsible for catalyzing the DNA cleavage and joining reactions [36], similarly to retroviral integrases and the IS4 family of insertion sequences [37,38]. This catalytic domain is located in the central part of the protein. Though, as with other DDE/D structures, a small insertion domain interrupts the middle portion [39]. The catalytic domain is flanked on both sides by regions creating the 'dimerization and DNA-binding domain', which is responsible for bringing two proteins together to form a dimer, and also for interacting with a short sequence part of the transposon terminal inverted repeats (TIRs), as well as with the target DNA [39]. A cysteine-rich domain (CRD) is found at the C-terminal end of the protein, which contains a nuclear localization signal [40]. The CRD was proposed to be indispensable for transposon DNA binding and DNA breakage [41], and it was shown that CRDs of two interacting transposases induce the formation of an asymmetric protein dimer structure associated with a single (the left) TIR of the transposon [39]. As opposed to this, a subsequent study challenged these results by claiming that the CRD is not required for PB DNA transposition, as CRD-deficient transposases, including a domesticated PGBD5 transposase, were found to be functional [42]. A possible resolution for this issue could lie within the evolution and structure–function variability of the CRD in active and domesticated PB transposases. A comparative analysis of various cysteine-rich domains revealed significant structural differences among distantly related PB elements. As a functional consequence, it was shown that the CRD of the domesticated *piggyMac* does not bind DNA [43].

Although numerous investigations have been published on addressing the structure and function of the insect PB transposase, our knowledge about the potential role of the N-terminal region is still very limited. This ~100 amino-acid-long portion is predicted to have a highly disordered structure, and the dominantly acidic nature of the residues makes it unlikely that it contributes to DNA binding [39]. As one of the most common functions of disordered protein regions is establishing inter- or intramolecular interactions, which often provide the basis for allosteric regulation [44], it is thus conceivable that similar modulation of the PB protein is achieved by protein partners interacting with the N-terminal part of the transposase. Therefore, this study aimed to understand how this N-terminal disordered region (NTDR) of the PB transposase contributes to PB transposition. By

deleting this N-terminal segment, we could detect a significant drop in overall transposition efficiency; however, the extent of the reduction showed strong cell type dependency. We also provide evidence that the excision step of transposition is heavily influenced by NTDR deletion. When the NTDR was replaced with a disordered segment that contained the same amino acids in a randomized order, wild-type transposition efficiency was not restored; in fact, the impairment was even more pronounced, indicating the importance of the protein sequence, not only the structure or the amino acid composition of this region. With a systematic bioinformatic analysis, we also show that the NTDR is specific for the *piggyBac* superfamily of DNA transposons and that it can be found even in domesticated, transposase-derived proteins that lost their ability to mobilize DNA. These include the five human *piggyBac*-derived sequences (PGBD1-5), which have been shown to have lost their mobilizing activity [22]. These results support the hypothesis that the NTDR of PB has an important regulatory function in transposition and that this regulatory function has likely been preserved during the evolution of PB-derived genes.

2. Results

2.1. Deletion of NTDR Significantly Decreases PB Transposition in a Cell Type-Specific Manner

We began our investigations using the mammalian codon-optimized version of the *T. ni* PB transposase (also called mPB, [13]) and its two-component vector system (Figure 1A). To test the influence of the NTDR on transposition, we deleted 100 amino acids from the N-terminus of the mPB transposase. This peptide region is predicted to have an intrinsically disordered segment (Figure 1B). The length of the deletion was determined by our earlier predictions showing that several PB elements have at least a 100 amino acid long disordered region at their N-termini. The resulting mutant transposase (mPBdel) was co-transfected with a PB transposon vector carrying a puromycin resistance gene cassette into HEK-293 cells. At 48 h post-transfection, transposon excision was checked by a diagnostic PCR, whereas determining the colony numbers after two weeks of puromycin selection provided information about the overall transposition efficiency of the mutant. The wild-type mPB transposase was used in parallel transfections as a positive control. In contrast, experiments with a catalytically inactive mutant (mutPB, see Figure 1B) served as a negative control to estimate the background of random integration. These assays revealed the overall proficiency of mPBdel to excise PB transposons (Figure 1C); however, the transposition rate dropped significantly, to approximately 55% of the level detected for the wild-type mPB transposase (see the colony assays, Figure 1D).

To address if the effect of NTDR deletion on mPB transposition can also be detected in other cell types, we performed similar experiments in HeLa cells. We measured a significant decrease in the overall transposition activity of the mPB-del mutant (Figure 2A,C). However, the magnitude of reduction was much higher than in HEK-293 cells: the colony numbers detected for the mPBdel reached only 20% of that measured for the wild-type PB transposase. We also tested the mPBdel activity in MCF-7 cells to further analyze transposition's cell type specificity. The tendency was similar in that the efficiency of transposition efficiency was reduced to approximately 35% of the wild-type transposase (Figure 2B,D). These results indicated that the disordered region plays an essential role in transposition: its deletion significantly reduces the transposition rate to varying extents in different human cell types.

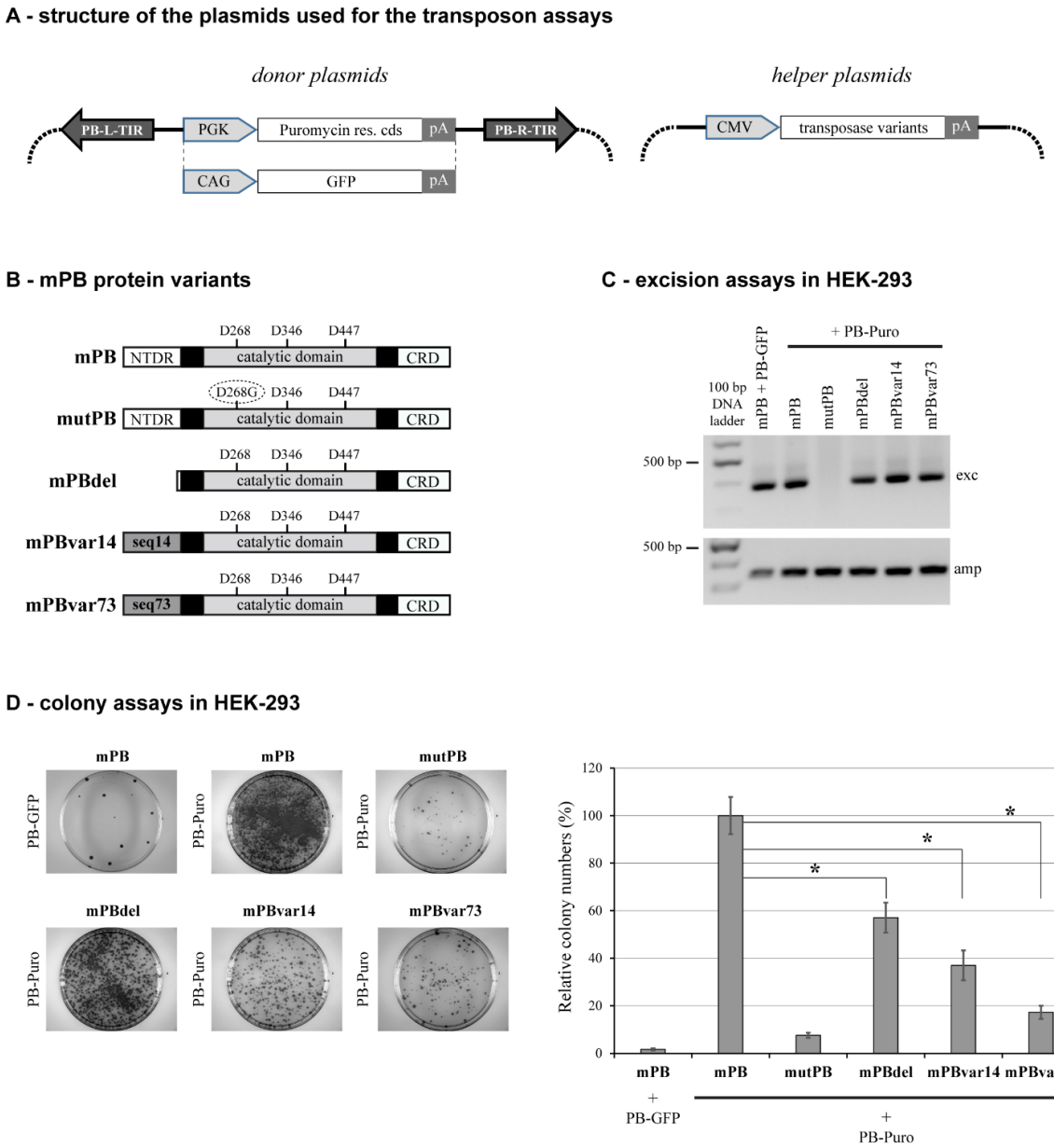


Figure 1. Testing the activities of *piggyBac* transposase mutants in HEK-293 cells. (**A**) Schematic overview of the donor and helper plasmids used for the excision and the colony assays. PB-L-TIR/PB-R-TIR: *piggyBac* left/right terminal inverted repeats; PGK: phosphoglycerate kinase promoter; CAG: cytomegalovirus-actin-globin artificial hybrid promoter; CMV: cytomegalovirus promoter; Puromycin res. cds: Puromycin resistance coding sequence; GFP: green fluorescent protein; pA: polyadenylation signal. (**B**) A schematic representation of the *piggyBac* transposase protein variants. In the catalytic domain, the positions of the three important aspartic acids (D) are shown; for the catalytic mutant (mutPB), the 268 position is mutated to glycine (G) as indicated. The two black boxes represent the two distinct regions of the 'dimerization and DNA-binding domain'. NTDR: N-terminal disordered region; CRD: cysteine-rich domain; seq14/seq73: two scrambled protein sequences designed to have similar disordered characteristics as the original N-terminal region of mPB while preserving the amino acid composition. (**C**) A representative result of the diagnostic PCRs detecting the occurrence of transposon excision. The upper product ('exc', 381 bp) is derived from the excised and repaired donor plasmids. The lower product ('amp', 340 bp) is a control amplicon from the ampicillin resistance gene on the backbone of the transfected plasmids. (**D**) On the left, representative images of plates are shown containing resistant colonies after puromycin selection (digital photos of Petri dishes with diameter of 10 cm). The graph on the right presents the quantification

of colony numbers for each assay condition. All values are expressed as percentages of the 'mPB+PB-Puro' control reaction. Mean values of at least four independent biological experiments are shown, and error bars represent standard deviations, *: $p < 0.05$. For excision and colony assays, the combinations of donor and helper plasmids are indicated for each reaction.

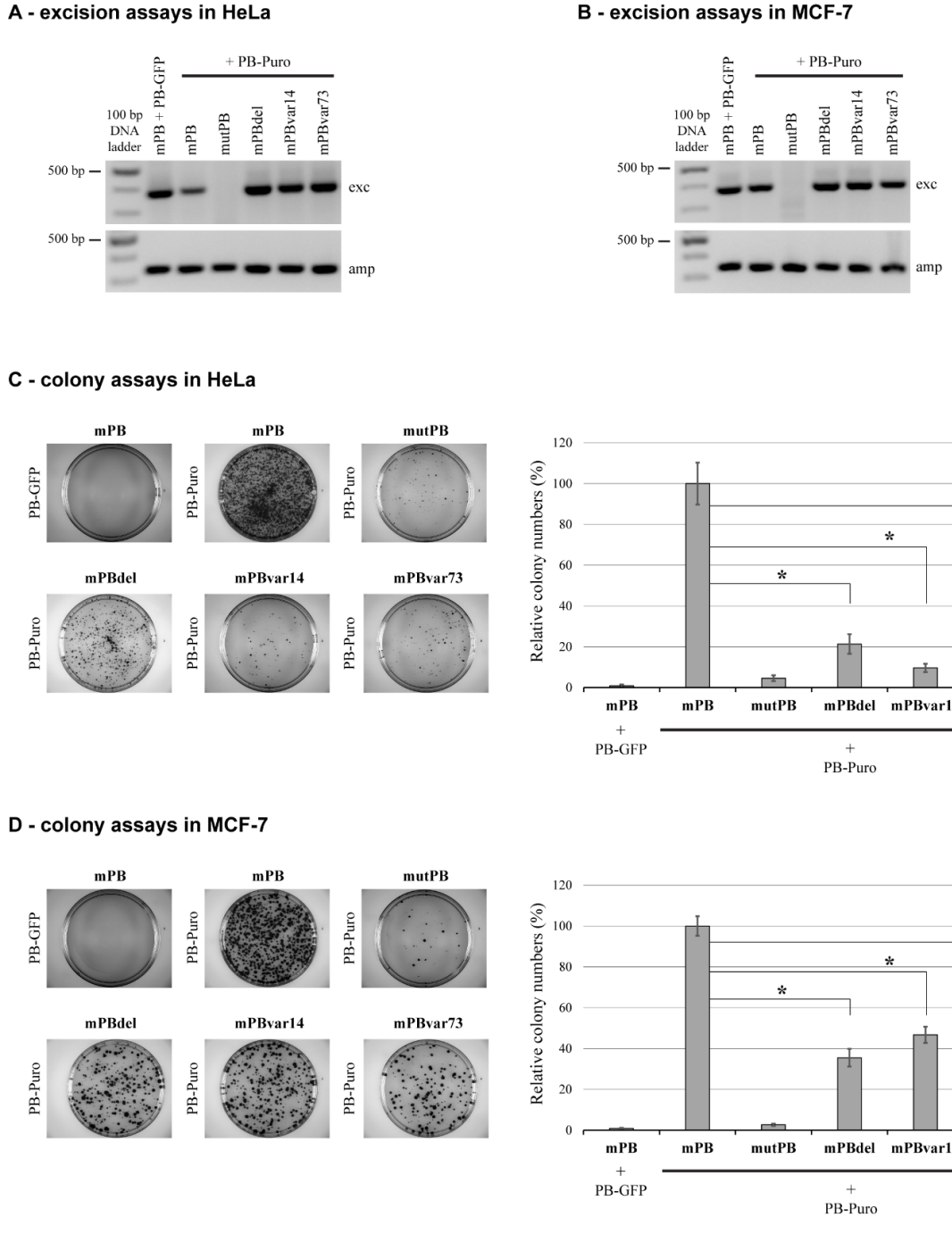


Figure 2. Testing the activities of *piggyBac* transposase mutants in HeLa and MCF-7 cells. (**A**,**B**) show representative images of excision PCRs in HeLa and MCF-7 cells, respectively. The upper band ('exc', 381 bp) is detected when transposon excision and plasmid repair have occurred, whereas the lower control band ('amp', 340 bp) is detected in case of successful plasmid transfection. (**C**,**D**) show representative images and quantification of colony assays in HeLa and MCF-7 cells, respectively (digital photos of Petri dishes with diameter of 10 cm). Shown values are means of at least six (HeLa) or three (MCF-7) independent biological replicates. Error bars represents standard deviations, *: $p < 0.05$.

2.2. The Sequence of NTDR Is Important for Its Proper Function

To determine whether the amino acid sequence is also important for the functional influence (in addition to the disordered feature of the region), we replaced the original NTDR with segments bearing similar disordered characteristics but having a randomized amino acid sequence. This approach allowed the preservation of the amino acid composition but disrupted any function that may be linked to specific sequence motifs. We tested two scrambled NTDR PB mutants named mPBvar14 and mPBvar73. We performed the transposition assays with these variants in all three cell lines described above to evaluate if these sequences can complement the NTDR deletion phenotype. These experiments showed the inability of these disordered protein segments to rescue wild-type transposition efficiency; in most cases, it was even further reduced as compared to the mPB-del activity (Figures 1 and 2) except for mPBvar14 in MCF-7 cells, where the transposition activity was slightly higher than for the mPBdel mutant (Figure 2D). The results support the hypothesis that sequence conservation of the PB transposase NTDR is important for its functionality. In addition, since the deletion or the replacement of this disordered segment did not abolish transposition, it indicates that this region contributes to the "fine-tuning" of the reaction, most likely via interacting with cellular partners.

2.3. The Effect of NTDR Deletion on the Hyperactive PB Variant

To further investigate the functional role of the NTDR, we deleted this region also from a recently developed hyperactive PB variant (hyPB, [18]) (Figure 3A). In line with the previous findings for mPB, transposon excision was always detected (Figure S1). However, in contrast to the results for the mPB transposase, no significant changes were detected between transposition efficiencies of the hyPB and the hyPBdel variants in HEK-293 or HeLa cells (Figure 3B,C). On the other hand, in MCF-7 cells, we detected an approximately 40% reduction in the overall transposition rate for the hyPBdel variant, again indicating cell type-specific regulation of transposition (Figure 3D). When the NTDR was replaced with either of the previously generated, randomized sequence disordered segments (denoted as hyPBvar14 and hyPBvar73), we detected a substantial reduction of transposition in all three cell lines examined, with the most prominent effect seen in HeLa cells with the var73 sequence (Figure 3B–D).

There is only one amino acid difference in the NTDR of hyPB compared to mPB. To determine how much this I30V mutation contributes to the transposition efficiency, we mutated this amino acid position back to the original isoleucine found in the mPB sequence (the mutant is denoted as hyPB^{30I}, see Figure 3A). This "back mutation" showed a similar cell type-specific effect on transposition as the hyPBdel: no significant change or a slight tendency of increase was detected in the HEK-293, and HeLa cell lines, respectively, whereas a more minor but significant decrease was seen in the MCF-7 cell line (Figure 3B–D). All these results further supported a cell type-specific role of this N-terminal region in the transposition function. They indicated the importance of sequence constraint in the NTDR of PB transposase.

2.4. NTDR Deletion Has a Higher Impact on the Excision Step of Transposition

In an attempt to distinguish the influence of PB NTDR on the distinct steps of transposition, we used a specific reporter cell line to assess the deletion of this region during the excision and on the integration step. This HeLa-derived cell line contains one copy of a construct where a puromycin resistance gene is interrupted by a PB transposon carrying a neomycin resistance cassette (Figure 4A). Suppose a functional PB transposase is expressed in this cell line. In that case, it could tracelessly remove the transposon resulting in puromycin resistance, thereby quantitatively scoring the efficiency of the excision step of the reaction. In addition, if a puromycin-neomycin double selection is applied, one can screen for the complete transposition reactions ending with functional integration events. We compared the activity of the mPB transposase and the mPBdel variant in this reporter system and detected a significant reduction, by approximately 70%, in the

transposon excision by mPBdel (Figure 4B, upper panels, compare the 'exc' values for mPB and mPBdel). The overall transposition efficiency (number of the combined excision and integration events) also drops significantly due to the NTDR deletion. However, the ratio of integrations among the excised transposon population does not decrease: its value is higher for the mPBdel variant (Figure 4B, upper right graph). These results indicate that the excision step during PB transposition is influenced more significantly by the removal of NTDR than the subsequent step of genomic integration.

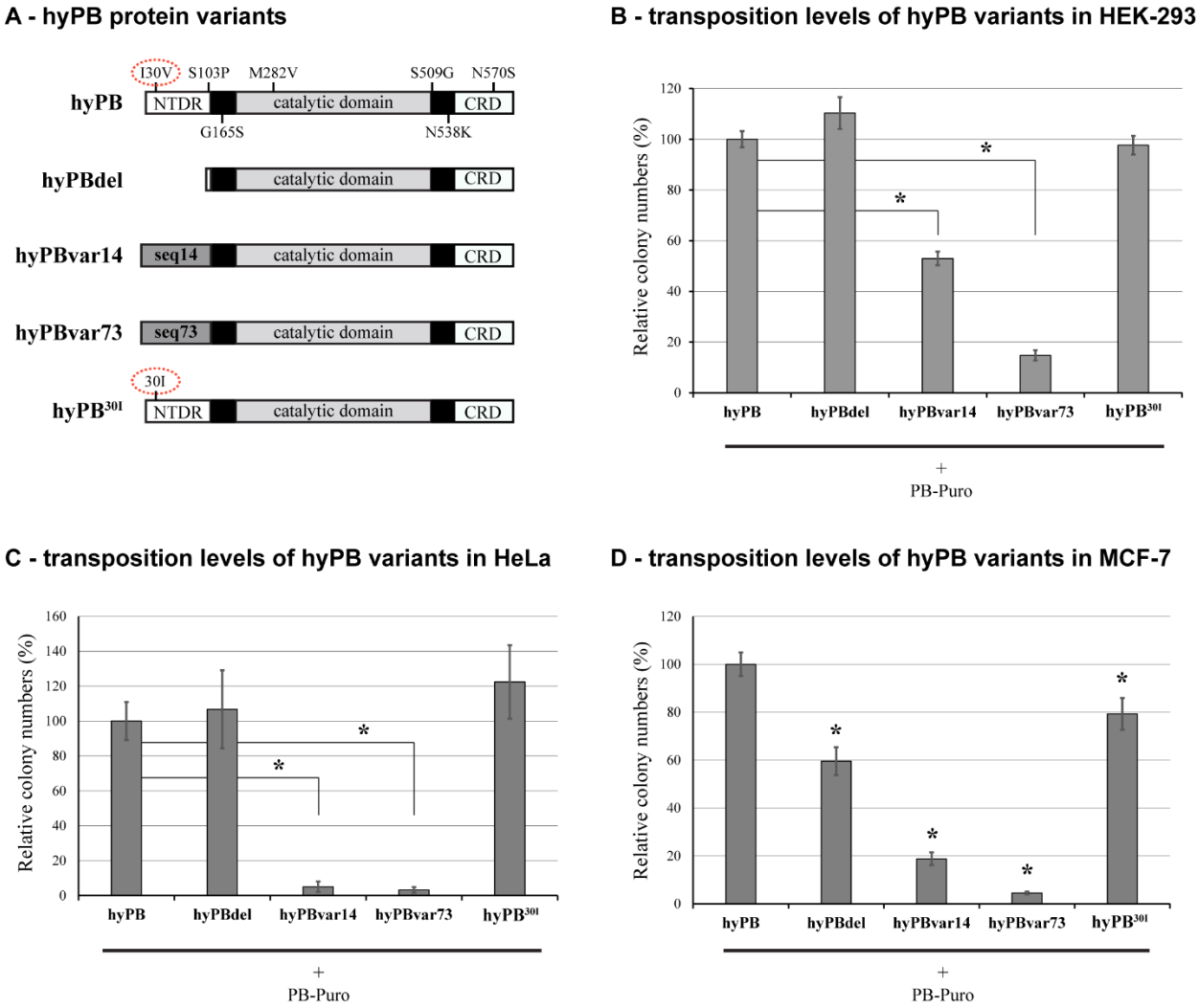


Figure 3. Testing the activities of the hyperactive *piggyBac* transposase mutants in three cell lines. (**A**) A schematic representation of the hyPB transposase protein variants. The hyPB differs from the mPB protein in 7 amino acid positions indicated on the upper drawing. Only one of these positions (indicated in red dashed lined ovals) is affected by the NTDR deletion (I30V) and tested for its influence on hyPB activity (30I). The domain structure of the transposase is indicated in Figure 1B. (**B**–**D**) panels show quantifications of colony assays in HEK-293, HeLa, and MCF-7 cells. Shown values are means of at least three independent biological replicates. Error bars represents standard deviations, *: $p < 0.05$.

We also performed similar experiments on this HeLa reporter cell line with the hyPB variants. It was revealed that in such limited substrate conditions (acting on a single transposon copy in the genome), the hyPB transposase was more effective than the mPB version, both in the absolute numbers of excision events, as well as in the integration rates, as expected. On the other hand, as compared to the mPBdel mutant, the NTDR deletion of hyPB resulted in a more severe (reaching 90%) reduction in excision efficiency. However,

the integration rate did not decrease for the hyPBdel mutant: in fact, it also showed a tendency to increase, similarly to the mPBdel mutant (Figure 4B, graphs on the right). Again, these results further support the previous observation that the excision step of transposition is more significantly affected by a deletion of the NTDR in the PB transposase.

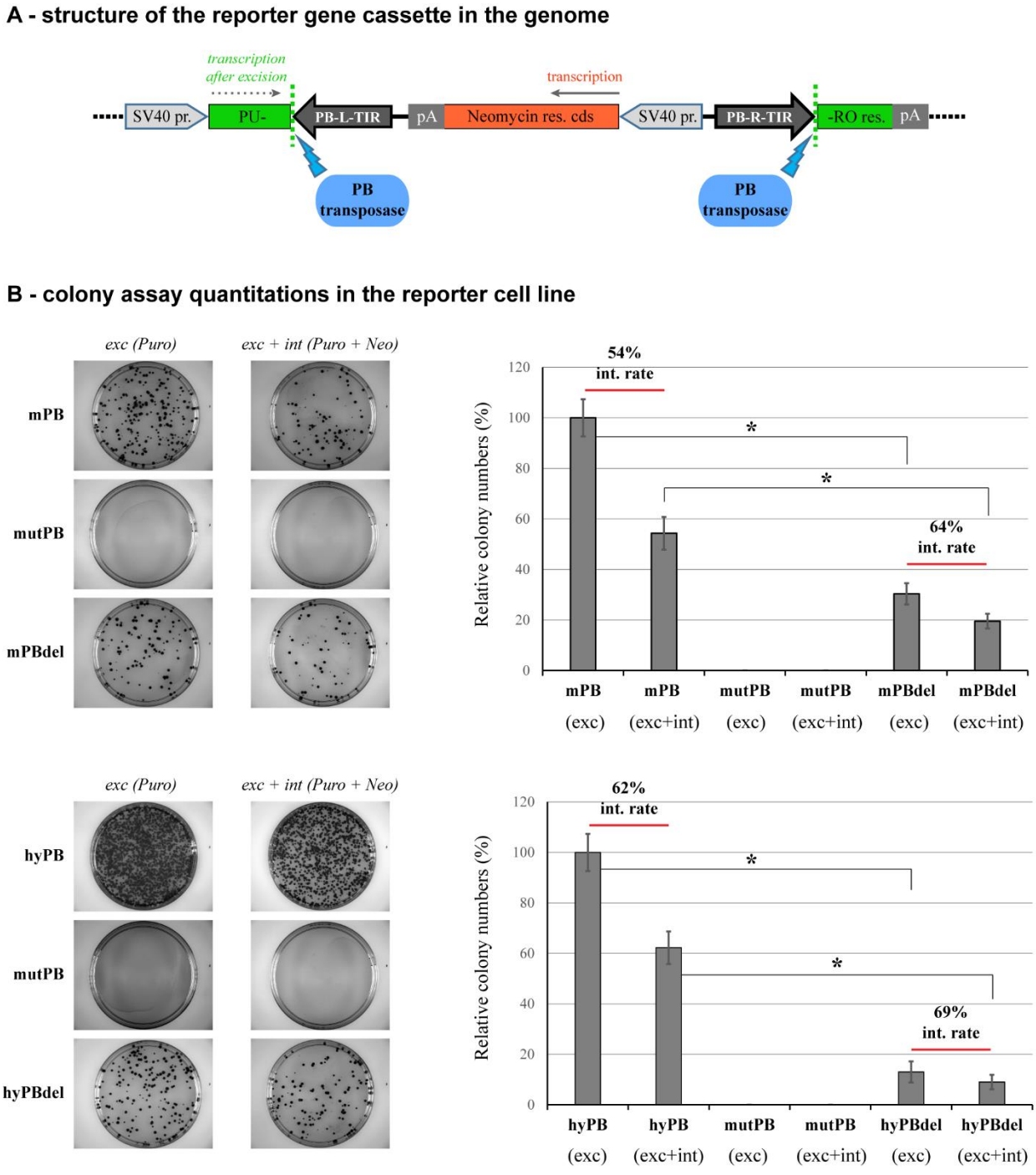


Figure 4. Testing the activities of the mPB and the hyPB transposase variants in a HeLa reporter cell line containing a single copy of a transposon substrate. (**A**) Structure of the reporter gene cassette in the genome. A puromycin resistance gene (marked in green) is interrupted with a *piggyBac* transposon carrying a neomycin resistance gene cassette (marked in red). If an active transposase is expressed in the cells (depicted by a blue oval), it may seamlessly excise the transposon unit, thereby restoring puromycin resistance. SV40 pr.: simian virus 40 early promoter; PB-L-TIR/PB-R-TIR: *piggyBac* left/right terminal inverted repeats; pA: polyadenylation signal. (**B**) On the left, representative

images of plates are shown after selection with antibiotics (digital photos of Petri dishes with diameter of 10 cm). Puromycin treatment selects only for transposon excisions ('exc'), whereas puromycin and neomycin double treatment selects for transposon excision and integration events ('exc+int'). On the right, the graphs show the quantification of colony numbers for various assay conditions. All values are expressed as percentages of the mPB (upper graph) or hyPB (lower graph) control reactions; the experiment is carried out with the catalytically inactive transposase mutant for negative control. Integration rates ('int. rate') for double selections are calculated as the percentage of mere excision events (puromycin selection) for each transposase variant. Mean values of at least three independent biological experiments are shown, and error bars represent standard deviations; *: $p < 0.05$.

2.5. Conservation of the NTDR in Domesticated PB-Derived Proteins

Seeing that the NTDR is necessary for the proper function of mPB, we asked whether the presence of such a longer disordered region at the N-terminus is a characteristic feature of other *piggyBac*-related proteins. Using the IUPred2 program [45,46] to predict the disorder tendencies of the five domesticated human *piggyBac*-derived sequences (PGBD1-5) [22,31], we found that all of these proteins contain a long stretch of structurally disordered regions upstream of the transposase domains, even PGBD1, which has gained additional N-terminal protein domains during evolution (Figure 5A). Inspired by this, we conducted a further systematic analysis of all the currently available sequences of this superfamily of DNA transposons. This included active transposases and domesticated PB proteins that are proven or predicted to have become domesticated. This analysis revealed that more than 95% of PB proteins contain a disordered stretch of >10 amino acids upstream of the transposase catalytic domain and can be considered NTDR (Table S1). We conclude that such a protein region is a widespread feature of these transposase-originated proteins, even those that have acquired additional protein domains during evolution. Based on their sequences and their domain structures, the analyzed proteins could be classified into distinct groups: most of these sets could be defined and named after the five representative domesticated human PB proteins, PGBD1-5; in addition, proteins that do not fit into those categories form an extra group (these are almost exclusively active transposases, including the *T. ni* insect transposase). Most of the members belonging to the *piggyBac1* and the *piggyBac4* groups contain extra domains: the majority of the former group having acquired a SCAN and a KRAB domain during evolution which is located upstream of the transposase domain [47]. As opposed to this, the *piggyBac4* proteins contain a Zn-ribbon domain downstream of the catalytic domain of the transposase. Despite these extra domains, all these PB proteins still have a conserved NTDR upstream of the catalytic domain of the transposase, indicating its functional relevance even for the domesticated functions. However, it is important that the evolutionary most ancient *piggyBac5* group contains the most members where such NTDR cannot be recognized (Table S1).

It is also important to note that the Anchor algorithm, designed to identify potential interaction sites within disordered sequences [48], shows a high probability of the NTDR regions containing such regions (Figure 5A, blue lines). This underlines the relevance of the experimental results, where the scrambled NTDR sequences could not restore the transposase activity of mPB. In agreement with this observation, the scrambled variants do not have a pronounced disordered binding site, according to the ANCHOR prediction (Figure S2).

Finally, we addressed the question of whether an NTDR is also present in other DNA transposases from different superfamilies. We performed a disorder prediction for the *Tol2* transposase from the hAT superfamily and the *Setmar* protein from the Tc1/Mariner superfamily. In contrast to the PB proteins, they do not show an N-terminal concentration of disordered region; the predictions indicated the absence of such long disordered regions throughout the entire proteins (Figure 5B). These results supported the previous observation that the presence of a long NTDR upstream of the catalytic domain of the transposase is a characteristic feature of the PB superfamily.

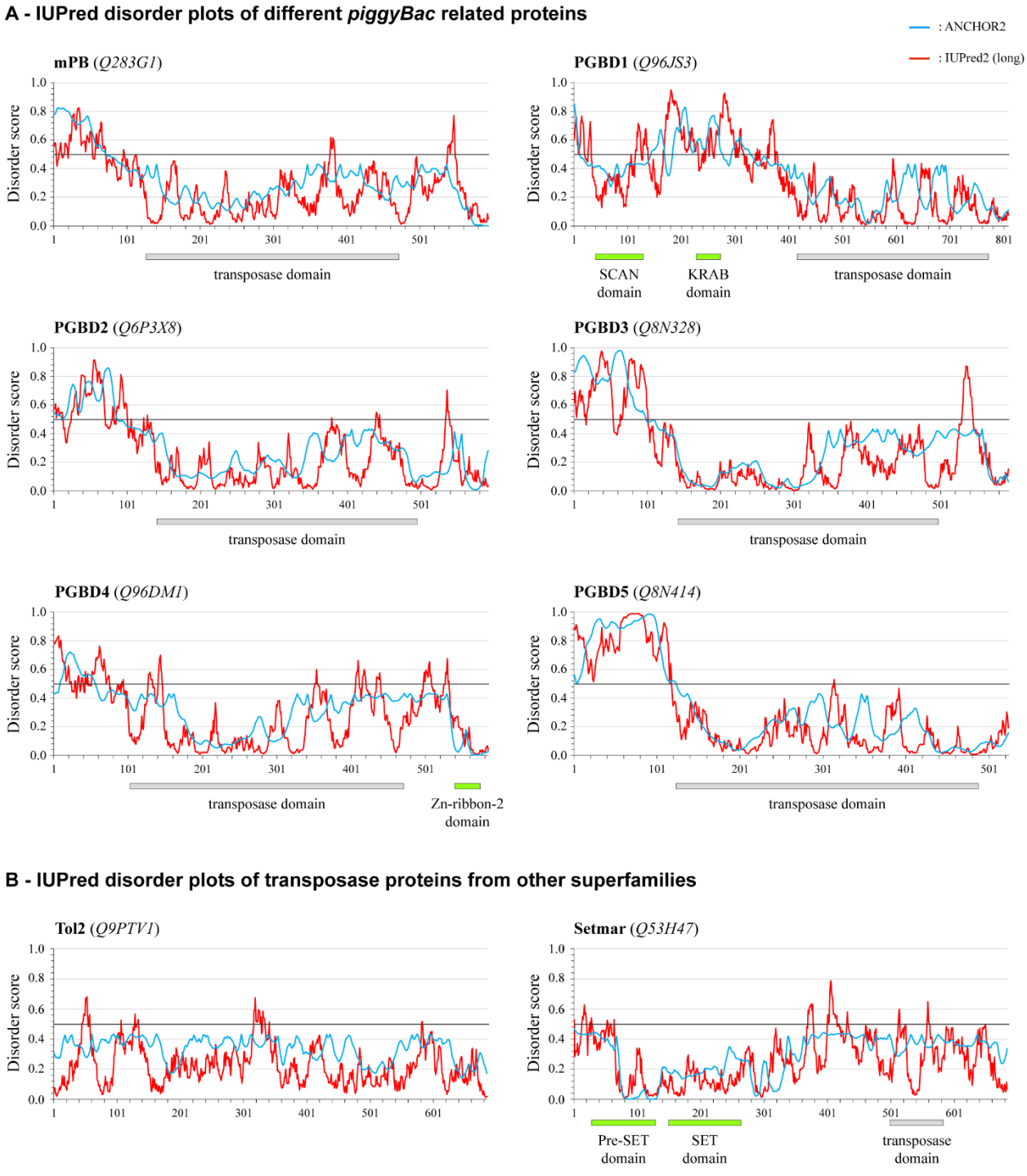


Figure 5. Prediction of disordered properties of various transposase proteins using the IUPred prediction software. (**A**) Disordered plots of the mPB transposase and the five human *piggyBac*-derived domesticated protein sequences (PGBD1-5). Amino acids numbered from the N-terminal of the proteins are shown on the *x*-axis; disorder scores plotted on the *y*-axis, calculated using the ANCHOR2 (blue line) or the IUPred2 (red line) algorithms. A peptide region is considered structurally disordered if the score values are above 0.5, and the horizontal solid black line indicates this threshold. (**B**) Disordered plots of non-*piggyBac* transposase proteins: *Tol2* belongs to the hAT superfamily, whereas *Setmar* belongs to the Tc1/Mariner superfamily. Known domains (transposase or others) are indicated under the graph of a given protein; UniProt identifiers of proteins are shown in parenthesis.

3. Discussion

Although various domains of the PB transposase have been mapped and functionally well characterized, the N-terminal disordered region has not been examined, and its function remains elusive [39]. In this study, we investigated the role of this NTDR by performing transposition reactions with various deletion mutants in different cell lines. Testing the mPB transposase lacking the NTDR, we could detect a drastic and significant drop in transposition efficiency in a strong cell type-dependent manner. However, transposition was not completely abolished, indicating that the presence of the NTDR is not essential for transposition. Interestingly, suppose the NTDR was replaced with disordered protein segments with the same amino acids in a randomized order. In that case, this could not complement the deletion phenotype, but it further lowered the transposition rate in most cases. These results clearly showed that the sequence of the NTDR also carries functional relevance, possibly through interaction with other molecules, as suggested by the Anchor predictions.

As opposed to the results with mPB, deletion of the NTDR in the hyperactive hyPB transposase [18] did not recapitulate the negative phenotype in most cell lines, whereas the scrambled mutants always severely lowered transposition efficiency. The results indicate that the 7 "hyperactive" mutations could collectively compensate for the lack of NTDR in a cell type-specific manner while simultaneously sensitizing the protein to the presence of a dysfunctional disordered segment. The inhibitory effects of cell-type specific factors cannot be ruled out at this point, but so far, no reliable information on those is available. The hyPB was initially developed in a yeast system, with subsequent testing in mouse embryonic stem cells. Although significantly outperforming the wild-type mPB variant, substantial differences in the activity were detected even between those cell types [18]. In a later systematic study, the robustness and the clear advantage of the hyPB were demonstrated in human embryonic stem cells. In contrast, in other cell types, it showed significant variability, not always providing the expected "hyperactivity" [49]. Such cell type-dependent variability was demonstrated here with the hyPB NTDR deletion mutant. We could show that the only "hyperactive" mutation (I30V) positioned in the NTDR provides only a minor contribution to the hyperactive phenotype and, therefore, to the lack of transposition impairment of the NTDR-deleted hyPB. Nevertheless, further systematic studies are needed to investigate the exact structural role of these mutations and to decipher how and to what extent they can override the molecular interaction(s) of the NTDR sequence.

Our results with a single copy reporter cell line revealed that the excision step of transposition is more severely affected by the NTDR deletion, both for the mPB and the hyPB transposase variants. Previous studies already indicated that the two stages of transposition are kinetically distinct: the excision efficiency of the hyPB showed a 17-fold increase compared to the wild-type mPB variant, whereas the integration efficiency was elevated "only" by 9-fold [18]. A definite proof was provided by developing an integration deficient PB transposase, showing the structural background for the selective inhibition of the integration step [50]. Future studies would be needed to see how the NTDR deletion affects this variant. Such experiments would help focus on the selective function of the disordered region on the PB excision reaction.

Our bioinformatic analyses provided evidence that such a disordered protein region is present at the N-terminal part of the transposase domain in other *piggyBac*-related proteins. It was revealed that the position of this region is well-conserved among the members of the *piggyBac* superfamily, even among the domesticated proteins that were shown to have lost their ability for transposition, including the five domesticated human PGBD proteins [22]. It is intriguing that regardless of the acquisition of other non-transposase domains during evolution (as in the case of PGBD1 or PGBD4), the position of the disordered region is strictly kept upstream of the DDE_Tnp_1_7 transposase domain among *piggyBac* proteins; however, selected members from other DNA transposon superfamilies do not show this structural feature. This special structural constellation of the *piggyBac*-related sequences would further point to functional conservation, but the question remains: what would be

the role of NTDR in PB proteins that do not show excision activities anymore [22,47]? A possible explanation could be that certain protein–protein interactions may well serve the new domesticated functions, and understanding the role of the NTDR in PB transposition would therefore be helpful to reveal the endogenous function of the co-opted proteins.

So, what could be revealed from these experiments concerning the potential function of the N-terminal disordered region? The most likely conclusion is that it is involved in various protein–protein interactions that modulate the transposition reaction. The PB transposase does not seem to require many protein partners, indicated by the lack of strict host specificity and the wide range of organisms where this transposon system can be used. However, there are certain protein groups that the transposase possibly interacts with: these include the DNA repair factors to suppress genotoxic effects caused by double-stranded breaks generated during transposition or certain chromatin-associated factors that determine the distinct integration profile of the transposon [23,51]. Cell cycle regulating proteins may also be good candidates as such interactions were shown to be important for controlling other DNA transposases, connecting their activities to cell proliferation [52]. Our preliminary attempt to validate selected DNA repair or cell cycle factors as potential *piggyBac* regulators failed. The results indicated that a future systematic screen is required to identify *bona fide* NTDR interacting partners. The flexibility of a disordered region could provide a platform to interact with several protein partners or even different parts of the same protein, promoting different functions on the same partner, referred to as the "moon-lighting function" of disordered proteins [44]. Previous studies showed that the lack of disorder-to-order transition, alternatively called the "fuzziness" of intrinsically disordered proteins [53,54], could provide the structural flexibility to perform unrelated functions on different substrates, as proven by the versatile actions of the thymosin-β4 protein [55]. Another example showed that a disordered segment could have opposing functions on the same target protein. As a result of two independent stochastic actions, a random-coil fragment of the dihydropyridine receptor can activate or inhibit its ryanodine receptor target during muscle contraction [56]. Based on such examples, one can hypothesize that the NTDR could enhance the excision step with specific protein partner(s). Subsequently, it may also promote substrate integration during PB transposition.

Another possibility is that the NTDR can modulate interactions between individual PB transposase proteins or even mediate intramolecular connections to initiate "higher order PB assembly synapses", as predicted by Chen et al. [39], and the CRD region of the *piggyBac* protein (Figure 1B) may be a good candidate for that [41]. However, our results on the cell-type specific behavior of the NTDR mutants point to the role of interaction with other protein partners, whose cell-type specific expression profiles may lie behind the observed differences. Nevertheless, it would be interesting to see if any revealed functions are still conserved in the domesticated PGBD proteins or if new functions arose for the NTDR in those cases. Another proposed function of disordered protein segments is that they may facilitate protein evolution via exon shuffling by structurally separating the newly combined domains [57]. One can argue that this could be the case for mammalian PGBD1 proteins [47]. However, it does not exclude other functional role(s) of the conserved NTDR. Be that as it may, further experimental studies are required to decipher the potential function(s) of the N-terminal disordered region in *piggyBac* transposition, which results would help elucidate the domesticated functions of various co-opted *piggyBac*-derived protein sequences.

4. Materials and Methods

4.1. Plasmid Constructs

Structures of donor and helper plasmids used for the transposon assays are shown in Figure 1A. Donor plasmids contain *piggyBac* transposon units carrying either a PGK promoter-driven puromycin resistance gene or a CAG promoter-driven GFP gene [49]; the helper plasmids contain a CMV promoter-driven appropriate *piggyBac* transposase variant expression cassette (see the figure legend for more details). The codon-optimized mPB [13]

and the hyperactive hyPB [18] transposases were used as positive controls, and the D268G catalytic mPB mutant generated earlier [22] was used as a negative control. All other expression plasmids used in these studies were generated by the Gibson assembly method and verified by Sanger sequencing; primers used for the assemblies are listed in Table S2. For the NTDR deletion mutants mPBdel and hyPBdel, amino acids 3–100 from the N-terminal part of the proteins are removed. For variants mPBvar14, hyPBvar14, mPBvar73, and hyPBvar73, the disordered sequences '14' (VNHPSHDLSSSETIVDADETDLQLDWKSDLELHDTCNVDEETISSQVNLKGEIDGVPLTQDSIGRQFPSASILSDVEGIQSEISLREESEDQSREASH) and '73' (SVATATLNIKDSDHGPVEDIQERARTDSCSSVELDEDDEQIELDHGVIDSSHLQLQLDWEGNLELPQTFSISKHSSEVSDTGQLRSSSNVEIIDEPES) were used to replace the amino acids 3–100 of the original transposases, respectively. These sequences were generated from the original sequence by randomizing the amino acids, resulting in the same amino acid composition and similar disorder characteristics. Coiled-coil containing aggregation-prone sequences were filtered out. Since the I30V mutation of hyPB is localized in the NTDR, for comparative studies, this position was mutated "back" to the original isoleucine present in mPB, thus generating the hyPB30I variant. For the HeLa reporter cell line (see Figure 4), a single genomic copy of the PB transposon disrupting a puromycin resistance gene cassette was stably inserted using the *Frog Prince* transposon system [58].

4.2. Cell Culturing and Transfection Methods

HEK-293 (human embryonic kidney), HeLa (human cervical cancer), and MCF-7 (human breast cancer) cell lines were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% of fetal calf serum, 1% of L-glutamine, and 1% of penicillin–streptomycin (Thermo Fisher Scientific, Waltham, MA, USA).

Transfections were carried out in duplicates. For the lipid-based transfections, 5×10^5 cells were seeded onto 6-well plates. The next day, 500 ng transposon donor and 500 ng transposase expressing helper plasmid were co-transfected to the cells using the FuGENE® 6 reagent, according to the manufacturer's instructions (Roche Applied Science, Mannheim, Germany). At 48 h post-transfection, cells were harvested for further experiments or analyses.

4.3. Transposition Assays

Transposition assays were carried out essentially as described previously [22]. Briefly, for excision assays, plasmids were isolated from the transfected cells using a modified protocol of the QIAGEN plasmid Miniprep Kit, then 10 ng of the isolated plasmids were used as templates in a two-round nested PCR assay to detect plasmid copies that underwent transposon excision and DNA repair; the ampicillin sequence presenting in all donor construct was used as an assay control. PCR products were separated in a 2% agarose gel and visualized by ethidium bromide staining using a Universal Hood Gel Imager Model # 75S (BioRad, Hercules, CA, USA). PCR primers for excision analysis are listed in Table S2.

For colony assays, starting at 48 h post-transfection, 1% of the transfected cells were seeded onto cell culture Petri dishes and selected for two weeks using puromycin (Sigma-Aldrich) in the final concentration of 1 μg/mL. Surviving cells were fixed in ice-cold methanol and stained with 0,04% Crystal Violet (Sigma-Aldrich, St. Louis, MO, USA) in 25% methanol. To test the effects on distinct steps of transpositions, we used a specific HeLa-derived reporter cell line (see above and Figure 4). These cells were treated with 100 mg/mL neomycin and 1 μg/mL puromycin simultaneously for double antibiotic selection. After the selection procedure, surviving colonies were quantified with the Universal Hood Gel Imager Model # 75S, using the Quantity One 4.4.0 software (BioRad). For statistical analyses, the mean and standard deviation values of at least three independent biological replicates were compared, performing F-tests and subsequent two-sided Student's *t*-tests; statistically significant differences were accepted at $p < 0.05$.

4.4. Analyzing piggyBac Sequences and Disorder Predictions

We created a nonredundant *piggyBac* protein dataset by an advanced search in the UniProtKB database (Release 10 February 2021) with the following keywords: protein name: 'piggybac' OR protein name: 'dde_tnp_1_7' OR gene name: 'pgbd'. The resulted dataset of 4277 proteins was narrowed down: first, proteins that couldn't be classified, based on protein or gene name, as type 1, 2, 3, 4, or 5 were removed. Next, the dataset was narrowed down to have a maximum of 1-1 protein per species and type, the longest form was kept from each *piggybac* type. Minor corrections were done based on the Pfam database (http://pfam.xfam.org/) (accessed on 13 May 2021 and 18 January 2022), and proteins without transposase domain were removed. The protein classification was also corrected based on the extra domain content if needed. *Piggybac* type 1 proteins often contain a SCAN domain, while many *piggybac* type 4 proteins contain a Zn-ribbon domain. Some known and important *piggybac* proteins without classification were added to the list: '*piggyBat*' (*Myotis lucifugus*), not in UniProt, translated from DNA; '*piggyMac*' (*Paramecium tetraurelia*), A0DFJ7; piggybac-like protein Tpb2p (*Tetrahymena thermophila*), D2Z1K6; Transposase mPB (*Trichoplusia ni*), Q283G1. Based on its Zn-ribbon domain, *piggyBat* was classified as a type 4 *piggybac* protein.

For control datasets, we used 2 transposase families: Tc1 mariner transposase family and *Tol2* transposases. We carried out an advanced search in UniProtKB with the following keyword: gene name: 'setmar' (SETMAR: SET domain and mariner transposase fusion protein homolog) to find Tc1 mariner transposons and narrowed down the dataset as we did in the case of *piggybac* proteins. Such a comprehensive investigation in the case of *Tol2* transposases could not be completed because not enough candidates were found in Uniprot. A known *Tol2* transposase is Q9PTV1.

Structural disorder of transposase proteins was predicted by the IUPred algorithm (https://iupred2a.elte.hu/) (accessed on 20 May 2021 and 25 January 2022) [46], which is based on estimating the total pairwise inter-residue interaction energy gained upon folding of a polypeptide chain. The predictor returns a position-specific disorder score in the range of 0.0–1.0, and an amino acid with a score at least 0.5 is considered locally disordered. The mean disorder was computed as the average of residue scores. To characterize the disorder tendency of the whole protein or different parts of the protein (domains, 100 amino acid long regions before the transposase domains, regions before and after the transposase domains), we calculated the ratio of disordered residues within the given region. Domain boundaries are derived from Pfam database (http://pfam.xfam.org/) (accessed on 13 May 2021 and 18 January 2022).

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms231810317/s1.

Author Contributions: Conceptualization, T.I.O. and Á.T.; methodology, G.W., K.H. and A.P.; software, É.S.; validation, G.W., Á.T., Z.I. and T.I.O.; formal analysis, G.W. and T.I.O.; investigation, G.W. and É.S.; resources, Á.T., Z.I. and T.I.O.; data curation, Á.T. and T.I.O.; writing—original draft preparation, T.I.O.; writing—review and editing, G.W., É.S., K.H., A.P., Z.I., Á.T. and T.I.O.; visualization, G.W., É.S. and T.I.O.; supervision, T.I.O.; project administration, Á.T. and T.I.O.; funding acquisition, Á.T. and T.I.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the SA-81/2021 grant from the Eötvös Loránd Research Network (given to Á.T. and T.I.O.) and grants 2018-1.2.1-NKP-2018-00005, VEKOP-2.1.1-15-2016-00156 and VEKOP-2.3.3-15-2017-00014 (T.I.O.) and K-125340 (Á.T.) from the National Research, Development and Innovation Fund of Hungary.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank Kornélia Némethy, Maria Pappa, and Maximilian Amberger for excellent technical help and Péter Bálint for the advice on statistical analysis.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Canapa, A.; Barucca, M.; Biscotti, M.A.; Forconi, M.; Olmo, E. Transposons, Genome Size, and Evolutionary Insights in Animals. *Cytogenet. Genome Res.* **2015**, *147*, 217–239. [CrossRef] [PubMed]
2. Wells, J.N.; Feschotte, C. A Field Guide to Eukaryotic Transposable Elements. *Annu. Rev. Genet.* **2020**, *54*, 539–561. [CrossRef]
3. Munoz-Lopez, M.; Garcia-Perez, J.L. DNA transposons: Nature and applications in genomics. *Curr. Genom.* **2010**, *11*, 115–128. [CrossRef] [PubMed]
4. Volff, J.N. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **2006**, *28*, 913–922. [CrossRef] [PubMed]
5. Jangam, D.; Feschotte, C.; Betran, E. Transposable Element Domestication as an Adaptation to Evolutionary Conflicts. *Trends Genet.* **2017**, *33*, 817–831. [CrossRef]
6. Burns, K.H. Our Conflict with Transposable Elements and Its Implications for Human Disease. *Annu. Rev. Pathol.* **2020**, *15*, 51–70. [CrossRef]
7. Babarinde, I.A.; Ma, G.; Li, Y.; Deng, B.; Luo, Z.; Liu, H.; Abdul, M.M.; Ward, C.; Chen, M.; Fu, X.; et al. Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res.* **2021**, *49*, 9132–9153. [CrossRef]
8. Judd, J.; Sanderson, H.; Feschotte, C. Evolution of mouse circadian enhancers from transposable elements. *Genome Biol.* **2021**, *22*, 193. [CrossRef]
9. Cosby, R.L.; Chang, N.C.; Feschotte, C. Host-transposon interactions: Conflict, cooperation, and cooption. *Genes Dev.* **2019**, *33*, 1098–1116. [CrossRef]
10. Cosby, R.L.; Judd, J.; Zhang, R.; Zhong, A.; Garry, N.; Pritham, E.J.; Feschotte, C. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **2021**, *371*, eabc6405. [CrossRef]
11. Rubin, G.M.; Spradling, A.C. Genetic transformation of Drosophila with transposable element vectors. *Science* **1982**, *218*, 348–353. [CrossRef] [PubMed]
12. Ding, S.; Wu, X.; Li, G.; Han, M.; Zhuang, Y.; Xu, T. Efficient transposition of the piggyBac (PB) transposon in mammalian cells and mice. *Cell* **2005**, *122*, 473–483. [CrossRef]
13. Cadinanos, J.; Bradley, A. Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res.* **2007**, *35*, e87. [CrossRef]
14. Ivics, Z.; Hackett, P.B.; Plasterk, R.H.; Izsvak, Z. Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell* **1997**, *91*, 501–510. [CrossRef]
15. Kawakami, K.; Shima, A. Identification of the Tol2 transposase of the medaka fish Oryzias latipes that catalyzes excision of a nonautonomous Tol2 element in zebrafish Danio rerio. *Gene* **1999**, *240*, 239–244. [CrossRef]
16. Kawakami, K.; Shima, A.; Kawakami, N. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 11403–11408. [CrossRef] [PubMed]
17. Mates, L.; Chuah, M.K.; Belay, E.; Jerchow, B.; Manoj, N.; Acosta-Sanchez, A.; Grzela, D.P.; Schmitt, A.; Becker, K.; Matrai, J.; et al. Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat. Genet.* **2009**, *41*, 753–761. [CrossRef]
18. Yusa, K.; Zhou, L.; Li, M.A.; Bradley, A.; Craig, N.L. A hyperactive piggyBac transposase for mammalian applications. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1531–1536. [CrossRef]
19. Hudecek, M.; Izsvak, Z.; Johnen, S.; Renner, M.; Thumann, G.; Ivics, Z. Going non-viral: The Sleeping Beauty transposon system breaks on through to the clinical side. *Crit. Rev. Biochem. Mol. Biol.* **2017**, *52*, 355–380. [CrossRef]
20. Henssen, A.G.; Henaff, E.; Jiang, E.; Eisenberg, A.R.; Carson, J.R.; Villasante, C.M.; Ray, M.; Still, E.; Burns, M.; Gandara, J.; et al. Genomic DNA transposition induced by human PGBD5. *Elife* **2015**, *4*, e10565. [CrossRef] [PubMed]
21. Beckermann, T.M.; Luo, W.; Wilson, C.M.; Veach, R.A.; Wilson, M.H. Cognate restriction of transposition by piggyBac-like proteins. *Nucleic Acids Res.* **2021**, *49*, 8135–8144. [CrossRef] [PubMed]
22. Kolacsek, O.; Wachtl, G.; Fothi, A.; Schamberger, A.; Sandor, S.; Pergel, E.; Varga, N.; Rasko, T.; Izsvak, Z.; Apati, A.; et al. Functional indications for transposase domestications—Characterization of the human piggyBac transposase derived (PGBD) activities. *Gene* **2022**, *834*, 146609. [CrossRef] [PubMed]
23. Yusa, K. piggyBac Transposon. *Microbiol. Spectr.* **2015**, *3*, MDNA3-0028. [CrossRef]
24. Tipanee, J.; VandenDriessche, T.; Chuah, M.K. Transposons: Moving Forward from Preclinical Studies to Clinical Trials. *Hum. Gene Ther.* **2017**, *28*, 1087–1104. [CrossRef] [PubMed]
25. Woodard, L.E.; Wilson, M.H. piggyBac-ing models and new therapeutic strategies. *Trends Biotechnol.* **2015**, *33*, 525–533. [CrossRef]
26. Kaji, K.; Norrby, K.; Paca, A.; Mileikovsky, M.; Mohseni, P.; Woltjen, K. Virus-free induction of pluripotency and subsequent excision of reprogramming factors. *Nature* **2009**, *458*, 771–775. [CrossRef] [PubMed]

27. Woltjen, K.; Michael, I.P.; Mohseni, P.; Desai, R.; Mileikovsky, M.; Hamalainen, R.; Cowling, R.; Wang, W.; Liu, P.; Gertsenstein, M.; et al. piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature* **2009**, *458*, 766–770. [CrossRef] [PubMed]
28. Sandoval-Villegas, N.; Nurieva, W.; Amberger, M.; Ivics, Z. Contemporary Transposon Tools: A Review and Guide through Mechanisms and Applications of Sleeping Beauty, piggyBac and Tol2 for Genome Engineering. *Int. J. Mol. Sci.* **2021**, *22*, 5084. [CrossRef] [PubMed]
29. Sarkar, A.; Sim, C.; Hong, Y.S.; Hogan, J.R.; Fraser, M.J.; Robertson, H.M.; Collins, F.H. Molecular evolutionary analysis of the widespread piggyBac transposon family and related "domesticated" sequences. *Mol. Genet. Genom.* **2003**, *270*, 173–180. [CrossRef] [PubMed]
30. Pagan, H.J.; Smith, J.D.; Hubley, R.M.; Ray, D.A. PiggyBac-ing on a primate genome: Novel elements, recent activity and horizontal transfer. *Genome Biol. Evol.* **2010**, *2*, 293–303. [CrossRef] [PubMed]
31. Bouallegue, M.; Rouault, J.D.; Hua-Van, A.; Makni, M.; Capy, P. Molecular Evolution of piggyBac Superfamily: From Selfishness to Domestication. *Genome Biol. Evol.* **2017**, *9*, 323–339. [CrossRef]
32. Mitra, R.; Li, X.; Kapusta, A.; Mayhew, D.; Mitra, R.D.; Feschotte, C.; Craig, N.L. Functional characterization of piggyBat from the bat Myotis lucifugus unveils an active mammalian DNA transposon. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 234–239. [CrossRef]
33. Cary, L.C.; Goebel, M.; Corsaro, B.G.; Wang, H.G.; Rosen, E.; Fraser, M.J. Transposon mutagenesis of baculoviruses: Analysis of Trichoplusia ni transposon IFP2 insertions within the FP-locus of nuclear polyhedrosis viruses. *Virology* **1989**, *172*, 156–169. [CrossRef]
34. Fraser, M.J.; Smith, G.E.; Summers, M.D. Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses. *J. Virol.* **1983**, *47*, 287–300. [CrossRef] [PubMed]
35. Mitra, R.; Fain-Thornton, J.; Craig, N.L. piggyBac can bypass DNA synthesis during cut and paste transposition. *EMBO J.* **2008**, *27*, 1097–1109. [CrossRef] [PubMed]
36. Keith, J.H.; Schaeper, C.A.; Fraser, T.S.; Fraser, M.J., Jr. Mutational analysis of highly conserved aspartate residues essential to the catalytic core of the piggyBac transposase. *BMC Mol. Biol.* **2008**, *9*, 73. [CrossRef] [PubMed]
37. Nesmelova, I.V.; Hackett, P.B. DDE transposases: Structural similarity and diversity. *Adv. Drug Deliv. Rev.* **2010**, *62*, 1187–1195. [CrossRef]
38. Hickman, A.B.; Chandler, M.; Dyda, F. Integrating prokaryotes and eukaryotes: DNA transposases in light of structure. *Crit. Rev. Biochem. Mol. Biol.* **2010**, *45*, 50–69. [CrossRef]
39. Chen, Q.; Luo, W.; Veach, R.A.; Hickman, A.B.; Wilson, M.H.; Dyda, F. Structural basis of seamless excision and specific targeting by piggyBac transposase. *Nat. Commun.* **2020**, *11*, 3446. [CrossRef]
40. Keith, J.H.; Fraser, T.S.; Fraser, M.J., Jr. Analysis of the piggyBac transposase reveals a functional nuclear targeting signal in the 94 c-terminal residues. *BMC Mol. Biol.* **2008**, *9*, 72. [CrossRef]
41. Morellet, N.; Li, X.; Wieninger, S.A.; Taylor, J.L.; Bischerour, J.; Moriau, S.; Lescop, E.; Bardiaux, B.; Mathy, N.; Assrir, N.; et al. Sequence-specific DNA binding activity of the cross-brace zinc finger motif of the piggyBac transposase. *Nucleic Acids Res.* **2018**, *46*, 2660–2677. [CrossRef] [PubMed]
42. Helou, L.; Beauclair, L.; Dardente, H.; Arensburger, P.; Buisine, N.; Jaszczyszyn, Y.; Guillou, F.; Lecomte, T.; Kentsis, A.; Bigot, Y. The C-terminal Domain of piggyBac Transposase Is Not Required for DNA Transposition. *J. Mol. Biol.* **2021**, *433*, 166805. [CrossRef] [PubMed]
43. Guerineau, M.; Bessa, L.; Moriau, S.; Lescop, E.; Bontems, F.; Mathy, N.; Guittet, E.; Bischerour, J.; Betermier, M.; Morellet, N. The unusual structure of the PiggyMac cysteine-rich domain reveals zinc finger diversity in PiggyBac-related transposases. *Mob. DNA* **2021**, *12*, 12. [CrossRef] [PubMed]
44. Tompa, P.; Schad, E.; Tantos, A.; Kalmar, L. Intrinsically disordered proteins: Emerging interaction specialists. *Curr. Opin. Struct. Biol.* **2015**, *35*, 49–59. [CrossRef]
45. Meszaros, B.; Erdos, G.; Dosztanyi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **2018**, *46*, W329–W337. [CrossRef]
46. Erdos, G.; Dosztanyi, Z. Analyzing Protein Disorder with IUPred2A. *Curr. Protoc. Bioinf.* **2020**, *70*, e99. [CrossRef] [PubMed]
47. Raskó, T.; Pande, A.; Radscheit, K.; Zink, A.; Singh, M.; Sommer, C.; Wachtl, G.; Kolacsek, O.; Inak, G.; Szvetnik, A.; et al. A novel gene controls a new structure: PiggyBac Transposable Element-derived 1, unique to mammals, controls mammal-specific neuronal paraspeckles. *Mol Biol Evol.* **2022**, msac175. [CrossRef]
48. Meszaros, B.; Simon, I.; Dosztanyi, Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **2009**, *5*, e1000376. [CrossRef] [PubMed]
49. Kolacsek, O.; Erdei, Z.; Apati, A.; Sandor, S.; Izsvak, Z.; Ivics, Z.; Sarkadi, B.; Orban, T.I. Excision efficiency is not strongly coupled to transgenic rate: Cell type-dependent transposition efficiency of sleeping beauty and piggyBac DNA transposons. *Hum. Gene Ther. Methods* **2014**, *25*, 241–252. [CrossRef] [PubMed]
50. Li, X.; Burnight, E.R.; Cooney, A.L.; Malani, N.; Brady, T.; Sander, J.D.; Staber, J.; Wheelan, S.J.; Joung, J.K.; McCray, P.B., Jr.; et al. piggyBac transposase tools for genome engineering. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E2279–E2287. [CrossRef]

51. Li, M.A.; Pettitt, S.J.; Eckert, S.; Ning, Z.; Rice, S.; Cadinanos, J.; Yusa, K.; Conte, N.; Bradley, A. The piggyBac transposon displays local and distant reintegration preferences and can cause mutations at noncanonical integration sites. *Mol. Cell Biol.* **2013**, *33*, 1317–1330. [CrossRef] [PubMed]
52. Walisko, O.; Izsvak, Z.; Szabo, K.; Kaufman, C.D.; Herold, S.; Ivics, Z. Sleeping Beauty transposase modulates cell-cycle progression through interaction with Miz-1. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 4062–4067. [CrossRef] [PubMed]
53. Sigalov, A.B.; Zhuravleva, A.V.; Orekhov, V.Y. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie* **2007**, *89*, 419–421. [CrossRef] [PubMed]
54. Goyal, S.; Gupta, G.; Qin, H.; Upadya, M.H.; Tan, Y.J.; Chow, V.T.; Song, J. VAPC, an human endogenous inhibitor for hepatitis C virus (HCV) infection, is intrinsically unstructured but forms a "fuzzy complex" with HCV NS5B. *PLoS ONE* **2012**, *7*, e40341. [CrossRef] [PubMed]
55. Tantos, A.; Szabo, B.; Lang, A.; Varga, Z.; Tsylonok, M.; Bokor, M.; Verebelyi, T.; Kamasa, P.; Tompa, K.; Perczel, A.; et al. Multiple fuzzy interactions in the moonlighting function of thymosin-beta4. *Intrinsically Disord. Proteins* **2013**, *1*, e26204. [CrossRef] [PubMed]
56. Haarmann, C.S.; Green, D.; Casarotto, M.G.; Laver, D.R.; Dulhunty, A.F. The random-coil 'C' fragment of the dihydropyridine receptor II-III loop can activate or inhibit native skeletal ryanodine receptors. *Biochem. J.* **2003**, *372*, 305–316. [CrossRef] [PubMed]
57. Schad, E.; Kalmar, L.; Tompa, P. Exon-phase symmetry and intrinsic structural disorder promote modular evolution in the human genome. *Nucleic Acids Res.* **2013**, *41*, 4409–4422. [CrossRef] [PubMed]
58. Miskey, C.; Izsvak, Z.; Plasterk, R.H.; Ivics, Z. The Frog Prince: A reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic Acids Res.* **2003**, *31*, 6873–6881. [CrossRef] [PubMed]

Biologia Futura (2023) 74:17–28
https://doi.org/10.1007/s42977-023-00154-7

REVIEW

One locus, several functional RNAs—emerging roles of the mechanisms responsible for the sequence variability of microRNAs

Tamás I. Orbán[1]

Received: 24 November 2022 / Accepted: 8 February 2023 / Published online: 27 February 2023

Abstract
With the development of modern molecular genetics, the original "one gene-one enzyme" hypothesis has been outdated. For protein coding genes, the discovery of alternative splicing and RNA editing provided the biochemical background for the RNA repertoire of a single locus, which also serves as an important pillar for the enormous protein variability of the genomes. Non-protein coding RNA genes were also revealed to produce several RNA species with distinct functions. The loci of microRNAs (miRNAs), encoding for small endogenous regulatory RNAs, were also found to produce a population of small RNAs, rather than a single defined product. This review aims to present the mechanisms contributing to the astonishing variability of miRNAs revealed by the new sequencing technologies. One important source is the careful balance of arm selection, producing sequentially different 5p- or 3p-miRNAs from the same pre-miRNA, thereby broadening the number of regulated target RNAs and the phenotypic response. In addition, the formation of 5', 3' and polymorphic isomiRs, with variable end and internal sequences also leads to a higher number of targeted sequences, and increases the regulatory output. These miRNA maturation processes, together with other known mechanisms such as RNA editing, further increase the potential outcome of this small RNA pathway. By discussing the subtle mechanisms behind the sequence diversity of miRNAs, this review intends to reveal this engaging aspect of the inherited "RNA world", how it contributes to the almost infinite molecular variability among living organisms, and how this variability can be exploited to treat human diseases.

Keywords RNA interference · RNAi · MicroRNA · Arm selection · isomiR

Introduction

One of the fundamental ideas that revolutionized biological science and was also considered to lay down the bases of molecular biology was the famous "one gene—one enzyme" hypothesis formulated by George Beadle and Edward Tatum in the beginning of the 1940s (Beadle and Tatum 1941). They proposed that enzymes carry out the metabolic functions in the cells, and all enzymes are determined by unique DNA segments called genes. With the advancement of biochemistry and genetics, and after the formulation of the "central dogma" of molecular biology (Crick 1958), the hypothesis was refined as the "one gene—one polypeptide" statement; yet, it still turned out to be a simplified version of genetic complexity. Soon it was revealed that at the RNA level, the "message" is often a combinatorial output of the DNA sequence of a particular gene. The discovery of splicing and alternative splicing provided the first evidence that several mRNA isoforms can be generated from a single gene, potentially coding for several protein species (Berget et al. 1977; Chow et al. 1977; Baralle and Giudice 2017; Shenasa and Hertel 2019). Another important discovery was the phenomenon of RNA editing, showing that the nucleotide sequence of the transcribed mRNA can be functionally modified, which also results in the alteration of the encoded polypeptide sequence (Benne et al. 1986; Powell et al. 1987). Since these original observations, follow-up investigations revealed several other RNA modification pathways, further supporting the view that gene activity indeed produces a complex pool of transcripts, rather than a single RNA with

✉ Tamás I. Orbán
orban.tamas@ttk.hu; orbant@biomembrane.hu

[1] Institute of Enzymology, Research Centre for Natural Sciences, Eötvös Loránd Research Network, Magyar Tudósok Körútja 2, Budapest 1117, Hungary

a single determined function (reviewed in (Li and Mason 2014)).

The next important breakthrough in the RNA field was the discovery of small regulatory RNAs, later identified as microRNAs (miRNAs) (Lee et al. 1993; Wightman et al. 1993). Since their initial description, miRNAs were found to be important posttranscriptional molecular modulators, and apart from fungi, they represent an intracellular regulatory network of the RNA interference (RNAi) machinery in all other organisms examined so far (Lee et al. 2004; Ameres and Zamore 2013). Considering their transcription and maturation, the original assumption again was that one miRNA gene produces a single small RNA product of a discrete length. However, especially with the unforeseen developmental pace of next generation sequencing (NGS) technologies, more and more experimental data suggested that this is not the case and the miRNA repertoire is far more complex than previously thought. Once again in the history of RNA biochemistry, the "one gene—one RNA product" hypothesis needed to be revised by the data showing that the activity of miRNA loci produce numerous small RNAs, often a large population of isomiRs (Gebert and MacRae 2019; Trabucchi and Mategot 2019; Bofill-De Ros et al. 2020). In this review I attempt to present the molecular mechanisms behind this complexity: I will address the flexibility of the once underestimated pre-miRNA arm selection regulation, as well as the mechanisms resulting in 5' and 3' heterogeneity of mature miRNAs, producing numerous 5' and 3' isomiRs from a single miRNA gene. I will also discuss the functional consequences and the medical relevance of these carefully regulated miRNA maturation processes, and how the better understanding of these could contribute to the solution of current public health problems. The aim is to give an appreciative view on the unprecedented sequence diversity of small RNA species produced from a single miRNA gene, and through that, to demonstrate the complexity that is present in the existing "RNA world" of today.

The life of miRNAs—maturation and a "wobbling" mechanism of target recognition

miRNAs are the key actors in one of the RNA interference pathways, and have been evolved in parallel with the short interfering RNAs (siRNAs) and the Piwi-associated small RNAs (piRNAs) (Ghildiyal and Zamore 2009). The latter two represent two pathways that have still kept the common ancient function of protecting the organisms from invasive genetic elements such as viruses and transposons. siRNAs are the main defense guards against outside invaders, namely viruses, and represent a "cellular immune system" which is the major security line in organisms that lack an adaptive immune system (such as lower invertebrates or most plants) (Rosendo Machado et al. 2021; Kong et al. 2022). piRNAs, however, are found exclusively in the animal kingdom, and control the "resident aliens" of the genomes, eliminating endogenous viruses and transposons, especially in the germline (Onishi et al. 2021). Although recent studies revealed various functions of piRNAs in somatic cells, they are still predominantly considered as guardians of the germline, protecting stem cells and especially the gametes (Peng and Lin 2013; Wang et al. 2022). miRNAs, on the other hand, have evolved for a different functionality: they represent an elaborate network of posttranscriptional regulators, controlling endogenous mRNAs and thereby fine-tuning cellular gene expression patterns (Gebert and MacRae 2019). They are present in most eukaryotic groups of the tree of life, having a common origin in plants and animals, and current molecular evidence suggest that the loss of this pathway in some species (such as in fungi) rather represents a secondary loss during evolution (Moran et al. 2017).

microRNAs are short, ~ 20–24 nucleotide long, single-stranded RNA molecules encoded by the genome, and they may be relatively numerous in a given organism: in the human genome, for instance, approximately 2000 miRNAs have already been described, although some still require further validation (http://mirbase.org/). miRNAs can be regulated by their own promoter or they can be located in introns (or even in exons) of other, predominantly protein coding genes (Bartel 2018). Interestingly, in mammalian genomes they often form clusters where several miRNAs can be transcribed simultaneously (Altuvia et al. 2005; Kim and Kim 2007; Chan et al. 2012; Ramalingam et al. 2014; Kabekkodu et al. 2018). After transcription of the primary-miRNAs (pri-miRNA), their maturation requires two subsequent endonucleolytic cleavage steps (Fig. 1A). In animal cells, the first one occurs in the nucleus by the "Microprocessor" complex, containing the RNase III type enzyme Drosha, and its partner protein DGCR8 ('DiGeorge Syndrome Critical Region 8'). It generates a precursor-miRNA (pre-miRNA) with a hairpin-like structure that is transported to the cytoplasm where the second cleavage takes place by another endonuclease called Dicer. The so formed, usually mismatch containing short double-stranded RNAs are incorporated into an Argonaute (Ago) protein containing, RNA-induced silencing complex (RISC), where further processing eliminates one strand (the "passenger" strand), resulting in a single stranded miRNA-containing functional RISC (Iwakawa and Tomari 2022). The effector function of these ribonucleoprotein complexes is then elicited by the Watson–Crick base pairing complementarity between the miRNA and the target RNA, initiating RNA decay or translation inhibition (Huntzinger and Izaurralde 2011). However, in contrast to siRNAs and piRNAs, often only a small portion of the miRNA has complementarity to the target RNA. In most cases, the 2–8

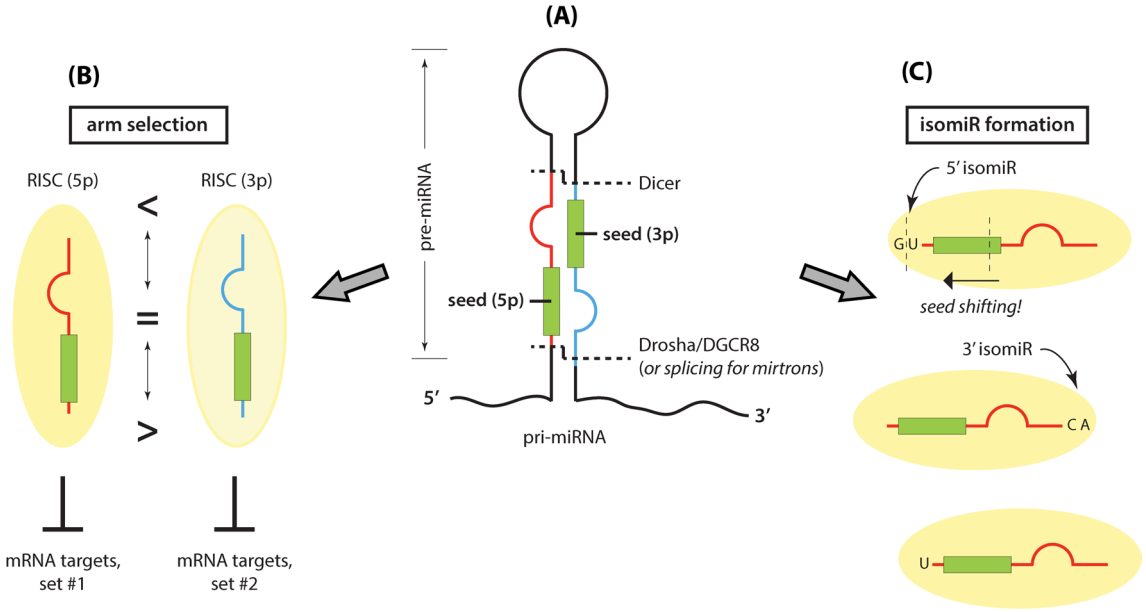


Fig. 1 MicroRNA maturation processes **A** Structure and processing of the pre-miRNA formed on the primary transcript (pri-miRNA). The staggered lines indicate the two consecutive endonucleolytic cleavages: the first is performed by the Drosha/DGCR8 (Microprocessor) complex in the nucleus or substituted by the spliceosome for the Drosha-independent mirtrons. The second cleavage occurs in the cytoplasm by Dicer, the other RNase III type enzyme involved in the maturation. Red and blue lines indicate the 5p or 3p arms, green rectangles depicting the seed sequences. **B** If they are functionally selected, the different arms are incorporated into distinct RNA-induced Silencing Complexes (RISCs). The ratio of such complexes may vary in cell types, or in different developmental or physiological states, and they target different sets of RNA molecules, although individual targets may overlap. **C** RISCs formed from a given miRNA locus may differ in their isomiR contents. As examples, distinct 5' and 3' isomiR-containing RISCs are shown above the one with the "canonical" or "reference" miRNA-containing complex. Note that for 5' isomiRs, a difference in target selection may occur due to "seed shifting" (depicted by an arrow)

nucleotides of the 5' sequence, the so-called "seed region" provides the base-pairing interactions but several validated examples have been shown where varying portions from the 3' end of the miRNA also contribute to target recognition which makes the mere bioinformatic target prediction a very difficult task (Berezikov 2011; Pasquinelli 2012). The lack of a long and strict sequence constraint also explains why one miRNA can often bind to several target mRNA molecules (in animals, usually in the 3' untranslated regions, 3' UTRs), and why a given mRNA may contain multiple miRNA-binding sites. The result of this is that similarly to the regulation by transcription factors, miRNAs and their targets represent a complex network of interactions, resulting in a post-transcriptional fine-tuning of gene expression (Arora et al. 2013; Gebert and MacRae 2019; Trabucchi and Mategot 2019).

For many years, it was believed that miRNAs are produced exclusively by the above outlined, well-defined "canonical" biogenesis pathway, with occasional differences only in certain organisms. However, later scientific data challenged this paradigm by the discovery of the Drosha- and Dicer-independent alternative maturation pathways (Winter et al. 2009; Miyoshi et al. 2010). The most prominent of these is certainly the mirtron pathway, with the spliceosome substituting for the Drosha/DGCR8 complex (Fig. 1A) (Berezikov et al. 2007; Okamura et al. 2007; Ruby et al. 2007). Although it was first considered to be a speciality of invertebrate genomes with a short average length of introns, later studies confirmed the existence of mirtrons also in higher organisms, including mammals (Havens et al. 2012; Schamberger et al. 2012; Sibley et al. 2012). The emergence of alternative pathways increased the potential sources of miRNA precursors; however, it soon turned out that the diversity of miRNA sequences is further increased by previously underestimated mechanisms acting not only on the precursors but also on the mature miRNA species.

miRNA arm selection—the "guide" or the "passenger"?

After cleaving the terminal loop of pre-miRNA by Dicer, an important step is the selection of the appropriate strand (referred to as "arm") of the short double-stranded RNA which will be acting as a guide for the Ago protein in the RISC (Fig. 1A, B). A similar process occurs during siRNA generation (also catalyzed by Dicer) therefore in early studies, the selection mechanism was investigated in parallel in

the two RNAi pathways. Analyzing a high number of siRNA or miRNA containing complexes, a clear bias between the strands was found and was shown to be associated with the thermodynamic properties of the arm sequences. It was revealed that the strand having the less stable 5' end is favorable over the other, and it is selected for incorporating into the Ago-complex by unknown mechanisms (Khvorova et al. 2003; Schwarz et al. 2003). This "thermodynamic asymmetry rule" seemed to explain the formation of several mature miRNAs, and the selected arm was named as the "guide" strand, as opposed to the other "passenger" strand (or miRNA*) which was considered to be functionless and therefore destined to be degraded (Hutvagner 2005). Although this rule was clearly helpful in designing effective siRNAs or shRNAs for genetic knock-down experiments, large-scale sequencing data already indicated the functional presence of miRNA* species for some miRNA loci (Ruby et al. 2006), including their presence in human embryonic stem cells (Morin et al. 2008). Meanwhile, a systematic study on miRNA expression profiles from several species indicated that the thermodynamic selection rule is not universal and for various loci, both arms can be expressed simultaneously. Moreover, if expression abundance is considered for defining the dominant arm, than a tissue-specific total switch from miRNA to miRNA* was also revealed for a number miRNA genes (such as for mouse mmu-miR-30e-5p and -3p arms) (Ro et al. 2007). Analyzing NGS data, another group also identified an "outlier" from the rule: in some cancer samples, they detected an almost 1000-fold expression difference of the hsa-miR-223* species, and they proposed that this pattern is likely to be used as a biomarker for leukemia (Kuchenbauer et al. 2008). Nevertheless, the thermodynamic-based selection rule seemed to be working for most examined miRNA loci, still supporting the "one miRNA gene—one mature small RNA product" hypothesis.

But why would one consider the importance of the passenger strand at all? The first answer would be that even if the two strands in the pre-miRNA were fully complementary to each other, they would still be sequentially different: selecting the miRNA* would open up the possibility of targeting other mRNA species by the RISC, and thereby increasing the regulatory repertoire of a given miRNA locus. On the other hand, as shown in Fig. 1A, the two arms of the pre-miRNA stem-loop structure are very often not fully complementary and contain one or more bulges with varying length. This structural feature, however, provides an excellent platform for distinct sequence selection and evidence were soon given for the evolutionary independence of the two arms. In David Bartel's laboratory, when investigating the origin and the evolution of miRNAs, they found that both arms of the aqu-miR-2015 in sponges are expressed in a cell type-dependent manner, and this regulated arm switching was found to be important not only in sponge development but also in the evolution of sponge species (Grimson et al. 2008). In a parallel study, in Eric Lai's laboratory they revealed the high conservation and therefore the evolutionary importance of several miRNA* species, and their connection to the evolution of 3' UTR sequences of the targeted protein coding genes (Okamura et al. 2008). Several subsequent studies showed that the guide and the passenger arms show tissue-specific expression patterns in various vertebrate species, including humans, and these all indicated the active regulation of arm selection, despite the thermodynamic stability differences between the two strands (Cloonan et al. 2011; Yang et al. 2011; Marco et al. 2012; Schamberger et al. 2012; Zhou et al. 2012).

The widespread occurrence of arm switching was followed by studies where the phenomenon was connected to physiological regulations and to medical importance. An early example was the work on the expression from the hsa-miR-9/hsa-miR-9* locus in neuronal development: both strands were detected in the samples and the misregulation of their expression profiles was not only associated with Huntington disease but they were also found to target mRNAs from the same biochemical pathway, even sharing common targets (Packer et al. 2008). This was followed by various examples where the miRNA guide/passenger ratio was shown to influence pathways of medical relevance in human diseases, such as the association of the miRNA arm usage pattern with the prognosis and the therapy of macroglobulinemia (Roccaro et al. 2009). There were particularly numerous examples in different cancer types (Zhang et al. 2013; Kuo et al. 2015; Mitra et al. 2015a, 2015b), indicating that the functional expression of the miRNA* strand is much more frequent than previously thought (Table 1).

But if not only the thermodynamic rule then what else regulates pre-miRNA arm selection? Can those other mechanisms override the stability-based assortment? The first indication for a distinct and elaborate mechanism was revealed when the loading of miRNA duplexes into different Ago proteins was studied in Drosophila. In that model organism, two proteins, Ago1 and Ago2 can incorporate these double-stranded small RNA species: it was shown that Ago2 is sensitive to central bulges which can take the priority over thermodynamics and leads to the active selection of the formerly considered passenger strands (Okamura et al. 2009). Similarly in human cells, where four Agos compete for the targets, Ago3 preferentially selects for the passenger strands of certain miRNA duplexes, such as the let-7a-3p arm. It was also shown that hemin can selectively induce Ago3 expression in the K562 cell line, a widely used cellular model of myeloid leukemia, so the selective incorporation of let-7a-3p into this Argonaute protein may have relevance in that particular cancer type (Winter and Diederichs 2013). In line with these observations, an earlier study revealed that the human Ago2 has a clear bias for the first 5'-nucleotide

Table 1 Selected examples on the functional role of miRNA* species in cancers

miRNA* species (arm)	Main target genes	Cancer type	Reference
miR-146a* (-3p)	NTRK2 proto-oncogene	Thyroid cancer	Jazdzewski et al. (2009)
miR-199a* (-3p)	MET proto-oncogene; ERK2	Several primary and metastatic	Kim et al. (2008)
miR-126* (-5p)	Prostein	Prostate cancer	Musiyenko et al. (2008)
miR-18a* (-3p)	K-Ras proto-oncogene	Skin and colon cancer	Tsang and Kwok (2009)
miR-29c* (-5p)	DNMT1, DNMT3A	Mesothelioma	Pass et al. (2010)
miR-136* (-3p)	TRIAP1[a]		
miR-361* (-3p)	RASL10 proto-oncogene	Gastric cancer	Li et al. (2012)
miR-423* (-3p)	RAP2C proto-oncogene		

[a]*TP53 regulated inhibitor of apoptosis 1*

identity, selecting mature miRNAs with either uracil (U) or adenine (A) base in that position (Frank et al. 2010). In a systematic follow-up study, this U and A preference was also shown for all human Agos and the authors proposed that the miRNA arm selection asymmetry is determined by a combination of an "analog pattern" (the thermodynamic instability preference) and a "digital code" (the U or A preference at the 5' end) (Suzuki et al. 2015). Nevertheless, this model still does not give a reassuring answer for all experimental results, nor does it explain the difference in Ago3 selectivity described above (Nakanishi 2022).

In parallel with investigating the roles of Agos in arm selection, various studies revealed that asymmetry "sensors" already influence the process before the double-stranded small RNAs are loaded into the pre-RISC. In Drosophila, Dicer-2 and its binding partner, R2D2 are responsible not only for loading the siRNA duplexes into the Ago2-containing complex but also influencing strand selection; however, it was also shown that human Dicer is dispensable for the asymmetry (Betancur and Tomari 2012). Quite contrary, in Jennifer Doudna's lab, they provided evidence that human Dicer, in complexed with either the TRBP or the PACT protein, strongly influence miRNA arm selection of Agos (Noland and Doudna 2013), which aspect was further supported by the crystal structure of the Dicer-TRBP complex (Wilson et al. 2015). In addition, a recent study revealed that 3' uridylation of the pre-miRNA could also actively influence arm selection: in case of the pre-miR-324, it causes an alternative (re-positioned) Dicer cleavage which results in a switch from 5p to 3p usage in miRNA maturation (Kim et al. 2020). On the other hand, other mechanisms also exist which can ensure the strict usage of one miRNA arm. One prominent example is the Dicer-independent alternative maturation pathway: in such cases, illustrated by miR-451, the short stem length of the pre-miRNA excludes Dicer binding, and the Ago2 cleavage of the stem-loop structure already determines the exclusive presence of the 5p arm in the RISC-loading complex (Cheloufi et al. 2010; Cifuentes et al. 2010). Interestingly, this alternative pathway is also utilized for shRNA vector design to eliminate off-target effects caused by the expression of the passenger strand (Herrera-Carrillo et al. 2015).

Thorough investigations in recent years eventually changed our original model of pre-miRNA maturation and the exclusive arm usage determined by the thermodynamic stability differences between the two arms. From sequence constraints to the discriminating action of regulatory proteins, several mechanisms were uncovered which can influence miRNA strand selection and although the molecular details are not always completely understood, it is clear that this process is actively regulated in a cell-type dependent manner, presumably by at least the cell-type specific expression level of the identified regulatory factors. This shift in the model is also noticeable in the nomenclature (https://mirbase.org/), as we are not talking about guide and passenger (*) miRNAs any more but rather 5p or 3p species matured from a given miRNA locus. Although understanding this aspect already challenged the 'one locus—one miRNA' belief, the real paradigm shift came with the discovery of miRNA isomiRs.

Formation of isomiRs—substantially increasing the miRNA repertoire

After the discoveries that miRNAs are wide-spread posttranscriptional regulators in most eukaryotic genomes, scientist started annotating and placing miRNA gene sequences into databases (such as "mirbase", https://mirbase.org/), analogously to protein coding genes with their mature mRNA sequences. However, these annotation very often ran into discrepancies concerning the mature miRNA end sequences, detecting not only alternative ends and internal sequences but also non-templated 5' or 3' extensions, the latter ones being end nucleotides different from those present in the flanking genomic sequences (Azuma-Mukai et al. 2008). When several studies reported such miRNA sequence variations in various plant and animal species, the idea of these

representing merely sequencing errors were becoming less and less favorable (Ebhardt et al. 2009; Wu et al. 2009; Lee et al. 2010; Zhang et al. 2010; Wyman et al. 2011). It was even revealed that the tissue-specific distributions of these miRNA variations are significantly different, again arguing for the functional presence of these species, which were subsequently referred to as "isomiRs" (Cloonan et al. 2011; Li et al. 2011; Zhou et al. 2012). Their presence was first connected to the miRNA arm switching or the RNA editing phenomena but they were soon shown to represent a distinct miRNA modifying pathway (Ebhardt et al. 2009; Cloonan et al. 2011; Zhou et al. 2012).

Following many years of uncertainty, it finally became accepted that isomiRs exist and are functional, and they were eventually considered as "the hitherto overlooked repertoire of the microRNAome" which potentially influence target selection, miRNA stability, or affect miRNA loading into the RISCs (Burroughs et al. 2011; Neilsen et al. 2012). This again lead to a "paradigm shift", since from that on, the "reference miRNA" sequence became very hard to define, and a miRNA gene could not be considered as a locus producing a single RNA product with a defined length and sequence, but rather a locus producing a population of related small RNAs with considerable sequence variability (Fig. 1C). According to a recent classification, at least the following categories of isomiRs can be distinguished beyond the canonical (the "reference") miRNA species: (i) 5' isomiRs, having a different length at the 5' end; (ii) 3' isomiRs, having a different length at the 3' end; (iii) polymorphic isomiRs, with the same length but having internal sequence variations; and (iv) "mixed type" isomiRs, containing differences in any combinations of the first 3 types (Tomasello et al. 2021). However, several questions and problems arose, including the mechanisms leading to isomiR formation, their functional versatility, as well as the problem of how to reliably detect and measure the level of a single isomiR.

Concerning their production, templated and non-templated isomiRs should be discussed separately. Templated 5' and 3' isomiRs can be considered as alternative or imprecise cleavage products from the precursor transcripts, sometimes referred to as Drosha or Dicer "promiscuity" (Wu et al. 2009; Burroughs et al. 2011). As discussed above for miRNA arm selection, theses enzymes are sensitive to sequential and structural differences, and especially for Dicer, structural motifs in the precursor hairpins influence cleavage site selection and therefore the length diversity of the small dsRNA products (Starega-Roslan et al. 2011, 2015a, 2015b). For miR-203, for example, it was found that a "sliding-bulge structure" can cause the formation of two differently folded pre-miRNA structures which are then processed differently by Dicer (Ma et al. 2016). In addition, as TRBP and PACT proteins are important regulatory partners of Dicer, they were also shown to influence cleavage site selection, thereby the production of isomiRs (Lee et al. 2013; Wilson et al. 2015). Finally, it was shown that miRNA arm selection and isomiR production are interconnected: systematically analyzing NGS data it was revealed that the more stable the 5p or the 3p arm of a miRNA is, the more stable its corresponding isomiR profile is (Guo and Chen 2014). It is overall not surprising, as similar sequence motifs and the same processing enzymes are responsible for both miRNA maturation processes.

Analyzing the constantly growing volume of NGS datasets, a significant amount of non-templated isomiRs were also discovered. The majority of them represent 3' uridylated or adenylated species of various lengths which were connected to miRNA stability processes and decay intermediates: it was suggested that 3' uridylation marks miRNAs for degradation, whereas 3' adenylation is a sign for stability (Scheer et al. 2016). On the other hand, there were indications that 3' mono- and oligo-uridylation of pre-miRNAs could play opposing roles in the biogenesis of certain miRNAs (Heo et al. 2012), whereas another study revealed that terminal uridylation of mature miRNAs is physiologically significant and it is especially pervasive in the neonatal period (Jones et al. 2012). But this raised an important question: can an isomiR with a seemingly *genomic* 3' end sequence be in fact a non-templated isomiR? In principle yes, it could be a mono-uridylated miRNA, even if by coincidence the Dicer-cleaved mature miRNA 3' end is followed by a U nucleotide in the genomic sequence. For certain miRNA groups, such as mirtrons, these cases may be easier to test: the 5' and the 3' ends of a mirtronic pre-miRNA is determined the stringency of splicing, therefore they represent better candidates for studying posttranscriptional modifications, such as non-templated 3' uridylation (Westholm et al. 2012). However, for general miRNA molecules, these cases are both technically and biologically challenging to verify. And this also leads to a very important aspect of isomiR investigations: how can functional isomiRs be reliably identified?

The first problem is the methodology: although carefully optimized NGS methods can detect all isomiRs in a sample (Smith and Hutvagner 2022), this approach is rather expensive and cumbersome when a single locus is to be studied. Traditional RNA detection methods may be applied—but Northern blots, although detecting all isomiRs expressed above a certain threshold, cannot be used to identify new species. The more sensitive real-time qPCR methodologies can increase detectability of single known miRNA variants but the canonical assays cannot reliably distinguish among closely related isomiR sequences (Lee et al. 2010; Schamberger and Orban 2014). However, newly developed and optimized qPCR platforms could provide the desired selectivity in isomiR detection but they still require preliminary NGS data to start with, followed by careful adjustments

when a new miRNA locus is to be investigated (Honda and Kirino 2015; Franco et al. 2022).

With the technological improvements, the detection problem seems to be settled. Nevertheless, the real biological problem still remains: can the functional role of a detected isomiR be verified? How can we distinguish miRNA degradation products from *bona fide* isomiRs? As a hint, there are examples where global changes in the NGS-determined isomiR profiles are associated with a certain phenotype or a disease. As examples, these include the drastic change in the mosquito vector isomiR distribution after Dengue virus infection (Etebari et al. 2015), the dysregulation of 5'-isomiR levels in connection to the Alzheimer disease (Wang et al. 2016), or the global changes in isomiR profiles in certain cancer types (Li et al. 2012; Kuo et al. 2015; Saiselet et al. 2015). However, these results cannot be attributed to single miRNA variants and the real challenge is to assign distinct function(s) to different individual isomiRs. For animal miRNAs, where target selection is strongly based on the seed sequence, the shift at the 5' terminus can clearly influence silencing function (Fig. 1C); however, the potential effect of changes in the length or the sequence of the 3' end could be more problematic to validate. If a new population of isomiRs is to be examined, only a rigorous pipeline of analysis could exclude potential decay intermediates and provide evidence for functional importance of the detected miRNA variants. These include investigating at least (i) the tissue-specific expression patterns, (ii) the association with Ago proteins, (iii) in vitro functional activities (in luciferase assays or on predicted targets), and (iv) the potential evolutionary conservations, although this latter one may exclude species-specific isomiRs (Tan et al. 2014). Despite such methodological challenges, many examples are already known where the concrete functions of selected isomiRs have been proven, which provides evidence that this process of miRNA regulation is actively regulated and physiologically important (Table 2).

These examples underline the relevance isomiR formation and distribution in human physiology and diseases, and the most prominent examples come from cancer biology. An earlier attempt to classify 32 cancer types using isomiR profiling was surprisingly successful: even if a "binary" (present or absent) approach was applied, this method could label different tumor types with high (90%) sensitivity and with a very low (3%) false discovery rate (Telonis et al. 2017). A recent meta-analysis of available cancer datasets and publications not only supported this approach but could identify specific miRNA isomiRs that can be used as biomarkers for certain tumor types (Zelli et al. 2021). A further deep analysis of isomiR datasets could differentiate even among breast cancer subtypes (Nersisyan et al. 2022), whereas a recent study provided evidence that the heterogeneous nuclear ribonucleoprotein C (hnRNPC) induced 5' isomiR shift of miR-21-5p could lead to liver cancer (Park et al. 2022). All these studies show that isomiR-based tumor profiling seems to be a very encouraging approach for cancer diagnostics but isomiR-based tumor targeting also holds promise for future therapy applications.

Nowadays, after many years of intensive research, it is widely accepted that miRNA isomiRs represent an actively regulated and physiologically important layer of gene regulation. The mechanisms behind isomiR formation have, in general, been revealed: beside the subtle regulation of the main miRNA processing enzymes (Drosha and Dicer), the action of terminal nucleotidyl transferases, exoribonucleases and RNA editing enzymes are also responsible for the existence of certain isomiR groups (Tomasello et al. 2021). On the other hand, the regulation of cell-type specific isomiR distribution is still not completely understood and this aspect is currently in the focus of intensive investigations (Zelli et al. 2021; Aparicio-Puerta et al. 2023). Nevertheless, isomiRs should definitely be considered in miRNA target predictions (Ahmed et al. 2014), as well as in the development of new diagnostic tools and in planned therapy applications (Nikolova et al. 2021; Scheper et al. 2022).

Table 2 Selected examples of miRNA isomiRs with experimentally verified functions

miRNA locus	IsomiR type	Function(s)	Reference
hsa-miR-101-3p	5' isomiRs	Selective and age specific silencing in brain tissues	Llorens et al. (2013)
hsa-miR-142-3p	5' isomiRs	Hematopoietic lineage regulation; selective viral mimicry	Manzano et al. (2015)
hsa-miR-124-3p	5' isomiRs	Roles in human retina development	Karali et al. (2016)
hsa-miR-140-3p	5' isomiRs	Tumor suppression in breast cancer	Salem et al. (2016)
hsa-miR-34-5p; hsa-miR-449-5p	5' isomiRs	Controlling multiciliogenesis in airway epithelial cells	Mercey et al. (2017)
hsa-miR-122-5p	3' isomiRs	Influencing Hepatitis C virus replication	Yamane et al. (2017)
mmu-miR-27-3p	3' isomiRs	Differences in regulation of metabolic genes	Ma et al. (2019)

Conclusions for future biology

After more than a decade of intensive research, RNA scientists started to appreciate the unforeseen complexity in the hidden layers of small RNA regulation. From the once described "interesting but seemingly unique examples", we have moved to discover the complex regulation of miRNAs and certainly reached "beyond the one-locus-one-miRNA" paradigm (Telonis et al. 2015). The development and the continuous refinement of new sequencing technologies provide an unprecedented resolution to reveal regulatory mechanisms even at the level of single autonomous cells (Smith and Hutvagner 2022). By revealing the subtle molecular mechanisms of miRNA arm selection and isomiR formation, we have started to understand both the evolutionary and the physiological importance of the elicited posttranscriptional fine-tuning of gene regulation. The discussed phenomena influencing the formation of dynamic miRNA populations have become important for diagnostic tools in several human diseases, presenting useful biomarkers especially in various cancer types, and furthermore also providing potential therapeutic targets for future biomedical research. In addition to the already described functional roles, with the constantly developing detection technologies, new important aspects of miRNA arm selection and isomiR formation can and should also be closely investigated, such as the population- and gender-specific expression profiles which were already described earlier (Loher et al. 2014). Moreover, with the current challenges of new viral threats to human populations, this unexplored population-dependency of miRNA and isomiR repertoire may have a strong influence on the response to viral infections, and therefore this medical aspect of small RNAs should not be neglected either (Rotival et al. 2020).

And the journey has certainly not stopped yet, as similar complexities are expected to be deciphered not only in the other RNAi pathways but most likely in other, RNA-related regulatory mechanisms. All these provide evidence that members of the former "RNA world" are not simple relics of our evolutionary past but they are still active contributors to the current life on Earth. Moreover, they should be considered and utilized also for human medical applications, especially in times of unexpected threats of new potent human infecting viruses.

Acknowledgements The author would like to thank Eszter Horváth for critical reading of the manuscript. MicroRNA-related research in the author's lab is supported by the PC-II-12/2022 grant from the Hungarian Academy of Sciences.

Author contribution Conceptualization and manuscript preparation were done by T.I.O.

Funding Open access funding provided by ELKH Research Centre for Natural Sciences.

Declarations

Conflict of interest The author declares that he has no conflict of interest.

References

Ahmed F, Senthil-Kumar M, Lee S, Dai X, Mysore KS, Zhao PX (2014) Comprehensive analysis of small RNA-seq data reveals that combination of miRNA with its isomiRs increase the accuracy of target prediction in *Arabidopsis thaliana*. RNA Biol 11(11):1414–1429. https://doi.org/10.1080/15476286.2014.996474

Altuvia Y, Landgraf P, Lithwick G, Elefant N, Pfeffer S, Aravin A, Brownstein MJ, Tuschl T, Margalit H (2005) Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33(8):2697–2706. https://doi.org/10.1093/nar/gki567

Ameres SL, Zamore PD (2013) Diversifying microRNA sequence and function. Nat Rev Mol Cell Biol 14(8):475–488. https://doi.org/10.1038/nrm3611

Aparicio-Puerta E, Hirsch P, Schmartz GP, Fehlmann T, Keller V, Engel A, Kern F, Hackenberg M, Keller A (2023) isomiRdb: microRNA expression at isoform resolution. Nucleic Acids Res 51(D1):D179–D185. https://doi.org/10.1093/nar/gkac884

Arora S, Rana R, Chhabra A, Jaiswal A, Rani V (2013) miRNA-transcription factor interactions: a combinatorial regulation of gene expression. Mol Genet Genom 288(3–4):77–87. https://doi.org/10.1007/s00438-013-0734-z

Azuma-Mukai A, Oguri H, Mituyama T, Qian ZR, Asai K, Siomi H, Siomi MC (2008) Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. Proc Natl Acad Sci U S A 105(23):7964–7969. https://doi.org/10.1073/pnas.0800334105

Baralle FE, Giudice J (2017) Alternative splicing as a regulator of development and tissue identity. Nat Rev Mol Cell Biol 18(7):437–451. https://doi.org/10.1038/nrm.2017.27

Bartel DP (2018) Metazoan MicroRNAs. Cell 173(1):20–51. https://doi.org/10.1016/j.cell.2018.03.006

Beadle GW, Tatum EL (1941) Genetic control of biochemical reactions in Neurospora. Proc Natl Acad Sci U S A 27(11):499–506. https://doi.org/10.1073/pnas.27.11.499

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, Tromp MC (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. Cell 46(6):819–826. https://doi.org/10.1016/0092-8674(86)90063-2

Berezikov E (2011) Evolution of microRNA diversity and regulation in animals. Nat Rev Genet 12(12):846–860. https://doi.org/10.1038/nrg3079

Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC (2007) Mammalian mirtron genes. Mol Cell 28(2):328–336. https://doi.org/10.1016/j.molcel.2007.09.028

Berget SM, Moore C, Sharp PA (1977) Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A 74(8):3171–3175. https://doi.org/10.1073/pnas.74.8.3171

Betancur JG, Tomari Y (2012) Dicer is dispensable for asymmetric RISC loading in mammals. RNA 18(1):24–30. https://doi.org/10.1261/rna.029785.111

Bofill-De Ros X, Yang A, Gu S (2020) IsomiRs: Expanding the miRNA repression toolbox beyond the seed. Biochim Biophys Acta Gene Regul Mech 1863(4):194373. https://doi.org/10.1016/j.bbagrm.2019.03.005

Burroughs AM, Ando Y, de Hoon MJ, Tomaru Y, Suzuki H, Hayashizaki Y, Daub CO (2011) Deep-sequencing of human Argonaute-associated small RNAs provides insight into miRNA sorting and reveals Argonaute association with RNA fragments of diverse origin. RNA Biol 8(1):158–177. https://doi.org/10.4161/rna.8.1.14300

Chan WC, Ho MR, Li SC, Tsai KW, Lai CH, Hsu CN, Lin WC (2012) MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach. Genomics 100(3):141–148. https://doi.org/10.1016/j.ygeno.2012.06.007

Cheloufi S, Dos Santos CO, Chong MM, Hannon GJ (2010) A dicer-independent miRNA biogenesis pathway that requires ago catalysis. Nature 465(7298):584–589. https://doi.org/10.1038/nature09092

Chow LT, Gelinas RE, Broker TR, Roberts RJ (1977) An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. Cell 12(1):1–8. https://doi.org/10.1016/0092-8674(77)90180-5

Cifuentes D, Xue H, Taylor DW, Patnode H, Mishima Y, Cheloufi S, Ma E, Mane S, Hannon GJ, Lawson ND, Wolfe SA, Giraldez AJ (2010) A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. Science 328(5986):1694–1698. https://doi.org/10.1126/science.1190809

Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, Barbacioru C, Steptoe AL, Martin HC, Nourbakhsh E, Krishnan K, Gardiner B, Wang X, Nones K, Steen JA, Matigian NA, Wood DL, Kassahn KS, Waddell N, Shepherd J, Lee C, Ichikawa J, McKernan K, Bramlett K, Kuersten S, Grimmond SM (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome Biol 12(12):R126. https://doi.org/10.1186/gb-2011-12-12-r126

Crick FH (1958) On protein synthesis. Symp Soc Exp Biol 12138–12163

Ebhardt HA, Tsang HH, Dai DC, Liu Y, Bostan B, Fahlman RP (2009) Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. Nucleic Acids Res 37(8):2461–2470. https://doi.org/10.1093/nar/gkp093

Etebari K, Osei-Amo S, Blomberg SP, Asgari S (2015) Dengue virus infection alters post-transcriptional modification of microRNAs in the mosquito vector Aedes aegypti. Sci Rep. https://doi.org/10.1038/srep15968

Franco S, Pluvinet R, Sanchez-Herrero JF, Sumoy L, Martinez MA (2022) Rapid and accurate quantification of isomiRs by RT-qPCR. Sci Rep 12(1):17220. https://doi.org/10.1038/s41598-022-22298-7

Frank F, Sonenberg N, Nagar B (2010) Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. Nature 465(7299):818–822. https://doi.org/10.1038/nature09039

Gebert LFR, MacRae IJ (2019) Regulation of microRNA function in animals. Nat Rev Mol Cell Biol 20(1):21–37. https://doi.org/10.1038/s41580-018-0045-7

Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. Nat Rev Genet 10(2):94–108. https://doi.org/10.1038/nrg2504

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature 455(7217):1193–1197. https://doi.org/10.1038/nature07415

Guo L, Chen F (2014) A challenge for miRNA: multiple isomiRs in miRNAomics. Gene 544(1):1–7. https://doi.org/10.1016/j.gene.2014.04.039

Havens MA, Reich AA, Duelli DM, Hastings ML (2012) Biogenesis of mammalian microRNAs by a non-canonical processing pathway. Nucleic Acids Res 40(10):4626–4640. https://doi.org/10.1093/nar/gks026

Heo I, Ha M, Lim J, Yoon MJ, Park JE, Kwon SC, Chang H, Kim VN (2012) Mono-uridylation of pre-microRNA as a key step in the biogenesis of group II let-7 microRNAs. Cell 151(3):521–532. https://doi.org/10.1016/j.cell.2012.09.022

Herrera-Carrillo E, Harwig A, Berkhout B (2015) Toward optimization of AgoshRNA molecules that use a non-canonical RNAi pathway: variations in the top and bottom base pairs. RNA Biol 12(4):447–456. https://doi.org/10.1080/15476286.2015.1022024

Honda S, Kirino Y (2015) Dumbbell-PCR: a method to quantify specific small RNA variants with a single nucleotide resolution at terminal sequences. Nucleic Acids Res 43(12):e77. https://doi.org/10.1093/nar/gkv218

Huntzinger E, Izaurralde E (2011) Gene silencing by microRNAs: contributions of translational repression and mRNA decay. Nat Rev Genet 12(2):99–110. https://doi.org/10.1038/nrg2936

Hutvagner G (2005) Small RNA asymmetry in RNAi: function in RISC assembly and gene regulation. FEBS Lett 579(26):5850–5857. https://doi.org/10.1016/j.febslet.2005.08.071

Iwakawa HO, Tomari Y (2022) Life of RISC: formation, action, and degradation of RNA-induced silencing complex. Mol Cell 82(1):30–43. https://doi.org/10.1016/j.molcel.2021.11.026

Jazdzewski K, Liyanarachchi S, Swierniak M, Pachucki J, Ringel MD, Jarzab B, de la Chapelle A (2009) Polymorphic mature microRNAs from passenger strand of pre-miR-146a contribute to thyroid cancer. Proc Natl Acad Sci USA 106(5):1502–1505. https://doi.org/10.1073/pnas.0812591106

Jones MR, Blahna MT, Kozlowski E, Matsuura KY, Ferrari JD, Morris SA, Powers JT, Daley GQ, Quinton LJ, Mizgerd JP (2012) Zcchc11 uridylates mature miRNAs to enhance neonatal IGF-1 expression, growth, and survival. PLoS Genet 8(11):e1003105. https://doi.org/10.1371/journal.pgen.1003105

Kabekkodu SP, Shukla V, Varghese VK, Jeevitha DS, Chakrabarty S, Satyamoorthy K (2018) Clustered miRNAs and their role in biological functions and diseases. Biol Rev Camb Philos Soc 93(4):1955–1986. https://doi.org/10.1111/brv.12428

Karali M, Persico M, Mutarelli M, Carissimo A, Pizzo M, Singh Marwah V, Ambrosio C, Pinelli M, Carrella D, Ferrari S, Ponzin D, Nigro V, di Bernardo D, Banfi S (2016) High-resolution analysis of the human retina miRNome reveals isomiR variations and novel microRNAs. Nucleic Acids Res 44(4):1525–1540. https://doi.org/10.1093/nar/gkw039

Khvorova A, Reynolds A, Jayasena SD (2003) Functional siRNAs and miRNAs exhibit strand bias. Cell 115(2):209–216

Kim YK, Kim VN (2007) Processing of intronic microRNAs. Embo J 26(3):775–783. https://doi.org/10.1038/sj.emboj.7601512

Kim S, Lee UJ, Kim MN, Lee EJ, Kim JY, Lee MY, Choung S, Kim YJ, Choi YC (2008) MicroRNA miR-199a* regulates the MET proto-oncogene and the downstream extracellular signal-regulated kinase 2 (ERK2). J Biol Chem 283(26):18158–18166. https://doi.org/10.1074/jbc.M800186200

Kim H, Kim J, Yu S, Lee YY, Park J, Choi RJ, Yoon SJ, Kang SG, Kim VN (2020) A mechanism for microRNA arm switching regulated by uridylation. Mol Cell 78(6):1224–1236. https://doi.org/10.1016/j.molcel.2020.04.030

Kong X, Yang M, Le BH, He W, Hou Y (2022) The master role of siRNAs in plant immunity. Mol Plant Pathol 23(10):1565–1574. https://doi.org/10.1111/mpp.13250

Kuchenbauer F, Morin RD, Argiropoulos B, Petriv OI, Griffith M, Heuser M, Yung E, Piper J, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Hansen CL, Marra MA, Humphries RK (2008) In-depth characterization of the microRNA transcriptome in a leukemia progression model. Genome Res 18(11):1787–1797. https://doi.org/10.1101/gr.077578.108

Kuo WT, Su MW, Lee YL, Chen CH, Wu CW, Fang WL, Huang KH, Lin WC (2015) Bioinformatic Interrogation of 5p-arm and 3p-arm specific miRNA expression using TCGA datasets. J Clin Med 4(9):1798–1814. https://doi.org/10.3390/jcm4091798

Lee RC, Feinbaum RL, Ambros V (1993) The C elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75(5):843–854. https://doi.org/10.1016/0092-8674(93)90529-Y

Lee R, Feinbaum R, Ambros V (2004) A short history of a short RNA. Cell 116(2 Suppl):S89-92

Lee LW, Zhang S, Etheridge A, Ma L, Martin D, Galas D, Wang K (2010) Complexity of the microRNA repertoire revealed by next-generation sequencing. RNA 16(11):2170–2180. https://doi.org/10.1261/rna.2225110

Lee HY, Zhou K, Smith AM, Noland CL, Doudna JA (2013) Differential roles of human Dicer-binding proteins TRBP and PACT in small RNA processing. Nucleic Acids Res 41(13):6568–6576. https://doi.org/10.1093/nar/gkt361

Li S, Mason CE (2014) The pivotal regulatory landscape of RNA modifications. Annu Rev Genom Hum Genet. https://doi.org/10.1146/annurev-genom-090413-025405

Li SC, Liao YL, Chan WC, Ho MR, Tsai KW, Hu LY, Lai CH, Hsu CN, Lin WC (2011) Interrogation of rabbit miRNAs and their isomiRs. Genomics 98(6):453–459. https://doi.org/10.1016/j.ygeno.2011.08.008

Li SC, Liao YL, Ho MR, Tsai KW, Lai CH, Lin WC (2012) miRNA arm selection and isomiR distribution in gastric cancer. BMC Genom 13(Suppl):1S13. https://doi.org/10.1186/1471-2164-13-S1-S13

Llorens F, Banez-Coronel M, Pantano L, del Rio JA, Ferrer I, Estivill X, Marti E (2013) A highly expressed miR-101 isomiR is a functional silencing small RNA. BMC Genom. https://doi.org/10.1186/1471-2164-14-104

Loher P, Londin ER, Rigoutsos I (2014) IsomiR expression profiles in human lymphoblastoid cell lines exhibit population and gender dependencies. Oncotarget 5(18):8790–8802. https://doi.org/10.18632/oncotarget.2405

Ma H, Wu Y, Niu Q, Zhang J, Jia G, Manjunath N, Wu H (2016) A sliding-bulge structure at the Dicer processing site of pre-miRNAs regulates alternative Dicer processing to generate 5'-isomiRs. Heliyon 2(9):e00148. https://doi.org/10.1016/j.heliyon.2016.e00148

Ma M, Yin Z, Zhong H, Liang T, Guo L (2019) Analysis of the expression, function, and evolution of miR-27 isoforms and their responses in metabolic processes. Genomics 111(6):1249–1257. https://doi.org/10.1016/j.ygeno.2018.08.004

Manzano M, Forte E, Raja AN, Schipma MJ, Gottwein E (2015) Divergent target recognition by coexpressed 5'-isomiRs of miR-142-3p and selective viral mimicry. RNA 21(9):1606–1620. https://doi.org/10.1261/rna.048876.114

Marco A, Macpherson JI, Ronshaugen M, Griffiths-Jones S (2012) MicroRNAs from the same precursor have different targeting properties. Silence 3(1):8. https://doi.org/10.1186/1758-907X-3-8

Mercey O, Popa A, Cavard A, Paquet A, Chevalier B, Pons N, Magnone V, Zangari J, Brest P, Zaragosi LE, Ponzio G, Lebrigand K, Barbry P, Marcet B (2017) Characterizing isomiR variants within the microRNA-34/449 family. FEBS Lett 591(5):693–705. https://doi.org/10.1002/1873-3468.12595

Mitra R, Lin CC, Eischen CM, Bandyopadhyay S, Zhao Z (2015a) Concordant dysregulation of miR-5p and miR-3p arms of the same precursor microRNA may be a mechanism in inducing cell proliferation and tumorigenesis: a lung cancer study. RNA 21(6):1055–1065. https://doi.org/10.1261/rna.048132.114

Mitra R, Sun J, Zhao Z (2015b) microRNA regulation in cancer: one arm or two arms? Int J Cancer 137(6):1516–1518. https://doi.org/10.1002/ijc.29512

Miyoshi K, Miyoshi T, Siomi H (2010) Many ways to generate microRNA-like small RNAs: non-canonical pathways for microRNA production. Mol Genet Genom 284(2):95–103. https://doi.org/10.1007/s00438-010-0556-1

Moran Y, Agron M, Praher D, Technau U (2017) The evolutionary origin of plant and animal microRNAs. Nat Ecol Evol 1(3):27. https://doi.org/10.1038/s41559-016-0027

Morin RD, O'Connor MD, Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18(4):610–621. https://doi.org/10.1101/gr.7179508

Musiyenko A, Bitko V, Barik S (2008) Ectopic expression of miR-126*, an intronic product of the vascular endothelial EGF-like 7 gene, regulates prostein translation and invasiveness of prostate cancer LNCaP cells. J Mol Med (berl) 86(3):313–322. https://doi.org/10.1007/s00109-007-0296-9

Nakanishi K (2022) Anatomy of four human Argonaute proteins. Nucleic Acids Res 50(12):6618–6638. https://doi.org/10.1093/nar/gkac519

Neilsen CT, Goodall GJ, Bracken CP (2012) IsomiRs–the overlooked repertoire in the dynamic microRNAome. Trends Genet 28(11):544–549. https://doi.org/10.1016/j.tig.2012.07.005

Nersisyan S, Zhiyanov A, Engibaryan N, Maltseva D, Tonevitsky A (2022) A novel approach for a joint analysis of isomiR and mRNA expression data reveals features of isomiR targeting in breast cancer. Front Genet. https://doi.org/10.3389/fgene.2022.1070528

Nikolova M, Naydenov M, Glogovitis I, Apostolov A, Saare M, Boggavarapu N, Salumets A, Baev V, Yahubyan G (2021) Coupling miR/isomiR and mRNA expression signatures unveils new molecular layers of endometrial receptivity. Life (basel). 11(12):1319. https://doi.org/10.3390/life11121391

Noland CL, Doudna JA (2013) Multiple sensors ensure guide strand selection in human RNAi pathways. RNA 19(5):639–648. https://doi.org/10.1261/rna.037424.112

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC (2007) The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. Cell 130(1):89–100. https://doi.org/10.1016/j.cell.2007.06.028

Okamura K, Phillips MD, Tyler DM, Duan H, Chou YT, Lai EC (2008) The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. Nat Struct Mol Biol 15(4):354–363. https://doi.org/10.1038/nsmb.1409

Okamura K, Liu N, Lai EC (2009) Distinct mechanisms for microRNA strand selection by Drosophila Argonautes. Mol Cell 36(3):431–444. https://doi.org/10.1016/j.molcel.2009.09.027

Onishi R, Yamanaka S, Siomi MC (2021) PiRNA- and siRNA-mediated transcriptional repression in *Drosophila* mice and yeast: new

insights and biodiversity. EMBO Rep 22(10):e53062. https://doi.org/10.15252/embr.202153062
Packer AN, Xing Y, Harper SQ, Jones L, Davidson BL (2008) The bifunctional microRNA miR-9/miR-9* regulates REST and CoREST and is downregulated in Huntington's disease. J Neurosci 28(53):14341–14346. https://doi.org/10.1523/JNEUROSCI.2390-08.2008
Park S, Yang HD, Seo JW, Nam JW, Nam SW (2022) hnRNPC induces isoform shifts in miR-21-5p leading to cancer development. Exp Mol Med 54(6):812–824. https://doi.org/10.1038/s12276-022-00792-2
Pasquinelli AE (2012) MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nat Rev Genet 13(4):271–282. https://doi.org/10.1038/nrg3162
Pass HI, Goparaju C, Ivanov S, Donington J, Carbone M, Hoshen M, Cohen D, Chajut A, Rosenwald S, Dan H, Benjamin S, Aharonov R (2010) hsa-miR-29c* is linked to the prognosis of malignant pleural mesothelioma. Cancer Res 70(5):1916–1924. https://doi.org/10.1158/0008-5472.CAN-09-3993
Peng JC, Lin H (2013) Beyond transposons: the epigenetic and somatic functions of the Piwi-piRNA mechanism. Curr Opin Cell Biol 25(2):190–194. https://doi.org/10.1016/j.ceb.2013.01.010
Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J (1987) A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. Cell 50(6):831–840. https://doi.org/10.1016/0092-8674(87)90510-1
Ramalingam P, Palanichamy JK, Singh A, Das P, Bhagat M, Kassab MA, Sinha S, Chattopadhyay P (2014) Biogenesis of intronic miRNAs located in clusters by independent transcription and alternative splicing. RNA 20(1):76–87. https://doi.org/10.1261/rna.041814.113
Ro S, Park C, Young D, Sanders KM, Yan W (2007) Tissue-dependent paired expression of miRNAs. Nucleic Acids Res 35(17):5944–5953. https://doi.org/10.1093/nar/gkm641
Roccaro AM, Sacco A, Chen C, Runnels J, Leleu X, Azab F, Azab AK, Jia X, Ngo HT, Melhem MR, Burwick N, Varticovski L, Novina CD, Rollins BJ, Anderson KC, Ghobrial IM (2009) microRNA expression in the biology, prognosis, and therapy of Waldenstrom macroglobulinemia. Blood 113(18):4391–4402. https://doi.org/10.1182/blood-2008-09-178228
Rosendo Machado S, van der Most T, Miesen P (2021) Genetic determinants of antiviral immunity in dipteran insects—compiling the experimental evidence. Dev Comp Immunol. https://doi.org/10.1016/j.dci.2021.104010
Rotival M, Siddle KJ, Silvert M, Pothlichet J, Quach H, Quintana-Murci L (2020) Population variation in miRNAs and isomiRs and their impact on human immunity to infection. Genome Biol 21(1):187. https://doi.org/10.1186/s13059-020-02098-w
Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. Elegans. Cell 127(6):1193–1207. https://doi.org/10.1016/j.cell.2006.10.040
Ruby JG, Jan CH, Bartel DP (2007) Intronic microRNA precursors that bypass Drosha processing. Nature 448(7149):83–86. https://doi.org/10.1038/nature05983
Saiselet M, Gacquer D, Spinette A, Craciun L, Decaussin-Petrucci M, Andry G, Detours V, Maenhaut C (2015) New global analysis of the microRNA transcriptome of primary tumors and lymph node metastases of papillary thyroid cancer. BMC Genom. https://doi.org/10.1186/s12864-015-2082-3
Salem O, Erdem N, Jung J, Munstermann E, Worner A, Wilhelm H, Wiemann S, Korner C (2016) The highly expressed 5'isomiR of hsa-miR-140–3p contributes to the tumor-suppressive effects of miR-140 by reducing breast cancer proliferation and migration. BMC Genom. https://doi.org/10.1186/s12864-016-2869-x
Schamberger A, Orban TI (2014) 3' IsomiR species and DNA contamination influence reliable quantification of microRNAs by stem-loop quantitative PCR. PLoS ONE 9(8):e106315. https://doi.org/10.1371/journal.pone.0106315
Schamberger A, Sarkadi B, Orban TI (2012) Human mirtrons can express functional microRNAs simultaneously from both arms in a flanking exon-independent manner. RNA Biol 9(9):1177–1185. https://doi.org/10.4161/rna.21359
Scheer H, Zuber H, De Almeida C, Gagliardi D (2016) Uridylation earmarks mRNAs for degradation… and more. Trends Genet 32(10):607–619. https://doi.org/10.1016/j.tig.2016.08.003
Scheper M, Romagnolo A, Besharat ZM, Iyer AM, Moavero R, Hertzberg C, Weschke B, Riney K, Feucht M, Scholl T, Petrak B, Maulisova A, Nabbout R, Jansen AC, Jansen FE, Lagae L, Urbanska M, Ferretti E, Tempes A, Blazejczyk M, Jaworski J, Kwiatkowski DJ, Jozwiak S, Kotulska K, Sadowski K, Borkowska J, Curatolo P, Mills JD, Aronica E, Consortium ME (2022) MiRNAs and isomiRs: serum-based biomarkers for the development of intellectual disability and autism spectrum disorder in tuberous sclerosis complex. Biomedicines. 10(8):1838. https://doi.org/10.3390/biomedicines10081838
Schwarz DS, Hutvagner G, Du T, Xu Z, Aronin N, Zamore PD (2003) Asymmetry in the assembly of the RNAi enzyme complex. Cell 115(2):199–208
Shenasa H, Hertel KJ (2019) Combinatorial regulation of alternative splicing. Biochim Biophys Acta Gene Regul Mech. 11–12:194392. https://doi.org/10.1016/j.bbagrm.2019.06.003
Sibley CR, Seow Y, Saayman S, Dijkstra KK, El Andaloussi S, Weinberg MS, Wood MJ (2012) The biogenesis and characterization of mammalian microRNAs of mirtron origin. Nucleic Acids Res 40(1):438–448. https://doi.org/10.1093/nar/gkr722
Smith CM, Hutvagner G (2022) A comparative analysis of single cell small RNA sequencing data reveals heterogeneous isomiR expression and regulation. Sci Rep 12(1):2834. https://doi.org/10.1038/s41598-022-06876-3
Starega-Roslan J, Krol J, Koscianska E, Kozlowski P, Szlachcic WJ, Sobczak K, Krzyzosiak WJ (2011) Structural basis of microRNA length variety. Nucleic Acids Res 39(1):257–268. https://doi.org/10.1093/nar/gkq727
Starega-Roslan J, Galka-Marciniak P, Krzyzosiak WJ (2015a) Nucleotide sequence of miRNA precursor contributes to cleavage site selection by Dicer. Nucleic Acids Res 43(22):10939–10951. https://doi.org/10.1093/nar/gkv968
Starega-Roslan J, Witkos TM, Galka-Marciniak P, Krzyzosiak WJ (2015b) Sequence features of Drosha and Dicer cleavage sites affect the complexity of isomiRs. Int J Mol Sci 16(4):8110–8127. https://doi.org/10.3390/ijms16048110
Suzuki HI, Katsura A, Yasuda T, Ueno T, Mano H, Sugimoto K, Miyazono K (2015) Small-RNA asymmetry is directly driven by mammalian argonautes. Nat Struct Mol Biol 22(7):512–521. https://doi.org/10.1038/nsmb.3050
Tan GC, Chan E, Molnar A, Sarkar R, Alexieva D, Isa IM, Robinson S, Zhang S, Ellis P, Langford CF, Guillot PV, Chandrashekran A, Fisk NM, Castellano L, Meister G, Winston RM, Cui W, Baulcombe D, Dibb NJ (2014) 5' isomiR variation is of functional and evolutionary importance. Nucleic Acids Res 42(14):9424–9435. https://doi.org/10.1093/nar/gku656
Telonis AG, Loher P, Jing Y, Londin E, Rigoutsos I (2015) Beyond the one-locus-one-miRNA paradigm: microRNA isoforms enable deeper insights into breast cancer heterogeneity. Nucleic Acids Res 43(19):9158–9175. https://doi.org/10.1093/nar/gkv922
Telonis AG, Magee R, Loher P, Chervoneva I, Londin E, Rigoutsos I (2017) Knowledge about the presence or absence of miRNA isoforms (isomiRs) can successfully discriminate amongst 32

TCGA cancer types. Nucleic Acids Res 45(6):2973–2985. https://doi.org/10.1093/nar/gkx082

Tomasello L, Distefano R, Nigita G, Croce CM (2021) The microRNA family gets wider: the isomiRs classification and role. Front Cell Dev Biol. 9:668648. https://doi.org/10.3389/fcell.2021.668648

Trabucchi M, Mategot R (2019) Subcellular heterogeneity of the microRNA machinery. Trends Genet 35(1):15–28. https://doi.org/10.1016/j.tig.2018.10.006

Tsang WP, Kwok TT (2009) The miR-18a* microRNA functions as a potential tumor suppressor by targeting on K-Ras. Carcinogenesis 30(6):953–959. https://doi.org/10.1093/carcin/bgp094

Wang S, Xu Y, Li M, Tu J, Lu Z (2016) Dysregulation of miRNA isoform level at 5' end in Alzheimer's disease. Gene 584(2):167–172. https://doi.org/10.1016/j.gene.2016.02.020

Wang X, Ramat A, Simonelig M, Liu MF (2022) Emerging roles and functional mechanisms of PIWI-interacting RNAs. Nat Rev Mol Cell Biol. https://doi.org/10.1038/s41580-022-00528-0

Westholm JO, Ladewig E, Okamura K, Robine N, Lai EC (2012) Common and distinct patterns of terminal modifications to mirtrons and canonical microRNAs. RNA 18(2):177–192. https://doi.org/10.1261/rna.030627.111

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell 75(5):855–862. https://doi.org/10.1016/0092-8674(93)90530-4

Wilson RC, Tambe A, Kidwell MA, Noland CL, Schneider CP, Doudna JA (2015) Dicer-TRBP complex formation ensures accurate mammalian microRNA biogenesis. Mol Cell 57(3):397–407. https://doi.org/10.1016/j.molcel.2014.11.030

Winter J, Diederichs S (2013) Argonaute-3 activates the let-7a passenger strand microRNA. RNA Biol 10(10):1631–1643. https://doi.org/10.4161/rna.26424

Winter J, Jung S, Keller S, Gregory RI, Diederichs S (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. Nat Cell Biol 11(3):228–234. https://doi.org/10.1038/ncb0309-228

Wu H, Ye C, Ramirez D, Manjunath N (2009) Alternative processing of primary microRNA transcripts by Drosha generates 5' end variation of mature microRNA. PLoS ONE 4(10):e7566. https://doi.org/10.1371/journal.pone.0007566

Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, Krouse MA, Webster PJ, Tewari M (2011) Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. Genome Res 21(9):1450–1461. https://doi.org/10.1101/gr.118059.110

Yamane D, Selitsky SR, Shimakami T, Li Y, Zhou M, Honda M, Sethupathy P, Lemon SM (2017) Differential hepatitis C virus RNA target site selection and host factor activities of naturally occurring miR-122 3 variants. Nucleic Acids Res 45(8):4743–4755. https://doi.org/10.1093/nar/gkw1332

Yang JS, Phillips MD, Betel D, Mu P, Ventura A, Siepel AC, Chen KC, Lai EC (2011) Widespread regulatory activity of vertebrate microRNA* species. RNA 17(2):312–326. https://doi.org/10.1261/rna.2537911

Zelli V, Compagnoni C, Capelli R, Corrente A, Cornice J, Vecchiotti D, Di Padova M, Zazzeroni F, Alesse E, Tessitore A (2021) Emerging role of isomiRs in cancer: state of the art and recent advances. Genes (basel) 12(9):1447. https://doi.org/10.3390/genes12091447

Zhang W, Gao S, Zhou X, Xia J, Chellappan P, Zhou X, Zhang X, Jin H (2010) Multiple distinct small RNAs originate from the same microRNA precursors. Genome Biol 11(8):R81. https://doi.org/10.1186/gb-2010-11-8-r81

Zhang Y, Yang P, Sun T, Li D, Xu X, Rui Y, Li C, Chong M, Ibrahim T, Mercatali L, Amadori D, Lu X, Xie D, Li QJ, Wang XF (2013) miR-126 and miR-126* repress recruitment of mesenchymal stem cells and inflammatory monocytes to inhibit breast cancer metastasis. Nat Cell Biol 15(3):284–294. https://doi.org/10.1038/ncb2690

Zhou H, Arcila ML, Li Z, Lee EJ, Henzler C, Liu J, Rana TM, Kosik KS (2012) Deep annotation of mouse iso-miR and iso-moR variation. Nucleic Acids Res 40(13):5864–5875. https://doi.org/10.1093/nar/gks247