dc_67_10

Relics from the Past: Molecular Biology and Genetic Applications of Resurrected DNA Transposons in Vertebrates

Thesis for the Degree of

Doctor of the Hungarian Academy of Sciences

Zoltán Ivics

Max Delbrück Center for Molecular Medicine

Berlin
2010

Table of Contents

ACKNOWLEDGEMENTS

First and foremost I am grateful to my wife and long-term colleague Dr. Zsuzsanna Izsvak with whom I have been working side-by-side for the past nineteen years. This time has been an adventure to “boldly go where no man has gone before”.

I am grateful to all members of the “Transposition” and “Mobile DNA” groups at the Max Delbrück Center for Molecular Medicine for their dedicated work. Their spirit, innovative ideas as well as hard benchwork provide the basis of past and future discoveries.

Lastly, I thank my parents for their unconditional love and support.

Work in the author’s laboratory has been supported by EU FP5, FP6 and FP7 research grants, as well as financial support from the Deutsche Forschungsgemeinschaft, the Bundesministerium für Forschung und Bildung and the Volkswagen Stiftung.

LIST OF PUBLICATIONS THAT FORM THE BASIS OF THIS THESIS

I. **Ivics, Z**., Izsvák, Zs., Minter, A. and Hackett, P.B. (1996). Identification of functional domains and evolution of Tc1-like transposable elements. *Proc. Natl. Acad. Sci. USA* 93:5008-5013.

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

III. Walisko, O., Schorn, A., Rolfs, F., Devaraj, A., Miskey, C., Izsvák, Z. and **Ivics, Z**. (2008). Transcriptional activities of the *Sleeping Beauty* transposon and shielding its genetic cargo with insulators. *Mol. Ther*. 16:359-69.

IV. Zayed, H., Izsvák, Zs., Khare, D., Heinemann, U. and **Ivics, Z**. (2003). The DNA-bending protein HMGB1 is a cellular cofactor of *Sleeping Beauty* transposition. *Nucleic Acids Res*. 31:2313-2322.

V. Walisko, O., Izsvák, Z., Szabó, K., Kaufman, C.D., Herold, S., and **Ivics, Z.** (2006). Sleeping Beauty transposase modulates cell-cycle progression through interaction with Miz-1. *Proc. Natl. Acad. Sci. USA* 103:4062-4067.

VI. Jursch, T., Izsvák, Z. and **Ivics, Z.** (2010). Regulation of DNA transposition by CpG methylation and chromatin structure in human cells. *Mobile DNA*. (to be submitted)

VII. Vigdal, T., Kaufman, C., Izsvak, Z., Voytas, D. and **Ivics, Z.** (2002). Common physical properties of DNA affecting target site selection of *Sleeping Beauty* and other Tc1/*mariner* transposable elements. *J. Mol. Biol*. 323:441452.

VIII. Miskey, Cs., Izsvák, Zs., Plasterk, R.H. and **Ivics, Z**. (2003). The *Frog Prince*: a reconstructed transposon from *Rana pipiens* with high activity in vertebrates. *Nucleic Acids Res*. 31:6873-6881.

IX. Miskey, C., Papp, B., Mátés, L., Sinzelle, L., Keller, H., Izsvák, Z. and **Ivics, Z**. (2007). 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*. 27: 4589-600.

X. Sinzelle, L., Kapitonov, V.V., Grzela, D.P., Jursch, T., Jurka, J., Izsvák, Z. and **Ivics, Z.** (2008). Transposition of a reconstructed *Harbinger* element in human cells and functional homology with two transposon-derived cellular genes. *Proc. Natl. Acad. Sci. USA* 105:4715-20.

XI. Zayed, H., Izsvák, Zs., Walisko, O. and **Ivics, Z.** (2004). Development of hyperactive *Sleeping Beauty* transposon vectors by mutational analysis. *Mol. Ther.* 9:292-304.

XII. Kaufman, C.D., Izsvák, Z., Katzer, A. and **Ivics, Z**. (2005). *Frog Prince* transposon-based RNAi vectors mediate efficient gene knockdown in human cells. *Journal of RNAi and Gene Silencing* 1:97-104.

XIII. Grabundzija, I., Irgang, M., Mátés, L., Belay, E., Matrai, J., Gogol-Döring, A., Kawakami, K., Chen, W., Ruiz, P., Chuah, M.K., VandenDriessche, T., Izsvák, Z. and **Ivics, Z**. (2010). Comparative analysis of transposable element vector systems in human cells. *Mol. Ther*. (in press)

XIV. **Ivics, Z**., Katzer, A., Stüwe, E.E., Fiedler, D., Knespel, S. and Izsvák, Z. (2007). Targeted *Sleeping Beauty* transposition in human cells. *Mol. Ther.* 15:1137–1144.

Other publications that significantly contributed to this thesis

Izsvák, Zs., **Ivics, Z.** and Hackett, P.B. (1995). Characterization of a Tc1-like transposable element in zebrafish (*Danio rerio*). *Mol. Gen. Genet*. 247:312-322.

Plasterk, R.H., Izsvák, Zs. and **Ivics, Z**. (1999). Resident Aliens: The Tc1/*mariner* superfamily of transposable elements. *Trends Genet*. 15:326-332.

Izsvák, Zs., Khare, D., Behlke, J., Heinemann, U., Plasterk, R.H. and **Ivics, Z**. (2002). Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in *Sleeping Beauty* transposition. *J. Biol. Chem*. 277:34581-34588.

Walisko, O. and **Ivics, Z**. (2006). Interference with cell cycle progression by parasitic genetic elements: *Sleeping Beauty* joins the club. *Cell Cycle*. 5:1275-80.

Walisko, O., Jursch, T., Izsvák, Z. and **Ivics. Z.** (2008). Transposon-host cell interactions in the regulation of *Sleeping Beauty* transposition. IN (Volff, J.-N. and Lankenau, D.-H. eds) Genome Dynamics and Stability vol. 4: *Transposons and the Dynamic Genome*. Springer-Verlag Berlin Heidelberg, Germany, pp 109-132.

Voigt, K., Izsvák, Z. and **Ivics, Z.** (2008). Targeted gene insertion for molecular medicine. *J. Mol. Med.* 86:1205-19.

Sinzelle, L., Izsvák, Z. and **Ivics, Z**. (2009). Molecular domestication of transposable elements: From detrimental parasites to useful host genes. *Cell. Mol. Life Sci*. 66:1073-93.

Mátés, 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**. and Izsvák, Z. (2009). Molecular evolution of a novel hyperactive *Sleeping Beauty* transposase enables robust stable gene transfer in vertebrates. *Nat Genet*. 41:753-61.

ABBREVIATIONS

AAV	Adeno Associated Virus
Ac	Activator
ASLV	Avian Sarcoma Leukosis Virus
ASV	Avian Sarcoma Vurus
ATP	adenosine 5´-triphosphate
Cdk	cyclin-dependent kinase
cDNA	complementary DNA
CHO	Chinese hamster ovary
CMV	cytomegalovirus
DBD	DNA-binding domain
DDD	amino acid sequence containing three aspartic acids
DDE	amino acid sequence containing two aspartic acids and one glutamic acid
DNA	deoxyribonucleic acid
DR	direct repeat
E. coli	*Escherichia coli*
EN	endonuclease
ENU	ethylnitrosourea
env	envelope
ERV	endogenous retrovirus
FACS	fluorescence-activated cell sorting
FP	*Frog Prince*
gag	group-specific antigen
GFP	green fluorescent protein
HA	hemagglutinin
HDR	homology-dependent repair
HIV	human immunodeficiency virus

HMG	high-mobility group protein
HSC	hematopoietic stem cell
HTH	helix-turn-helix
Hsmar1	*Homo sapiens mariner* type 1
IAP	intracisternal A particle
IHF	integration host factor
IN	integrase
IR	inverted repeat
IRES	internal ribosome entry site
IR/DR	inverted repeats containing direct repeats
IS	insertion sequence
kbp	kilo base pairs
kDa	kilo Daltons
L1	LINE1
LINE	long interspersed nuclear element
LTR	long terminal repeat
MBP	maltose binding protein
MLV	murine leukemia virus
Myr	million years
mRNA	messenger ribonucleic acid
NHEJ	nonhomologous end joining
NLS	nuclear localization signal
OPI	overproduction inhibition
ORF	open reading frame
PCR	polymerase chain reaction
PEC	paired end complex
pol	polymerase
polyA	poly-adenylation signal
Pol II	RNA polymerase II

Pol III	RNA polymerase III
PR	protease
RACE	rapid amplification of cDNA ends
RAG	recombination-activating gene
RNA	ribonucleic acid
RNAi	RNA interference
RSS	recombination signal sequence
RT	reverse transcriptase
SA	splice acceptor
SB	*Sleeping Beauty*
SD	splice donor
sem	standard error of the mean
shRNA	short hairpin RNA
SINE	short interspersed nuclear element
SV40	simian virus 40
Tc1	transposon of *Caenorhabditis elegans* 1
TE	transposable element
Tn	transposon
TSD	target site duplication
Ty	yeast retrotransposon
UTR	untranslated region
V(D)J	variable (diversity) joining
vs.	*versus*
X-Gal	4-Cl-5-Br-3-indolyl-β-galactosidase
ZF	zinc finger
ZFN	zinc finger nuclease

1 INTRODUCTION

1.1 Discovery of transposable elements

Transposable genetic elements (“jumping genes”) were first discovered by Barbara McClintock in the 1940s. She found that certain spontaneous mutations in enzymes required for the productions of the purple anthocyanin pigment in maize are due to “controlling” elements that could apparently move from site to site in different chromosomes. This idea of jumping genes ran contrarily to the traditional view of the age that genomes are stable and static entities. The possibility that pieces of DNA can “jump around” in a genome was viewed by biologists with much skepticism. Therefore, McClintock’s observations were thought to be rare phenomena and not of general interest. In fact, at a historic Cold Spring Harbor meeting 1951. McClintock's work greeted with "stony silence." It took nearly 30 years until McClintock’s conclusions from the 1940s were confirmed by findings in bacteria (insertion sequences) and *Drosophila melanogaster* (hybrid dysgenesis), and another ten years until she was rewarded for the discovery of transposable elements with the Nobel prize in 1983 (Fig. 1).

Figure 1. Barbara McClintock – Nobel Prize 1983

With the great advances of the molecular biology in the 1970s, it turned out that McClintock’s discovery was just the tip of an iceberg. Mobile element were found to be widespread not only in maize but in all kingdoms of living organisms from bacteria to humans. It turned out that transposable elements are indeed so abundant that they form a major fraction of the eukaryotic genome [1]. However, most researchers still assumed that repetitive DNA elements do not have any function: they are useless, selfish DNA sequences [2]. The term “junk DNA” coined by Sozumu Ohno repelled mainstream research from studying repetitive elements for many years [3]. As Doolittle and Sapienza termed in *Nature* in 1980: transposons’ “only «function» is survival within genomes”…“thus no phenotypic or evolutionary function need to be assigned to them”.

This view started to change in the 1990s, when it became evident that transposons are important integral components of eukaryotic genomes with deep impacts on the host evolution. It turned out that they interact with the surrounding genomic environment, and increase the ability of the organism to evolve [4].

Since their discovery, transposable elements have been broadening the scope of many fields of modern biology ranging from evolutionary genetics to gene therapy. There are numerous aspects of viewing transposable elements as subjects of scientific investigation. Transposons are of interest for *genome annotators*, for *structural and evolutional geneticists* who investigate the role of mobile elements in chromosome/genome dynamics and their different contributions to host evolution. The ongoing studies of *molecular biologists* are continuously increasing our understanding of the mechanism transposition. Moreover, *experimental geneticists* use transposons routinely for insertional mutagenesis, gene tagging, germline transformations, gene trapping, and gene therapy. Their experimental model organisms range from bacteria to mammals. Due to the discovery of a variety of different prokaryotic and eukaryotic transposons, they are now routinely used as genetic tools in functional biology. Thus, repetitive elements are relevant to a wide scale of genetic studies, and transposons begin to be viewed as genomic treasure [5, 6].

1.2 Classification of transposable elements

Discrete DNA sequences that possess an intrinsic capability to change their genomic locations are called **transposable elements** (TEs). TEs are distinguished whether their movement relies exclusively on DNA intermediates or includes an RNA stage. Transposons that move exclusively through a DNA intermediate are referred to as **DNA elements**. Mobile elements that move through an RNA intermediate (**RNA elements or retroelements**) are transcribed, reverse transcribed and integrate as double stranded cDNA. These elements include **retroviruses** and the **retrotransposons**. DNA elements can be found in both prokaryotic and eukaryotic organisms, whereas RNA elements are restricted to eukaryotes.

1.2.1 RNA elements

Based on their structural properties and evolutionary relationships those transposable elements that can mobilize themselves through an element-derived RNA intermediate are grouped to those with long terminal repeats (LTR-retrotransposons and retroviruses) and those without (non-LTR retrotransposons).

A common feature of ***LTR-retrotransposons*** and ***retroviruses*** is that their coding region is flanked by LTRs (Fig. 2C). These sequences contain important control sequences e.g. promoter, enhancer and polyadenylation (polyA) signals. The coding sequences are divided into at least two open reading frames (ORFs). The first ORF encodes the group-specific antigen (*gag)* protein, required for the assembly of the RNA transcript into cytoplasmic particles. The second ORF constitutes the *pol* gene, encoding a polyprotein, which consists of a protease (PR), a reverse transcriptase (RT) and an integrase (IN). The difference between retroviruses and LTR-retrotransposons is that retroviruses not only possess the capability to move between DNA molecules like other transposons, but they can leave their host cells too and integrate into new genomes. Nevertheless, retroviruses and LTR-retrotransposons are derived from a common progenitor [7].

LTR-retroelements can be subdivided into three families based on homologies within the RT gene. The first two groups are named after their founding members found in yeast and *Drosophila*, *Ty1/copia* and *Ty3/gypsy* [8]. The *Ty3/gypsy* elements form two subfamilies based on the presence or absence of a third ORF, *env*, encoding for envelope-like proteins. Retroviruses cluster into the third family of LTR-elements; they always possess a completely functional *env* gene for their viral life cycle. Many retroviruses, for example human immunodeficiency virus type 1 (HIV-1), contain additional proteins [9]. Endogenous retroviruses (ERVs) appear to have been recently active in the mammalian genome. LTR-retrotransposons are widely destributed in eukaryotes, and make up about 8% of the human genome. Retroviruses were for long thought to be restricted to vertebrate genomes until it was shown that the *gypsy* retrotranspsoson is indeed an infectious retrovirus of *Drosophila*

melanogaster [10]. Transposition occurs through reverse transcription of the retrotransposon RNA, and integration of the resultant cDNA into a new location by the integrase protein.

The most abundant transposable elements in mammalians are ***non-LTR retrotransposons*** represented by the long interspersed nuclear elements (LINEs) and the short interspersed nuclear elements (SINEs). Although LINEs are especially abundant in mammals (they make up 26% of the human X chromosome alone) [11], they have also been found in protozoan, insects, reptiles and plants [12]. The major LINEs in humans (LINE1 or L1) are 6 kbp long and contain two ORFs (Fig. 2A). These encode for a nucleic acid binding protein and an enzyme with endonuclease (EN) and RT activity, respectively [13]. EN generates a single-stranded nick in the target DNA, and RT uses the nicked DNA to prime reverse transcription from the 3'-end of the L1 RNA [14]. Because reverse transcription is frequently incomplete, the majority of L1s is truncated, and thus nonfunctional. Consequently, even though L1 has about 5 x 10^5 copies in the human genome, thereby making up about 17% of human genomic DNA [11], the vast majority of these elements are inactive: in humans there are only 30-100 potentially active copies of L1 [15].

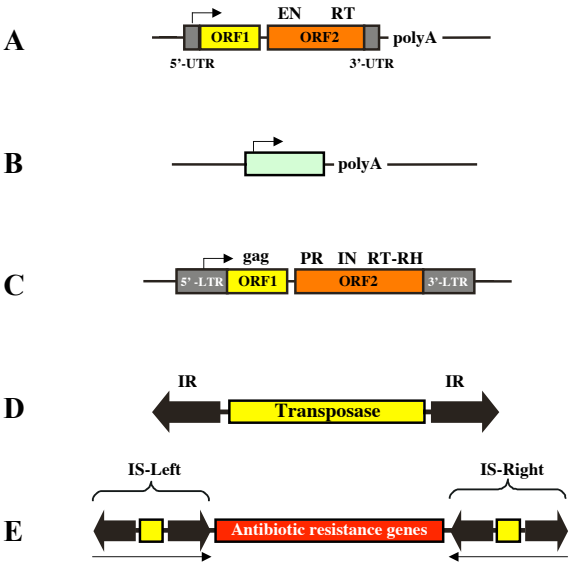


Figure 2. Structures and organization of the main types of transposable elements. (A) Non-LTR retrotransposon. The element consists of a 5' untranslated region that has promoter activity (arrow pointing towards the downstream genes), which is required to drive transcription of the element-encoded genes. ORF1 encodes a nucleic acid binding protein. ORF2 encodes an endonuclease (EN) and a reverse transcriptase (RT). The element has a polyA tail. **(B)** A typical SINE. The element is a small, RNA-derived pseudogene, which is transcribed from an RNA polymerase III promoter within the element (arrow). The element has a polyA tail. **(C)** LTR-retrotransposon. The element consists of long terminal repeats (LTRs) similar to those of retroviruses. The LTRs flank two open reading frames. ORF1 encodes the group specific antigen (gag), ORF2 encodes a protease (PR), an integrase (IN), and a reverse transcriptase-RNaseH (RT-RH) function. **(D)** DNA transposon. The central transposase gene (yellow box) is flanked by terminal inverted repeats (IRs, shown as black arrows). The IRs contain the binding sites for the transposase and sequences that are required for transposase-mediated cleavage. **(E)** Composite bacterial transposon. The element consists of antibiotic resistance genes (red box) flanked by two copies of an insertion sequence (IS) element that contains the transposase gene (yellow boxes). The arrows underneath indicate the inverted orientation of the IS elements.

SINEs are short (about 100–400 bp) retrotransposable elements that encode no proteins; therefore, all of them are non-autonomous (Fig. 2B), and thought to use the enzymatic machinery of LINEs for transposition [16, 17]. The vast majority of known SINEs are derived from tRNA

sequences, with the exception of the human *Alu* element, which is derived from the 7SL component of the signal recognition particle [18]. *Alu* elements were originally identified as repetitive DNA elements in human DNA renaturation curves, and contain a recognition site for the restriction enzyme *Alu*I. *Alu* elements are represented in the human genome with >1 x 10^6 copies which make up about 11% of the total genome. *Alu* is the only active SINE in humans. Full-length *Alu*s are 280bp long, contain promoter sequences for RNA polymerase III (Pol III) [19] and a polyA tail (Fig. 2B). The transcripts of Pol III-transcribed *Alu*s terminate at Pol III termination signals fortuitously present in the 3' flanking DNA. Rarely, RNA polymerase II (Pol II)-derived host gene transcripts can also be *trans*-mobilized by functional LINE proteins. These transposition products are named processed pseudogenes. They lack promoters, introns and end in a polyA tail. Only short target site duplications flanking these sequences provide evidence that these integrants are in fact transposition products.

1.2.2 DNA elements

These TEs can loosely be defined as sequences of DNA that can excise and insert into a variety of sites of a target DNA without the need to be reverse transcribed to cDNA. The simplest DNA elements are the **insertion sequences** (ISs) that were first characterized from bacteria in the late 1960s. Since then, approximately 800 ISs were identified (http://www-is.biotoul.fr). ISs are short (<2.5 kbps) and carry no genetic information except that necessary for their mobility. Thus, they are composed of a single gene coding for the transposase enzyme responsible for moving the element and of terminal inverted repeats (IRs) flanking it at both ends (Fig. 2D). The IRs are often called terminal inverted repeats (TIRs) or inverted terminal repeats (ITRs). The IRs contain the recombinationally active nucleotides at the very tips and specific recognition sequences for the transposase enzyme within.

Though most of the ISs are prokaryotic, a significant number of eukaryotic IS has also been documented. The largest and best-known group of these is the *Tc1/mariner* like elements that are structurally the closest to bacterial ISs (see more detail in the next section)

[20]. Another well-characterized member of the eukaryotic ISs is *P* element from *Drosophila melanogaster*.

Recently, new families of DNA elements have been identified from eukaryotes. The *Helitron* elements lack IRs and move by the rolling circle mechanism, similarly to the replication of plasmids, and together with their descendants they represent 2% of the *Arabidopsis thaliana* and the *Caenorhabditis elegans* genomes [21]. *Polintons* are a newly discovered, self-synthesizing, complex family of DNA transposons that possibly derived from ancient LTR retroelements [22]. The elements are very large (~20 kb), with IRs of several hundred bp in length and encode several proteins: i) a protein-primed DNA-dependent polymerase, ii) an ATPase, iii) a protease (not always), iv) an integrase, and 4-6 additional ORFs for proteins with unknown function. The 300,000 DNA transposon fossils in human add up to around 3% of the genome [11].

It became evident in the 1960s that genes responsible for antibiotic resistance in bacteria can move between DNA molecules in a process analogous to the movement of ISs [23]. It was suggested that mobile elements that carry one or more genes that encode other functions in addition to those related to transposition should be called **transposons** (today, the term „transposon“ is used in a wider sense, however: authors call all DNA elements, including ISs, transposons). Since these elements carry additional DNA they are usually larger than ISs (approximately 2.5-7 kbp). In some of these elements, called composite transposons, there are two complete ISs flanking a functional gene (Fig. 2E). This element can move as one functional unit, but also one or both of the bordering ISs can mobilize itself independently. There is a characteristic feature that distinguishes eukaryotic TEs from ISs and transposons in bacteria: the presence of a large number of inactive transposon copies that can in many cases be mobilized in *trans* by a limited number of active transposases [24].

1.2.2.1 The *Tc1/mariner* superfamily of transposons

When David Hirsch and Scott Emmons discovered the *Tc1* transposable element in 1983 as a repeat sequence in the genome of *Caenorhabditis elegans* [25], they probably did not realize how large the iceberg was of which they had found the tip. We now know that homologs of *Tc1* and those of the related *mariner* transposon found in *Drosophila mauritiana* [26], are probably the most widespread DNA-transposons in nature, and can be found in fungi, plants, ciliates and animals, including nematodes, arthropods, fish, frogs and humans. Together with related *pogo* transposons [27, 28], *Tc1* and *mariner* elements are members of a large superfamily of transposable elements, the *Tc1*/*mariner* superfamily [29-31], so named after its two best studied members. *Tc1*/*mariner* elements are about 1300-2400 bp in length and contain a single gene encoding a transposase enzyme which is flanked by IRs. Although quite divergent in primary sequence (about 15% amino acid identity between the transposases of the different families [30]), members of the *Tc1*/*mariner* superfamily are probably monophyletic in origin (Fig. 3) [30, 32], and have similar structures and molecular mechanisms of transposition. As shown in Fig. 3, a more remote similarity exists between the above mentioned transposons and several bacterial IS elements, LTR-retrotransposons and retroviruses [33]. The recombinase proteins encoded by these diverse genetic elements are all related and contain a signature of three acidic amino acids (DDE or DDD, Fig. 2) with a characteristic spacing [32, 33].

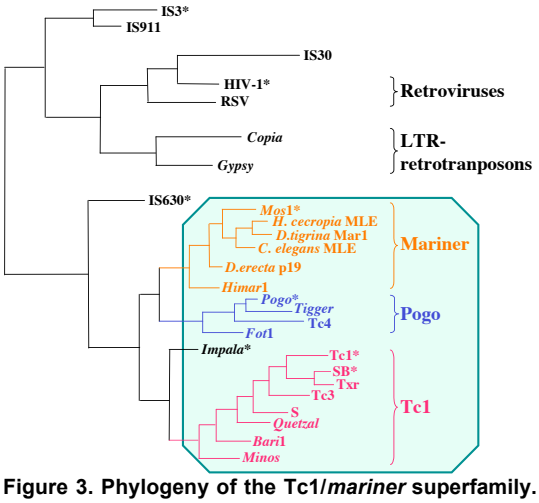


Figure 3. Phylogeny of the Tc1/*mariner* superfamily. DDE-containing recombinases are grouped into two major clusters: a DNA-transposon group and a retroelement group. Bacterial IS elements are DNA-transposons, but certain elements such as *IS3*, *IS911* and *IS30* are grouped together with the retroelement group, whereas the position of *IS630* is close to the *Tc1*/*mariner* superfamily (green box) in the phylogenetic tree. The *Tc1*, *mariner* and *pogo* transposon families are probably monophyletic.

1.2.2.1.1 Structural and functional components of *Tc1/mariner* transposons

1.2.2.1.1.1 The transposase

As discussed above, transposons are very diverse genetic entities; however, their enzymes carry out similar chemical reactions e.g. hydrolysis for strand cleavage and transesterification

for strand transfer. The similar activities of TEs are manifested in the remarkable overall structural similarity of the transposition proteins.

Both the transposases of ISs and transposons and INs of retroelements show structural similarities for their functional organization. Most of them can be divided into topological distinct functional domains. Partial proteolysis experiments revealed that the transposon-specific DNA-binding domains are generally localized in the N-terminal part, whereas the catalytic domain responsible for the strand cleavage and transfer is located in the C-terminal of the transposase protein (Fig. 4) [34, 35]. One possible explanation for this characteristic arrangement in prokaryotic elements is that during translation the N-terminal part of the premature transposase protein can fold independently of the C-terminal catalytic domain, and interact with its specific transposon binding sites close to the point of synthesis. This hypothesis is reinforced by the observation that the presence of the C-terminal part of some bacterial transposases decreases the affinity of IR binding [36]. This arrangement can facilitate that the transposase is going to act on the transposon that produced it (a phenomenon called *cis*-preference) [37].

The DNA-binding domain

It is a key feature of all transposases that they recognize their specific transposon ends. TEs that move by transposon-specific transposases possess recognition sequences in their IRs. The majority of ISs has simple, 10-40 bp long IRs, while others exhibit long and complex IRs. Most transposon ends are composed of two functional parts. The 2-3 terminal base pairs of the ends are the recombinationally active sequences involved in the cleavage and the strand transfer reactions. The other functional part is situated within the IRs and it ensures the sequence-specific positioning of the transposase on the transposon ends [38, 39]. ISs have single transposase binding sites whereas for example *Mu* and *Tn7* have complex, asymmetric recognition sites [40, 41]. The bi-functionality of the transposon ends is reflected in the arrangement of the transposase on its cognate transposon. Due to the flexibility of the transposase, the N-terminal region of the enzyme attaches to the inner segment of IRs while

the C-terminal contacts the external ends. The sequence-specific DNA-binding of both eukaryotic and prokaryotic transposases is often carried out by a helix-turn-helix (HTH) motif. This domain can be simple as it is the case of IS transposases [42], or can be complex and bipartite as found in *Ac*, *Mu* or in *Tc* transposases [29, 43]. The catalytic C-terminal domains of transposases are also involved in DNA-binding, however, this activity is not sequence specific and contributes to the correct positioning of the transposon end into the catalytic pocket [44].

The overall domain structure of the transposase is conserved in the entire *Tc1/mariner* superfamily [20]. Specific substrate recognition is mediated by an N-terminal, bipartite DNA-binding domain of the transposase (Fig. 4) [45-47]. This DNA-binding domain has been proposed to consist of two HTH motifs, similar to the paired domain of some transcription factors in both amino acid sequence and structure [47-49]. The modular paired domain has evolved versatility in binding to a range of different DNA sequences through various combinations of its subdomains (PAI+RED) [50]. The nucleotide sequences recognized by the composite paired domain are degenerate, the DNA-binding specificity is relaxed [51]. The origin of the paired domain is not clear, but phylogenetic analyses indicate that it might have been derived from an ancestral transposase [52].

The first of these HTH motifs, similar to the paired domain of some transcription factors (Fig. 4) [48, 49], has been crystallized in complex with double-stranded DNA corresponding to the termini of *Tc3* transposons in *C. elegans* [53]. The crystal structure indeed showed a HTH fold, and a dimer of transposase subunits bringing together the two DNA ends. The paired-like domain is followed by a second HTH motif embedded in a homeo-like DNA-binding domain (Fig. 4). Secondary structure predictions indicate that *mariner* transposases might also contain such a bipartite DNA-binding domain consisting of two HTH motifs (Fig. 4). *Pogo* and certain bacterial transposases [54] contain "solo" HTH motifs (Fig. 4). We found that a GRPR-like sequence between the two HTH motifs is conserved in *Tc1/mariner* transposases (Fig. 4). The GRPR motif is characteristic to homeodomain proteins [55], and mediates interactions with DNA in the Hin invertase of *Salmonella* [56] and in the recombination activating gene (RAG1) recombinase that evolved

from a DNA transposable element [57-59], and has been "domesticated" (as discussed in section 1.5.2.1.1) [60] to carry out recombination reactions to generate immunoglobulin and T-cell receptor gene diversity in jawed vertebrates [61-63]. The relatedness of DNA-binding by *Tc1* transposase and RAG1 recombinase is further supported by DNA sequence similarities between their binding sites [64]. Members of the retroviral integrase family carry a combined motif of a zinc-binding domain [65] and an HTH motif (Fig. 4) that resembles the *Tc3* paired-like structure [66].

The catalytic domain

The second major domain of the transposase has been referred to as the catalytic domain, because it is responsible for the DNA cleavage and joining reactions of transposition. The majority of known transposases and INs possess a well-conserved triad of amino acids, known as the aspartat-aspartat-glutamat, in short the DDE motif (actually, more of a signature than a "motif" in a usual sense) in their C-terminal catalytic domain (Fig. 4) [67]. The DDE motif is found in a large group of recombinases, including retrotransposon and retrovirus integrases, bacterial IS element transposases [33] and RAG1 [33, 68, 69].

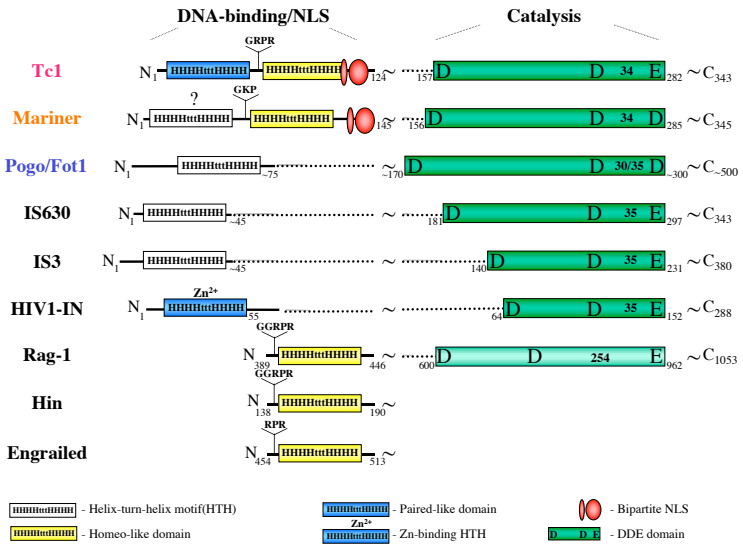


Figure 4. Functional domains of Tc1/*mariner* transposases. Modular structure of Tc1/*mariner* transposases and topology of their major functional domains in comparison with other DDE recombinases and DNA-binding proteins. Tc1/*mariner* transposases have an N-terminal DNA-binding domain followed by a nuclear localization signal and a C-terminal catalytic domain. DDE recombinases have a DDE signature-containing core catalyzing polynucleotidyl transfer reactions. The catalytic core acquired different DNA-binding modules during evolution to give rise to a diverse family of recombinases.

Structural analyses of HIV-1 INs and mutational studies revealed that the DDE triad lies in the heart of the catalytic domain of transposases and INs [47, 70]. These amino acids play essential role in catalysis by coordinating, in general, two divalent cations necessary for activity. Retroviral INs were shown to be able to coordinate Ca^{++}, Zn^{++} and Mn^{++} ions, but

the biologically relevant cation is thought to be Mg^{++} [71, 72]. One metal ion acts as a Lewis acid, and stabilizes the transition state of the penta-coordinated phosphate, the other one acts as a general base and deprotonizes the incoming nucleophil during transesterification and strand transfer [44].

The C-terminal half of *Tc1/mariner* transposases was initially proposed to be the catalytic domain based on the presence of the characteristic DDE (or DDD in the case of *mariner* and *pogo*) motif (Fig. 4). Site-directed mutagenesis of these positions in the *Tc3* transposase confirmed that these three amino acids are essential for all catalytic activities [73]. Interestingly, a change of the exceptional third D of *mariner*, turning the DDD into the canonical DDE, inactivates the transposase [74]. This is most easily explained by assuming that the catalytic role of either aspartic or glutamic acid is similar, but that the precise spatial position within the transposase fold requires the presence of the correct residue.

The crystal structure of the *Mos1 mariner* transposase from *D. melanogaster* has recently been solved [75]. The structure contains a dimer of transposase and four DNA duplexes. Two of these duplexes are recognized by the N-terminal DNA-binding domain of the transposase and are held in position in the catalytic domains as if they have just been cleaved (Fig. 5). In striking contrast to the anti-parallel orientation of transposon ends in the *Tn5* synaptic complex, these duplexes are approximately parallel. The N-terminal domain of the transposase (residues 1–112) comprises two HTH motifs linked by a minor groove binding motif. Residues 113–125 form a linker between the DNA-binding domain and the catalytic domain (residues 126–161 and 190–345). Residues 162–189 form a clamp loop extending out from the catalytic domain making key interactions

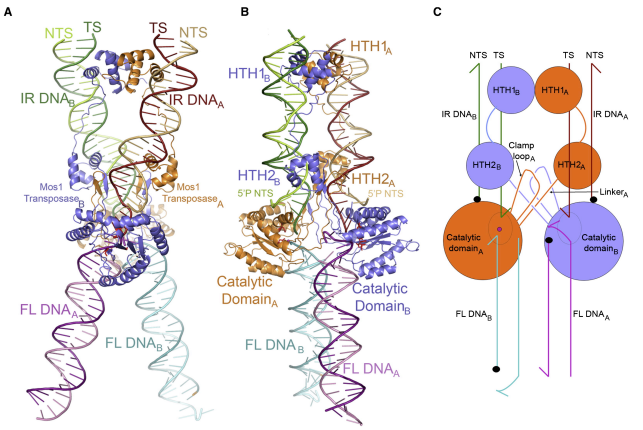


Figure 5. Architecture of the *Mos1* paired end complex. (A and B) Orthogonal views of the PEC crystal structure. Transposase monomer A is colored orange and monomer B blue. The two major-groove DNA-binding motifs contain HTH1 (residues 24–55) and HTH2 (residues 89–110). The minor-groove binding motif comprises residues 63–71. The two DNA duplexes bound by the DNA-binding domains are labeled IR DNA and the two extra DNA duplexes are labeled FL DNA. **(C)** Schematic diagram of the structure. An arrow indicates the 3' end of each DNA strand and a black dot indicates the 5' phosphate of the NTS. The purple sphere indicates the metal ion in active site A.

with the linker of the other transposase monomer in the complex (Fig. 5). Two additional DNA duplexes are bound by the catalytic domains in positions that could represent binding sites for DNA flanking the transposon. The catalytic domain has an RNaseH-like fold. In addition to *Mos1* and other DDE-containing transposases and integrases [70], crystallographic analyses of the catalytic domains of proteins whose functions are not obviously related to transposition, such as RNAaseH [76] or RuvC [77] have revealed a remarkably similar overall fold.

The emerging picture reinforces the notion of a common structural motif that catalyses polynucleotidyl transfer reactions in diverse biological contexts [65, 70], and that the different specificities in binding to DNA might have evolved by the apparent acquisition of different DNA-binding domains, and combinations thereof, in the evolution of DDE recombinases [33].

1.2.2.1.1.2 The transposon inverted repeats

Tc1/mariner elements have a roughly uniform size of approximately 1.6-1.7 kb, indicating a natural selection in genomes for this particular size. The transposons are bracketed by IRs that contain binding sites for the transposase. IRs vary in length and contain transposase-binding sites in different numbers and patterns in the *Tc1/mariner* family (Fig. 6). *Tc1* and *mariner* elements are the simplest and have repeats of less than 100 bp and a single binding site per repeat (Fig. 6) [47, 78]. *Tc3* elements have IRs of more than 400 bp in length, each of which contains two binding sites, but the internal pair is not required for transposition (Fig. 6) [79]. A third subgroup of the *Tc1/mariner* superfamily is named IR/DR, and has a pair of transposase-binding sites at the ends of the 200-250 bp long IRs (Fig. 6) [80]. The binding sites contain short, 15-20 bp direct repeats (DRs). This structure can be found in several elements whose inverted repeats are not significantly similar at the DNA sequence level, such as *Minos* and *S* elements in flies [81, 82], *Quetzal* elements in mosquitos [83], *Txr* elements in frogs [84] and at least three *Tc1*-like transposon subfamilies in fish [49], including *Sleeping Beauty*, a reconstructed transposon of the salmonid subfamily (Fig. 6) [85]. There

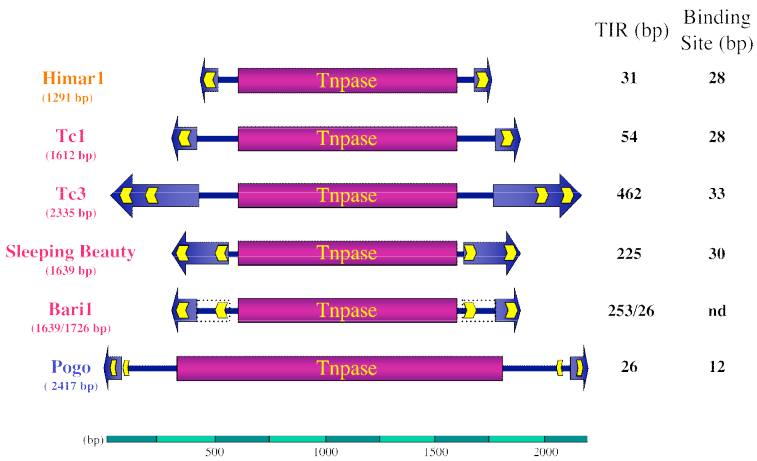


Figure 6. Structure of Tc1/*mariner* transposons. The central transposase genes (*tnpase*) are flanked by terminal inverted repeats (TIR) that contain binding sites for the transposase. TIRs come in different lengths and contain binding sites in different numbers and patterns in the *Tc1*/*mariner* superfamily. Dotted lines in *Bari* elements indicate that certain versions of these transposons have long inverted repeats. Actual or putative transposase binding sites are indicated as yellow arrows near the ends of the elements.

are two types of *Bari* elements in *Drosophila*; those that have short IRs similar to *Tc1*, and those that have IR/DR structure [86]. However, both types of *Bari* element have two putative transposase-binding sites flanking their transposase genes (Fig. 6). This suggests that it is not the long IRs *per se*, but the multiple binding sites for the transposase that are essential for the mobility of these elements.

Similar to the DNA-binding domains, the approximately 30 bp binding sites for *Tc1*-like transposases have a bipartite structure in which the 5'-part of the binding site is recognised by the homeo-like domain, whereas 3'-sequences interact with the paired-like domain of the transposase [47]. The binding sites for *mariner* transposase are also around 30 bp in length, supporting the hypothesis that these transposases also have bipartite DNA-binding domains. In contrast, *pogo* elements have binding sites of 12 bp within their short inverted repeats [87], consistent with the predicted single HTH motif in their DNA-binding domains. These binding sites are repeated either in direct or in inverted orientation at the ends of the element (Fig. 6), but it has not been determined whether they are required for the mobility of *pogo* elements. Taken together, the *Tc1*/*mariner* superfamily contains some simply structured elements in which the transposase gene is flanked by a pair of transposase binding sites, and more sophisticated ones with multiple binding sites that might impose some control over the timing and specificity of the transposition reaction.

1.3 Modes of transposition

The sum of molecular events involved in the movement of a transposable element from one chromosomal location to another is defined as **transposition.** There are two types of transposition reaction distinguished by whether the TE is replicated during the process or not (Fig. 7).

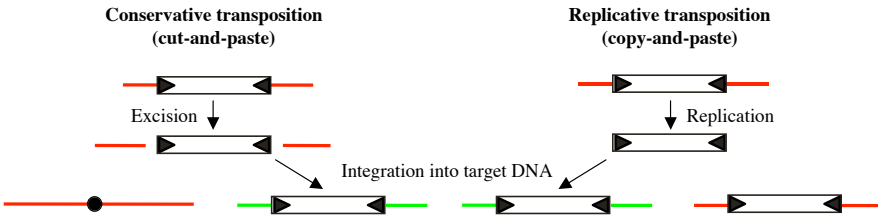


Figure 7. Schematic representation of the two major mechanisms of transposition. During conservative transposition, the element is excised from the donor DNA (red line), and integrates into a new target DNA (green line). The broken donor DNA has to be repaired by host factors, and this process can result in a small "footprint" (black dot) that marks the former presence of the element in that site. Replicative transposition requires amplification of the element either by replication or by copying of the element through transcription followed by reverse transcription. The amplified element gets inserted elsewhere in the genome.

During the vast majority of **replicative** (copy-and-paste) transposition events, the transposon does not get excised from its donor locus, but instead a copy of it is produced that subsequently inserts elsewhere in the genome (Fig. 7). Thus, replicative transposition leads to an increase in the copy number of the transposon within a genome. If the new copy is produced by transcription and subsequent reverse transcription of transposon sequences, the process is referred to as retrotransposition. The movement of retroviruses and retrotransposons is always of the replicative type, because it is the cDNA copy, not the original transposon, which is transposed. However, replicative transposition is not restricted to retroelements. For example, the *IS6* family and *Tn3* [42], and the complex DNA transposon, bacteriophage *Mu* [35], can also follow the replicative mode of transposition.

In **non-replicative** (also called conservative) transposition, the element is excised from a genomic locus and integrates to another through a so-called "cut-and-paste" mechanism (Figs. 7 and 8). In non-replicative transposition, the genetic information of the element is carried by DNA. The bacterial *IS10* [88], *Tn7* [89] and eukaryotic transposons including the *P* element [90], members of the *Tc1/mariner* family and the maize transposon *Ac/Ds* discovered by McClintock all use the cut–and-paste mechanism for their transposition [20].

In cut-and-paste transposition, amplification is not inherent to the transposition process itself; nevertheless, the copy numbers of DNA transposons also increase over time. Transposon amplification can occur when transposition takes place in the S-phase of the cell cycle. If a transposon is excised from an already replicated segment of the DNA, and reintegrates into a chromosome that has not been replicated, the process results in an increase by one copy of the transposon. If this event is followed by meiosis, two of the four germ cells have one more transposon copies compared to its parental cell [91]. Another way of increasing in copy number of non-replicative transposons was described for the *P* element [92] and the *Mos1 mariner* transposon [93]. After the excision of these elements the resulting gap in the donor chromosome can be sealed by a process called template-directed gap repair. This host repair mechanism uses the sister chromatid, the homologous chromosome or an ectopic site for refilling the gap created by the excised element.

1.3.1 The biochemistry of cut-and-paste transposition

Central to all transposition reactions are the excision and integration of a polynucleotide, therefore transposons execute polynucleotide transfer reactions. The transposase protein and the inverted repeats together engage in a series of molecular events that lead to the excision of the element from its DNA context and reintegration into a different locus, a process termed cut-and-paste transposition (Fig. 8). The transposition process can arbitrarily be divided into at least four major steps: 1) binding of the transposase to its sites within the transposon IRs; 2) formation of a synaptic complex in which the two ends of the elements are paired and held together by transposase subunits; 3) excision from the donor site by single-, or double-strand DNA cleavage; 4) reintegration at a target site and processing of the transposition product by host-encoded enzymes (Fig. 8) [44]. All transposition reactions involve DNA breakage and joining; the nature of the emerging transposition products depends on which strand of the DNA is cleaved and joined.

1.3.1.1 Transposon excision

The key process of all transposon excision is the exposure of the 3´-OH groups of the transposon ends, which will later be used at the strand transfer reaction for integration (Fig. 9) [94]. In the case of phage *Mu* and retroviral transposition the DNA cleavage involves only a single strand cut at each transposon ends. The vast majority of transposases, however, cleave both DNA strands of the corresponding transposon. During the excision of bacterial cut-and-paste elements, it is the first nick that generates the 3'-OH groups at the transposon ends. On the contrary, transposases of eukaryotic cut-and-paste transposons first generate a 5'-P on the transposon ends and the 3'-OH groups are exposed only as a result of the second strand cut [95]. In case of retroviruses, this process operates on the double-stranded cDNA of the element, and results in the cleavage of only two bases from the 3'-end of the cDNA [96].

Every DNA strand cleavage in all transposition reactions is a transposase- or integrase-catalyzed, Mg^{++}-dependent hydrolysis of the phosphodiester bonds of the DNA backbone, executed by a nucleophilic molecule. All the DDE recombinases catalyze similar chemical reactions [97], which begin with a single-strand nick that generates a free 3'-OH group. In the case of the first strand cleavage the nucleophilic molecule is H_2O [94]. During cut-and-paste transposition, nicking of the element is followed by the cleavage of the complementary DNA strand too. To catalyze second

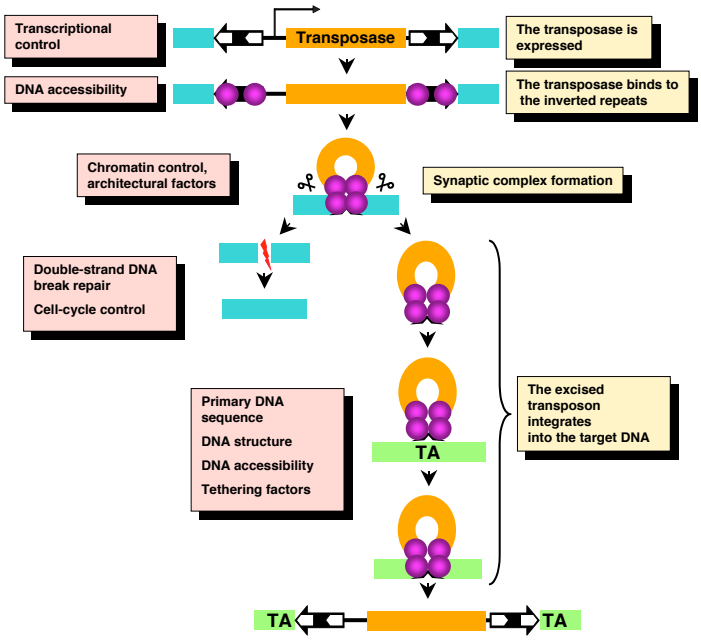


Figure 8. General mechanism and regulation of DNA transposition. The transposable element consists of a gene encoding a transposase (orange box) bracketed by terminal inverted repeats (solid black arrows) that contain binding sites of the transposase (white arrows) and flanking donor DNA (blue boxes). Transcriptional control elements in the 5'-UTR of the transposon drive transcription (arrow) of the transposase gene. The transposase (purple spheres) binds to its sites within the transposon inverted repeats. Excision takes place in a synaptic complex, and separates the transposon from the donor DNA. The excised element integrates into a new site (TA for *Tc1*/*mariner* transposons) in the target DNA (green box) that will be duplicated and will be flanking the newly integrated transposon. On the right, the various steps of transposition are shown. On the left, mechanisms and host factors regulating each step of the transposition reaction are indicated.

strand cleavage, DDE recombinases developed versatile strategies [98]. This cleavage can occur at different positions relative to the transposon ends. The position of 5'-cleavege of the second strand required for the liberation of the element occurs directly opposite to the 3'-cleavage site in V(D)J recombination [99] and for the bacterial *Tn10* element [100] (thereby generating blunt ended products). For *Tn7* the cleavage occurs three nucleotides toward the 5'-end of the element [44]. In case of the *Tc1/mariner* elements the non-transferred strand is cleaved a few nucleotides within the transposon (Fig. 9) (two nucleotides for the *Tc1* and *Tc3* elements [73, 78], and three nucleotides inwards the element in case of *mariner* [78, 101]). The double-strand DNA breaks (DSBs) generated by transposon excision are repaired either by the non-homologous end joining pathway (NHEJ), or by template-dependent gap repair [92, 93]. NHEJ generates transposon "footprints" (Fig. 9) that are therefore identical to the first or last 2-4 nucleotides of the transposon in *Tc1/mariner* transposition [101, 102]. In V(D)J recombination, the single-strand nick is converted into a DSB by a transesterification reaction in which the free 3'-OH attacks the opposite strand, thereby creating a hairpin intermediate [99, 103]. *Tn5* and *Tn10* transposons also transpose *via* a hairpin intermediate, with the difference that the hairpin is on the transposon and not on flanking DNA [100, 104].

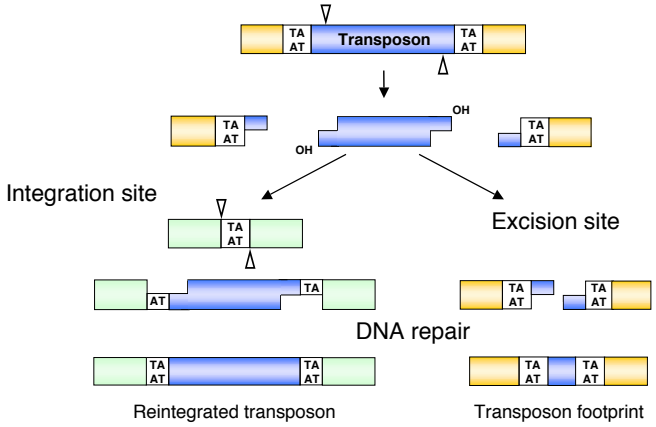


Figure 9. Cut-and-paste transposition of Tc1/*mariner* transposons. The element (black box) is removed from its original site with staggered cuts, which leaves some transposon nucleotides at the site of excision. The excised element reintegrates elsewhere in the genome at a TA target dinucleotide. Repair of the single stranded gaps of the integration site results in the duplication of the target TA. The excision site is predominantly repaired by non-homologous end joining, which leaves behind a transposon footprint.

1.3.1.2 Transposon integration and target site selection

The second step of the transposition reaction is the transfer of the exposed 3'-OH transposon tip to the target DNA molecule by transesterification (Fig. 9). Similarly to the initial DNA cut, the strand transfer is done by a nucleophilic attack. In this case, the 3'-OH groups of the already liberated transposon ends serve as a nucleophil that couples the element to

the target, without previous target DNA cleavage. As a result, the transposon ends are covalently attached to staggered positions: one of the transposon ends joining to one of the target strand, the other end joining to a displaced position of the target strand. Similarly to the initial strand cleavage, the strand transfer reaction does not need an external energy source, which suggest that it is the energy of the target phosphodiester bond that is used for the new transposon-target joint [94]. Although the initial excision and the strand transfer reactions are isoenergetic, many transposons such as *Tn7* and the *P* element, need molecules with high-energy bonds (ATP and GTP, respectively) for transposition *in vitro*. However, these molecules do not serve as an energy source, rather they only play regulatory roles [90, 105]. The final steps of transposition reaction are performed by host proteins. Due to the staggered way of insertion during the strand transfer step, there are short, single stranded gaps flanking the new integrant (Fig. 9). Host DNA repair factors then repair these gaps generating characteristic short direct repeats, also called target site duplications (TSDs), the hallmarks of transposition.

Most TEs do not integrate randomly into target DNA, and display some degree of specificity in target site utilization [106]. There is a wide spectrum of specificity in target site selection, hereby defined as the mechanism by which the specific DNA sequences of target sites are chosen. For example, the bacterial *Tn7* element is highly specialized to insert into a single sequence motif in the *E. coli* genome (discussed in more detail later in section 1.4.5) [106], whereas several other transposons, such as *Tn5*, can integrate at several locations even within a single gene [107]. Target selection may depend on primary DNA sequence and chromatin structure, which can influence target site utilization by modulating the accessibility of DNA. For some elements, such as *Tn7* and the *Ty1*, *Ty3* and *Ty5* retrotransposons in yeast, either element- or host-encoded accessory proteins play a role to locate a potential target area (discussed in more detail later in section 1.4.5) [108-111]. In other systems, including the bacterial transposon *Tn10* and the *Tc1* and *Tc3* transposons in *Caenorhabditis elegans*, target site selection is primarily determined by the transposase itself [112, 113]. Sequences responsible for target site selection of *Tn10* and retroviruses have been mapped

to the core catalytic domain of the transposase (or integrase) [112, 114], containing an evolutionarily conserved catalytic domain, the DDE domain.

The DDE domain is shared by a large group of recombinase proteins, including the *Tc1*/*mariner* superfamily, some bacterial IS/Tn elements, retroviruses, and the RAG1 immunoglobulin gene recombinase (see section 1.2.2.1.1.1) [20]. Several members of this family integrate fairly randomly, yet not all possible sites are utilized within a genome with equal frequencies. Despite the implication that the conserved catalytic domain is responsible for locating the target site, no common pattern of integration can be recognized on the sequence level. Therefore, assuming that there might be common features of target selection in the DDE family, it is an attractive hypothesis that structural properties of the target DNA will be among them.

The secondary structure of DNA is likely an important factor in the transposable element's insertional bias [106]. Indeed, secondary structural features influence integration of certain DNA transposons, including the bacterial elements *Tn3* [115], *Tn5* [107, 116, 117], *Tn7* [118] and *Tn10* [119, 120], *P* elements in *Drosophila* [121], retroviruses [122-125] and other retroelements [126, 127]. The insertional specificity for all of these groups is believed to exist because DNA at the site of integration forms an unusual or perturbed structure that allows better recognition by the transposition complex [118, 124, 126]. However, statistical significance of the structural features of transposon integration sites has not been considered in any of the previous analyses. *Tc1/mariner* elements insert into TA dinucleotides [20], and a transposon within this family, the *Himar1* element from *Haematobia irritans*, has already been implicated as having a structural preference for sequences in addition to the canonical TA [128]. However, the nature of these structural determinants and their relationship to the insertion site preferences of other *Tc1/mariner* transposons is unknown.

1.4 Regulation of transposition

De novo transposition events only become evolutionarily manifested, i. e. they only survive, if they can be stably transmitted to the next generation. Hence, restriction of transposition

events to the germline is thought to ensure that new transposon insertions are inherited by the next generation, and to avoid evolutionarily unproductive events in somatic tissues. A prototypic example for confining transposition events to the germline is provided by the *P* element transposon in *D. melanogaster*. Expression of the active *P* element transposase protein is restricted to the germline by tissue-specific, selective splicing of the transposase messenger RNA [129]. For the *I* factor, a *Drosophila* non-LTR LINE retrotransposon, expression of the ORF1 protein, which is essential for transposition, is limited to germline cells, where transposition occurs at high frequencies [130]. Expression of the intracisternal A-particle (IAP) LTR-retrotransposon was also shown to be restricted to the male germline in mice [131]. Similar, expression of RNA and proteins associated with the human L1 non-LTR retrotransposon preferentially occurs in germ cells [132-134]. Although the germline appears to be an attractive environment in which the products of transposition can be passed on to future generations, this strategy can be counteracted by protective mechanisms evolved by the host. For example, *Tc1*/*mariner* elements in the nematode *C. elegans* are active in the soma but silenced in the germline [135, 136] by RNA interference [137], which most likely protects the genome from heritable, transposition generated defects.

One definitive consequence of a completed transposition event is that a copy of the transposable element has been inserted into a new location somewhere in the host genome. This inherent quality of mobile DNAs to insert themselves into the host DNA constitutes a potential threat to the overall integrity of the host genome (by mechanisms discussed later in section 1.5.2). Therefore, it is of existential interest of the element to minimize damage to the integrity of the host genome. This is because an insertion event that in the worst case kills the host organism will consequently not be beneficial for the transposable element either, since its fate is intimately linked to the host. Therefore, transposons and their hosts have coevolved, and developed strategies that reduce the negative effects on the host but ensure proliferation of the element. On the molecular level, mobility of DNA-based transposable elements can be regulated by imposing constraints on transposition. One important form of transpositional control is represented by regulatory “checkpoints”, at which certain molecular requirements have to be fulfilled for the transpositional reaction to proceed. These

requirements can operate at any of the four different stages of transposition discussed in section 1.3.1, and can be brought about by both element-encoded and host-encoded factors (Fig. 8). Taken together, there is a great variety of mechanisms which put a limit on transpositional activity. The outcome of this regulation is that transposable elements move at very low frequencies in natural populations.

1.4.1 Transcriptional control of transposition

Expression of factors required for the transposition process is a limiting step in the transposition reaction, and therefore constitutes a major checkpoint in the transposition process (Fig. 8). Transposase expression from endogenous promoters requires host-encoded factors such as RNA polymerases and accessory proteins, and hence represents an important interface between the transposable element and the host organism. In general, endogenous promoters appear to drive transposase expression rather inefficiently, which is exemplified by the bacterial insertion sequences *IS911*, *IS21*, *IS30* and the maize *Spm* and *Ac* transposons [138-142]. Endogenous promoters are located in the IRs, which in addition contain the transposase binding-sites. This permits regulation of promoter activity by the transposase or its truncated derivatives, because transposase binding at the IRs my partially block access of transcription factors to the promoter. This has been shown for example for the bacterial insertion sequence *IS911* [140] and for the eukaryotic *P* element transposon [143]. For the *P* element, the transposase was shown to prevent assembly of the Pol II complex *in vitro* [143].

DNA methylation is an effective means of transcriptional gene regulation, and it was suggested to serve as a host-defense mechanism to restrict transposable element mobility [144]. Indeed, transcription of the transposase gene can be negatively affected by DNA methylation of the promoter region. For example, *IS10* and *Tn5* elements contain DNA methylation sites close to or overlapping the endogenous promoters; methylation of these sites leads to inefficient transcription [145, 146]. Similar findings were also obtained for eukaryotic transposable elements: transcription of the maize *Ac* transposon is abolished

when *Ac* DNA is methylated [147], the promoter of the LINE-1 (L1) retrotransposon is repressed when methylated [148-150], and the mouse endogenous retrovirus IAP gets transcriptionally activated in cells deficient in DNA methyltransferase [151]. Hypomethylation of DNA in *Arabidopsis* leads to activation of both DNA transposons and retrotransposons [152, 153]. In addition to transcriptional silencing, methylation of cytosine residues leads to deamination and thus results in rapid sequence divergence.

The activity of promoter sequences located within transposable elements are frequently found to be regulated by host-encoded proteins, which most likely allows to impose spatial and temporal regulation of recombinase expression. For example, for the L1 retrotransposon, a relatively large set of host-encoded proteins has been identified to bind to the 5´-untranslated region (UTR), which contains an internal promoter [154]. A particularly significant group of proteins are the high-mobility group (HMG)-box transcription factors. Two members of the HMG-box family of transcription factors have been found to interact with the 5´-UTR of L1: SOX11, which is a positive regulator of L1 transcription upon overexpression [155] and SOX2, which represses L1 protein expression [156]. Furthermore, the Ying Yang-1 (YY-1) transcription factor was reported to bind to the 5´-UTR and contribute to transcriptional regulation of L1 [157, 158], whereas the runt-domain transcription factor RUNX3 was shown to decrease L1 transcription [159].

1.4.2 Control of synaptic complex assembly during transposition

The mobility of transposable elements is restricted by the requirement to form nucleoprotein complexes, which are a prerequisite for the execution of the chemical steps needed for a successful transposition event. Such structures, also called synaptic complexes or transpososomes, contain the DNA of the transposable element, the element-encoded recombinase(s), in some cases the target DNA and in many cases host-encoded protein factors that aid the formation and the stability of these complexes. Formation of such higher-order nucleoprotein complexes, which contain all DNA sites and protein components needed for the transposition reaction, protects cells from aberrant transposition events. Therefore,

synaptic complex formation is an important checkpoint in transposition (Fig. 8). For example, transposition of the bacterial transposon *Tn5* and that of *P* elements and *mariner* elements in *Drosophila* can be regulated by repressor proteins, which are truncated or point mutant versions of the transposase polypeptide [160]. Such defective transposases can compete with wild-type transposase for binding sites located in the transposon ends. There are examples emerging where host-encoded proteins contribute to this process, making transposition a joint work of two partners. For example, *Mu*, a temperate phage discovered in *E. coli*, carries a linear, double-stranded genome, which can transpose into the bacterial genome. The structural organization of *Mu* is rather complex: each transposon end carries multiple MuA transposase binding sites, which show unequal orientation and spacing. At an early step of *Mu* transposition *in vitro*, when the left and the right transposon ends are brought together, the bacterial HU protein, a sequence-independent DNA-binding and bending protein, is required in addition to the MuA transposase [161, 162]. HU binds to the left *Mu* end [163], where it is thought to play an architectural role, which promotes synaptic complex assembly. In addition to the MuA transposase binding sites, the *Mu* transposon carries a transpositional enhancer sequence, which contains an integration host factor (IHF)-binding site. Like HU, IHF is a host-encoded, DNA-binding and bending protein [164], which stimulates transposition *in vitro* by acting on the enhancer sequence through the introduction of a sharp bend [165] that plays an important role in synaptic complex assembly. IHF was also found to modulate *Tn10* transposition both *in vitro* [166] and *in vivo* [167] by binding adjacent to the outside end of *Tn10*. The IHF-bent end of *Tn10* then wraps around the transposase to form an activated synaptic complex [168].

The concept of transpositional regulation through formation of nucleoprotein complexes can also be found in vertebrate systems. For example, the V(D)J recombination system requires host-encoded factors for synaptic complex assembly and successful excision. During V(D)J recombination, immunoglobulin and T-cell receptor genes are assembled from pre-existing gene segments, which are separated by so-called recombination signal sequences (RSSs) (also discussed later in section 1.5.2.1.1). Each RSS consists of conserved heptamer and nonamer regions, which are separated by a

relatively non-conserved spacer region of 12 or 23 bp (12-RSS and 23-RSS). Binding and cleavage by the RAG1/2 recombinase is more efficient at the 12-RSS than at the 23-RSS, but both binding and cleavage of the 23-RSS can be enhanced by the HMGB1/2 proteins [169]. RAG1 physically interacts with HMGB1/2 [170], which facilitates binding of RAG1 to the 23-RSS by bridging the distance in the 23-RSS between the heptamer and the nonamer region through a sharp bent in the DNA [169]. HMGB1 has also been found to enhance integration of avian sarcoma virus (ASV) [171]. Another member of the HMG protein family, HMGI(Y) has been found to be associated with the viral preintegration complex of HIV-1 and murine leukemia virus (MLV), a large nucleoprotein complex competent for integration [172, 173]. Binding of HMGI(Y) to the viral cDNA aids to compact the viral DNA to form an integration-competent complex [173].

From the *Tc1*/*mariner* transposon superfamily, the best chracterized system in terms of synaptic complex assembly is the *Mos1 mariner* element originally discovered in *Drosophila mauritiana* [174]. The *Mos1* transposase binds differentially to the imperfect IRs of the element [175] with a higher affinity to the right end, where cleavage preferentially occurs. The *Mos1* transposase first cleaves the nontransferred strand (the 5'-end of the transposon) within a single-end complex before the two ends are juxtaposed to form a paired-end complex (PEC), in which cleavage of the transferred strand (the 3'-end of the transposon) occurs [176]. Transposase mutants that result in a reduced transposase-transposase interaction mobilize the element with reduced activity, suggesting that transposase-transposase interactions are required to form a catalytically active PEC [175]. Whereas crystallographic studies of the *Mos1* transposase suggest that the PEC contains the two DNA ends together with a transposase dimer [75, 177, 178], biochemical studies proposed that the two ends are brought together by a transposase tetramer [179, 180].

1.4.3 Regulation of transposition by chromatin

Formation of a catalytically active synaptic complex requires expression of the recombinase specific for the element and a DNA topology, which makes the element accessible for the

protein machinery required for catalysis. The eukaryotic genome is typically organized into either of two types of chromatin: euchromatin, a relatively relaxed chromatin structure, in which the DNA is packed less tightly and heterochromatin, a more inaccessible and highly condensed fraction of the genome. Heterochromatic regions carry characteristic features, which distinguish them from euchromatic DNA, such as dense cytosine-methylation (5-Me-C), hypo-acetylation of lysine residues in the N-terminal tails of histone H3 and H4 and methylation of specific lysine residues such as lysine 9 in histone H3. In contrast to euchromatin, which is largely composed of unique (protein coding) sequences, the DNA sequence of heterochromatin is usually repetitive and gene poor [181]. One class of repetitive sequences found in heterochromatic regions of different genomes are transposable elements, and therefore it is believed that the accumulation of transposable DNA sequences in heterochromatic regions provides a “safe” place, where the deleterious potential of these elements can be kept on leash [182]. Indeed, there is a strong correlation between chromatin structure and the activity of transposable elements. For example, insertion of reporter genes in or in close proximity to heterochromatin results in silencing of gene expression [183-186], suggesting that heterochromatin represses transcription of genes located within or nearby. Thus, recruiting transposable DNAs into heterochromatic regions may provide efficient silencing of transcription of element-encoded proteins, and thus provides genome stability. In addition to its repressive function on transcription, heterochromatin also exerts a repressive influence on recombination [187, 188]; hence, containing repeated sequences in heterochromatic regions may prevent irregular recombination and genome instability.

1.4.4 Regulation by cell-cycle and DNA repair processes

The gap phases G1 and G2 provide important checkpoints in the cell-cycle of proliferating cells, at which the presence of damaged DNA is detected, and sufficient time is allocated for repair prior to the onset of DNA replication and mitosis. DSBs represent the most hazardous type of DNA damage that, if left unrepaired, can lead to genomic instability. DSBs can be introduced by exogenous agents such as ionizing-radiation or chemicals, but also by

endogenous cellular processes such as DNA transposition or V(D)J recombination (Fig. 8). Different pathways have evolved that allow the efficient repair of DSBs. In NHEJ, the broken ends of DNA are rejoined without a requirement for a homologous template. In contrast, for homology-dependent repair (HDR), extensive homology is required between the region with the DSB and a template (usually a sister chromatid or a homologous chromosome). The two pathways act at different stages of the cell-cycle; NHEJ acting primarily during G1/early S [189] and HDR being active in late S/G2 [190].

V(D)J recombination is strictly dependent on the NHEJ pathway for the repair of RAG-mediated DSBs [191], because only these repair products yield new, potentially contiguous reading frames of immunoglobulin and T cell receptor genes. To confine V(D)J recombination to the G1-phase of the cell-cycle, the RAG2 protein, that constitutes the recombinase (transposase) together with RAG1, accumulates during G1, declines before the cell enters S-phase, and remains low throughout the rest of the cell-cycle [192]. Phosphorylation of RAG2 by the cyclin A-cdk2 complex shortens the half-life of the protein at the G1/S boundary, whereas overexpression of $p27^{Kip1}$, a negative regulator of cyclin A-cdk2, results in elevated levels of RAG2, G1-arrest and increased recombination [193]. Hence, cell-cycle regulated protein stability of RAG-2 confines V(D)J recombination to G1.

Many viruses have developed strategies to modulate the host cell-cycle machinery and cellular self-destruction mechanisms to maximize the chance for successful infection and the production of virus progeny. For example, the *vpr* accessory gene of HIV-1 blocks cellular proliferation at the G2 phase in various eukaryotic cells including T cells [194, 195], experimental cell lines such as HeLa or 293 cells [195, 196] or even yeast [197], suggesting that the molecular mechanisms leading to *vpr*-induced G2 arrest are highly conserved. In addition to changes in the state of phosphorylation and subcellular compartmentalization of key cell-cycle regulatory proteins, *vpr*-induced herniations in the nuclear envelope and defects in the nuclear lamina have been proposed to contribute to the cell-cycle arrest [198]. *Vpr* might also delay or prevent apoptosis of infected cells, thereby maximizing viral expression, and increasing the amount of virus each infected cell produces [199].

1.4.5 Avoiding insertional damage to host cell genes by site-specific transposition

Unlike viruses, transposons do not possess envelope genes, and hence lack an extracellular phase in their life-cycle. This makes their fate closely linked to the fate of the host cell, and may result in integration patterns less mutagenic to the cell. The higher the gene density of a genome, the higher the chance for transposable elements to insert into coding sequences, resulting in potentially fatal consequences to the cell. Significant fractions of genomes with a small proportion of coding regions and extensive intergenic regions can be composed of transposon-derived sequences (e.g., 45% of the human genome), in contrast to organisms having a small genome with high gene densitiy, such as yeast. Thus, another form of transposon regulation that evolved, especially in small genomes, is site-selective insertion of transposons into “safe” places in the genome (Fig. 8). *Ty* LTR-retrotransposons in *Saccharomyces cerevisiae* are structurally and functionally related to retroviruses. Integration of *Ty1*, *Ty3* and *Ty5* retrotransposons is tethered to certain sites in the genome by host proteins. The *Ty1* element shows a strong insertion preference for genes transcribed by Pol III. 90% of *Ty1* insertions can be found about 1 kb upstream of tRNA genes [200]. A second preferred integration area of *Ty1* is found upstream of the 5S RNA genes that are also transcribed by Pol III [201]. Targeting of this site by *Ty1* elements may thus depend on the same factors as targeting of the tRNA genes. Indeed, components of the Pol III transcription machinery were found to be required for targeting of *Ty1* [202]; however, other factors such as chromatin components, physical properties of DNA or subnuclear localization of the target may as well specify integration sites.

Ty3 integrates one or two base pairs upstream of Pol III transcription start sites. TFIIIB and TFIIIC are important factors for assembly of Pol III complexes at transcription start sites of Pol III-transcribed genes, and are also involved in the recruitment of *Ty3* [110]. Though TFIIIB is sufficient to target *Ty3*, TFIIIC orientates binding of TFIIIB to the TATA box [203], and weakly interacts with *Ty3* IN [204]. The *Ty5* element interacts with the host protein Sir4p [205], which targets insertions to heterochromatic regions of the genome such as telomers and silent mating locus [111]. Interaction of *Ty5* IN with Sir4p is mediated by its

targeting domain, a 6-amino-acid motif at the C-terminus of *Ty5* IN. Mutations within this domain abolish interaction between IN and Sir4p, and result in random integration of *Ty5* retrotransposons. Concordantly, random integration of *Ty5* is observed in cells deficient in Sir4p [205].

Targeting of a specific genomic site may be specified by primary DNA sequence recognized by specific DNA-binding domains. In addition, physical properties of the DNA such as kinks due to protein binding, triplex DNA or altered/abnormal DNA structures due to base composition may cause preferential binding of proteins or protein complexes at certain sites. For the bacterial transposon *Tn7*, both sequence- and structure-specific binding apply. The *Tn7* transposon encodes five different proteins: TnsABCD and E. Depending on proteins involved in the transposition process, either a particular DNA structure found during conjugation or a specific site in the bacterial genome is targeted [206]. During bacterial conjugation, TnsE seems to recognize DNA structures with recessed 3'-ends during lagging strand DNA synthesis, and directs integration of the transposon to this site. TnsD binds to a specific DNA sequence called *attTn7* in the 3'-end of the bacterial glutamine synthetase (*glmS*) gene in the bacterial genome, followed by insertion of the transposon several base pairs downstream of *glmS*. Binding of TnsD creates DNA distortion probably responsible for recruitment of TnsC, which in turn interacts with TnsAB promoting insertion of *Tn7* at *attTn7*. Importantly, *Tn7* inserts into the human homologue of *glmS* in *Escherichia coli* and test tube reactions [207], but *Tn7* transpositional activity in human cells has not been reported. Another, particularly interesting feature of the bacterial *Tn7* element that it does not insert into DNA that already contains a copy of *Tn7*, a phenomenon called target immunity. Target immunity helps to avoid multiple copies of the element in the same DNA molecule, which might result in deleterious recombination between the two elements [206].

The eukaryotic microorganism *Dictyostelium discoideum* has a highly compact genome of 34 Mb with 76% coding regions and a surprisingly high transposon load of 10%. Transposons in *D. discoideum* have developed two strategies to avoid genotoxic insertion into coding sequences. One of these strategies is nested integrations of transposons forming

clusters. For example, the DIRS LTR-retrotransposon family shows no initial target site selectivity, but can be found in few clusters, made up of several copies of themselves [208], located in centromeric and telomeric regions of chromosomes. The other strategy is targeted integration into “safe” regions of the genome free from protein-coding sequences. This strategy is primarily used by non-LTR retrotransposons that insert up- and downstream of tRNA genes [209]. The non-LTR retrotransposons collectively called TRE (tRNA gene-targeting retrotransposable elements) can be divided into two groups: TRE5 elements preferentially integrate about 50 bp upstream of tRNA genes, whereas TRE3 elements favour integration 100-150 bp downstream to tRNA genes. An *in vivo* assay using a reporter gene tagged with a tRNA coding region showed targeted integration of TRE5 in the same manner as in a genomic context, indicating that targeted insertion of TRE5 is dependent on interactions with Pol III transcription factors [210]. Indeed, the ORF1 protein encoded by the TRE5 element was recently shown to interact with TFIIIB, suggesting a role of this interaction in targeting integration into tRNA genes [211]. Altogether, these observations suggest a general model wherein interactions between transposase/IN and DNA-bound proteins mediate insertional target choice. In sum, the existence of transposable elements with natural targeting abilities raises promise that recombinase/transposase/IN proteins with target-selective insertion properties can be engineered.

1.5 DNA elements in natural hosts

1.5.1 The evolutionary life-cycle of DNA transposons

Phylogenetic relationships between very closely related *Tc1/mariner* elements are often inconsistent with those of their hosts [49, 212]. For instance, the closest relatives of a *mariner* subfamily in humans can be found in insects, worms and in a hydra species [213, 214]. It has been suggested that “horizontal transfer” accounts for the spreading of elements across distantly related phyla (Fig. 10) [84, 215]. Because TEs are not infectious, it is not exactly known how they can invade new genomes. Potential vectors of horizontal transmission include viruses, external and intracellular parasites [216-218]. Once a

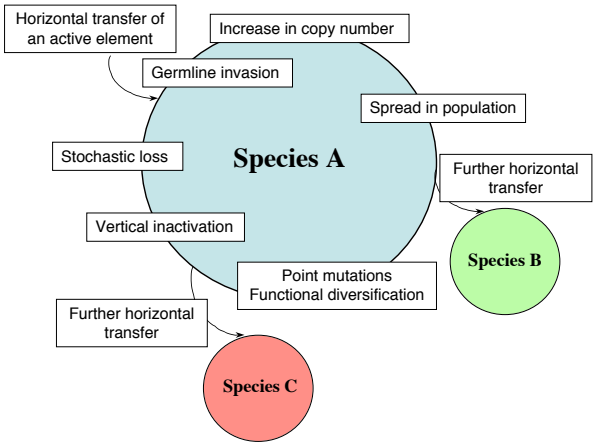


Figure 10. Evolutionary life-cycle of Tc1/*mariner* elements in natural hosts. The main events of the life-cycle are depicted (for details see text). The cycle was proposed to describe the evolution of *mariner* elements (Hartl *et al*., 1997), but is probably also valid for other DNA elements. Horizontal transfer of active transposons into new species can occur before or after functional diversification. Modified after Hartl *et al*., 1997 and Lampe *et al*., 2001.

transposon is transferred to a new host, it has to colonize its germline to persist in a population or, ultimately, in the entire species. At this initial phase, transposons can explosively amplify themselves (Fig. 10) [219]. However, transposons are not selected for function, and thus mutations may accumulate in them in a time-proportional manner (neutral evolution), resulting in partially or completely inactive transposon copies. This process is termed vertical inactivation (Fig. 10) [220]. In parallel, mutated transposase copies might become dominant negative regulators of transposition. Thus, with time, the rate of propagation slows down and finally, due to random genetic drift, transposons start to be extinct from their host genomes. The phenomenon is known as “stochastic loss” (Fig. 10) [31]. Therefore, in order to survive, transposons have to be horizontally transferred to new germlines and start their life cycle over again (Fig. 10). DNA transposons are believed to be transferred horizontally more often than retroelements, possibly because the endurance of DNA intermediates of transposition within cells offers a better chance for hitchhiking transfer vectors [221]. Indeed, in some retrotransposition reactions the RNA intermediate is directly reverse transcribed into the integration site [222], thereby offering little chance to be horizontally transferred.

Due to the above mechanisms, *Tc1/mariner* transposons are extraordinarily widespread in nature, but the vast majority of these elements are defective in all eukaryotic genomes. The active invertebrate *Tc1*/*mariners* were isolated from *Caenorhabditis elegans* (Tc1, [25] and *Tc3* [223]), from the *Drosophila* genus (*Mos1*, [224] and *Minos* [81]) and from the earwig *Forficula auricularia* (*Famar1*, [225]). The active *Himar1* element is a majority rule consensus of cloned genomic copies obtained from the horn fly *Haematobia irritans* [78]. However, extensive search for active vertebrate transposons has so far failed to yield an active vertebrate *Tc1*/*mariner*-like transposon.

1.5.2 Impact of transposons on host genomes: Mutations, genome size and the evolution of novel gene functions

When first colonizing new genomes, TEs are only parasitic sequences, however, over evolutionary times they can become integral components of genomes. Their effects can result directly from transpositional activity and TE-induced mutations, or because TEs represent a rich enzymatic and regulatory diversity that can result in the co-option of their sequences and enzymatic activities by the host [226].

A general feature of TEs is that they can replicate independently of the cellular replication cycle, and new copies can emerge at new locations in the genome. Thus, mobile elements can cause insertional mutagenesis if they land within a gene [227], but they can also lead to altered gene expression and genetic recombination. For example, insertions close to or within genes may lead to misexpression due to transcriptional up- or downregulation, insertions into introns may result in altered splicing patterns, whereas insertions into exons may give rise to loss-of-function mutations. For example, bacterial IS elements were identified as DNA insertions in the *E. coli gal* operon which cause highly polar mutations. In maize, the *Mutator* system can increase the mutation frequency by 50-fold over background. In *Drosophila*, it is estimated that 50-70% of all mutations are due to transposition, and the DNA transposon *Tc1* is the main cause of mutation in the nematode *Caenorhabditis elegans.* In addition to direct insertion into exon sequences of a gene, TEs can effect gene expression and regulation by integrating into non-transcribed or non-translated regions of genes [228-230]. For example, insertion between the core promoter and adjacent enhancer regions would increase the distance between these regions, and thus negatively affect promoter activity. In addition to insertional mutagenesis, cut-and-paste TEs can alter gene function by excising. It is because after the DNA break the host repair can rarely reproduce the sequence, as it existed before the integration. The excision can result in addition of new sequences or deletion of host sequences [231]. Another damaging aspect of TEs is that repeated, dispersed copies of homologous sequences can promote secondary rearrangements due to increased irregular recombination events that can lead to genome

instability [232], which can result in deletions, duplications and inversions. This potential of dispersed transposon copies to promote homologous recombination can be even more damaging to the genome than a *de novo* insertion.

TEs can also induce large-scale changes in the whole genome size. There is greater than 80,000-fold difference in size of the smallest and the largest eukaryotic genomes, however, the genome size is not correlated with organism complexity, which gives rise to the C-value paradox. In many plant and animal species and also in humans, abundant TEs account for the C-value paradox [11, 233]. A striking example of TE-induced genome restructuring is the programmed somatic excision of interstitial DNA segments in ciliates. These cells contain a macro- and a micronucleus. The genome of the transcriptionally active macronucleus consists of segments of the micronucleus, which is rearranged during development. The process involves extensive DNA excision and rejoining. TEs are major components of the eliminated DNA, and it has been proposed that invasion of these TEs contributed to the evolution of the nuclear excision process [234].

About 1 in 600 mutations in humans is estimated to arise from retrotransposon-mediated insertion. The major causative agent of endogenous genomic insertions is L1 [227]. An average human being has 80-100 retrotransposition-competent L1s, which belong to a particular subfamily of these elements in the human genome. Results also suggest that a relatively small number of very active L1s comprise the bulk of L1 activity [227]. A current estimate for transpositional frequencies in humans is that about 1 in 8 individuals harbor a new L1 insertion [14]. New, disease-causing insertions of L1 in humans were in fact the first retrotransposition events detected in mammals. These insertions occurred in the blood clotting protein Factor VIII, dystrophin, APC and β-globin genes.

Alu elements continue to amplify at a rate of about one insertion every 200 new births. New insertion events can lead to genetic disorders including hemophilia, neurofibromatosis, cholineserase deficiency, breast cancer and leukemia [232]. *Alu* element insertion is estimated to contribute to about 0.1% of human genetic diseases. The large number of *Alu* elements within the human genome also provides ample opportunity for

homologous recombination events between disperesed *Alu* repeats. These events can result in deletion or duplication of exons in a gene, and other chromosomal abnormalities. This mode of mutagenesis is estimated to account for 0.3% of human genetic diseases, including Fabry disease, Duchenne's muscular dystrophy, ADA deficiency and a variety of cancers.

One of the properties that distinguish TEs from other mutagens is that they are regulated both by themselves and by the host. Self-restraint of transposons has probably evolved to decrease the extent of damage to the host. However, the costs and benefits of TE movement can change during evolutionary times. Indeed, increased rates of transposition can even be selected when the host population is under stress [235]. If the tight regulation of TEs breaks down due to stress, their activity can potentially produce host variants with enhanced fitness. There are many examples of this phenomenon in plant evolution [236].

1.5.2.1 Transposons as a creative force

Transposable elements cannot only do harm, but also represent a creative force. In *Drosophila*, telomere maintenance is not brought about by telomerase, but by repeated transposition of two non-LTR retrotransposons, HeT-A and TART, into chromosome ends. The acquisition of new transposon insertions can donate regulatory elements to genes, or even lead to the evolution of new genes. L1 elements can carry non-transposon sequences into new places, a process that can contribute to "exon shuffling" and thus to gene evolution [14]. This is because L1 transcription can read through the native transcription termination site of the element into flanking genomic sequences. It is estimated that about 0.5-1% of the human genome may have been generated by L1-mediated transduction of 3'-flanking sequences [237]. The L1 retrotransposition machinery can also mediate reverse transcription and genomic insertion of host gene mRNAs, resulting in processed preusogenes. Some of these insertions can give rise to functional processed genes.

1.5.2.1.1 “Domesticated”, transposase-derived cellular genes

Our perception of the selfish nature of TEs has considerably evolved during the past two decades as a result of increasing numbers of studies that described the capacity of these elements as an important force in the evolution of gene regulation and in the creation of genetic novelty. Indeed, the literature describes several examples of TEs that donated promoters or enhancer sequences to host genes, as well as their contribution to provide alternative splice sites, polyadenylation sites and *cis*-regulatory sequences (reviewed in [16, 17]).

Another consequence of the intimate relationship between transposon and host genome is the creation of chimeric genes, which can in some cases give rise to a functional protein. In *Drosophila*, one particular insertion of *P* element has been shown to produce a chimeric gene encoding the DNA-binding domain of the *P* element and a functional domain of the target host gene [18]. Several genetic processes that lead to the formation of chimeric genes have been higlighed in plants. As an example, the alternative transposition of the maize *Ac/Ds* element from the *hAT* superfamily that involves the 5’- and 3’-ends of different elements has been shown to provoke the fusion of the coding sequence of two genes generating a functional chimeric gene and subsequently a new phenotype [19]. In rice, 3000 chimeric elements called Pack-MULEs that had captured >1000 gene fragments from different chromosomal loci have been detected [20]. However, the origins and the roles of these chimeric proteins remain enigmatic. Similarly, such transposon-induced rearragements of large-scale duplication and shuffling of coding sequences have been reported for other *Mutator* elements, *Helitrons* and *CACTA* transposons (reviewed in [21]).

The great contribution of TEs on the evolution of a protein coding region was fully appreciated recently with large-scale *in silico* studies performed on the vast amount of sequences available from model organisms, including human [22, 23]. Indeed, it has been reported that TEs or TE fragments have contributed to at least 4% of human protein-coding genes [24, 25]. The majority of TEs were found to be distinct exons recruited into coding regions by splicing. Thus, it appears that in many instances, TEs and host genome have

evolved a mutually beneficial relationship that balance TE survival and the evolutionary interest of the host.

The most striking beneficial contribution of TEs is illustrated by an evolutionary process referred to as “molecular domestication”, by which a TE-derived coding sequence gives rise to a functional host gene. Thus, domesticated genes represent stable functional components of the genome. Such transposon-derived genes were first identified as domesticated *P* elements in *Drosophila* [26] and further extended to plant and animal genomes, including human [27-29]. Preliminary sequence analysis of the human genome identified 47 TE-derived genes with a likely origin in up to 38 different transposon copies [2]. For instance, domesticated genes are known to have derived from almost all superfamilies of DNA transposons with the exception of *CACTA* and *Merlin* superfamilies.

Several criteria have been proposed to determine strong cases of DNA transposon-derived genes [30]. In contrast to the repetitive nature of TEs, domesticated genes exist as single copies in the genome, and orthologs are detectable in distanly related species. Structurally, these genes are devoid of the molecular hallmarks of transposition such as flanking IRs and TSDs. The protein products of domesticated genes are phylogenetically linked to transposon-encoded proteins. They assume important biological roles *in vivo* but, in general, they have lost their capacity to mediate transposition.

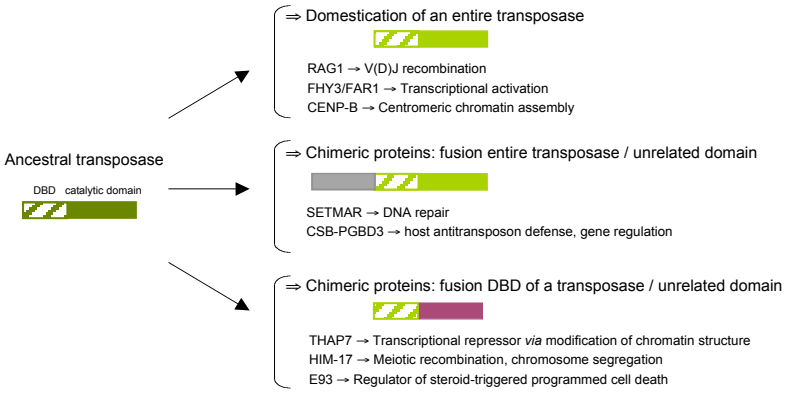


Figure 11. Structural diversity of domesticated proteins. Classical transposase proteins contain a DNA-binding domain (DBD) (hatched green rectangle) and a catalytic domain (green rectangle). Domestication events of a transposase can give rise to diverse structural proteins: domestication of an entire transposase gene, chimeric genes formed by an entire transposase domain and an additional functional domain, and chimeric genes formed by the DBD of a transposase and an additional functional domain. For each of these three cases, some domesticated proteins and their respective functional role(s) are provided as examples.

The increasing number of newly discovered domesticated genes clearly highlights their structural diversity. Some of these genes have emerged from the entire coding sequence of the transposase or exist as chimeric genes, in which the entire coding sequence

of the transposase has been fused to a preexisting functional domain (Fig. 11). Furthermore, the structural diversity is reinforced by the fact that many domesticated genes have retained only the DNA-binding domain of the ancestral transposon-encoded protein (Fig. 11).

V(D)J recombination, a site-specific recombination reaction in the immune system of jawed vertebrates is incontestably the most spectacular example that TEs can derive complex and crucial functions in the host. In this process, that occurs during lymphocyte development, preexisting V (variable), D (diversity), and J (joining) gene segments are rearranged to generate a large repertoire of T-cell surface receptor (TCR) and immunoglobulin molecules necessary for the recognition of diverse pathogens. The recombination event involves the *cis*-acting RSS sequences that flank each receptor gene segment and the RAG1 and RAG2 recombinase proteins (Fig. 12). Site-specificity of the recombination reaction is defined by the binding of RAG1 to the RSS. Typically, V(D)J recombination is subdivided into two stages, a cleavage phase and a joining phase (reviewed in [31]). The complex formed by the RAG1 and RAG2 proteins introduces DSBs in the DNA between the heptamer of the RSS and the neighboring coding DNA *via* a nick-hairpin mechanism. The reaction results in the formation of two hairpins at the coding end and two

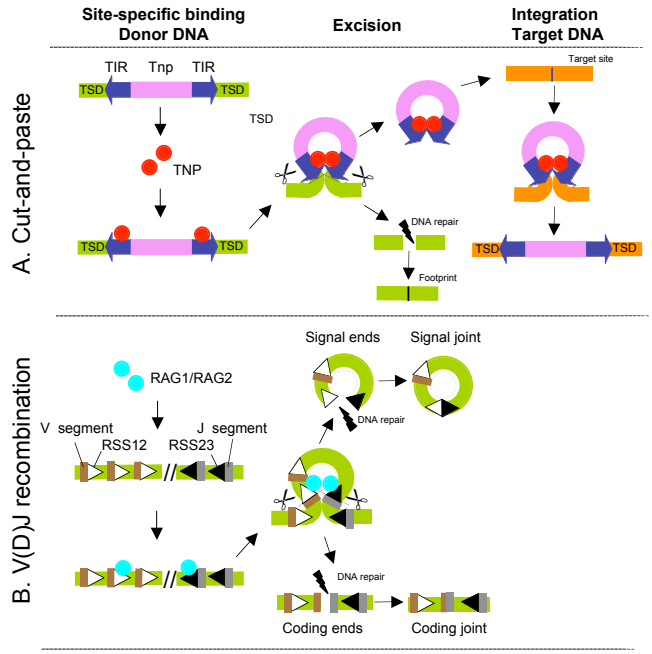


Figure 12. Functional homology between classical cut-and-paste transposition and V(D)J recombination. (**A**) Scheme of the classical cut-and-paste transposition process. An autonomous transposon consists of a coding region for the transposase (Tnp, pink rectangle), flanked on both ends by terminal inverted repeats (TIRs) (blue arrows). The TIRs are flanked by target site duplications (TSDs), characteristic to each transposon family. The transposase protein (red sphere) specifically binds to its recognition sequences at each end of the transposon. The transposase excises the transposon by cleaving the DNA at the ends of the TIRs following formation of a synaptic complex. The cellular DNA repair machinery seals the excision site, and generates a transposon footprint of different length characteristic to each transposon family. The transposase recognizes a target site, and integrates the transposon into the target DNA, upon which the target site gets duplicated. (**B**) Schematic representation of a VJ recombination reaction. The brown and grey bars indicate V and J coding segments, respectively. Each J segment is associated with an RSS23 (black triangles) and each V segment with an RRS12 (open triangles). Recombination initially requires specific binding of the RAG1/RAG2 recombinase to a 12/23 RSS pair. RAG1/RAG2 form a synaptic complex, in which the two DNA strands immediately adjacent to each RSS are cleaved and processed by a nick-hairpin mechanism. The double-stranded breaks in the coding DNA are repaired to give rise to coding joints. Signal ends are joined together to generate signal joints which are lost from the cell.

blunt signal ends by a transesterification mechanism. After opening of the hairpins, repair factors of the NHEJ pathway join the two coding DNA segments together to generate the mature receptor gene (coding joints), as well as the signal ends (signal joints) which are lost from the cell.

Mechanistically, the V(D)J recombination reaction shares significant similarities with the excision step of the cut-and-paste transposition process by which the transposon is excised from the donor-site DNA *via* double-strand breaks (Fig. 12) [32]. Moreover, V(D)J recombination produces a hairpin intermediate formed at the ends of the broken donor DNA similar to that described in *Hermes* transposition [33]. *In vitro*, purified RAG proteins have the capacity to transpose a piece of DNA flanked by two RSSs into a target DNA [32, 34]. In addition, RAG transposition events can occur at low frequencies in yeast and mammalian cells [35-37]. RAG-mediated transposition predominantly produces 5-bp TSDs upon insertion (reviewed in [38]).

The link between DNA transposons and V(D)J recombination has also been emphasized with the analysis of the structural features of the V(D)J recombination components [38]. The C-terminal domain of RAG1 including the DDE catalytic triad, the structure of the RSSs as well as the characteristic TSDs strongly support that RAG1 and the RSSs originate from a formerly active *Transib* transposon. Recently, a novel transposon called *N-RAG-TP* identified from the sea slug *Aplysia california* was found to encode a protein similar to the N-terminal part of RAG1 in vertebrates, which further supports the emergence of the V(D)J recombination machinery from transposons [39].

1.6 Transposons as genetic tools

Genome sequences of many model organisms of developmental or agricultural importance are becoming available. The tremendous amount of sequence data is fuelling the next phases of challenging research: annotating all genes with functional information, and devising new ways for the experimental manipulation of vertebrate genomes. TEs are known to be efficient carriers of foreign DNA into cells. Importantly, the transposase gene can be

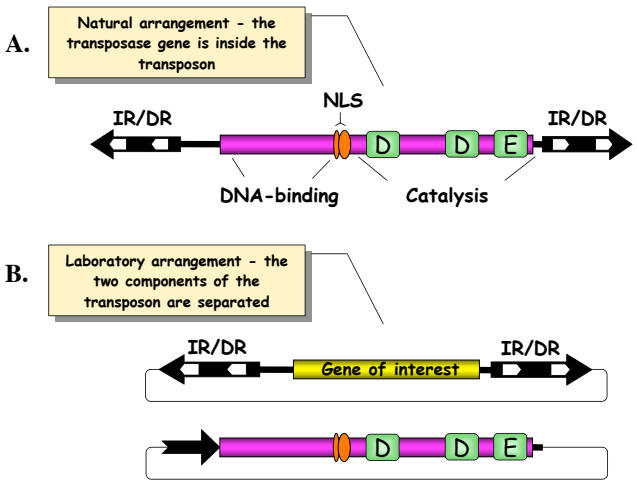


Figure 13. Transposon-basaed gene transfer system. (A) Structure of the transposon. The central transposase gene (purple box) is flanked by terminal inverted repeats (IR, black arrows) that contain binding sites for the transposase (white arrows). The transposase consists of an N-terminal DNA-binding domain, a nuclear localization signal (NLS) and a catalytic domain characterized by the DDE signature. **(B)** Gene transfer vector system. The transposase coding region can be replaced by a gene of interest (yellow box) within the transposable element. This transposon can be mobilized if a transposase source is provided in cells; for example, the transposase can be expressed from a separate plasmid vector containing a suitable promoter (black arrow).

physically separated from the IRs, and replaced by other DNA sequences (Fig. 13). These transposase-deficient elements can be mobilized if the transposase is provided *in trans*; thus, it is possible to stably integrate a desired DNA molecule into the genome using transposable elements as transgene vectors in a controlled manner [238, 239]. This represents the basis of utilizing transposable elements as transgene vectors; essentially any DNA of interest can be cloned between the IRs, and mobilized by supplying the transposase function in cells (Fig. 13).

P element and *Tc1* transposon-based vectors have been extremely valuable in exploring gene function in the invertebrate model organisms *Drosophila melanogaster* and *Caenorhabditis elegans*, respectively [240, 241]. However, efficiently manipulating vertebrate genomes with TEs was until recently not feasible. This is because, unfortunately, vertebrate model organisms seem to lack active, endogenous DNA transposons like *P* and *Tc1*; the only exception so far is the *Tol2* element in the medaka fish (*Orysias latipes*) [242]. To address this problem, a variety of invertebrate TEs, including *Tc1*/*mariner*s, were adopted for gene transfer in vertebrates. However, invertebrate transposons tend to have moderate activity in vertebrates [243], most likely due to restricting activities, or to the lack of specific cofactors (e.g. [244]). Molecular reconstruction of *Sleeping Beauty* (see later in section 3.1.1) represents a milestone in transposon-based technologies that expanded our abilities in genome manipulations, including insertional mutagenesis, transgenesis and gene therapy, in vertebrate organisms.

1.6.1 Insertional mutagenesis

Alongside with computational approaches and gene expression studies, mutational analysis is the most straightforward way of identifying gene function. One approach of creating mutants is to target and disrupt a gene of interest by homologous recombination; also referred to as reverse genetics. However, in spite of our growing acquaintance with protein domains, protein-protein interactions and molecular structures, our knowledge is yet inadequate to reliably predict the biological process that will be affected by knocking out a particular gene.

Another approach of obtaining mutant phenotypes is to introduce loss-of-function mutations into genomes of model organisms in a random and genome-wide fashion, termed forward genetics. Mutagenesis efforts have been carried out mainly based on X-ray irradiation and chemicals. However, it turned out that X-ray irradiation can cause a variety of chromosomal rearrangements affecting several genes simultaneously, which makes the identification of functions of individual genes difficult. Ethylnitrosourea (ENU) is a potent chemical mutagen that primarily introduces point mutations into DNA [245]. Two large-scale mutagenesis screens have been performed in zebrafish (*Danio rerio*) using ENU [246, 247], and it is routinely used in functional genetic analyses of the mouse genome [248]. The major advantages of ENU are easy use and highly efficient mutagenic rates in high-throughput screens. Nonetheless, a common disadvantage of these mutagenesis approaches is the time consuming and labor intensive molecular identification of the affected genes by positional cloning. While in some cases mutant phenotypes implicate certain signal transductional or developmental processes or genes, such a candidate gene approach can only be used in a fraction of the mutants. There are >20.000 genes in mammals [11], which necessitates the development of methods for rapid identification and functional annotation of genes.

An alternative approach of introducing mutations into the genome is insertional mutagenesis. Discrete pieces of foreign DNA can be harnessed to disrupt host gene function by creating random insertions in the genome. As opposed to chemical mutagenesis, inserting DNA fragments into genes simultaneously provide a molecular tag, which can be used to

rapidly identify the mutated allele. Viral and non-viral technologies have been devised to facilitate the penetration of transgenes through biological membranes. Non-viral methods, including naked DNA injection, electroporation, liposomes, “gene-guns” can be useful to introduce DNA into the cells, but chromosomal integration of the introduced DNA is still very inefficient. Moreover, a common drawback of the integration created by these techniques is the concatamerization of the foreign DNA at the insertion locus. Such events can facilitate chromosomal rearrangements [249], aberrant splicing, heterochromatin formation, gene silencing [250], and can interfere with cloning. The above problems can be circumvented by using retroviruses. The overt advantage of using viruses as vehicles for delivering DNA into cells is their capability to penetrate membranes and to catalyze the integration of single copies of the proviral DNA into chromosomes. However, retroviruses have pronounced preferences for their sites of integration [251], thereby limiting the spectrum of mutations. Moreover, retroviral vectors have limited packaging size and, due to their long terminal repeats, they can induce gene silencing [250] and ectopic reporter gene expression. Additionally, the observations coming from mutagenesis screens in zebrafish suggest that virus-based techniques are labor-intensive, and achieving high-throughput requires a large facility for screening [252]. Therefore, as an alternative approach to viruses, techniques of transposon-based whole-genome manipulation launched a new wave of research in functional genomics.

Cut-and-paste DNA elements have been routinely used for studying bacterial, fungal and plant genes in forward genetic screens. Similarly to retrovirus-based methods transposons can be utilized for insertional mutagenesis, followed by the easy identification of the mutant gene. However, DNA transposons have several advantages compared to the above approaches. For example, unlike proviral insertions, transposons can be remobilized *in trans*. Thus, instead of performing time-consuming microinjections, it is possible to generate *de novo* transposon insertions by simply crossing stocks transgenic for the two component of the transposon system (transposon and transposase). This scenario is especially useful when transposition events are directed to the germline of the experimental animal in order to mutagenise germ cells. Also, transposase expression can be directed to

particular tissues or developmental stages by a variety of specific promoters. Furthermore, remobilization of a mutagenic transposon out of its insertion site can be used to isolate revertants and, if transposon excision is associated with a deletion of flanking DNA, it can be used to generate deletion mutants. Since transposon are composed of DNA and can be maintained in plasmids, they are much safer and easier to work with than highly infectious retroviruses. Furthermore, timing of transposase activity is feasible by supplying the transposase in the form of DNA, mRNA or protein in the desired experimental phase.

When transposons are used in insertional mutagenesis screens, transposon vectors often comprise three major classes of constructs to identify the mutated genes rapidly (Fig. 14). These contain a reporter gene, which should be expressed depending on the genetic context of the integration. These vectors are only expressed if they land in-frame in an exon or close downstream to a promoter of an expressed gene. In *polyA traps*, the marker gene lacks a polyA signal, but contains a splice donor (SD) site. Thus, when integrating into an intron, a fusion transcript can be synthesized comprising the marker and the downstream exons of the trapped gene. *Gene traps* (or exon traps) lack promoters, but are equipped with a splice acceptor (SA) preceding the marker gene. Reporter activation occurs if the vector is integrated into an expressed gene, and splicing between the reporter and an upstream exon takes place. The gene trap and polyA trap cassettes can be combined. In that case, the marker of the polyA trap part is amended with a promoter so that the vector can also trap downstream exons, and both upstream and downstream fusion transcripts of the trapped gene can be obtained [253]. The above constructs also offer the possibility to visualize spatial and temporal expression patterns of the mutated genes by using *LacZ* or fluorescent proteins as markers.

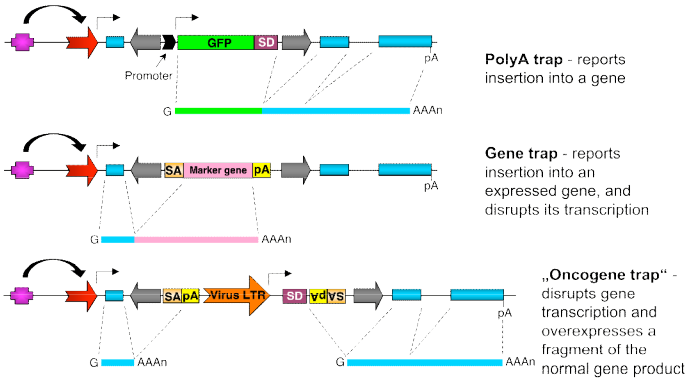


Figure 14. Transposon-based gene trapping vectors. On top, a hypothetical transcription unit is depicted with an upstream regulatory element (purple box), a promoter (red arrow), three exons (blue boxes) and a polyadenylation signal (pA). Major classes of transposon-based trapping constructs and spliced transcripts are shown below. Transposon inverted repeats are indicated by black arrows, different promoters are depicted as green arrows, SD and SA represent splice donor and slice acceptor sites, respectively.

The *Minos* transposase has been shown to mobilize nonautonomous *Minos* elements in mice by transposase expression in the oocytes using ZP3 [254] and in the lymphocytes using CD2 promoters [255]. *PiggyBac* has also been used in coinjection experiments in mice [256]. The activity of *Tol2* element has already been demonstrated in mouse embryonic stem (ES) cells [257] and *in vivo* in the mouse liver [258]. *SB* transposition is efficient in cells of different vertebrate classes in tissue culture [259, 260] and in somatic as well as germline tissues of fish [261-263], frogs [264, 265], mice [243, 266-270] and rats [271, 272] *in vivo*. Therefore, *SB* is a valuable tool for functional genomics in several model organisms [239, 273, 274]. In the mouse system, the classical way to set up mutagenesis screens is to generate double transgenic mouse lines were generated bearing chromosomally present transposons and an either ubiquitously [267-269, 275] or male germline-specifically [243] expressed transposase gene (Fig. 15). Segregating the transposition events by mating the founder males to wild-type females (Fig. 15) revealed that up to 80% of the progeny can carry transposon insertions [275], and a single sperm of a founder can contain, on average, two insertion events [267]. Additionally, subsequent studies elegantly showed that the germline of such a founder can harbor approximately 10,000 different mutations [268].

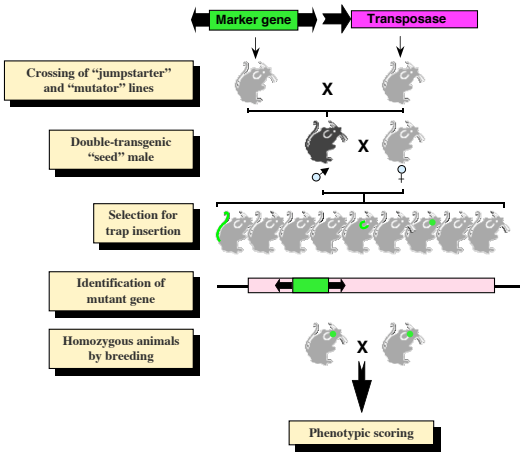


Figure 15. *In vivo* germline mutagenesis of the mouse with transposable elements. Breeding of "jumpstarter" and "mutator" stocks induces transposition in the germline of double-transgenic "seed" males. The transposition events that take place in germ cells are segregated in the offspring. Animals with transposition events need to be bred to homozygosity in order to visualize the phenotypic effects of recessive mutations. Mutant genes can easily be cloned by different PCR methods making use of the inserted transposon as a unique sequence tag.

All the vectors used in vertebrate insertional mutagenesis to date are versions of gene trapping insertional mutagenic constructs (Fig. 14), equipped with elevated mutagenicity and other useful properties. The mutagenicity of gene trap vectors is higher than that of simple insertional vectors, and they enable easy identification of the mutagenized gene by RT-PCR of composite transcripts made up by sequences of the insertional vector and the endogenous gene. Indeed, transposition of gene trap transposons identified mouse genes with ubiquitous and tissue-specific expression patterns, and mutant/lethal phenotypes

were easily obtained by generating homozygous animals [268, 269]. Similarly to the GAL4/UAS system in *Drosophila,* a conditional, tetracycline-regulated system has been shown to be applicable to TE-mediated insertional mutagenesis in mice [276].

As an alternative to the loss-of-function approaches, targeted over- and/or misexpression has been shown to be efficient in somatic tissues of mice using *SB*. Viral enhancer-promoter elements incorporated into *SB* vectors (Fig. 14) were shown to be useful to induce cancer in experimental animals [277, 278]. These screens can also capitalize on TEs with an intronic preference of insertion, such as members of the *Tc1* family. In order to devise customized screens for cancer development, a current approach is pointing towards establishing mouse lines conditionally expressing the transposase [279]. One approach is to express the transposase from tissue-specific promoters. The second is to generate a Cre recombinase-inducible transposase allele, and take advantage of the many existing Cre strains to induce mutagenesis in specific tissues in mice [279].

1.6.2 Transgenesis

The other major field of applications of transposon-based technologies is somatic and germline transgenesis. Transposon-based technologies can be exploited for gene transfer in cultured cells (Fig. 16). Once integrated, transposase-deficient nonautonomous transposons are stable in the absence of the transposase. Transposons can be harnessed to integrate plasmid-based siRNA expression cassettes into chromosomes to obtain stable knockdown cell lines by RNA interference (see also later in section 3.2.2) [280, 281]. Also, TEs hold potentials for generating transgenic model organisms, or animals of agricultural and biotechnological importance.

Classical methods to express foreign genes in vertebrates rely on microinjection of nucleic acids into oocytes or fertilized eggs. Two main drawbacks of these approaches are the low rates of genomic integration, and that the injected DNA generally integrates as a concatemer. Both drawbacks can be circumvented utilizing transposition-mediated gene delivery, as it can increase the efficiency of chromosomal integration and

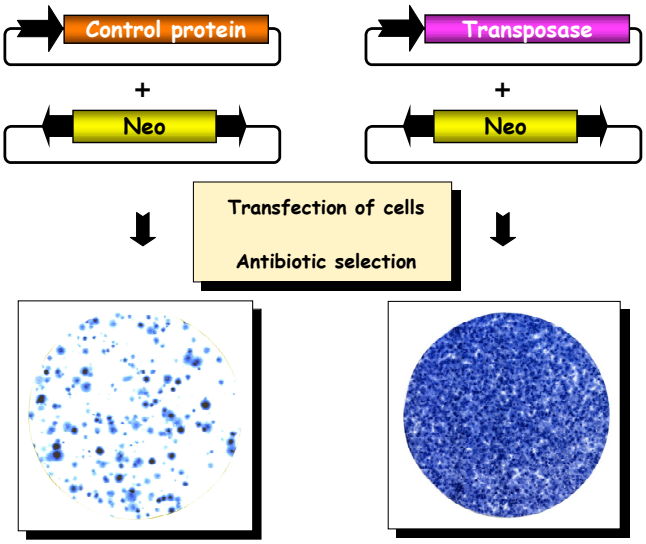


Figure 16. Transposition in tissue culture. The transposon containing a selectable antibiotic resistance gene (neo) is transfected either with or without a transposase-expressing helper plasmid. Transfected cells are placed under antibiotic selection. The dramatic increase in the number of resistant cell colonies in the presence of transposase is the result of transposition of the element from the plasmid vector into chromosomes.

facilitates single-copy insertion events. Single units of expression cassettes are presumably less prone to transgene silencing than the concatemeric insertions created by classical methods. Retroviral vectors are also useful tools for the same purpose, but their integration pattern is potentially more mutagenic, due to their preference for the 5'-end of transcription units (reviewed in [282]). In case of transgenesis, a single-copy insertion away from endogenous genes is clearly desired. The insertional spectrum of *Tc1*/*mariner* elements satisfies this need the best, as these elements integrate randomly at the genome level, and do not show pronounced bias for integration into genes. Another particular problem concerning transgenesis is that founders that develop from the injected oocytes or eggs are predominantly mosaic for the transgene, because integration generally occurs relatively late during embryonic development. Therefore, in order to potentiate successful transmission of the transgene through the germline to the next generation, it is necessary to shift the window of integration events as early as possible. This can be facilitated by co-injection of engineered transposons with transposase mRNA. This method has been employed to generate transgenic zebrafish with *Tc3* [283], *Mos1* [284], *Tol2* [285] and *SB* [286] transgenic *Xenopus* with *SB* [264] and *Tol2* [287] and transgenic mice with *SB* [288-290]. The far end on the scale of transposition-based somatic gene transfer is human gene therapy. Indeed, a large body of work has already been done in mice investigating possibilities of transposon-based human gene therapy.

1.6.3 Transposons as vectors for gene therapy

Considerable effort has been devoted to the development of gene delivery strategies for the treatment of inherited and acquired disorders in humans. For effective gene therapy it is necessary to: 1) achieve delivery of therapeutic genes at high efficiency specifically to relevant cells, 2) express the gene for a prolonged period of time, 3) ensure that the introduction of the therapeutic gene is not deleterious. There are several methods and vectors in use for gene delivery for the purpose of human gene therapy [291]. These methods can be broadly classified as viral and non-viral technologies, and all have advantages and limitations, none of them providing a perfect solution.

Adapting viruses for gene transfer is a popular approach, but genetic design of the vector is restricted due to the constraints of the virus in terms of size, structure and regulation of expression. In addition, safety, immunogenicity and production issues hamper clinical progress [292, 293]. For example, onco-retroviral and lentiviral vectors are efficient at integrating foreign DNA into the chromosomes of transduced cells, and have enormous potential for life-long gene expression [294]. However, there are several other considerations including safety [295]; preferential integration of retroviral and lentiviral vectors into expressed genes [296] poses the risk of inadvertent oncogene activation and congruent development of cancer. In addition, the requirement of cell replication for integration limits the use of retroviral vectors to dividing cell types. Adenovirus vectors have been shown to be capable of in vivo gene delivery of transgenes to a wide variety of both dividing and non-dividing cells, as well as mediating high level transgene expression. However, adenoviruses lack the ability to integrate the transferred gene into chromosomal DNA, and their presence in dividing cells is short-lived. Whereas early generation adenoviral vectors still contained residual viral backbone genes that contributed to inflammatory immune responses, toxicity and short-term expression, the latest generation adenoviral vectors (so-called gutless of helper-dependent adenoviral vectors) do not contain any residual viral genes and hence have a significantly improved safety and expression profile compared to early generation adenoviral vectors [297, 298]. Nevertheless, even these latest generation adenoviral vectors

still activate the innate immune system, particularly in larger animals and in patients [299] by virtue of their interaction with antigen-presenting cells [300]. Although long-term transgene expression has been achieved in mouse models using gutless adenoviral vectors, expression is typically transient in larger animal models. Hence, repeated vector administration would be required to boost expression levels, but the induction of a humoral (and possibly also cellular) [301] immune response against the capsid proteins precludes vector readministration. Adeno Associated Virus (AAV) vectors have several potential advantages to be explored, including the ability to transduce both dividing and non-dividing cells and the potential for stable transgene expression, even in large preclinical animal models, including non-human primates. Limitations of AAV include low maximal insert size, preferential integration into genes, and the induction of chromosomal rearrangements at the site of insertion [302]. Moreover, AAV administration in patients has been associated with the induction of a possible cellular immune response directed against the processed AAV capsid antigens [303], leading to transient and acute hepatotoxicity and precluding long-term transgene expression [304, 305].

Problems associated with virus vectors have led to an emphasis on development of non-viral methods [306-309]. DNA condensing agents, liposomes, microinjection, electroporation and “gene guns” might be easier and safer to use than viruses. Advantages of non-viral systems include their reduced immunogenicity, no strict limitation of the size of therapeutic expression cassette and improved safety/toxicity profiles. In addition, non-viral vectors are easier and less expensive to manufacture; for example, plasmid-based vectors can be produced in bacteria such as *E. coli*. However, non-viral approaches have been suffering from inefficient delivery, lack of chromosomal integration and resulting transient transgene expression. Recent advances indicate that efficient, long-term gene expression can be achieved by non-viral vectors based on transposable elements.

TEs represent nonviral vector systems that possess the capacity to stably integrate into the genome, and thus provide long-lasting expression of transgene constructs in cells. *SB* is the most thouroughly studied vertebrate transposon to date, and it has been shown to

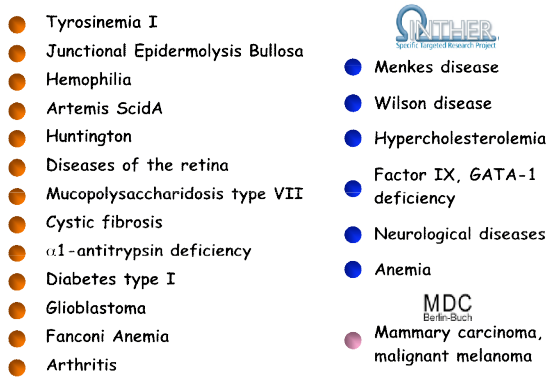


Figure 17. Current preclinical gene therapy experiments using *Sleeping Beauty*

provide long-term transgene expression in preclinical animal models (see [310] for a recent review). Since, unlike viruses, transposons are not infectious, they have to be actively delivered into the cell. Various methods for non-viral DNA delivery including hydrodynamic injection, electroporation, microinjection and complexing of the transposon components with polyethylene-imine (PEI), have been tested in conjunction with transposable element vectors (reviewed in [310]). Alternatively, transposon vectors can be delivered into cells by coupling the integration machinery of the transposable element to the cell infection machinery of a virus. Transposon-virus hybrid vectors delivering the components of the *SB* transposon system into cells by infection of adenovirus [311] or herpes simplex virus [312] have been developed.

The past couple of years have seen a steady growth in interest in applying the *SB* system for the treatment of a number of conditions including haemophilia A and B [266, 313, 314], junctional epidermolysis bullosa [315], tyrosinemia I [316], glioblastoma [317], Huntington disease [318] and type 1 diabetes (Fig. 17) [319]. In addition, important steps have been made towards *SB*-mediated gene transfer in the lung for potential therapy of α-1-antitrypsin deficiency, cystic fibrosis and a variety of cardiovascular diseases (Fig. 17) [320, 321]. Thus, the establishment of non-viral, integrating vectors generated considerable interest in developing efficient and safe vectors for human gene therapy [322-324].

1.6.3.1 The genotoxic risk of integrating gene therapy vector systems

About 23 % of gene therapy clinical trials have used retroviral and lentiviral vectors based on the murine leukemia virus (MLV), the avian sarcoma-leukosis virus (ASLV) or the human immunodeficiency virus (HIV) (http://www.wiley.co.uk/genmed/clinical/). However, with any vector that integrates into chromosomes in a nearly random manner comes the potential risk

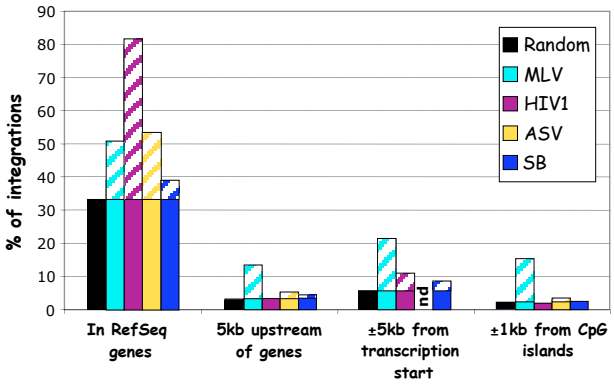


Figure 18. Genomic insertion preferences of integrating vector systems.

of insertional activation or inactivation of cellular genes [325]. MLV has been shown to have a strong tendency to insert into transcription start sites of genes [326], whereas HIV exhibits a bias towards insertions into transcription units but without bias to transcription start sites (Fig. 18) [327]. ASLV shows the weakest preference for insertion into active genes in this group, but still at a frequency higher than that of random integration (Fig. 18) [328].

Integration of the vector into a gene or its regulatory elements can knock out the gene, alter its spatio/temporal expression pattern or lead to truncation of the gene product (Fig. 19). Such genotoxic effects can have devastating consequences for the cell and the whole organism, including the development of cancer [325]. Such unfortunate events were observed in clinical trials using an MLV-based vector for gene therapy of X-linked severe combined immunodeficiency (SCID-X1). 9 out of 11 patients could be cured upon *ex vivo* transfer of a gene construct encoding the γ chain of the common cytokine receptor (γc) into autologous CD34+ bone marrow cells [329]. However, several years after the gene therapy treatment, two patients developed T-cell leukaemia. In both patients, development of the leukaemia was due to insertion of the transgene close to the promoter region of the LIM domain only 2 (*LMO2*) gene [330], and deregulated cell proliferation driven by retrovirus enhancer activity on the *LMO2* promoter. Since then, the number of severe adverse events in this

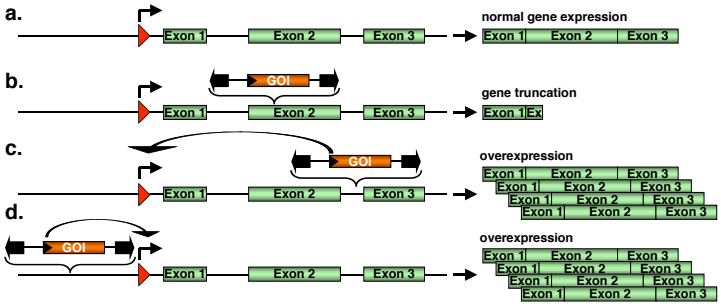


Figure 19. Possible mutagenic consequences of transgene integration in or close to a transcription unit. (a) The figure depicts a hypothetical transcription unit with a promoter (red arrow) and three exons. Normal gene expression results in physiological levels of the correctly spliced protein. **(b)** A gene of interest (GOI) carried by an integrating vector inserts into an exon, thereby resulting in a truncated gene product. The black arrows flanking the GOI represent retroviral long terminal repeats or transposable element terminal inverted repeats. **(c)** Transgene insertion occurs in an intron. An enhancer linked to the GOI upregulates transcription of the endogenous gene, resulting in overexpression and/or ectopic expression. **(d)** Transgene insertion occurs upstream of the targeted gene. An enhancer linked to the GOI upregulates transcription of the endogenous gene, resulting in overexpression and/or ectopic expression.

particular clinical trial has grown to four [331], and yet a new case has been reported in a separate SCID-X1 trial [332]. These incidents very drastically underscored the peril of insertional mutagenesis upon transgene integration. Taken together, potential genotoxic effects elicited by integrating viral vector systems give rise to serious risk for patients undergoing gene therapy. Targeted integration of the therapeutic gene to a “safe” site in the human genome would prevent possible hazards to the host cell and organism due to the problems mentioned above.

2. AIMS

2.1 Relics from the past: molecular biology of resurrected transposons and transposase-derived cellular genes in vertebrates

In this aim we set out to investigate the molecular biology of DNA transposons in vertebrate genomes. For this purpose we aimed at molecular reconstruction of active elements, mainly from the *Tc1/mariner* superfamily. The main question that we addressed was, first and foremost, can we reactivate „dead“ elements at all? Can we collect sufficient sequence as well as phylogenetic information of these transposons that would be required for their functional reactivation? Second, once we had these elements, we aimed at using them as experimental systems to address fundemental questions related to their mechanism of transposition, regulation by both element- as well as host cell-encoded factors and mechanisms and their interaction with their cellular environment. Finally, by using the reconstructed transposons as molecular references for transpositional activities, we set out to probe the functional properties of domesticated, transposase-derived cellular genes.

2.2 DNA transposons as a gene delivery platform for genetic manupilations in vertebrates

In this aim we set out to establish the resurrected transposons as molecular tools for genomic manipulations of vertebrate genomes. We specifically aimed at developing hyperactive transposase mutants as well as transposon-based vector system of enhanced utility for gene transfer in vertebrate cells. We further aimed at validating transposon vectors for delivering shRNA cassettes into cells for stable knockdown of gene expression. Finally, we aimed at improving the safety profile of transposon-based gene vectors by employing chromatin insulators to shield transcriptional activities of transgene cassette and by investigating technologies for target-selected transposon insertion for the purpose of human gene therapy.

3 RESULTS and DISCUSSION

3.1 Relics from the past: molecular biology of resurrected transposons and transposase-derived cellular genes in vertebrates

3.1.1 Molecular reconstruction of *Sleeping Beauty*, a *Tc1*-like transposon in fish, and its transposition in human cells (Papers I and II)

As discussed above, despite their wide distribution, all *Tc1/mariner* transposons isolated to date from vertebrates are transpositionally inactive. To address this problem, an ancestral *Tc1*-like element was reactivated from fish genomes. The molecular resurrection procedure involved the systematic removal of inactivating mutations by mutagenesis of an inactive transposase sequence (Fig. 20). The active element is a majority rule consensus sequence of several dead genomic copies of transposons from different fish species. Therefore, the engineered element, which was called *Sleeping Beauty* (*SB*), represents an archetypical sequence that was presumably active 10-15 million years ago [85].

The *Sleeping Beauty* transposase is a typical transposase encoded by *Tc1/mariner* transposons: an N-terminal, paired-like (PAI + RED) DNA-binding domain consisting of two HTH motifs is followed by a classical catalytic domain containing the DDE signature (Fig. 21). Partially overlapping with the RED subdomain in the transposase is a nuclear localization signal (NLS in Fig. 21), flanked by phosphorylation target sites of casein kinase II [49]. Phosphorylation of these sites is a potential checkpoint in the regulation of transposition. The NLS

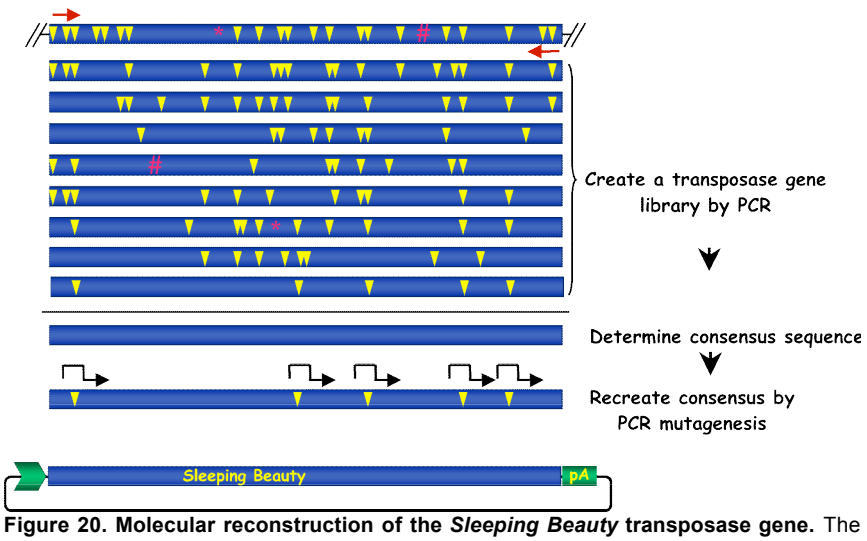


Figure 20. Molecular reconstruction of the *Sleeping Beauty* transposase gene. The strategy of first constructing an open reading frame for a salmonid transposase and then systematically introducing amino acid replacements into this gene is illustrated. Amino acid residues are typed black when different from the consensus, and their positions within the transposase polypeptide are indicated with arrows. Translational termination codons appear as asterisks, frame shift mutations are shown as #.

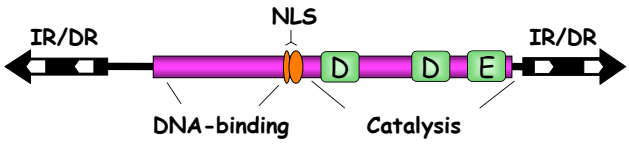


Figure 21. Schematic representation of *Sleeping Beauty*, a Tc1/*mariner* transposon. The terminal inverted repeats (IR/DR, black arrows) contain one or two binding sites for the transposase (white arrows). The element contains a single gene encoding the transposase (purple box). The N-terminal part of the transposase contains a DNA binding domain, followed by a nuclear localization signal (NLS). The C-terminal part of the protein is responsible for catalysis, including the DNA cleavage and rejoining reactions. The DDE amino acid triad is a characteristic signature of the Tc1-like transposases, *mariners* have DDD.

indicates that these transposons, unlike murine retroviruses, can take advantage of the receptor-mediated transport machinery of host cells for nuclear uptake of their transposases (Fig. 22). A characteristic GRPR-like motif (GRRR) between the two HTH motifs is similar to an AT-hook [45], responsible for minor groove interactions in the Hin invertase of *Salmonella* [56] and in the RAG1 recombinase of V(D)J recombination (Fig. 4) [62]. Within the catalytic domains of *Tc1*-like transposases, a conserved glycine-rich subdomain can be found [85]. The function of this subdomain is unknown.

The transposase gene is flanked by 200-250 bp long IRs that carry a pair of transposase-binding sites at the ends of each IR characterized by short, 15-20 bp DRs (Fig. 21). This special organization of inverted repeat is termed IR/DR [20, 80], and can be found in numerous elements in the *Tc1* transposon family, including the *Minos*, *S*, *Paris* and *Bari* elements in various *Drosophila* species [20, 81, 82, 86], *Quetzal* elements in mosquitos [83], at least three *Tc1*-like transposon subfamilies in fish [49] and *Txr*, *Eagle*, *Froggy* and *Jumpy* transposons in *Xenopus* [84, 333], as discussed in section 1.2.2.1.1.2. The spacing of about 200 bp between the outer and inner binding sites is conserved in all elements within the IR/DR group, but the actual DNA sequences are not similar, suggesting convergent evolution of the IR/DR-type repeats. The IR/DR group significantly differs from *Tc1* or the *mariner* elements that are more simple and have repeats of less than 100 bp and a single transposase binding site per repeat. All four binding sites within the IR/DR structure are required for *SB* transposition [259]. The four binding sites are not identical, the outer ones are longer by two base pairs. The inner DRs are more strongly bound by the transposase than the outer DRs [334, 335], and replacement of the outer DRs

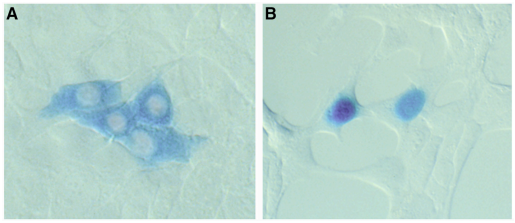


Figure 22. Identification of a nuclear localization signal in the *Sleeping Beauty* transposase. (A) Cytoplasmic localization of β-galactosidase. **(B)** Nuclear localization of β-galactosidase fused to the NLS of the *SB* transposase.

with inner DR sequences was found to abolish transposition [334]. This suggests that the unequal strengths of transposase binding and the positions of the DRs within the inverted repeats are required for ordered assembly of transposase-DNA complexes at the ends of the transposon that has a fundamental effect on the outcome of the transposition reaction. The IRs are not identical either; the left IR contains a sequence motif called the HDR, which resembles the 3'-half of the transposase binding sites [45]. A construct containing two left IRs transposes better than the wild-type transposon, but another version that has two right IRs has very poor mobility, indicating that the left and right IRs are functionally distinct [45].

3.1.1.1 The molecular mechanism of *Sleeping Beauty* transposition

3.1.1.1.1 Transcriptional activities of the *Sleeping Beauty* transposon (Paper III)

Some of the 5'-untranslated regions (UTRs) upstream of the initiation codon of the transposase gene contain promoter motifs [336], suggesting that they might have functions associated with control of transposition activity. However, previous studies did not reveal an internal promoter in the *Tc1* element; instead they showed that the elements are transcribed by read-through transcription from *C. elegans* genes [137]. The left IR is separated from the transposase coding sequence by a 160-bp stretch of DNA (Fig. 23) with no apparent function in the transposition reaction [337]. To assess the potential of the 5'-UTR (including the left IR and about 160 bp DNA of unknown function) of *SB* to drive transcription, the transposase gene was replaced by a luciferase reporter gene at the ATG start codon of the coding region, and transcriptional activities were measured in transient transfection experiments in human HeLa cells. Transcription driven by the 5'-UTR of *SB* is about 18-fold higher than transcription of a promoter-less sequence, about 4.6-fold higher than transcription driven by a TATA-box minimal promoter, and about 2.5-fold higher than transcription driven by the 5'-UTR of the closely related *Frog Prince* (*FP*) transposon (discussed later in section 3.1.5) (Fig. 23). The 5'-UTR drives expression of the *SB* transposase at a level sufficient to detect *SB* transposition in a colony-forming transposition assay in HeLa cells (Fig. 23). To test for directionality in promoter activity, the orientation of the 5'-UTR of *SB* was reversed relative to

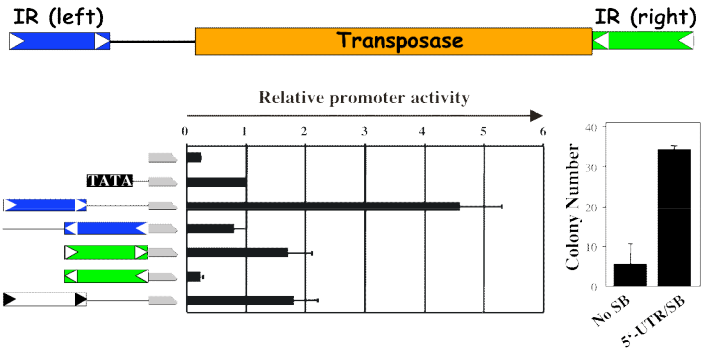


Figure 23. The UTRs of the *SB* transposon exhibit moderate, directional promoter activities. Transcriptional activities residing within the *SB* transposon. Promoter activities were determined by transient luciferase assays in HeLa cells. Activity of a minimal promoter (TATA-box) control was arbitrarily set to value 1. Transposon sequences flanking the transposase gene were placed in front of a luciferase reporter gene in two possible orientations (in the case of the 5'-UTR, the luciferase gene precisely replaces the transposase coding region). Blue box: left IR/DR of *SB*; green box: right IR/DR of *SB*; white box: left IR/DR of *Frog Prince*; small triangles in the boxes: transposase binding sites; black lines connecting the IR/DRs and the luciferase gene represent transposon sequences directly upstream of the transposase coding regions. (***right***) The 5'-UTR of SB can drive transposase expression at a level sufficient for the detection of chromosomal transposition events in cultured cells. A neo-tagged SB transposon plasmid (pT/Neo) was cotransfected together with an SB expression construct in which the transposase is expressed from the 5'-UTR of the transposon or with an empty cloning vector. The difference in numbers of G418-resistant cell colonies is evidence for transposition.

the luciferase gene, resulting in significant reduction of luciferase expression down to the activity of the TATA-box minimal promoter (Fig. 23).

In the natural arrangement of *SB* transposon components, the transposase coding region is followed directly by the right IR (Fig. 23). Thus, the 3'-UTR practically consists of the right IR. Since the IRs of the SB transposon share a significant sequence similarity, we included the right IR of SB in the promoter analysis. As shown in Fig. 23, the right IR can drive expression towards the inside of the element, but at lower efficiency than the 5'-UTR. In addition, similar to the 5'-UTR, the right IR appears to be unable to drive expression towards the outside of the element (Fig. 23). Convergent transcription of *SB* transposons raises the possibility for the formation of transposon-specific double-stranded RNA molecules that may serve as triggers of transposon regulation by RNA interference: an idea that remains to be tested by future investigations. The *SB* transposase physically interacts with HMG2L1, an HMG-box DNA-binding domain-containing protein that shares structural similarity with lymphocyte enhancer-binding factor 1 (LEF-1), sex-determining region Y (SRY) and SRY-related HMG-box protein 4 (SOX4) transcription factors [338].

In vivo interaction of the *SB* transposase with hemagglutinin (HA)-tagged HMG2L1 (HMG2L1/HA) was investigated using co-immunoprecipitation with an anti-HA antibody, blotting and hybridization with an antibody against the *SB* transposase. *SB* transposase was precipitated in lysates coexpessing HMG2L1, but not in lysates expressing HA-tagged SETMAR (a transposase-derived human protein [101]) used as a control (Fig. 24). Physical

interaction of the *SB* transposase with HMG2L1 suggests that this interaction may contribute to the regulation of *SB* transposition.

HMG-box transcription factors specifically bind their target DNA through their HMG-box domains, and regulate transcription of target genes (for review see [338]). Based on its predicted role in transcriptional regulation and its potential to interact with the *SB* transposase, we hypothesized that HMG2L1 may regulate transcription of the transposase gene. To investigate potential physical interaction of HMG2L1 with transposon DNA, *in vivo* chromatin immunoprecipitation (ChIP) was used following cotransfection of cells with plasmid DNA containing the 5'-UTR of *SB* and a vector expressing HMG2L1-HA. After chemical cross-linking, HMG2L1-bound DNA was precipitated using anti-HA antibody coupled to agarose beads, and amplified using a diagnostic PCR. As shown in Fig. 24, PCR products were only recovered in the presence of HMG2L1/HA and the 5'-UTR of *SB*, and were highly enriched in antibody-treated samples. These results suggest an interaction between HMG2L1 and SB transposon DNA *in vivo*.

Transcription from the 5'-UTR of *SB* is upregulated by the host-encoded factor HMG2L1 [339]. As shown in Fig. 24, expression of HMG2L1 upregulated transcription from the 5'-UTR of *SB* 10-15-fold, independent of the orientation of the 5'-UTR with regard to the luciferase reporter gene. Induction of transcription by HMG2L1 is specific to *SB* transposon DNA, since HMG2L1 failed to induce transcription of the promoterless and TATA-box minimal promoter-containing control constructs (Fig. 24). Specificity for the *SB* 5'-UTR is further evidenced by finding that HMG2L1 also failed to induce transcription from the right IR in either orientation and the 5'-

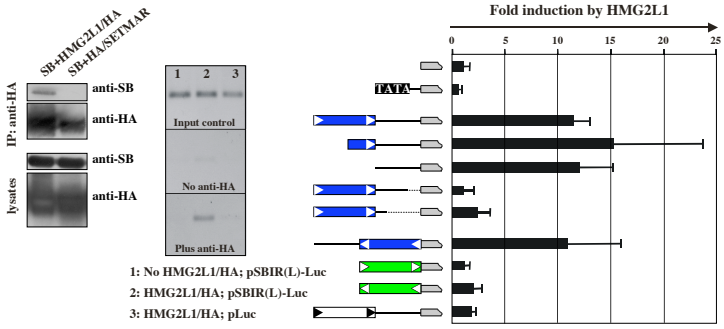


Figure 24. The human HMG2L1 protein physically interacts with functional components of *SB*, and mediates transcription from the 5'-UTR of the *SB* transposon. (***left***) *In vivo* interaction between the *SB* transposase and HMG2L1. Immunoblots of total extracts of HeLa cells coexpressing *SB* and HMG2L1/HA, or HA/SETMAR as control, were hybridized with anti-SB, and anti-HA antibodies, following immunoprecipitation (IP) with an anti-HA antibody. The Western blots on the cell lysates show proper expression of the test proteins. (***middle***) HMG2L1 interacts with the 5'-UTR of the *SB* transposon. Protein-DNA interaction was determined by *in vivo* chromatin immunoprecipitation. (***right***) Transcriptional regulation of 5'-UTR sequences by HMG2L1 in transient transfections in HeLa cells. Data show fold induction of transcription in the presence of HMG2L1 as compared to values measured in the absence of exogenously introduced HMG2L1.

UTR of the *FP* transposon. A 65-bp deletion immediately upstream of the luciferase coding region completely abrogated induction of transcription by HMG2L1, whereas deletions of IR sequences had no apparent effect on transcriptional activation (Fig. 24).

In summary, we identified HMG2L1 as a component for transcription of the transposase gene, and the above data suggest that its interaction with the *SB* transposase plays a role in this process. To determine the biological relevance of this interaction, the *SB* transposase was coexpressed with HMG2L1, and transcriptional activities associated with the 5'-UTR of *SB* were measured in transient luciferase reporter assays. Coexpression of the *SB* transposase with HMG2L1 not only abolished HMG2L1-mediated transcriptional activation, but apparently had a repressing effect on transcription by the 5'-UTR (Fig. 25). Furthermore, when the transposase was expressed in the absence of exogenously introduced HMG2L1, a considerable reduction in promoter activity became evident (Fig. 25), probably due to interactions of the *SB* transposase with endogenous HMG2L1 protein. To test whether transcriptional repression by the transposase requires primary binding of the transposase to its binding sites, transcriptional activities of a reporter construct lacking the left IR (which contains the binding sites) but retaining the ~160-bp intervening sequence between the left IR and the ATG codon of the transposase coding region were tested. Lack of the transposase binding sites did not affect the ability of the transposase to antagonize HMG2L1-induced transcription (Fig. 25). We conclude that transposase expression in the context of the naturally occurring transposable element is subject to negative feedback regulation, with the transposase acting as a transcriptional repressor. This model postulates a sensitive

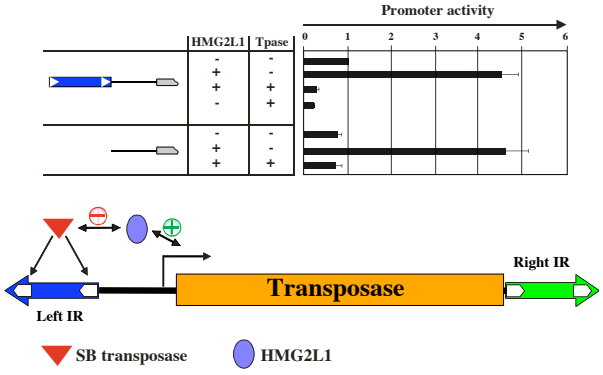


Figure 25. The *SB* transposase is a negative regulator of HMG2L1-mediated promoter activity. (***top***) Promoter activity of the SB 5'-UTR sequence in the presence of HMG2L1 and *SB* transposase. Values obtained in the presence of empty expression vectors only were arbitrarily set to value 1. (***bottom***) A model for transcriptional regulation of the *SB* transposase gene. In the wild-type, natural transposon, the central transposase gene (orange box) is flanked by untranslated regions (UTRs) that include the left and right inverted repeats (IRs, blue and green arrows, respectively) that contain binding sites for the transposase (white arrows). Arrow indicates the direction of transcription that is initiated within the 5'-UTR. HMG2L1 upregulates, whereas *SB* transposase downregulates transcription from the 5'-UTR.

balance in the regulation of transposase expression that is calibrated by transposase concentrations in the cell: low concentrations allow more transposase to be made, whereas high concentrations lead to shutting off transposase expression.

In addition to the control of transposase expression, interaction between the transposase and HMG2L1 might possibly regulate transcription of yet unknown cellular target genes, thereby affecting transposition. HMG2L1 has been shown to negatively regulate the Wnt/beta-catenin signalling pathway [340]; thus, it may be that the *SB* transposase/HMG2L1 interaction modulates transcription of Wnt/beta-catenin target genes, which in turn affects transposition. Future investigations will have to clarify if transposon regulation through such mechanism exists.

3.1.1.1.2 Specific DNA-binding by the *Sleeping Beauty* transposase

The paired-like DNA-binding domain of the *SB* transposase binds both DRs within the transposon IR/DR repeats (Fig. 26). Similar to the DNA-binding domain of the transposase, the binding sites also have a bipartite structure in which the 3'-part of the binding site is recognized by the PAI subdomain, whereas the 5'-sequences interact with the RED subdomain of the transposase [45]. Specificity of DNA-binding is predominantly determined by base-specific interactions mediated by the PAI subdomain [45]. The PAI subdomain also binds to the HDR motif within the left inverted repeat of *SB*, and mediates protein-protein interactions with other transposase subunits [45]. Thus, the PAI subdomain is proposed to have at least three distinct functions: interaction with both the DRs and the HDR motif, and transposase oligomerization. In cooperation with the main DNA-binding domain, the GRRR motif was shown to function as an AT-hook, contributing to specific substrate recognition [45]. Although part of the NLS is included in the

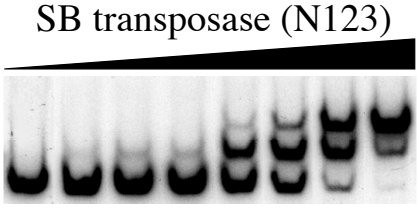


Figure 26. DNA-binding by the N-terminal DNA-binding domain of the *SB* transposase. Mobility shift analysis of the ability of the N-terminal DNA-binding domain (N123) of the *SB* transposase to bind to the transposon inverted repeats. Increasing amounts of transposase generate two complexes (complex 1 and complex 2) representing binding to one or two transposase binding sites within the terminal inverted repeats of the *SB* transposon.

RED subdomain, it does not appear to contribute to DNA recognition. Domain swapping experiments have shown that primary DNA-binding is not sufficient to determine specificity of the transposition reaction [45]. Zebrafish *Tdr1* elements are closely related to *SB*, but are not mobilized by the *SB* transposase. Comparison of the transposase binding site sequences of *SB* and *Tdr1* elements revealed main differences in the 5'-half of the DRs. This sequence is contacted by the RED subdomain, indicating that the function of the RED is to enforce specificity at a later step in transposition. Substrate recognition of *SB* transposase is therefore sufficiently specific to prevent activation of transposons of closely related subfamilies.

The spacing between the DRs is conserved in the IR/DR group, and decreasing the distance between the DRs has a negative effect on transposition [259]. The transposase does not bind the DRs with equal affinity, it preferentially binds the internal recognition sequences [334, 335]. Perhaps due to the two-base-pair difference in length, the helical phasing of the outer binding sites make transposase binding unfavored at these sites. The significance of this unequal affinity in binding is discussed in the next section.

3.1.1.1.3 Synaptic complex assembly and the role of multiple binding sites for the transposase

A uniform requirement among transposition reactions is the formation of a nucleoprotein complex, before the catalytic steps can take place. This very early step, synaptic complex assembly, is the process by which the two ends of the elements are paired and held together by transposase subunits (Fig. 8). *SB* transposition is controlled at the level of complex assembly [45]. The paired-like DNA-binding domain forms tetramers in complex with transposase binding sites [45]. The necessary factors that are required for synaptic complex assembly of *SB* include the complete IRs with four transposase binding sites, the HDR motif and tetramerization-competent transposase. These tetrameric complexes form only if all the four binding sites are present and they are in the in proper context. The HDR motif is important but not essential in transposition, and therefore can be viewed as a

transpositional enhancer that, together with the PAI subdomain of the transposase, stabilizes complexes formed by a transposase tetramer bound at the IR/DR. In contrast to *Mu* transposase, where the two specificities of binding to the enhancer and to the recombination sites are encoded in two distinct domains [341], the paired-like region of *SB* transposase combines these two functions in a single protein domain.

3.1.1.1.4 The role of HMGB1 in *Sleeping Beauty* transposition: Ordered assembly of synaptic complexes (Paper IV)

Differential interactions between the transposon and host-encoded factors may result in limitation of host range. The high mobility group protein HMGB1 is required for efficient *SB* transposition in mammalian cells [335]. HMGB1 is an abundant, non-histone, nuclear protein associated with eukaryotic chromatin, and has the ability to bend DNA [342]. *SB* transposition was significantly reduced in HMGB1-deficient mouse cells. This effect was complemented by expressing HMGB1 and HMGB2, but not with the more distantly related HMGA1 protein. Overexpression of HMGB1 in wild-type cells enhanced transposition, indicating that HMGB1 is a limiting factor of transposition. HMGs have low affinity to standard, B-form DNA, and interactor proteins need to guide them to certain sites [342]. *SB* transposase was found to interact with HMGB1 *in vivo*, and to form a ternary complex with the transposase and transposon DNA, suggesting that the transposase may actively recruit HMGB1 to transposon DNA via protein-protein interactions. Considering the significant drop of transposition activity in HMGB1-deficient cells, the role of HMGB1 in transposition is a critical one.

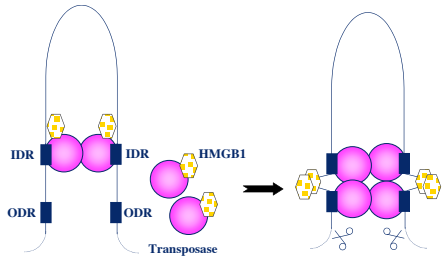


Figure 27. A proposed model for the role of HMGB1 in *Sleeping Beauty* synaptic complex formation. *SB* transposase (pink spheres) recruits HMGB1 (dotted hexagons) to the transposon inverted repeats. First, HMGB1 stimulates specific binding of the transposase to the inner binding sites (IDRs). Once in contact with DNA, HMGB1 bends the spacer regions between the DRs, thereby assuring correct positioning of the outer sites (ODRs) for binding by the transposase. Cleavage (scissors) proceeds only if complex formation is complete. The complex includes the four binding sites (black boxes) and a tetramer of the transposase.

HMGB1 was proposed to promote communication between DNA motifs within the transposon that are otherwise distant to each other,

including the DRs, the transpositional enhancer and the two IRs (Fig. 27). However, as mentioned above, physical proximity of the DRs is not sufficient for *SB* transposition; a highly specific configuration of functional DNA elements within the inverted repeats has a critical importance. As mentioned earlier, *SB* transposase preferentially binds the inner DRs within the transposon IRs. It was also found that HMGB1 enhances transposase binding to both DRs, but its effect is significantly more pronounced at the inner sites. It appears, therefore, that the order of events that take place during the very early steps of transposition is binding of transposase molecules first to the inner sites, and then to the outer sites. The pronounced effect of HMGB1 on binding of the transposase to the inner sites suggests that HMGB1 enforces ordered assembly of a catalytically active synaptic complex (Fig. 27). Indeed, interference with this sequence of events by replacing the outer transposase binding sites with the inner sites abolishes *SB* transposition [334]. This ordered assembly process probably controls that cleavage at the outer sites occurs only if all the previous requirements had been fulfilled. An assembly pathway similar to the one proposed for SB has been described for bacteriophage λ [343].

In summary, the IR/DR-type organization of inverted repeats introduces a higher level regulation into the transposition process. The repeated transposase binding sites, their dissimilar affinity for the transposase, and the effect of HMGB1 to differentially enhance transposase binding to the inner sites are all important for a geometrically and timely orchestrated formation of synaptic complexes, which is a strict requirement for the subsequent catalytic steps of transposition.

3.1.2 *Sleeping Beauty* transposase modulates cell-cycle progression through interaction with Miz-1 (Paper V)

Transposons have evolved together with the genome as an indwelling component. However, similar to viruses, transposons are, for pragmatic reasons, best viewed as molecular parasites that propagate themselves using resources of the host cell. We investigated differential gene expression in cells undergoing transposition, using microarray analysis. A

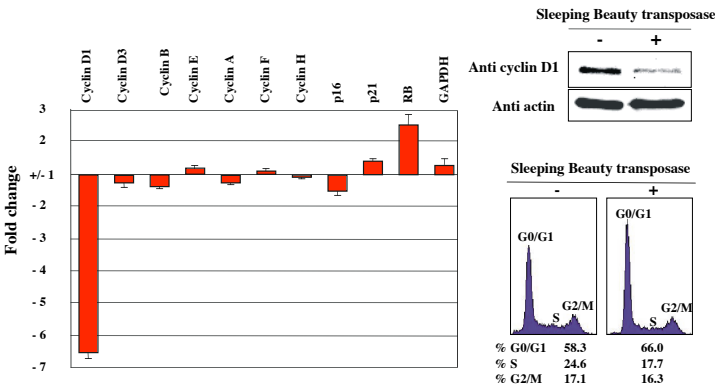


Figure 28. *Sleeping Beauty* transposase downregulates cyclin D1 expression and induces a slowdown of cell-cycle progression. (***left***) Differential expression of selected cell-cycle regulatory genes in human HeLa cells in response to the presence of the *SB* transposase. Shown are relative changes of transcript levels in transgenic cells stably expressing the *SB* transposase compared to wild-type cells. (***right top***) Cyclin D1 protein levels are reduced in *SB* transposase-expressing cells, as revealed by Western hybridization. (***right bottom***) Induction of a G1 slowdown by *SB* transposase in human HuH7 cells. Shown are FACScan profiles of *SB* transposase-expressing and control cells.

HeLa-derived transgenic cell line stably expressing the *SB* transposase, as well as control cells carrying only the integrated empty expression vector, were transfected with a GFP-tagged transposon plasmid, which can undergo transposition in transposase-expressing cells. In order to distinguish changes in cellular gene expression that are specific for transposition events from those that arose due to the presence of the transposase, we also included untransfected, transposase-expressing cells in the analysis.

Affymetrix HGU95A gene chips were used for hybridization, and Fig. 28 shows average changes in transcript levels of selected cell-cycle genes in the presence of either the transposase only or the transposase plus transpososon DNA, as compared to samples without the transposase. Transcriptional changes of most cyclins and cyclin-dependent kinase (cdk) inhibitors were within the range of the reference control gene GAPDH, and thus were not considered significant. In contrast, a handful of known cell-cycle regulatory genes did produce significant changes. Most significantly, in the presence of the transposase, there was a 4- to 7-fold decrease in cyclin D1 mRNA levels (Fig. 28). Interestingly, changes in cyclin D1 transcript levels did not seem to depend on the presence of transposon DNA, and were apparently due to the mere presence of the transposase. This observation indicates that the effect on cyclin D1 expression, and thus on the cell-cycle profile of transposase-expressing cells, is not associated with a cellular response to transposition-induced DNA damage. The effect of the *SB* transposase on cellular cyclin D1 levels was validated by Western hybridization (Fig. 28).

D-type cyclins are required for progression through the G1 phase of the cell-cycle [344]. Decreased cellular levels of cyclin D1 prevent cells from entering the S phase, resulting in cell-cycle arrest in the G1 phase [345]. Thus, because cells expressing the *SB* transposase have decreased levels of cyclin D1, they are expected to exhibit reduced growth associated with an increase in G1 cell population. Flow cytometry was applied to analyze the cell-cycle profiles of transposase-expressing and control cells. Cells expressing the transposase showed a 7- to 11% increase in the G0/G1 phase cell population as compared to control cells (Fig. 28). Accumulation of *SB* transposase-expressing cells in G0/G1 was accompanied by a decrease in S-phase cells, whereas the G2/M-population appeared to be largely unaffected (Fig. 28). These data suggest that the *SB* transposase can slow down the cell-cycle specifically in the G1 phase, at least in part due to downregulation of the cyclin D1 gene.

The Miz-1 transcription factor [346] was identified as an interactor of the *SB* transposase in a yeast two-hybrid screen [347]. Because Miz-1 was previously shown to act as a transcriptional regulator of the cyclin D1 gene [346], we wondered whether the effect of the *SB* transposase on cyclin D1 expression is mediated by Miz-1. We addressed this hypothesis by knocking down Miz-1 expression using RNA interference using a short hairpin RNA (shRNA). We investigated the effect of stable Miz-1 knockdown on the cell cycle profiles of SB transposase-expressing and control cells. As shown in Fig. 29, knockdown of endogenous Miz-1 abrogates the *SB* transposase-induced G1 slowdown in *SB* transposase-expressing cells, but does not influence the cell-cycle profiles of control cells. The effect of Miz-1 knockdown on cyclin D1 expression was also analyzed.

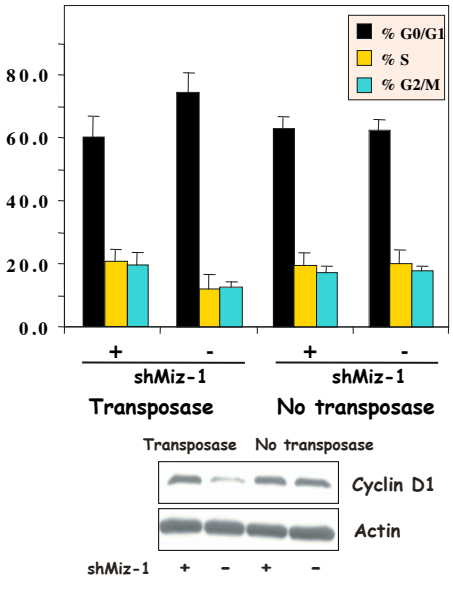


Figure 29. Expression of short hairpin RNA directed against human Miz-1 ablates *SB* transposase-induced G1 slowdown. (***top***) Percentages of cells, established in experiments done in triplicate, in different cell-cycle stages of exponentially growing, HuH7-derived, *SB* transposase-expressing and control cells, which have either the shMiz-1 construct or the empty vector stably integrated. (***bottom***) Cyclin D1 expression levels in *SB* transposase-expressing and control cells in the absence and presence of shMiz-1.

Whereas cyclin D1 expression is reduced in *SB* transposase-expressing cells in the absence of Miz-1 shRNA, cyclin D1 levels return to normal when the shRNA directed against Miz-1 is expressed in *SB* transposase-expressing cells (Fig. 29). We conclude that Miz-1 is required for transposase-mediated G1 slowdown as well as for downregulation of cyclin D1 expression.

The above results show that, mediated by the cell-cycle regulatory factor Miz-1, *SB* transposase downregulates the cyclin D1 promoter, resulting in slower cell growth. What can be the biological relevance of this process? To investigate whether a transposase-induced G1 slowdown has an impact on the efficiency of transposition, a quantitative transposition assay was performed in CHO-K1 cells that readily respond to serum-starvation in a reversible manner. Serum withdrawal resulted in an enrichment of a population of cells in the G1 phase, as determined by FACS analysis (Fig. 30). As shown in Fig. 30, there was an approximately 50% increase in transposition efficiency in serum-starved cells. We conclude that an artificially induced block in the G1 phase of the cell-cycle enhances *SB* transposition, suggesting that *SB*'s natural ability to slow down the cell-cycle is beneficial for the transposition process.

The likely biological significance of our finding is that by inducing a temporary G1 delay, the *SB* transposase potentiates the involvement of NHEJ to repair transposition-inflicted DNA damage [348]. Indeed, a delay in the G1/S transition and S phase progression by cell-cycle checkpoints is thought to facilitate DNA repair to avoid replication and subsequent propagation of potentially hazardous mutations. In eukaryotic cells, DSBs can be repaired by at least two pathways, HDR and NHEJ. The two pathways are complementary, but act at different

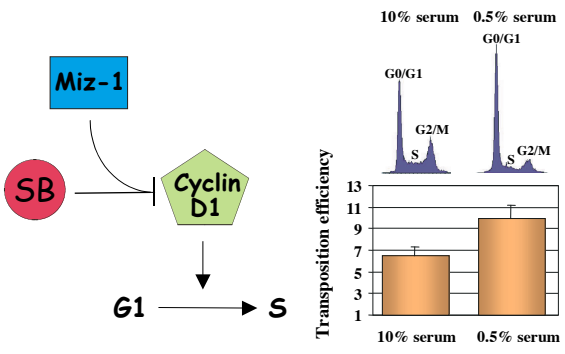


Figure 30. A temporary G1 arrest by serum-starvation enhances *Sleeping Beauty* transposition in CHO-K1 cells. (***left***) The *SB* transposase, through its interaction with Miz-1, downregulates cyclin D1 expression, which results in an inhibition of the G1/S transition of the cell-cycle. (***right***) *In vivo* transposition assay in G1-arrested (serum-starved) and exponentially growing CHO-K1 cells. Transposition efficiency represents fold increase in numbers of antibiotic-resistant cell clones in the presence *versus* in the absence of the transposase.

stages of the cell-cycle: NHEJ is preferentially active in the G1 and early S phases [189], whereas HDR is active in the late S and G2 phases [190]. Accordingly, there is increasing evidence for a correlation between the particular pathway used for the repair of transposon-induced DNA damage and the cell-cycle stage where recombination occurs. This is nicely illustrated by gene rearrangements through V(D)J recombination, which is tightly linked to the G1 phase of the cell-cycle and to NHEJ (as discussed in more detail in section 1.4.4) [191, 193]. It was recently shown that DSBs generated by *SB* transposition are preferentially repaired by the NHEJ pathway [348, 349]. Furthermore, the *SB* transposase physically interacts with the Ku DNA-binding subunit of DNA-PK, a key component of the NHEJ machinery [348]. Based on our findings, we propose a model in which *SB* transposase induces a cyclin D1-dependent G1 slowdown in proliferating cells through interaction with Miz-1, thereby ensuring that transposon-induced DNA damage is repaired by NHEJ. In nature, preferential use of NHEJ for the repair of transposon-induced DSBs might help avoid homologous recombination events between dispersed copies of transposable elements in the genome, thereby assisting the maintenance of genomic stability.

Other parasitic genetic elements have also developed versatile strategies to perturb the cellular machinery to maximize their chance for survival and propagation (Fig. 31). For example, infection by HIV-1 blocks cellular proliferation at the G2 phase, triggered by the HIV-1 gene product Vpr [350]. Herpes simplex virus [351], cytomegalovirus [352] and Epstein-Barr virus [353] slow down the G1/S transition phase to allow ample opportunity for expression of viral genes before the onset of cellular genomic replication (Fig. 31). For example, the Kaposi's sarcoma-

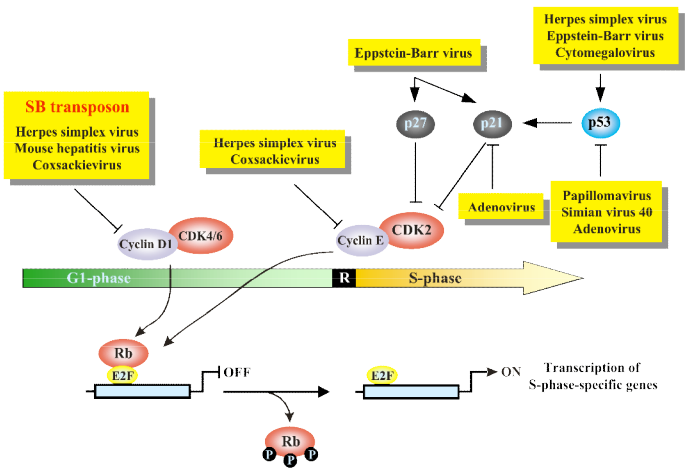


Figure 31. Model for the G1/S transition checkpoint and its regulation by selected viruses and the SB transposon. Cyclin D1-CDK4/6 and cyclin E-CDK2 kinases promote (arrows) the G1/S transition by phosphorylation (black circles) of Rb, which leads to the transcription of S-phase specific genes. The cyclin-CDK complexes are negatively regulated (perpendicular lines) by inhibitory proteins (grey circles). G1/S transition perturbations through interactions of various components of the G1/S transition machinery with gene products of selected viruses and the *SB* transposon are indicated.

associated herpesvirus K-bZIP protein physically associates with cyclin-CDK2, and downregulates its kinase activity. The result of this association is a prolonged G1 phase [354]. Intriguingly, mouse hepatitis virus replication was shown to induce a reduction in the amounts of cyclin-cdk complexes, resulting in insufficient phosphorylation of Rb, and an inhibition of the cell-cycle in the G1 phase (Fig. 31) [355]. Thus, overriding the normal cell-cycle program seems to be a shared strategy of parasitic genetic elements.

3.1.3 Regulation of *Sleeping Beauty* transposition by DNA CpG methylation (Paper VI)

The activity of transposable elements can be regulated by different means. CpG methylation is known to decrease or inhibit transpositional activity of diverse transposons (as discussed in section 1.4.3). However, very surprisingly, Yusa *et al*. showed that CpG methylation of the *SB* transposon produces elevated transpositional activity in mouse embryonic stem (ES) cells [356]. Chromatin immunoprecipitation experiments revealed that the hyperactive genomic donor sites have the characteristics of a heterochromatic structure. The *SB* transposase was found to colocalize with heterochromatin protein 1 (HP1), a well-established marker for heterochromatin, suggesting the transposase preferentially associates with heterochromatic DNA [357]. Based on these results, it was postulated that heterochromatin formation at the transposon donor site can upregulate *SB* transposition [356].

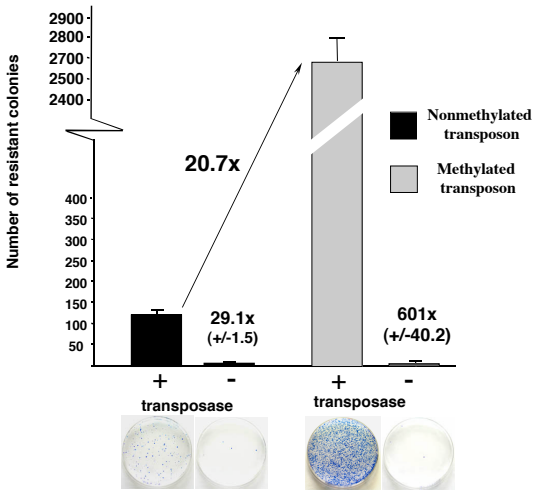


Figure 32. Effect of CpG methylation on *Sleeping Beauty* transposition. Transposon donor plasmids carrying a gene trap cassette harboring an antibiotic resistance gene were methylated at CpG sites in vitro, followed by transfection together with a transposase-expressing helper plasmid or a control plasmid into HeLa cells. Antiobiotic resistant cell colonies were counted and used to measure differences in transpositional efficiencies.

We addressed the question if transposition of other *Tc1/mariner* elements are also enhanced by CpG methylation. The members of this superfamily can be divided into two groups based on the size and structure of their IRs, as discussed in section 1.2.2.1.1.2 (Fig. 6). The first group – which contains *Tc1*, *Himar1* and *Hsmar1* and others – has IRs of short

length with only one transposase binding sites. The second group (the IR/DR group) – comprising e. g. *SB*, *FP* and *Minos* – has longer IRs (approximately 250 bp) with two transposase binding sites per IR. In order to investigate the unexpected response of *SB* transposition to CpG methylation, related transposable elements from both groups, i. e. *Tc1*, *Himar1*, *HsMar1*, *FP*, and *Minos* were tested for effects on transposition by CpG methylation and compared to *SB*. A significant increase of >20-fold in transposition of *SB*, *FP* and *Minos* was seen (Figs. 32 and 33), whereas transposons with simple repeats (*Tc1*, *Himar1* and *HsMar1*) showed no or nearly no difference in transposition between CpG-methylated or untreated transposons (Fig. 33).

At which step(s) of cut-and-paste transposition does the effect of CpG methylation manifest? The first molecular event that has to take place in order for transposition to proceed is binding of the transposase to ist binding sites within the transposon inverted repeats. Methylation of CpG sites can increase the DNA-binding affinities of several proteins; thus, a possible explanation for the methylation effect is a model in which *SB*, *FP*, and *Minos* transposases show increased binding to CpG-methylated DNA, whereas *Tc1*, *Himar1* and *Hsmar1* do not. By using an *in vivo* one-hybrid DNA-binding assay in cultured human cells we found that CpG methylation had no appreciable effect on the affinity of *SB* transposase to ist binding sites (not shown). Thus, a difference in DNA-binding by the transposase cannot explain the drastic effect of CpG methylation on transposition.

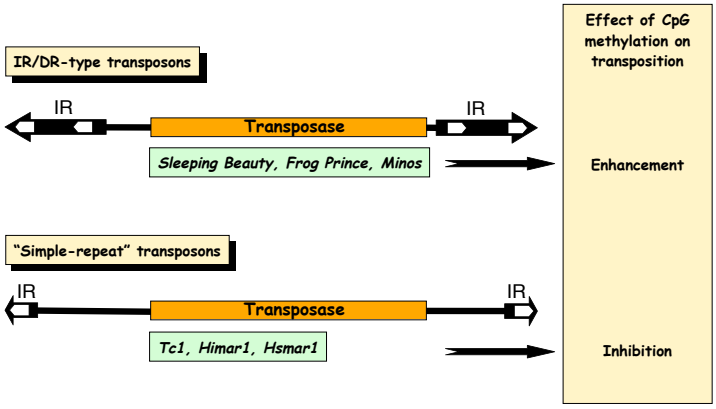


Figure 33. The enhancing effect of CpG methylation is specific for IR/DR-group transposons of the *Tc1/mariner* superfamily.

After transposase binding and synaptic complex formation, transposon excision is the first catalytic step of the transposition process. The group of IR/DR transposable elements showed increased excision after CpG methylation as compared to untreated transposon donor plasmids (not shown). Excision of the *FP* transposon was increased at least 16-fold by

CpG methylation. In contrast, the group of simple repeat elements showed reduced excision following CpG methylation.

CpG methylation of chromosomal DNA leads to formation of heterochromatin. To investigate chromatin packaging of CpG-methylated *versus* untreated transfected plasmids, a chromatin immunoprecipitation (ChIP) experiment was performed. Transfected plasmids were precipitated with either anti-acetylated histone H3 (AcH3) antibodies – which are used as marker for euchromatin –, or with antibodies against tri-methylated histone H3 lysine 9 (H3triMeK9) as hallmark for heterochromatin. We found that the fraction of plasmid in condensed chromatin is threefold increased by CpG-methylation (not shown). Furthermore, a quantification of the ChIP by transformation of the (still intact) plasmid-containing fractions into highly competent *E. coli* revealed that the CpG-methylated donor plasmids were equally precipitated by anti-AcH3 and anti-H4triMeK9 (53% *vs*. 47%), while the untreated donor plasmids where precipitated by 85% by anti-AcH3 and only 15% were found in the anti-H3triMeK9-fraction (not shown). We conclude that CpG methylation introduces a compact chromatin structure into transposon donor plasmids that enhances excision. Tight packaging of DNA and histones might bring DNA sequences and sites into close proximity. Since the CpG methylation effect is limited to transposons with an IR/DR structure, a compact chromatin structure might especially bring the inner and outer transposase binding sites of each IR together (Fig. 34). The proximity of these sites would probably assist the formation of transposase dimers as soon as they bind and encourage the formation of the whole synaptic complex subsequently. Thus, the data are

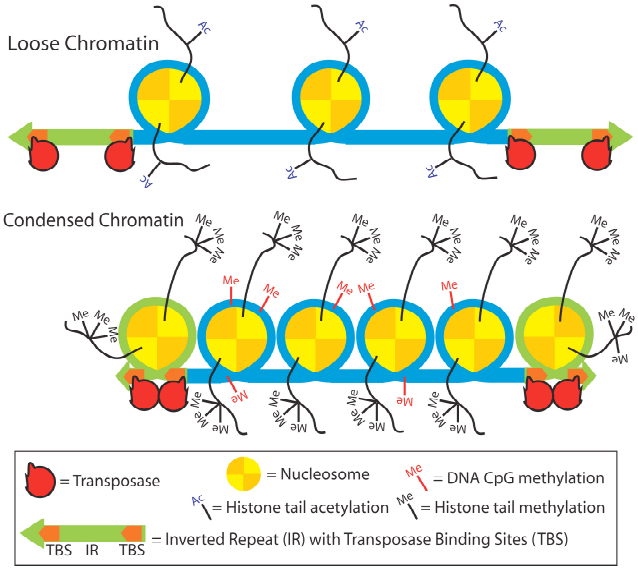


Figure 34. A model for the enhancing effect of a compact chromatin structure on *Sleeping Beauty* transposition. Euchromatin contains DNA wrapped around nucleosomes in a "beads-along-a-string"-like conformation (upper panel). Transposase subunits bound within the transposon inverted repeats (IRs) are separated by 166 bp DNA. Heterochromatin (lower panel), characterized by DNA CpG methylation and specific histone tail modifications, e.g. trimethylated lysine 9 of histone H3, features a higher histone:DNA ratio. Positioning of a nucleosome between the transposase binding sites (TBS) will shorten the distance between these sites, and could facilitate the formation of transposase dimers and subsequent assembly of the synaptic complex.

compatible with a model in which formation of heterochromatin at the transposon inverted repeats might facilitate the formation of a catalytically active synaptic complex (Fig. 34), and thereby enhances transposon excision. Thus, similarly to the effect of HMGB1, conformational changes of the excising transposon may greatly influence the efficiency of transposition.

3.1.4 Common physical properties of DNA affecting target site selection of *Sleeping Beauty* and other *Tc1/mariner* transposable elements (Paper VII)

In order to analyze *SB*'s insertion profile on the genomic level, transposon insertions were generated in human HeLa cells using an *in vivo* transposition assay [85], which is based on mobilization of a zeocin resistance gene (*zeo*)-marked SB element from extrachromosomal plasmids into chromosomes. The only level of selection in recovering transposition events in this assay was that the *zeo* gene within the integrated transposon has to be expressed. 138 insertion sites were identified and mapped on human chromosomes by computer analysis, using NCBI's human genome BLAST service. As shown in Fig. 35, although some chromosomes were hit more frequently than others, no clear preference is apparent for any chromosome, or for certain subchromosomal regions. The Y chromosome is not present in human HeLa cells, thus no hits were recorded. This observation indicates that most (if not all) chromosomes can serve as good targets for transposition. One insertion was found in the 3'-UTR region of a gene, 46 were mapped to intron sequences and one transposon landed in an exon. Thus, 48 out of the 138 integrations (35%) occurred in

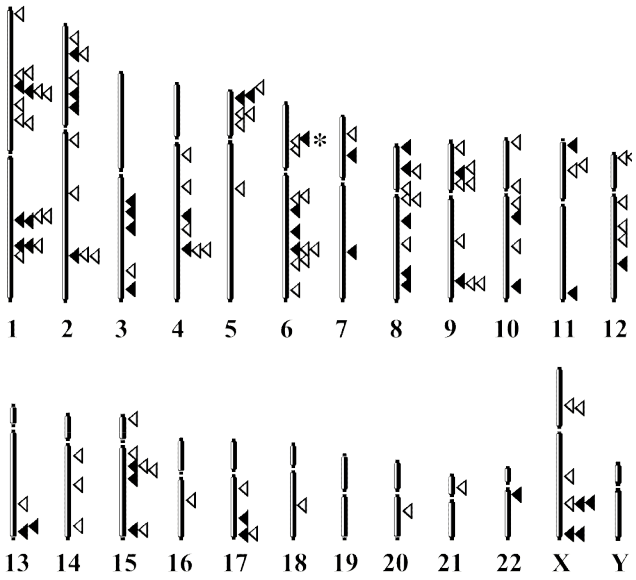


Figure 35. Mapping of *Sleeping Beauty* insertion sites on human chromosomes. Schematic representation of human chromosomes with 138 unique *SB* insertions. Insertion sites are marked with triangles, whereas filled triangles represent insertions in genes. Asterisk marks the single transposition event that occurred in an exon of a gene.

transcribed regions. Because about one third of the human genome is estimated to be transcribed [11], this frequency suggests no preference for or against insertion into genes. Such a subgenic distribution of *SB* is unlike that of *P* elements in *Drosophila*, which have the preference to insert into 5'-UTRs of genes, close to the transcriptional start site [358]. The predominant targeting of introns suggests that these sequences are hit more frequently either because their base composition makes them more attractive targets for the transposon, or because they tend to be significantly longer than exons or promoters [11], therefore representing a larger target into which a transposon can integrate. Eight insertions were found in repetitive sequences: five in other transposable elements such as *Alu*, L1 and MER1, and three in centromeric repeats. Three elements landed closer than 1 kb to a 5' region of a gene. Taken together, these results indicate a fairly random pattern of integration of SB elements in human chromosomes.

SB, like all other *Tc1*/*mariner* elements, integrates at TA dinucleotides, which occur approximately once every 20 basepairs, on average, in vertebrate genomes. We next investigated whether all TAs are equally good targets, or if there are other sequence determinants influencing *SB*'s target site selection. Integrated transposons were recovered from cells, and 71 chromosomal sequences flanking the integrated transposons were used to determine the DNA sequence of a consensus target site. All sequences were aligned at the canonical TA insertion site in the same orientation, relative to the transposon. We found six bases directly surrounding the insertion site forming a short, palindromic AT-repeat: ATA**TA**TAT, in which the central underlined TA is the insertion site (Fig. 36). Particularly conserved are the 5'- and 3'-most bases in the consensus, represented by an A and a T, respectively, in 66% and 70% of the target sites.

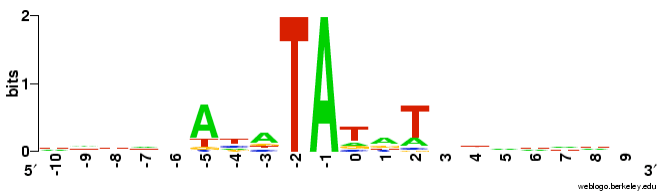


Figure 36. Consensus sequence of *Sleeping Beauty* insertion sites. Seqlogo analysis. Ten base pairs upstream and downstream of the TA target site were analyzed. The y-axis represents the strength of the information, with 2 bits being the maximum for a DNA sequence.

Having found a particular sequence into which *SB* preferentially integrates, we next asked whether integration sites have anything in common on the structural level. The

structural properties of DNA examined in this study included GC content, B-DNA twist, A-philicity, DNA bending and protein-induced deformability. B-DNA twist affects the tightness of the DNA coil and the ability of molecules to interact within the grooves of the DNA. These interactions allow DNA to serve as areas of binding for proteins [359]. A-philicity represents the propensity of DNA to form an A-DNA like double helix [360]. A-DNA has a wide and shallow minor groove that is believed to provide proteins easier access to form hydrogen bonds with bases within the DNA helix. Along with A-philicity, DNA bending can lead to changes in the width and depth of the major and minor grooves, affecting a protein's access to bases of the DNA [361]. Protein-induced deformability is the ability of DNA to change shape when in contact with a protein, which in turn affects the binding of other proteins or the action of the protein already bound [362].

Physical properties of a data set containing 58 sequences, each with 61 bases flanking the TA insertion site on each side, were analyzed to determine significant features of *SB* target sites other than the actual DNA sequence. Random DNA sequences from human chromosome 21 were analyzed for comparison with *SB* insertion sites. All random human sequences were also aligned at a TA dinucleotide, allowing us to evaluate physical properties of the DNA that were not due to the canonical TA target dinucleotide. At the 90% confidence level, all five physical properties deviated from the control data around the area of *SB* insertion (Fig. 37). Bendability deviated from the control data at the 95% confidence level, as evidenced by the number of positions that are significantly different from the

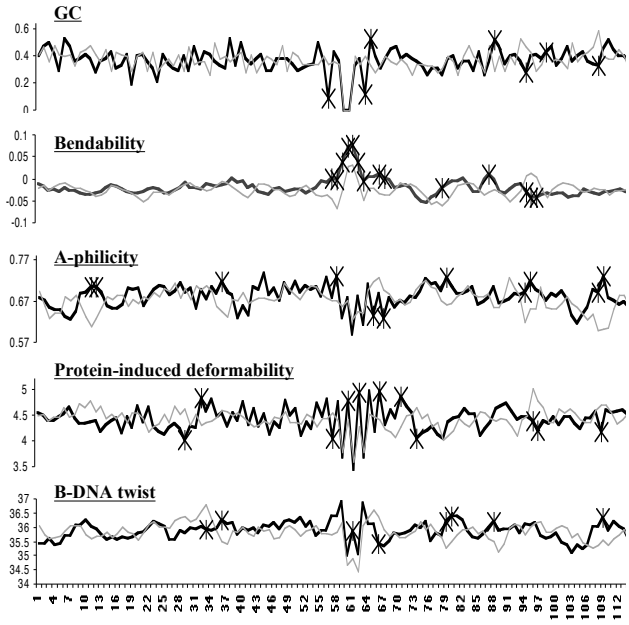


Figure 37. Physical properties of *Sleeping Beauty* insertion sites. *SB* insertion sites and random human sequences were compared for five different physical properties. The random sequences from human DNA were taken from chromosome 21 and aligned at a TA. In the GC profile the base composition of the insertion and random sequences is given. Lower values in the A-philicity chart mean that the sequence is more likely to form A-DNA. Black lines represent averaged sample data, gray lines represent averaged control data. Asterisks mark base pairs which were found significantly different at a confidence level higher than 95%. Base position 60 corresponds to the beginning of the TA insertion site.

control dataset. These significant positions appeared clustered in the immediate vicinity of the insertion sites (Fig. 37). A strong signal clustered around the transposon insertion sites was also observed for protein-induced deformability (Fig. 37). Areas of significance outside of the insertion site formed no discernible pattern. These data suggest that *SB* insertion sites have unique physical properties.

To directly test the predicition of increased bendability of the insertion target sequences, a DNase I digestion assay [361] was performed on the consensus integration sequences of *SB* and *Tc1*. The eight-base *SB* (ATATATAT) and ten-base *Tc1* (CACATATGTG) [113] consensus sequences were compared to two control sequences predicted to have low bendability (AAAAAAAA and AAATAAAA) [361]. The second "bad bender" sequence contains a central TA dinucleotide to reflect that of the *SB* and *Tc1* consensus sequences. These four sequences were flanked by identical sequences. The digestion parameters were such that, on average, DNase I cleaved each DNA molecule less than once, and therefore cleavage occured at the most favorable position, one that is the most bendable [361]. Thus, the more bendable a particular sequence is, the more often DNase I will digest there, and the more intense the resulting radioactive band will be. Quantitation of the digestion patterns showed that both the *SB* and *Tc1* oligos were digested more often within their consensus target sequences than were the control oligos (Fig. 38). These data confirm the computer predictions of increased bendability of DNA sequences at transposon insertion sites.

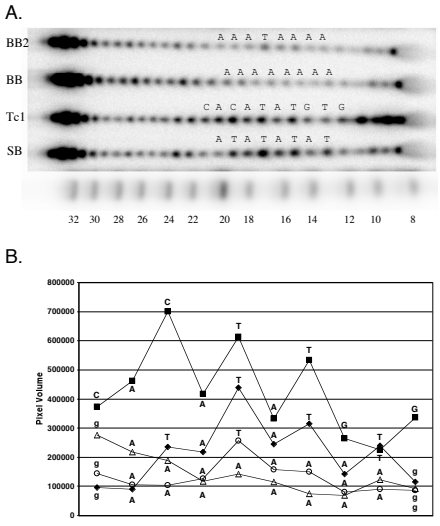


Figure 38. DNase I digestion assay of bendability. (A) Electrophoretic patterns of DNase I digestion of 32-bp oligos containing either the 8-bp *SB* or 10-bp *Tc1* consensus insertion sequences, or two different 8-bp sequences with low predicted bendability. The bands corresponding to the consensus sequence are labeled with the respective nucleotide, and the sizes of the Oligo Size Marker ladder bands are marked. **(B)** The bands were quantitated and graphed relative to each other. The uppercase letters indicate the core target sequence, while the lowercase letters indicate the identical flanking sequence. (◆) *SB* consensus, (■) *Tc1* consensus, (△) Bad Bender, (❍) Bad Bender2.

Transposase, like other DNA-binding proteins, likely forms hydrogen bonds with its DNA substrate. Previously, analysis of a large number of *P* element insertion sites in

Drosophila identified a 14-bp palindromic pattern of hydrogen bonding sites using a computer program called HbondView [121]. This graphical tool identifies patterns of bond donors or acceptors in the major groove of DNA sequences by converting a set of aligned DNA sequences into a display of potential hydrogen-bonding positions. We compared insertion sites of *Tc1/mariner* elements and random DNA for their respective propensities to form hydrogen bonds in their major grooves. Both *SB* and *Tc1* genomic insertions showed a symmetrical pattern (Fig. 39A and C). The transposon insertion sites have a 10-bp palindromic pattern, including the TA target plus four base pairs on each side. Because the non-insertion control data sets for both human and *C. elegans* genomic DNA lack such a pattern (Fig. 39B and D), these results indicate that in addition to structural features, a specific pattern of hydrogen-bonding sites at the target DNA contributes to target site selection of transposons. Such palindromic pattern and the symmetry of the consensus target site sequence together indicate that the target DNA is recognized by a dimeric or multimeric form of the transposase. Indeed, we have shown that *SB* transposase forms tetramers in solution, suggesting the involvement of a transposase tetramer in *SB* transposition [45].

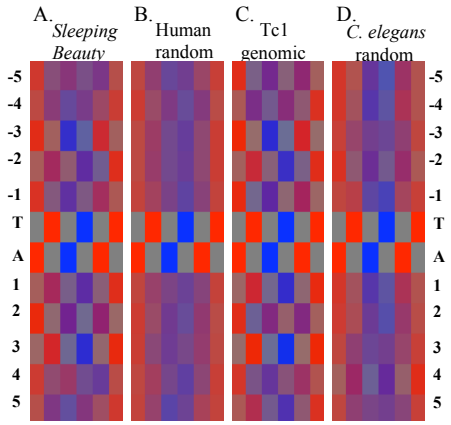


Figure 39. Hydrogen-bond analysis of insertion sites. Columns indicate six potential sites with which proteins can form hydrogen bonds. The rows indicate base pairs 3' (negative) and 5' (positive) of the TA insertion site. The colors of a given cell denote the type of hydrogen bond that can be formed: donor is red, acceptor is blue and sites that cannot hydrogen bond are gray. The analysis was performed with multiple insertions sites, and so the final color is determined by the percentages of hydrogen-bond donors, acceptors and non-hydrogen-bonding sites at a given position. **(A)** *SB* insertions in human DNA; **(B)** random human DNA; **(C)** *Tc1* insertions in *C. elegans* DNA; **(D)** random *C. elegans* DNA.

In summary, we have shown that target site selection of TEs is considerably more specific than it was assumed before, and that it is primarily determined on the DNA structural rather than on the sequence level. Our results indicate that a combination of particular physical properties (Figs. 36-38) generate a spatial optimum of the DNA for transposase interaction. Such a spatial optimum, together with a specific hydrogen-bonding capacity (Fig. 39) recruits the transposase with a substantial degree of specificity. The significance of our findings is supported by the observation that this pattern of structural preference is conserved in the *Tc1/mariner* family and in other, relatively randomly integrating transposons

in the DDE recombinase family such as the bacterial elements *Tn5* (data not shown), *Tn7* [118], *Tn10* [120], *Mu* bacteriophage [363], *IS231* [364] and retroviral integrases [123, 125]. Significantly, transposition by the RAG V(D)J recombinase is preferentially targeted to distorted DNA structures [365]. However, these factors cannot be the only determinants of target site selection, because the *Tc1* and *Tc3* elements have different insertion profiles in *C. elegans* [113]. Therefore, it appears that there exist at least two levels of selection that together determine how favorable a particular DNA sequence is for transposon insertion. Physical properties of the DNA primarily specify a set of sequences in a genome that are in a spatial optimum to receive a transposon insertion, whereas the ability of the transposase polypeptide to efficiently interact with such sequences specify a subset within these sites where insertions occur.

Compared to virus-based integrating vector systems, including retrovirus-, HIV- or AAV-based vectors, that were found to have a propensity for integrating into genes versus non-genic regions [293, 296, 326, 327, 366], the regional preferences associated with *SB*-mediated integration were much less pronounced (35% of *SB* insertions in RefSeq genes, versus 53% for ASV, 51% for MLV, 83% for HIV-1 [367] and 72% for AAV [366]. Importantly, in contrast to most integrating virus-based, microarray analyses revealed no correlation between the integration profile of *SB* and transcriptional status of targeted genes [367], suggesting that *SB* might be a safer vector for therapeutic gene delivery than most viruses that are currently used. Indeed, it is important to note that no dominant adverse effects associated with *SB* vector integration have been so far found in experimental animals [324]. Nevertheless, the genotoxic potential of *SB*-based vectors will have to be systematically assessed in the future, probably by applying high throughput, cell-based assays.

3.1.5 The *Frog Prince*: a reconstructed transposon from *Rana pipiens* with high activity in vertebrates (Paper VIII)

SB shows no host-restrictions in vertebrates, but the efficiency of transposition in cell lines derived from different species is variable [259]. Therefore, having a palette of different,

vertebrate-derived transposons with different host preference widens the potential of transposons as genomic tools in vertebrates.

Relatively high copy number of inactive transposable elements in genomes practically prohibits the isolation of functional transposase genes using nonselective methods. In search for potentially active transposase genes in vertebrates, we devised an open reading frame (ORF)-trapping method. The procedure is based on generating a pool of PCR products from genomic DNA using primers flanking the transposase gene sequences (Fig. 40). The 5'-primer contains the predicted translational initiation signal, and the 3'-primer lacks the stop codon. The PCR products are then cloned into an expression vector to generate fusion genes with *lacZ*. The recombinant plasmids are transformed into *E. coli*, and plated on X-gal-containing plates. Blue colonies can only arise if the cloned sequences are in frame with the *lacZ* gene, and do not contain a stop codon.

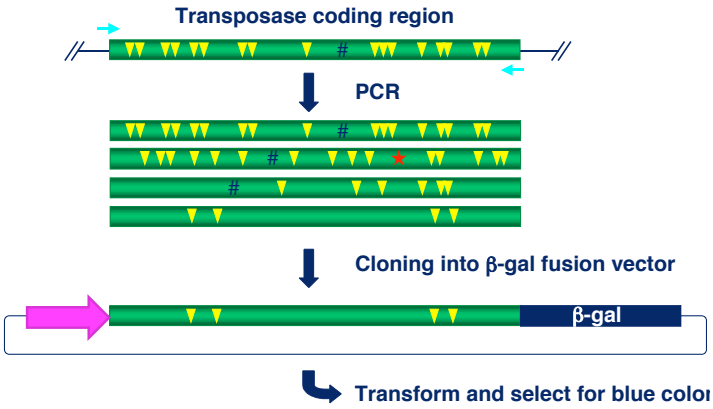


Figure 40. Strategy for trapping transposase ORFs from the *Rana pipiens* genome. Transposase genes (green boxes) are PCR-amplified from genomic DNA (arrows show primers). The vast majority of these genes are defective due to point mutations (yellow arrowhead), frameshifts (#) and premature translational stop codons (*). ORFs can be selected by cloning the PCR products in fusion with the *lacZ* gene driven by the CMV promoter, transformation into *E. coli*, and plating on X-gal-containing plates.

We applied the ORF-trap on genomic DNA from *Rana pipiens*, using PCR primers designed to the consensus sequence of *Txr* elements in *Xenopus laevis* [84]. Three resultant blue bacterial colonies indicated the presence of transposase-coding sequences that did not contain premature stop codons. The genomic copy number of the *R. pipiens* transposon was estimated by dot blotting. Assuming that the size of the *R. pipiens* haploid genome is 6.6 x 10^9 bps [368], we estimated that the transposase gene is represented about 8000 times per haploid genome. To assess what fraction of these elements contains intact ORFs, additional transposase coding regions were PCR-amplified from the *R. pipiens* genome, and cloned without selecting for ORFs. Seven transposase genes were sequenced and, to our surprise, we found that three of them contained ORFs. These results suggest that this transposon family is a relatively young component of the *R. pipiens* genome.

The ten transposase genes isolated above were aligned to generate a consensus sequence. The consensus *R. pipiens* transposase gene encodes a typical *Tc1*-like transposase containing an N-terminal DNA-binding domain composed of two predicted HTH motifs [20], a bipartite NLS [49], an AT-hook motif [45] and a catalytic domain with the DDE signature (Fig. 4). The ten transposase genes were about 99% identical to the consensus sequence, and one of them differed only in two nucleotides from the consensus, resulting in two amino acid substitutions in its ORF. One of these mutations was a T152S exchange in the first part of the catalytic domain of the transposase, and the other was an R315C substitution close to the C-terminus of the protein. Site-specific PCR mutagenesis was used to derive the sequence of the consensus *R. pipiens* transposase gene.

In order to derive the binding sites for the *R. pipiens*-type transposase, linker-mediated PCR was applied on genomic DNA to amplify the complete inverted repeats together with genomic flanking sequences. Alignments of five different clones revealed 214 bp long, perfect inverted repeats flanking the transposase genes. The *R. pipiens* transposons are typical IR/DR-type elements. The IR sequences together with the consensus transposase gene constitute the components of a novel transposable element system that we named *Frog Prince* (*FP*). To determine the phylogenetic position of *Frog Prince* among other *Tc1*-like transposase genes, amino acid sequences of *Tc1* from *C. elegans, Txr* and *Txz* from *X. laevis* [84], *Tdr1* [80] and *Tdr2* (*Tzf*) from zebrafish [49, 369], *SB* which represents the salmonid subfamily of fish elements [85] and the putative *FP* transposase were used to generate a phylogenetic tree (Fig. 41). The topology of the unrooted tree shows significant phylogenetic distance between the *SB* and the *FP* transposases, and displays that *FP* is most closely related to the *Txr* elements.

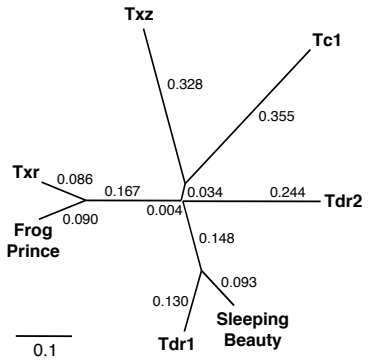


Figure 41. Phylogenetic position of *Frog Prince* among *Tc1*-like transposons. Numbers at the branches indicate the phylogenetic distances calculated by ClustalX.

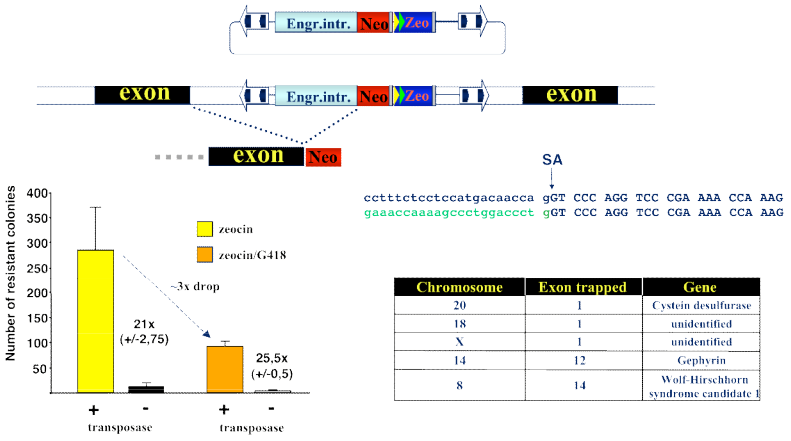
SB shows high transpositional activity in human cells [85]. Therefore, the initial tests for transpositional activity of the *Frog Prince* element were done in cultured HeLa cells, using a transposition assay established for *SB* [85]. The reconstructed consensus *R. pipiens*

transposase ORF (in pFV-FP) was transfected together with either the *Txr*-type (pTxr-neo) or with the *Frog Prince*-type (pFP-neo) substrate constructs. A 17-fold increase in colony number was detected when pFV-FP was cotransfected with its own substrate, pFP-neo (not shown). The significant sequence similarity between the *Xenopus* and *Rana* elements could still allow cross-mobilization between them, as it is the case among the hAT-superfamily elements *hobo* and *Hermes* [370]. Indeed, we observed a 5-fold increase in the number of G418-resistant cell colonies when pFV-FP was cotransfected with pTxr-neo (not shown). Thus, the *R. pipiens* transposase can cross-mobilize a *X. laevis* transposon, indicating that the two transposon families in these species diverged recently. In contrast, no cross-mobilization was observed between *FP* and *SB*. Taken together, the data demonstrate that we successfully derived an active transposon system from the *R. pipens* genome, and that *FP* can significantly increase the efficiency of transgene integration from plasmid-based vectors to the human genome.

Tc1/mariner elements transpose via a cut-and-paste mechanism (see section 1.3.1). During the first step of this process, the element is excised by a pair of staggered DSBs. The host DNA repair machinery seals the gap and, according to the number of the protruding nucleotides, a small insertion indicates the former presence of a transposon. *Tc1*/*mariner* elements generate footprints in the range of 2-4 base pairs [73, 78, 102, 371]. Primers flanking the transposons were used in a series of nested PCR to identify the footprints left behind by *FP* transposition in the donor plasmids. Sequencing of the PCR products revealed that *FP* transposition leaves a CTG or CAG triplet at the excision site, indicating that excision of *FP* generates 3-nucleotide-long overhangs. *Tc1*/*mariner* elements transpose into TA dinucleotides, which are duplicated and flank the integrated transposon [20]. Flanking sequences of three integrated *FP* transposons were isolated from individual G418-resistant HeLa clones. All three *FP* insertions were flanked by the expected TA dinucleotides, followed by different human genomic sequences. In sum, these data show that *Frog Prince* follows precise cut-and-paste transposition into various locations in the human genome.

High frequency, precise transposition into different genomic loci suggests that genome-wide gene trapping is feasible with *FP*. A prerequisite of successful transposon-based gene trapping is that the terminal IRs do not contain potential splice sites. To examine this possibility, an *FP*-based gene trap vector (pFP/GT-geo) containing an SA sequence of the mouse *engrailed-2* gene followed by the *lacZ-neo* (*geo*) fusion was constructed and used for transposition in HeLa cells. 26 out of 27 individual *neo*-resistant colonies were positive for β-galactosidase activity, indicating at least one successful gene trap event per clone (data not shown). *LacZ* fusion transcripts were identified from G418-resistant, β-galactosidase-positive cells with cRACE [372]. We identified a transcript in which splicing generated a fusion between an endogenous RNA and the marker exactly at the *engrailed-2* SA (Fig. 42). These data indicate that the inverted repeats of *FP* do not interfere with the desired splicing event between a splice donor sequence of an endogenous transcript and the *engrailed-2* SA within the transposon.

Next, we wanted to determine the efficiency of gene trapping and to identify the tagged genes. For this purpose, an *FP*-based donor plasmid (pFP/GT-neo) was constructed which contains *engrailed-2* intron sequences with the SA, a glycine bridge to allow proper folding of the marker in protein fusions [373], an ATG-less *neo* gene, a zeocin resistance gene (*zeo*) driven by dual eukaryotic/bacterial promoters and a plasmid



Chromosome	Exon trapped	Gene
20	1	Cystein desulfurase
18	1	unidentified
X	1	unidentified
14	12	Gephyrin
8	14	Wolf-Hirschhorn syndrome candidate 1

Figure 42. Gene trapping with *Frog Prince* in human HeLa cells. (***top***) The gene trap cassette contained a promoterless neo gene, followed by a zeocin resistance marker driven by dual eukaryotic/prokaryotic promoters (yellow and green arrowheads). Expression of the neo marker is dependent on transposition of the cassette into an intron of an expressed gene, and correct splicing of an upstream exon to the splice acceptor site donated by the *engrailed* sequences. (***bottom left***) Efficiency of gene trapping with *FP*. Numbers of antibiotic-resistant colonies are indicated on the *y*-axis. Zeocin selection was used to deduce the transpositional efficiency in the presence of the helper (pFV-FP) *vs*. the control (pCMV-β) plasmid (yellow column). Gene trapping efficiencies were determined by using zeocin/G418 double selection (orange column). Numbers next to the columns indicate the fold difference in numbers of colonies obtained in the presence *vs.* absence of the transposese. (***right***) Fusion transcript. On top, nucleotide sequences of the *engrailed-2/lacZ* junction in pFP/GT-geo are shown. Intron sequences are typed in lowercase, exon sequences are in uppercase. The arrow indicates the splice acceptor site (SA). Human transcript sequences (typed in green) are fused to the *engrailed-2* exon due to correct splicing at the SA. (***bottom right***) Gene trapping events identified by transposon rescue.

origin of replication (Fig. 42). All chromosomal transposition events give rise to zeocin-resistant cells. A subset of transformant cells will be G418-resistant, if the transposon inserted into an intron of an expressed gene in the proper orientation, and if splicing occurred in-frame with *neo*. The plasmid origin of replication within the element can be used to isolate the integrated transposon from genomic DNA by plasmid rescue. The number of zeocin/G418 double-resistant colonies was about one third of those resistant to zeocin alone, indicating that about 30% of all transposition events occurred in introns of expressed genes and in-frame splicing took place (Fig. 42). Five insertion sites of the *FP* gene trap transposons were identified. All of them mapped to introns of genes in different chromosomes, in the correct orientation (Fig. 42). Our results suggest that *FP* can potentially target a large fraction of genes in the human genome.

SB has varying transpositional activity in different vertebrate cell lines [259]. However, *SB* is a synthetic element of fish origin and *FP* was reconstructed from the genome of an amphibian. Thus, the same set of cell lines can provide different permissive environments to the two transposon systems. To test this hypothesis, we compared the activities of the two systems in cultured cell lines derived from two mammalian, an amphibian and two fish species with the standard transposition assay (Fig. 43). The vector backbones, the promoters, the poly-A signals and the transposon marker genes were identical in the constructs of the two systems. *FP* was found active in cell lines of representatives of major vertebrate taxa, and in some cell lines it has higher transpositional activity than *SB* (Fig. 43). In considering possible explanations for the significantly more efficient transposition of *FP* in zebrafish cells, the higher intrinsic transpositional activity of *FP* can presumably be ruled out, as the two systems were about equally

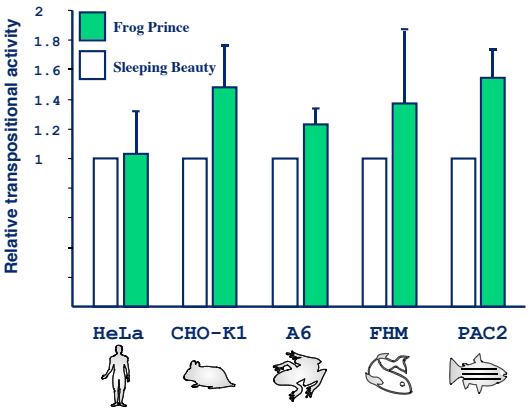


Figure 43. Activity of *Frog Prince* in comparison with *Sleeping Beauty* in diverse vertebrate species. The donor and helper plasmids of *FP* and *SB* were cotransfected in Hela (human), CHO-K1 (hamster), A6 (*Xenopus laevis*), FHM (fathead minnow) and PAC2 (zebrafish) cell lines. Transposition efficiencies were calculated by deriving ratios between the numbers of G418-resistant cell clones obtained in the presence *versus* in the absence of the transposases. Activities of *FP* (indicated by green columns) were compared to those of *SB* (white columns). Transpositional efficiency of *SB* was normalized to the value 1 for each cell line. The error bars show SEM.

active in human HeLa cells. More likely, the zebrafish cellular environment is more favorable for the *FP* system because of the absence of repressing activities that interfere with *SB*. Since transposases can mobilize inactive elements *in trans*, the ratio of inactive to active elements in eukaryotic genomes increases [31]. Some of these inactive copies might function as repressors either by dominant-negative complementation [160], or by competition with transposase in substrate binding. Similarly, thousands of dispersed endogenous elements might inhibit the activity of an exogenously supplied transposase by transposase titration [374]. *SB* is a fish transposon, and the zebrafish genome contains about 1000 copies of a *Tc1*-like transposon, *Tdr1* [80]. It is likely that the endogenous *Tdr1* elements can interfere with *SB* transposition, since they share over 80% sequence identity with *SB*, both on the DNA and protein sequence levels [80]. In contrast, *FP* is a phylogenetically distant element with only about 50% transposase sequence identity to either *SB* or *Tdr1* (not shown). Thus, the newly introduced *FP* system is perhaps immune or at least less vulnerable to the above inhibitory mechanisms in zebrafish cells.

In summary, the ability of *FP* to precisely integrate single copies of foreign DNA into various chromosomal loci in a variety of vertebrate genomes allows us to propose the usefulness of the *Frog Prince* system in transgenesis and insertional mutagenesis. We have tested the ability of *FP* to efficiently trap expressed genes in a simple cotransfection assay in human cells. Klinakis *et al*. (2000) showed that the *Tc1/mariner* element *Minos* could potentially tag all human genes in HeLa cells [375]. However, only about 1-10% of all *Minos* insertions was estimated to represent actual gene trap events. We estimate that approximately 30% of all selectable *FP* transposon insertions occur in genes. To our knowledge, such high gene trapping frequencies have not been seen with other vectors. It is yet to be determined whether the higher gene trapping efficiency of *FP* reflects a different insertion site preference as compared to the *Minos* element. In comparison with retroviruses that preferentially integrate into the 5'-regions of genes [296], the integration pattern of *Tc1*-like transposons is more random [376]. Therefore, transposon insertions are expected to produce a different mutational spectrum than retroviruses.

In the era of functional genomics, there is a sore need for developing efficient means to explore the roles of genes in different cellular functions. The availability of alternative transposon systems in the same species opens up new possibilities for genetic analyses. For example, *piggyBac* transposons can be mobilized in *Drosophila* in the presence of stably integrated *P* elements [377]. Because *P* element- and *piggyBac*-based systems show different integration site preferences [358, 377], the number of fly genes that can be insertionally inactivated by transposable elements can greatly be increased. *P* element vectors have also been used to introduce components of the *mariner* transposable element into the *D. melanogaster* genome by stable germline transformation. In these transgenic flies, *mariner* transposition can be studied without accidental mobilization of *P* elements [378]. We have shown that *Frog Prince* and *Sleeping Beauty* do not detectably interact in an *in vivo* transposition assay. Thus, *FP* can be used as a genetic tool in the presence of *SB*, and *vice versa*, which considerably broadens the utility of these elements. As an alternative transposon system, significantly different from any other active transposon, *Frog Prince* can expand our possibilities for transposon-mediated genetic manipulations in vertebrates.

3.1.6 The ancient *mariner* sails again: Transposition of the human *Hsmar1* element by a reconstructed transposase and activities of the SETMAR protein on transposon ends (Paper IX)

Mariner elements make up a diverse family of eukaryotic DNA transposons, present in a wide variety of genomes, including humans ([30, 213, 379] and references therein). Transposition results in the accumulation of hundreds or thousands of transposon copies over evolutionary time. However, most *mariner* copies appear to be dead remnants of once active transposons inactivated by mutations (see section 1.5.1). To date, only three *mariner* elements out of the hundreds of sequences that have been described have proven to be active. Two of these, *Mos1* and *Famar1*, are natural elements isolated from the genomes of *Drosophila mauritiana* [26] and the earwig *Forficula auricularia* [225], respectively. The active *Himar1* element is a majority rule consensus of cloned genomic copies obtained from the

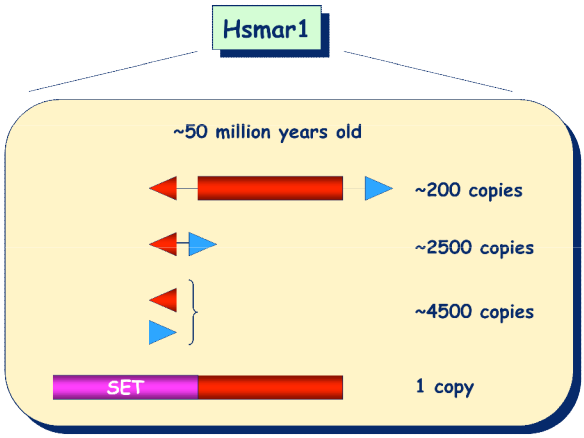


Figure 44. Components of the *Hsmar1* transposon family in the human genome. The panel summarizes the structures and copy numbers of *Hsmar1*-derived sequences. Red and blue arrowheads represent terminal inverted repeats, red box represents the transposase coding region. SETMAR is a chimeric protein made up by a histone methyltransferase (SET) domain and a particularly conserved transposase domain.

horn fly *Haematobia irritans* [78]. *Mos1* and *Himar1* have been used as molecular tools for genome manipulations in diverse species (reviewed in [20]). However, the utility of these invertebrate *mariner* transposons in mammalian genetics is hindered by their limited activity in mammalian cells [243].

Mariner elements are represented by two subfamilies in the human genome: *Hsmar1* [213] and *Hsmar2* [380]. The first *Hsmar1* element entered the primate genome lineage approximately 50 million years (Myr) ago, and transposition was ongoing until at least 37 Myr ago, producing about 200 *Hsmar1* copies (Fig. 44) [213]. However, none of the present copies encodes a functional transposase protein due to mutational inactivation. The *Hsmar1* transposon copies are accompanied by about 4500 copies of solo-IRs (containing a single inverted repeat) and about 2500 copies of an *Hsmar1*-related, paired-IR element, *MiHsmar1* (Fig. 44) [11, 213]. Such miniature inverted-repeat transposable elements (MITEs) are thought to have been generated by internal deletions of longer transposons; they make up the predominant fraction of DNA elements in flowering plants, and are often found in animal genomes [381].

We have previously reconstructed two functional transposable elements from vertebrate genomes: *Sleeping Beauty* (*SB*) from fish (section 3.1.1) [85] and *Frog Prince* (*FP*) from amphibians (section 3.1.5) [382]. Both gene reconstructions were based on the hypothesis that a consensus sequence, established from several cloned inactive copies, represents an active gene. Encouraged by our former success, we set out to

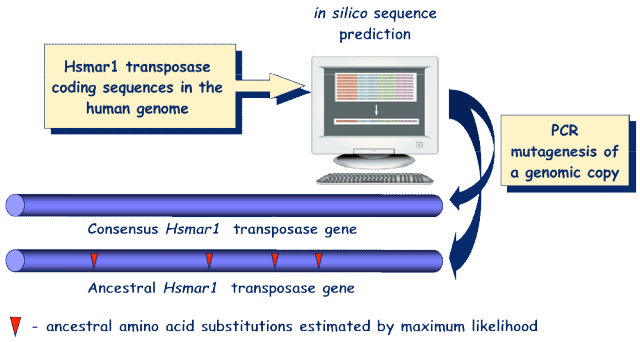


Figure 45. Phylogenetic reconstruction of the *Hsmar1* transposase gene. The ancestral, active transposase (blue box) differs from the consensus sequence in four amino acid positions (red arrowheads).

reconstruct an active *Hsmar1* transposable element from the human genome. We engineered the consensus sequence of the transposase gene [213] by site-directed mutagenesis of 21 codons of an *Hsmar1* ortholog obtained from the chimpanzee genome. Transposition activity was assessed in human HeLa cells, using a two-component transposition system similar to those established for *SB* and *FP* [85, 382]. There was no indication of *Hsmar1* transposition in these experiments (not shown), suggesting that the consensus of the *Hsmar1* transposase gene represents an inactive sequence.

The approach to gene sequence prediction based on consensus does not incorporate phylogenetic information. For example, an inactivating mutation in a transposable element may become overrepresented if that particular mutant was preferentially amplified over the active sequence. With the aim of reconstructing the ancient, active *Hsmar1* transposase gene that colonized the genome lineage of primates, we applied a statistically rigorous approach based on maximum likelihood that has been successfully used to reconstruct ancestral gene sequences [383]. First, human *Hsmar1* transposase-like amino acid sequences were obtained from the human genome by TBLASTN similarity searches. To infer the sequence of the last common ancestor of primate and invertebrate mariner transposases, a likely candidate for an active transposase protein capable of colonizing new hosts, invertebrate mariner transposase sequences of the *cecropia* subfamily were also included in the phylogenetic analysis. The evaluations of the reconstructed ancestral amino acid states at the node connecting the invertebrate *mariner* elements with the branch leading to the human sequences revealed the following four amino acid substitutions in the known consensus transposase protein sequence with posterior probabilities greater or equal to 0.9: C53R, P167S, L201V, A219C (Fig. 45). Independent inferences from human nucleotide sequences (as identified by BLASTN) or from chimpanzee amino acid sequences also identified these amino acids as the most likely ancestral states at these sites. Furthermore, inspection of these positions in an alignment of *cecropia*-type transposase sequences revealed that the predicted substitutions represent conserved amino acids within the subfamily, suggesting that these residues may be important for transposase activity.

The putative ancestral *Hsmar1* transposase gene was engineered by incorporating the four predicted amino acid substitutions into the framework of the consensus *Hsmar1* transposase, and the resulting protein was tested for transposition. Upon cotransfection of a *neo*-marked *Hsmar1* transposon with a vector expressing the modified transposase, a 23-fold increase in the number of antibiotic-resistant colonies was observed (not shown), suggesting that the resurrected protein efficiently catalyzes transgene integration from the donor plasmids into human chromosomes. Thus, it is likely that the inferred sequence, which was named *Hsmar1-Ra*, represents or is very similar to the sequence of the ancient *mariner* element that colonized the genome lineage of primates. We established molecular evidence of *Hsmar1* transposition by showing TTA or TAA triplets at transposon excision sites, corresponding to the 5'- and 3'-terminal nucleotides of *Hsmar1* transposons (not shown). This is consistent with footprint formation by the *Himar1* [78, 384] and *Mos1* [385] *mariner* transposons that predominantly generate 3-bp footprints. In order to obtain formal molecular proof for cut-and-paste *Hsmar1* transposition, 47 integration events were isolated from HeLa cells. All of the transposon insertions occurred at TA dinucleotides scattered on 16 human chromosomes (not shown). We found that the chromosomal distributions of the endogenous genomic copies and the *de novo* integrants overlap, and have a bias to larger chromosomes. 44% of the hits were identified in introns of genes, indicating a fairly random genomic distribution similar to that found with *SB* in human cells [376]. In sum, the results above indicate that we successfully reactivated the first vertebrate *mariner* transposon from the human genome.

It is widely believed that MITEs can only be mobilized by transposases supplied *in trans* [381], but only one such instance has been documented at the molecular level [386]. Although mechanisms of preferential transposition due to small size [128, 213], and transposition linked to the cellular process of DNA replication [387, 388] has been suggested, mechanisms of MITE mobilization and amplification are incompletely understood. The human genome contains about 2500 copies of an *Hsmar1*-related MITE (Fig. 44). The *MiHsmar1* elements have a consensus sequence of 80 bps containing 37-bp IRs, from which the first 30

bps are identical to the IRs of *Hsmar1* [213]. Copy number of *MiHsmar1* is at least an order-of-magnitude higher than that of the full-sized *Hsmar1* elements in the human genome [213].

One possible explanation that might account for their abundance is their small size, which could predispose them for efficient transposition [128, 213]. To test this hypothesis, transposon donor plasmids were constructed containing a zeocin resistance gene and either the consensus sequence of *MiHsmar1*, or the long versions of the transposon (*Hsmar1*-neo or the autonomous element) inserted into the same position in the coding region of the ß-lactamase (*bla*) gene (Fig. 46). In addition, long transposons with IRs identical to those of *MiHsmar1* were created. The insertions disrupt the *bla* reading frame that can only be restored if the transposons are excised by the transposase, and NHEJ repairs the plasmid producing the canonical footprints (Fig. 46). The donor plasmids harboring transposons of different length were transfected into HeLa cells together with vectors expressing either the *Hsmar1-Ra* or the *SB* (control) transposase; the reporter containing the autonomous *Hsmar1* element was transfected alone. Low molecular weight DNA purified from the cells was transformed into *E. coli*. Excision events were scored by counting amp^R/zeo^R colonies; selection with zeocin alone served to control overall plasmid recovery. *MiHsmar1* elements were excised from the donor plasmids two orders of magnitude more efficiently than any of the longer versions of the transposon (Fig. 46). These data indicate that reduced size is responsible for the elevated excision

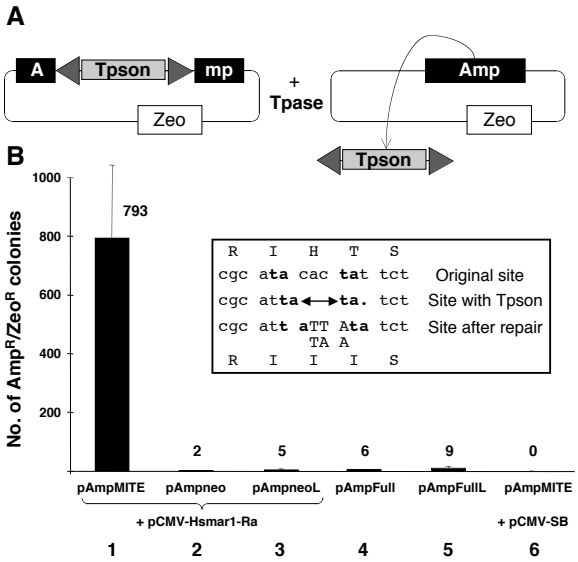


Figure 46. Excision of *Hsmar1* and *MiHsmar1* transposons. (A) Outline of the excision assay. A transposon insertion disrupts and inactivates the ampicillin resistance gene (Amp). Transposon excision followed by DNA repair can restore ampicillin resistance, which can be selected in *E. coli*. **(B)** Coding triplets of the original, the modified and the repaired excision sites are listed in the panel. Transposon footprints and amino acid sequences are in capitals. Various transposon donor plasmids (indicated on the *x* axis) were transfected into HeLa cells either with pCMV-Hsmar1-Ra (experiments 1-3) or alone (experiments 4 and 5) or with pCMV-SB (experiment 6) as a control. Indicator plasmids contain the following transposons in the ampicillin resistance gene: pAmpMITE, the consensus *MiHsmar1* sequence; pAmpneo, the neo-tagged *Hsmar1* element; pAmpFull, the autonomous *Hsmar1* transposon; pAmpneoL, the neo-tagged *Hsmar1* element with inverted repeat sequences that match those of *MiHsmar1*; pAmpFullL, the autonomous *Hsmar1* transposon with inverted repeat sequences that match those of *MiHsmar1*. The Amp^R/Zeo^R colonies represent excision events followed by canonical footprint formation at the excision site. The normalized numbers of the double-resistant colonies (shown above the columns) were obtained by dividing the numbers of Amp^R/Zeo^R colonies with the corresponding Zeo^R colony numbers.

frequency of *MiHsmar1*.

Despite their parasitic nature, there is increasing evidence that transposable elements are a powerful force in gene evolution (discussed in section 1.5.2.1). Indeed, about 50 human genes are derived from transposable elements [11], among them genes that are responsible for immunoglobulin gene recombination in all vertebrates [61]. One of these "domesticated", transposase-derived genes is *SETMAR*, a fusion gene containing an N-terminal SET domain fused in-frame to an *Hsmar1* transposase (Fig. 44) [213]. The *SETMAR* gene has apparently been under selection; the transposase open reading frame is conserved, and shows only 2.4% divergence from a consensus *Hsmar1* transposase gene sequence (*vs*. 8% average divergence between *Hsmar1* transposase genes) [213]. *SETMAR* is broadly expressed in humans [389], suggesting a housekeeping function. The SET domain can be found in histone methyltransferases that regulate gene expression by chromatin modifications [390]. The SETMAR protein has been shown to have histone H3 methyltransferase activity *in vitro*, and has been proposed to play a role in DSB repair [389]. Both the transposase domain of SETMAR as well as the full-length SETMAR protein were shown to bind to *Hsmar1* IR sequences [391, 392], and it was recently demonstrated that SETMAR can perform transposition reactions using precleaved transposon substrates *in vitro* [392]. Since the protein used in those experiments was lacking the SET domain, we addressed if the full-length, physiological form of SETMAR can also exhibit transposase-related activities.

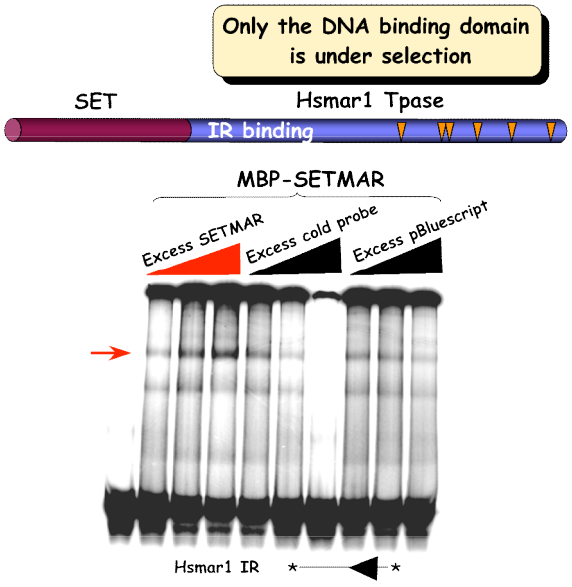


Figure 47. *In vitro* DNA binding activity of the SETMAR protein. A radioactively labeled DNA fragment of 83 bp containing the 5'-IR of *Hsmar1* (depicted below) was incubated with purified MBP-SETMAR protein, and DNA-protein complexes were visualized by EMSA. As compared to free, unbound probe (lane 1), increasing concentrations of MBP-SETMAR produce DNA-protein complexes. A substantial amount of the SETMAR protein formed aggregates at all concentrations used, which were unable to enter the gel during electrophoresis. Complex assembly was challenged with cold (unlabeled) probe as a specific competitor, or increasing amounts of pBluescript was used as nonspecific competitor DNA.

To test such possibility, SETMAR was expressed in *E. coli* and purified as an N-terminal fusion with the maltose binding

protein (MBP-SETMAR). Electrophoretic mobility shift experiments showed binding of MBP-SETMAR to the 5'-IR of *Hsmar1* in a sequence-specific manner, since binding was competed with cold specific DNA but not with excess non-specific DNA (Fig. 47).

Despite the high sequence similarity of SETMAR to the *Hsmar1*-Ra transposase, SETMAR contains several amino acid substitutions potentially compromising its catalytic functions. For example, the third D of the catalytically essential DDD triad in the transposase domain of SETMAR is replaced by an N. Even the conservative D to E mutation in this position abolished the catalytic activity of the *Mos1* transposase [74]. In order to test possible catalytic activity of SETMAR, an *in vitro* DNA cleavage assay was first applied. The assay is based on incubation of recombinant, purified protein with a double-stranded DNA substrate containing IR sequences of *Hsmar1*, followed by ligation of the 3'-end of a single-stranded oligonucleotide linker to phosphorylated 5'-ends of DNA exposed as a result of endonuclease activity, and PCR using substrate- and linker-specific primers (Fig. 48).

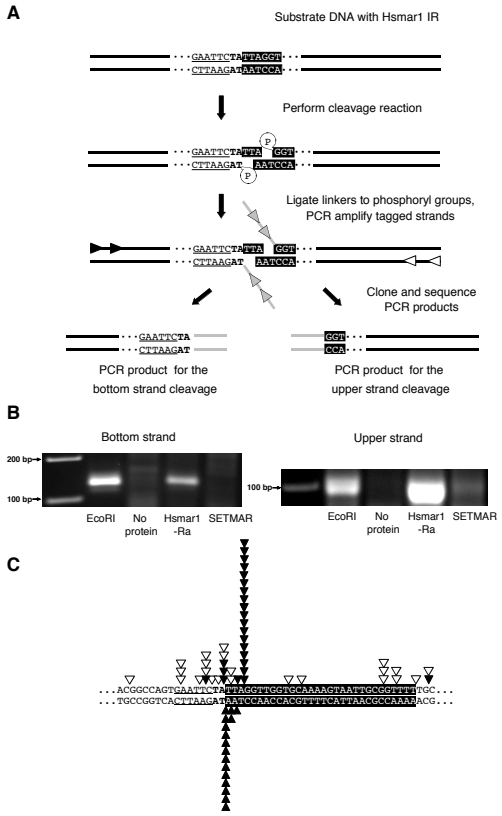


Figure 48. *In vitro* cleavage site analysis with ligation-mediated PCR. (**A**) Schematic overview of the assay is shown. A 281-bp long DNA fragment containing the 5'-IR of *Hsmar1* was used as the substrate for the reaction. The 5'-end of the IR is typed in black background, the TA nucleotides bordering the element are in bold, the neighboring *Eco*RI site is underlined. Cleavage of one or both strands of the substrate DNA by the purified proteins results in phosphoryl-terminated DNA ends, which can later be ligated to 3'-hydroxyl-terminated single-stranded oligonucleotide linkers (gray lines) with T4 RNA ligase. The linkers tag the cleavage sites, and enable amplification of the cleavage products by nested PCR. (**B**) Agarose gels with PCR products obtained with linker-specific primers (gray arrows) and primers specific for the upper or bottom strand of the substrate (black or white arrows, respectively). PCRs performed on *Eco*RI-digested DNA served as positive controls. (**C**) Distribution of cleavage positions identified after cloning and sequencing of the PCR products. Black triangles indicate cleavage sites introduced by MBP-Hsmar1-Ra, white triangles mark positions where MBP-SETMAR nicked the upper strand of the substrate DNA.

Purified MBP-SETMAR or MBP-*Hsmar1*-Ra proteins were incubated with a double-stranded DNA fragment containing the 5'-IR of *Hsmar1*. The substrate also contained an *Eco*RI recognition site, directly adjacent to the 5'-end of the *Hsmar1* IR (Fig. 48). Digestion of the probe with *Eco*RI served as a control for the assay. MBP-*Hsmar1*-Ra cleavage products were

identified for both strands of the *Hsmar1* transposon ends (Fig. 48). The most prominent cleavage site on the upper strand was three nucleotides inside the transposon DNA, whereas the lower strand was predominantly cut at the end of the transposon (Fig. 48). Therefore, the cleavage pattern of *Hsmar1*-Ra fully corresponds to the 3-bp transposon footprints that are generated after transposon excision. Cleavage products by MBP-SETMAR could only be identified for the upper strand (Fig. 48), suggesting that the SETMAR protein has very weak 3'-nicking activity, if any. These findings are in good agreement with the results of Liu *et al*. who found inefficient 3'-cleavage activity with the transposase domain of SETMAR [392]. Multiple SETMAR nicking sites were identified, none of them corresponding to the major 5'-cleavage site by *Hsmar1*-Ra (Fig. 48). Altogether, the data indicate that SETMAR is a defective transposase *in vitro*, and that it retains only a fraction of the biochemical activities of a transposase; namely, 5'-cleavage at transposon IR sequences.

In vivo catalytic activity of SETMAR was addressed by codelivery of an *Hsmar1* transposon plasmid and a SETMAR expression plasmid into mammalian cultured cells, recovery of extrachromosomal plasmids from the transfected cells, and PCR amplification using primers flanking the transposable element in the donor plasmid. In contrast to the strong, dominant footprint products generated by the *Hsmar1*-Ra transposase (Fig. 49A, lane 1), SETMAR generated only weak, smeary products (Fig. 49, lane 2), which were cloned and sequenced. As shown in Fig. 49B, 10 out of the 12 recovered products contained sequences either from one or both IRs. The majority (9/12) of the excision products contained deletions of pUC19 sequences flanking the transposon in the donor plasmid (Fig. 49B). The DNA ends were almost exclusively rejoined at 2-9 nucleotide long microhomologies, shared either between the left and right transposon

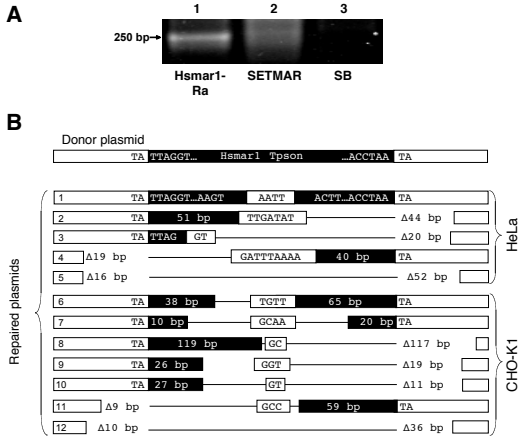


Figure 49. *In vivo* cleavage activity of SETMAR. (**A**) The agarose gel shows PCR products of excision assays on plasmid DNA from HeLa cells transfected with pHsmar1-neo and the helper plasmids indicated. The faint, smeary products around the expected size in lane 2 indicates loss of *Hsmar1* transposons from the donor plasmids catalyzed by the SETMAR protein. (**B**) The overall structures of footprints from HeLa and CHO-K1 cells are shown. The schematic view of the donor site is depicted above. Δ represents deletions, sequences in the white boxes are microhomologies at the junctions.

sequences, or between transposon and vector backbone sequences flanking the element (Fig. 49B). No canonical footprints were identified at the excision sites. These results are consistent with SETMAR-mediated nicking of *Hsmar1* elements *in vivo*. However, in contrast to the DNA lesions generated by the *Hsmar1*-Ra transposase that are predominantly repaired by NHEJ, the characteristics of transposon footprints by SETMAR could be best explained by the involvement of the HDR pathway.

The apparent inactivity of the confident majority-rule consensus *Hsmar1* transposase sequence implies that inactive *Hsmar1* copies were efficiently mobilized in the past by a transposase source *in trans*; thus, non-autonomous elements contributed more effectively to the spread of *Hsmar1* copies. One plausible mechanism that could explain such phenomenon is that transposase-producing *Hsmar1* copies suffered mutations within their IRs that compromised their ability to move. Almost all MITEs previously identified from different genomes are inactive, and thus their mechanisms of transposition and accumulation in eukaryotic genomes have been poorly understood. Although there are strong indications that MITEs are mobilized *in trans* by a corresponding transposase [e.g., *mPing*/*Pong* and *Stowaway*/*Osmar* mobilization in the rice genome [393, 394], *Tourist*/*PIF* interaction in the maize [395], or *in vitro* interaction between the *Arabidopsis* elements *Emigrant* and *Lemi1* [396]], this has only been experimentally demonstrated for MITE mobilization by the *impala* transposase in *Fusarium* [386]. The new *Hsmar1-Ra* transposon system provides a unique opportunity to investigate the origin and transpositional dynamics of these elements, and their contribution to primate genome evolution.

MITEs can accumulate to copy numbers far exceeding those of transposase-encoding DNA-transposons in different genomes [381]. The suggestion that MITEs might be preferentially mobilized due to their small size has been speculative. Here we show that *MiHsmar1* elements can be excised by two orders of magnitude more efficiently than their longer transposon versions. This phenomenon could have contributed to the prevalence of *Hsmar1*-related MITEs in primate genomes.

The Human Genome Project identified about 50 human genes derived from transposable elements [11]. However, to date there is no evidence for current transpositional

activity of any of these “domesticated” genes in humans. The only exceptions are the RAG genes, whose physiological function is to generate the immunoglobulin repertoire by a transposition-like process called V(D)J recombination (discussed in more detail in section 1.5.2.1.1). We provided experimental evidence that SETMAR, the product of a domesticated gene in the human genome derived from an *Hsmar1* transposase, retains its capacity to cleave *Hsmar1* transposon DNA *in vitro*. However, whereas the *Hsmar1*-Ra transposase efficiently cleaved both strands of DNA at transposon ends, thereby generating DSBs, the SETMAR protein only exhibited 5’ -cleavage activity, generating single-strand nicks. We showed that DNA damage inflicted by *Hsmar1*-Ra and SETMAR is processed differently by cells. Whereas transposon excision sites generated by the *Hsmar1*-Ra transposase predominantly contained the canonical, 3-bp footprints, SETMAR activity *in vivo* is associated with extended stretches of transposon sequences, deletions of flanking DNA and microhomologies at the junctions at the excision sites. The structure of these non-canonical footprints can be best explained by interrupted synthesis-dependent strand-annealing (SDSA) pathway of HDR, completed by an end-joining process generating microhomologies [348]. SDSA has been shown to play a role in the repair of transposon excision sites, and to be responsible for generating internally deleted versions of diverse transposable elements in animals and plants [92, 397-399], including the *Mos1 mariner* element [93] and *SB* [348]. Pathway choice in DSB repair can be influenced by the structure of the gap, the availability of repair factors and cell cycle phase [190, 348, 397]. Our observations for the lack of detectable 3’-cleavage activity of SETMAR suggests that the differential utilization of repair pathways by *Hsmar1*-Ra and SETMAR can be explained by the different structures of the cleavage sites: DSBs for *Hsmar1*-Ra and single-strand nicks for SETMAR. Single-stranded nicks have been shown to be potent triggers of HDR in mammalian cells [400]. For example, repair of DSBs generated by the RAG recombinase in V(D)J recombination is tightly linked to NHEJ, as discussed in section 1.4.4 [191]. However, nick-only RAG mutants have been shown to stimulate robust homologous recombination, and RAG-mediated nicking has been proposed to contribute to gene duplication events and chromosomal rearrangements [400]. Interestingly, some of the repair products obtained after SETMAR cleavage (Products 1, 2, 4,

6, 8-11 in Fig. 49B) resemble the structure of the *Hsmar1*-related solo-IRs and MITEs present in the human genome. Thus, interrupted SDSA repair events following *Hsmar1* transposon excision catalyzed by a *Hsmar1* or SETMAR transposase source could have played a role in the emergence and proliferation of the MITEs and solo-IRs.

Emergence of the *SETMAR* gene and the invasion of the ancient primate genome by the *Hsmar1* transposons took place within an overlapping evolutionary time window, between 40-58 myr ago [391]. Thus, it may be that the SETMAR protein played a role in regulating *Hsmar1* transposition. The 5'-UTR of the *Hsmar1* transposon has significant promoter activity, sufficient to drive transposase expression (not shown). Through its ability to bind to *Hsmar1* transposon IR sequences (Fig. 47), and to catalyze specific histone modifications [389], SETMAR could induce local chromatin changes at the *Hsmar1* transposase gene promoter, thereby regulating transposase expression.

SETMAR has likely been under selection in human cells for a function other than its residual nicking activity, but this function remains enigmatic. Cordeaux *et al*. (2006) have found that selection has been preserving the IR-binding activity of the SETMAR transposase [391]. Thus, the function of the SETMAR protein is likely associated with its ability to specifically recognize numerous genomic binding sites represented by the *Hsmar1* IRs. It is tempting to speculate, that some of these binding sites are conserved because targeted chromatin modifications by SETMAR at these genomic locations are required for normal cellular functions. Ongoing work will have to clarify the past and present functions of SETMAR, making use of the active *Hsmar1* transposon as an experimental system.

3.1.7 Transposition of a reconstructed *Harbinger* element in human cells and functional homology with two transposon-derived cellular genes (Paper X)

PIF/Harbinger is a superfamily of eukaryotic DNA transposons found in diverse genomes including plants and animals [395, 401-405]. Only few *PIF/Harbinger* elements have been reported to be active. The *P instability factor* (*PIF*) and its associated miniature inverted-repeat transposable element called *mPIF* were found to actively transpose in maize [395]. In

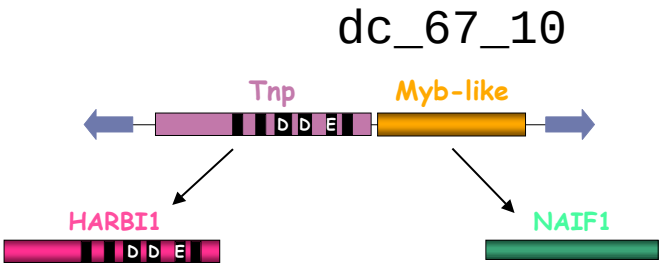


Figure 50. Schematic representation of *Harbinger3_DR* and similarities of transposon-encoded proteins to cellular factors. Structure of autonomous *Harbinger3_DR* elements. IRs are indicated by gray arrows. The transposase (Tnp) and the Myb-like protein gave rise to the domesticated vertebrate genes HARBI1 and NAIF1, respectively.

rice, the *mPing* element can be mobilized upon *trans*-activation by its autonomous partner *Pong* [393, 406].

Harbinger3_DR is one of the three families of *PIF/Harbinger* transposons described in the zebrafish genome [407]. The family contains five full-length elements predicted to be inactive due to mutations and about 1000 copies of a shorter element called *Harbinger3N_DR. Harbinger3N_DR* does not have coding capacity, but it shares most of its sequences including the IRs with *Harbinger3_DR*; therefore, these elements likely used the transpositional machinery of autonomous elements for propagation. *Harbinger3_DR* contains two genes flanked by short, 12-bp IRs and 3-bp TSDs. The first gene encodes a transposase, whereas the second gene encodes a protein of unknown function that contains a SANT/Myb/trihelix domain, and hence is referred to as the Myb-like protein (Fig. 50) [403, 405, 407]. This motif is characterized by three alpha helices and the conservation of three bulky aromatic residues, and might be involved either in a DNA-binding function similar to that observed in Myb-related transcriptional regulators (Myb-like domain) or in protein-protein interactions as described for chromatin remodeling factors (SANT-like domain) [408]. Both genes encoded by *Ping* and *Pong* elements were recently found to be required for *mPing* transposition [409].

Transposons can contribute to the emergence of new genes with functions beneficial to the host via an evolutionary process referred to as “molecular domestication” (as discussed in section 1.5.2.1.1). *PIF/Harbinger* transposons also contributed to the evolution of cellular genes. In *Drosophila*, the *DPLG1-7* genes were recruited from at least three dictinct *PIF*-like transposase sources [405]. In vertebrates, the *HARBI1* gene constitutes the only known example of domesticated genes derived from a *PIF/Harbinger* transposase (Fig. 50) [407]. HARBI1 is conserved in all studied jawed vertebrates, and is most similar to the *Harbinger3_DR* transposase with a 30-40% sequence identity. Since the putative catalytic

motifs including the DDE triad present in *PIF/Harbinger* transposases [403, 407] are preserved (Fig. 50), HARBI1 is expected to retain catalytic, transposase-related activities.

tBLASTn searches identified NAIF1 (nuclear apoptosis-inducing factor 1), also referred to as C9ORF90, as a protein closely related to the Myb-like protein (Fig. 50). NAIF1 was previously characterized as a single-copy gene conserved across vertebrates [410]. An alignment between the Myb-like transposon proteins and the fish, frog, bird and mammalian orthologs of NAIF1 revealed high homology between the N-terminal region of NAIF1 (spanning residues 1 to 92) and the N-terminal region of the Myb-like protein (spanning residues 1 to 90) with 36-38% of sequence identity (not shown). The position of the putative trihelix motif and the three bulky aromatic residues are conserved in NAIF1 (not shown), suggesting potential functional homology with the Myb-like protein. The NAIF1 and HARBI1 proteins are not detectable in the recently assembled genomes of the jawless vertebrates *Pertomyzon marinus* (lamprey), tunicates *Ciona intestinalis* and *Ciona savignyi* (sea squirts), and deuterostoma *Strongylocentrotus purpuratus* (sea urchin). Therefore, it appears that both proteins have emerged in a common ancestor of jawed vertebrates after its separation from jawless vertebrates some 500 million years ago. Phylogenetic analysis of the NAIF1 and HARBI1 proteins suggests that both highly conserved proteins have evolved in a similar mode, which may be due to their involvement in the same molecular pathway. Functional studies have shown that overexpression of human NAIF1 induced apoptosis, and that its N-terminal region was critical for its apoptosis-inducing function [410]. However, the physiological role of NAIF1 remains unknown.

Based on the consensus sequences established previously [407], transposon components projected to be sufficient for *Harbinger* transposon mobility, namely, a non-autonomous *Harbinger3N_DR* element and the coding sequences for both the transposase and the Myb-like protein were synthesized. The transposon components were used to set up a cell-based transposition assay similar to that established for *Sleeping Beauty* [85]. The system consisted of a transposon donor plasmid carrying an SV40 promoter-driven neomycin-resistance gene (neo) inserted into the consensus *Harbinger3N_DR* element [pHarb(SV40-neo) in Fig. 51] and two helper plasmids expressing the transposase and the

dc_67_10

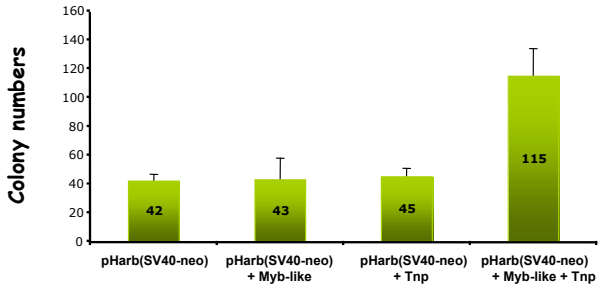


Figure 51. **Transposition of the reconstructed *Harbinger* transposon system in HeLa cells**. The numbers represent the mean values of the colony numbers in three independent assays. The error bars indicate SEM.

Myb-like protein [pFV4a(Tnp) and pFV4a(Myb-like) in Fig. 51]. The pHarb(SV40-neo) plasmid was transfected together with either pFV4a(Tnp) or pFV4a(Myb-like) or both in HeLa cells. Transposition, and its efficiency, was assessed from the numbers of G418-resistant colonies. Cotransfection of either the transposase- or the Myb-like protein-expressing plasmid together with the transposon donor construct did not increase colony numbers (Fig. 51). However, coexpression of both proteins produced neomycin-resistant colonies at a 2.7-fold higher rate than transfection with the donor plasmid alone, indicative of chromosomal transposition events (Fig. 51). Because HARBI1 was found to be the most closely related to the *Harbinger3_DR* transposase [407], the zebrafish ortholog of HARBI1 was also tested, and found to be deficient in transposition (not shown). Inactive transposase mutants might act as regulators of transposition; however, coexpression of HARBI1 together with the transposon components did not have any appreciable effect on *Harbinger* transposition (not shown).

In silico analysis of a large number of *Harbinger3_DR* and *Harbinger3N_DR* integration sites in the zebrafish genome revealed a 17-bp palindromic target site centered on the CWG triplet [407]. To investigate target site preferences of the reconstructed *Harbinger* element in human cells, we analyzed a total of 46 transposition events isolated from three independent transposition assays. 95% of the insertions (44/46) were flanked either by CAG or CTG trinucleotides. Sequence logo analysis of the 46 genomic

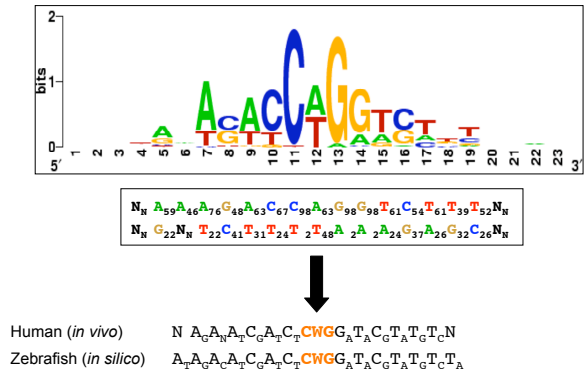


Figure 52. *Harbinger* transposon integration sites. WebLogo analysis of 23-bp insertion sequences. The most frequent nucleotides at each position and the alternative, frequently appearing nucleotides are indicated with their frequencies. (***bottom***) Alignment of consensus target sequences derived from *de novo* integration events of the reconstructed *Harbinger* system in human cells and from *in silico* studies in zebrafish.

integration sites revealed a 15-bp consensus sequence including the CWG target site (Fig. 52), which matches the zebrafish consensus in 15 out of the 17 base pairs (positions 1 and 17 being not conserved) (Fig. 52). Taking into account the alternative nucleotides at each position in the zebrafish consensus, each of the 46 integration sites retains at least 12 out of the 17 base pairs. Thus, our data demonstrate that the *Harbinger* transposon retains its target site specificity independent of the host genome. This highly selective target site choice by the reconstructed *Harbinger* element is unique even within the *PIF/Harbinger* transposon family. Thus, the *Harbinger* transposon system may serve as a useful experimental tool for investigating determinants of target site selection of mobile genetic elements, as well as for establishing technologies for site-specific transgene integration.

Since both the Myb-like protein and the transposase were required for the transposition process, their possible physical interaction was examined by coimmunoprecipitation. Myc-tagged Myb-like protein (Myb-like/Myc) and hemagglutinin-tagged transposase (Tnp/HA) were co-expressed in HeLa cells. The transposase was precipitated with an anti-HA antibody, and the immunoprecipitated proteins were analyzed for the presence of the Myb-like protein by immunoblotting with an anti-Myc antibody. As shown in Fig. 53A, the anti-HA antibody coprecipitated Myb-like/Myc (lane 4), indicating that the transposase and the Myb-like protein form a complex in cells. This interaction did not require transposon DNA (compare lanes 4 and 6 in Fig. 53A) and was specific, because the anti-HA antibody failed to coprecipitate either Myb-like/Myc when coexpressed with HA-tagged Jazz-*SB* transposase [411] or Myc-tagged Rep78 of AAV when coexpressed with Tnp/HA (Fig. 53B, lanes 3 and 4).

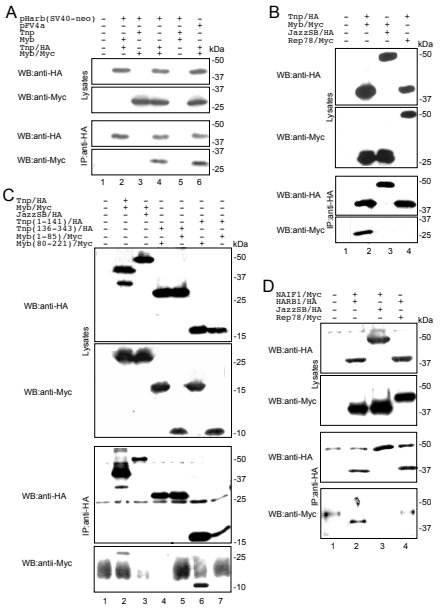


Figure 53. Physical interactions between the transposase and the Myb-like protein, and between HARBI1 and NAIF1 in human cells. **(A)** Interaction of the transposase (Tnp) with the Myb-like protein (Myb-like). Lysates and immunoprecipitates (IPs) were analyzed by Western blotting (WB) with anti-HA and anti-Myc antibodies. **(B)** Specificity of the interaction between Tnp and Myb-like protein. **(C)** Mapping of interaction domains for Tnp and Myb-like protein. (**D**) Physical interaction of HARBI1 with NAIF1.

Reciprocal experiments confirmed these results, since an anti-Myc antibody coprecipitated Tnp/HA when co-expressed with Myb-like/Myc, but not when co-expressed with Myc-tagged Rep78 (data not shown). In order to map the regions of both the Myb-like protein and the transposase that are essential for interaction, two deletion mutants were tested for each protein by coimmunoprecipitation. Myb-like(1-85) expresses the N-terminal region and Myb-like(80-221) lacks the N-terminal region of the Myb-like protein, whereas Tnp(1-141) contains the N-terminal 141 residues and Tnp(136-343) is restricted to the C-terminal 209 residues of the transposase. The anti-HA antibody coimmunoprecipitated Tnp(1-141) only when coexpressed with Myb-like(80-221) (Fig. 53C, lane 6). These data indicate that the interaction between the transposase and the Myb-like protein requires domains located in the N-terminal region of the transposase (amino acids 1-141) and the C-terminal region of the Myb-like protein (amino acids 80-221).

As a first step towards a functional analysis of HARBI1, we employed coimmunoprecipitation to assess its possible interaction with NAIF1 (Fig. 53D). Analysis of immunoprecipitates revealed efficient coprecipitation of Myc-tagged NAIF1 with HA-tagged HARBI1 (Fig. 53D, lane 2), suggesting that HARBI1 and NAIF1 associate with each other in cells. No immunoprecipitation was detected for cells coexpressing either NAIF1/Myc and HA-tagged Jazz-SB (lane 3) or Myc-tagged Rep78 and HARBI1/HA (lane 4), showing specificity of the interaction. These data provide evidence for a transposase/Myb-like protein interaction, and suggest that such interaction plays a role in transposition of *Harbinger3_DR*. Similar, HARBI1 interacts with NAIF1, suggesting functional parallels to the transposon components.

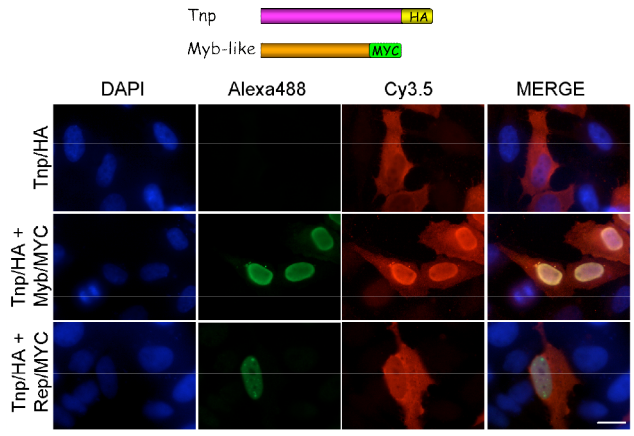


Figure 54. Subcellular localization of the transposase and the Myb-like protein. Colocalization assays of the full-length Tnp/HA and Myb-like/Myc proteins. From the left, the first panel shows DAPI staining, the second panel shows the green channel (Alexa488), the third panel shows the red channel (Cy3.5) and the last panel shows merged images. Scale bars=20 μm.

Having found that the Myb-like protein interacts with the transposase,

we examined the subcellular localization of both proteins. Red fluorescent protein-tagged Myb-like protein displayed specific localization to the nuclei of transiently transfected HeLa, whereas enhanced green fluorescent protein-tagged transposase was found to have cytoplasmic and nuclear distribution. Coimmunofluorescence was next applied to investigate potential effects of the transposase/Myb-like protein interaction on subcellular localization of both proteins. When Tnp/HA was expressed alone, it predominantly localized in the cytoplasm (Fig. 54, top). In contrast, when Tnp/HA and Myb-like/Myc were coexpressed, the transposase was enriched in the nucleus, where the Myb-like protein localized (Fig. 54, middle). Cotransfection of the Tnp/HA and Myc-tagged Rep78 showed Rep78 localization in the nucleus and intranuclear centers as expected for Rep [412], and a predominant transposase localization in the cytoplasm, similar to that observed in cells expressing transposase alone (Fig. 54, compare bottom and top panels).

Subcellular localization of HARBI1 and NAIF1 was investigated using the same experimental approach as described above. Both a physical interaction between HARBI1 and NAIF1 and their similarities to the transposon-encoded transposase and the Myb-like protein suggest that the two proteins may colocalize in cells. Indeed, cells coexpressing HARBI1/HA and NAIF1/Myc showed a dramatic relocalization of HARBI1 to produce a nuclear pattern characteristic of NAIF1/Myc (Fig. 55, top). In contrast, coexpression of Myc-tagged Rep78 with HARBI1/HA did not alter the subcellular localization pattern of HARBI1 (compare panels in Fig. 55). These results support the conclusion that NAIF1 promotes nuclear localization of HARBI1. In sum, both the Myb-like protein and NAIF1 are nuclear proteins that aid nuclear import of the transposase and HARBI1, respectively, an important step in biochemical reactions that involve DNA, including transposition.

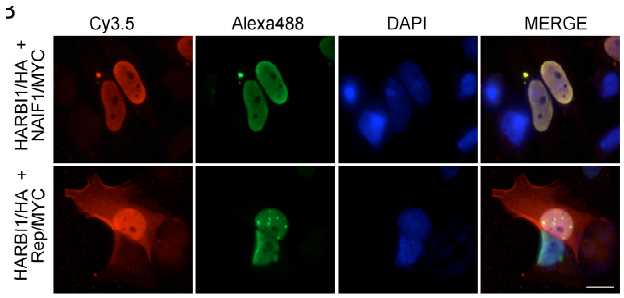


Figure 55. Subcellular localization of HARBI1 and NAIF1. Colocalization assays of HARBI1/HA and NAIF1/Myc.

Interaction of transposase molecules with the terminal regions of the transposon is a requirement for cut-and-paste transposition. In case of *Harbinger3_DR*, either one or both of

the two, transposon-encoded proteins is expected to have a DNA-binding function. The *PIF/Harbinger* transposases and the HARBI1 proteins have been predicted to contain a single HTH motif compatible with DNA-binding capacities [405]. Based on the presence of a putative Myb-like trihelix domain with a highly electropositive predicted surface charge (theoretical pI=10), the Myb-like protein is expected to have a DNA-binding activity [408].

In order to test the capacity of the transposase and the Myb-like protein to bind transposon DNA, EMSA was employed using MBP-tagged, purified proteins. MBP/Myb-like(1-85), MBP/Myb-like(80-221), MBP/Tnp(1-141) and MBP/Tnp(136-343) were incubated with a probe corresponding to the 5'-UTR of the *Harbinger3_DR* transposon including the left IR and flanking consensus target sequence. MBP/Myb-like(1-85) produced retarded bands, whereas MBP/Myb-like(80-221) did not (Fig. 56A), demonstrating that the trihelix motif is necessary and sufficient to bind DNA. No shift was observed for either MBP/Tnp(1-141) or MBP/Tnp(136-343), indicating that only the Myb-like protein has the capacity to bind transposon DNA (Fig. 56A). Increasing concentrations of MBP/Myb-like(1-85) in the binding reaction produced more slowly migrating complexes, indicating either the presence of multiple binding sites in the probe that became saturated, or multimerization of the protein upon DNA-binding (Fig. 56A). In order to map the binding sites of the Myb-like protein in the transposon, an overlapping series of double-stranded oligonucleotides covering the full consensus sequence of

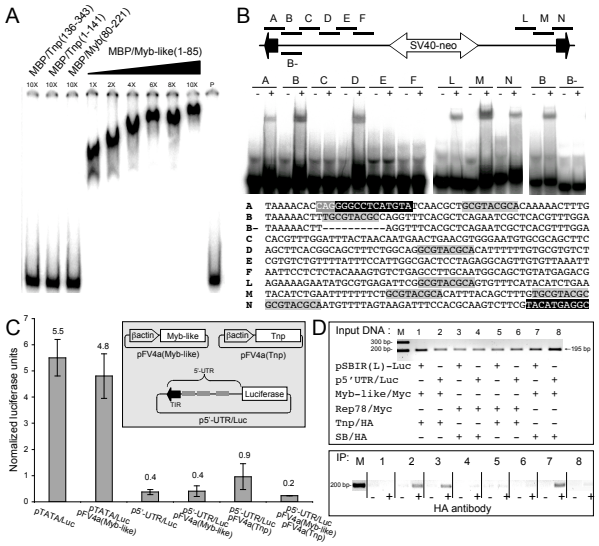


Figure 56. DNA-binding activities of the transposase and the Myb-like protein. (A) EMSA of MBP/Tnp(136-343) (1x=261 nM), MBP/Tnp(1-141) (1x=261 nM), MBP/Myb-like(80-221) (1x=291 nM) and increasing concentrations of MBP/Myb-like(1-85) (1x=320 nM) mixed with a 486-bp *Harbinger3_DR* transposon probe (depicted in Fig. 1A). **(B)** Mapping of the Myb-like protein binding sites. On the top, a schematic of the Harb(SV40-neo) element is shown with the relative positions of selected oligonucleotides used as probes in EMSA. Each reaction was performed with (+) and without (-) MBP/Myb-like(1-85) (600 nM). The sequences of the oligonucleotides are indicated with the TIRs highlighted in black (in probes A and N) and the 9-bp binding sites of the Myb-like protein highlighted in gray. **(C)** Luciferase reporter assay. The diagram represents reporter gene expressions (indicated on the *y* axis) in HeLa cells from the plasmids indicated below, in the absence or presence of pFV4a(Myb-like) and/or pFV4a(Tnp). The schematic representation of the luciferase reporter construct in p5'-UTR/Luc is depicted in the inset. **(D)** Chromatin immunoprecipitation (ChIP) assay. Transposase-complexed DNAs were precipitated using anti-HA antibody. PCR was performed with total DNA (Input DNA) and immunoprecipitated DNA (IP) using primers for the luciferase coding region generating a 195-bp product. M: size marker.

Harbinger3N_DR was tested for binding. Three binding sites were identified in both ends of the transposon sharing the 9-bp palindromic sequence motif 5'-GCGTACGCA (Fig. 56B). This sequence motif indeed contitutes the binding site of the Myb-like protein, because an oligonucleotide lacking the site was not shifted (compare probes B and B- in Fig. 56B). We conclude that the Myb-like protein binds six sites in the transposon ends via its trihelix motif. Since NAIF1 is predicted to have a trihelix motif similar to that described for the Myb-like protein, we tested its ability to bind DNA. Using the same probe as above, NAIF1 was found to bind to DNA, but no shift was observed for HARBI1 (not shown).

The SANT/myb/trihelix motif was found to function as a DNA-binding domain for a large number of transcription factors [408]; thus, the myb-like protein may play a role in transcriptional regulation of the transposase [407]. Transcriptional activation of the 5'-UTR of the transposase gene fused to a luciferase reporter was measured in an *in vivo* one-hybrid DNA-binding assay (Fig. 56C). The p5'-UTR/Luc and the control pTATA/Luc reporter plasmids were transfected into HeLa cells with or without pFV4a(Myb-like) and pFV4a(Tnp). The Myb-like protein apparently did not affect reporter expression (Fig. 56C), arguing against a role in transcriptional regulation of the transposase gene.

To investigate potential recruitment of the transposase into a complex formed by the Myb-like protein and transposon DNA, *in vivo* chromatin immunoprecipitation (ChIP) was used following cotransfection of cells with plasmid DNA containing the 5'-UTR of *Harbinger3_DR* and Tnp/HA with or without Myb-like/Myc (Fig. 56D). The 5'-UTR of *SB* together with HA-tagged *SB* transposase served as positive control for the assay. After cross-linking, transposase-complexed DNAs were precipitated using anti-HA antibody coupled to agarose beads, and amplified using a diagnostic PCR. As expected, *SB* transposon DNA was precipitated in an *SB* transposase-dependent manner irrespective of coexpressed proteins (Fig. 56D, lanes 1, 3, 5 and 7). Similar, precipitation of *Harbinger* transposon DNA required expression of the *Harbinger* transposase (compare lanes 2 and 8), but was only seen when the transposase was coexpressed with Myb-like/Myc (Fig. 56D, lane 2), but not when coexpressed with Rep78/Myc (Fig. 56D, lane 6). PCR products were only recovered in antibody-treated samples (Fig. 56D). Taken together, the results suggest that

the Myb-like protein contributes to *Harbinger* transposition by binding to the transposon DNA, and by recruiting the transposase to the transposon ends.

Our data are compatible with a transpositional model in which the two, transposon-encoded proteins contribute distinct functions to provide a transpositionally active complex (Fig. 57). The Myb-like protein promotes nuclear import of the transposase, and likely participates in forming a synaptic complex by directly binding to subterminal regions of the transposon and by recruiting the transposase to the transposon ends. Although quite unique among eukaryotic transposons, the requirement for multiple transposition factors is not without precedent. For example, transposition of the *En/Spm* element in maize was found to require two proteins, TnpA and TnpD, encoded by alternatively spliced transcripts derived from a single transcription unit [413]. The differential expression of the *Harbinger* element transposase and the Myb-like protein may contribute to the regulation of transposition: an intriguing concept subject of future investigations.

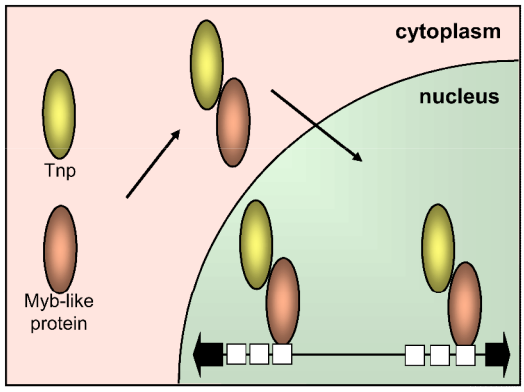


Figure 57. Model of the early steps of *Harbinger* element transposition. The transposase and the Myb-like protein form a complex in cells. The Myb-like protein promotes nuclear import of the complex, and binds to subterminal sequences in the transposon ends through its trihelix domain. The Myb-like protein recruits the transposase to transposon DNA.

The cellular functions of the vast majority of domesticated, transposon-derived genes remain largely enigmatic. We made steps towards functional characterization of the vertebrate HARBI1 and NAIF1 genes, and established functional homologies with the transposon-encoded proteins. Namely, similar to the interactions between the transposase and the Myb-like protein, NAIF1 interacts with HARBI1, promotes nuclear import of HARBI1, and acts as a DNA-binding protein. Thus, HARBI1 is expected to function in a DNA recombinational reaction together with NAIF1 as a cofactor. In functional analogy, NAIF1 might be a DNA-binding cofactor of HARBI1. Future investigations into the mechanism of *Harbinger* transposition and its regulation should facilitate novel discoveries regarding the cellular functions of NAIF1 and HARBI1.

3.2 DNA transposons as a gene delivery platform for genetic manipulations in vertebrates

3.2.1 Development of hyperactive *Sleeping Beauty* transposon vectors by mutational analysis (Paper XI)

Considerable effort has been devoted to the development of gene delivery strategies for the treatment of inherited and acquired disorders in humans (see section 1.6.3). A relatively new addition to the gene therapist's toolbox is TE-based gene vectors [85, 259]. Until very recently, transposon vectors were not available for genetic analyses in vertebrates. This is because the vast majority of elements currently residing in vertebrate genomes are transpositionally inactive [20, 31, 49, 220]. To address this problem, we reconstructed active transposons from vertebrate genomes (see previous sections). In particular, the *SB* system shows efficient transposition in a variety of vertebrate (including human) cell lines in tissue culture [102, 259] and in the mouse *in vivo*, both in somatic tissues [266] and in the germline [243, 267-269, 275].

Recent experiments from several laboratories have demonstrated some advantages of *SB* over the currently used viral and nonviral vectors, including stable, single copy integration [266, 414], use of simple plasmid vectors [259, 414], and long-term expression of integrated transgenes at therapeutic levels [266, 311, 315, 316, 320]. However, *in vivo* transformation rates with naked *SB* plasmids administered through the tail vein into mice were only around 5% in the liver [266]. Two immediately obvious areas where the efficiency of *SB*-mediated gene transfer could potentially be improved are the efficiency of vector delivery and the intrinsic transpositional activity of the element itself.

Previous improvements to transpositional activity have been made by manipulating either the transposon IRs or the transposase protein. For example, hyperactive *SB* vectors have been generated by reducing the length of vector DNA outside the transposon in donor plasmids [259], or by the introduction of site-specific mutations into the IRs [334]. However, even the improved vectors are subject to size-restrictions: transposition frequency of *SB* decreases with increasing the length of the transposon [259, 414, 415]. Large (>10 kb)

pieces of genomic DNA flanked by two identical copies of *Paris* elements have been mobilized in *Drosophila virilis* [416]. *Paris* is a *Tc1/mariner*-type transposon, indicating that mimicking such a naturally occurring arrangement could possibly extend the capacity of *SB* vectors to transpose large DNAs.

The "sandwich" element is a new transposable entity, in which DNA flanked by two copies of *SB* is mobilized. A requirement for such a transposon to work is the inability of the individual *SB* units to transpose on their own. The terminal nucleotides of *Tc1* element in *C. elegans* have been previously shown to be required for element excision [417]. Therefore, the terminal 5'-CA bases of the right IR of a neo-tagged *SB* transposon (pT/neo) were mutated to 5'-GC (Fig. 58A). To test the effect of these mutations on transposition, the activity of pT*/neo (with mutant right IR) transposon donor construct was compared to that of pT/neo (wild-type control), using the transposition assay described above. In the negative control, a plasmid expressing β-galactosidase (pCMVβ) replaced the transposase. The results showed that the CA→GC mutations completely abolish transposition of T*/neo (Fig. 58B), indicating that transposition of individual *SB* elements from sandwich constructs can be efficiently inhibited.

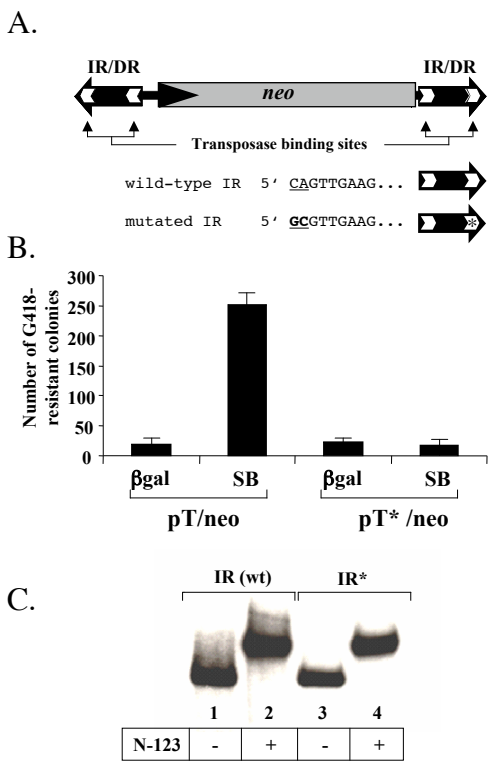


Figure 58. Mutations in the right inverted repeat of *Sleeping Beauty* interfere with transposition, but not with the binding capacity of the transposase. (A) Schematic representation of the SB transposon. The transposase gene is flanked by IR/DR-type inverted repeats (black arrows), which contain the binding sites for the transposase (white arrows at the ends of the IRs). Two base pair changes were introduced at the terminus of the right IR. The C in the sequence 5' CAGTTGAAG… is the first base of the transposon. **(B)** A transposon with mutant right IR cannot transpose. Efficiency of transposition is assessed as an increase in G418-resistant colony numbers in the presence (SB) versus in the absence (βgal) of transposase. Numbers are per $3x10^4$ transfected HeLa cells. The graphs show that a transposon that has the mutant IR (T*/neo) cannot be mobilized by the transposase. **(C)** Transposase can bind the mutant IR. Electrophoretic mobility shift assay using ^{32}P-radiolabeled wild-type (wt) or mutated (IR*) IR fragments as probes and N-123, a derivative of *SB* transposase containing the specific DNA-binding domain of the *SB* transposase within the N-terminal 123 amino acids.

All four binding sites within the IR/DR structure (Fig. 58A) are required for *SB* transposition [259]. Therefore, a further (suspected) requirement for the sandwich transposon to work is that the transposase should

be able to bind to all of its binding sites within the composite element. Both the wild-type and mutant IRs were radiolabeled, and their ability to be bound by the transposase examined in a mobility shift experiment, using N123 (the DNA-binding domain of *SB* transposase) (Fig. 58C). The results showed no difference between the wild-type and the mutant IR fragments in terms of binding to N123 (Fig 58C, lanes 2 and 4). These results therefore demonstrate that the induced mutations interfere only with the catalytic steps of transposition, but not with *SB* transposase binding.

Next, marker transgenes were cloned between two T* transposons (containing no transgene) in an inverted orientation, thereby forming a sandwich-like arrangement, from which the two individual *SB* elements cannot excise, but together define a new composite element (Fig. 59A). The structure of the sandwich transposon is therefore as follows: (intact left IR)-body of *SB* element-(disabled right IR) -- insert with a selection marker -- (disabled right IR)-body of *SB* element-(intact left IR). Earlier results showed that, although *SB* was able to transpose transgenes of up to ~10 kb, the efficiency of transposition significantly dropped as the elements got longer than 4 kb in length [259]. Therefore, a 4.7-kb piece of DNA was subcloned into the sandwich vector to yield a total transposon length of about 7.7 kb (construct pT/SA7.7 in Fig. 59B). An additional 4.5 kb piece of DNA containing the *lacZ* gene was subcloned into pT/SA7.7 to yield a total transposon length of 12.2 kb (construct pT/SA12.2 in Fig. 59B).

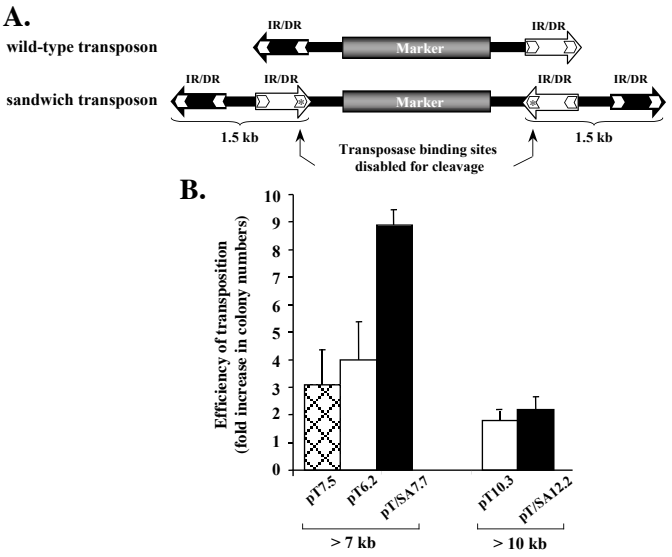


Figure 59. The sandwich *Sleeping Beauty* vector shows enhanced capacity to transpose long transgenes. (A) Outline of wild-type and sandwich *SB* transposons. In the sandwich vector, two complete *SB* elements flank a transgene to be mobilized in an inverted orientation. The individual *SB* units cannot transpose due to the mutations in their right IRs (asterisks). Only the full, composite element can transpose. The small, white arrows are the binding sites for the transposase within the left (large black arrows) and right (large white arrows) IR/DR repeats. **(B)** Comparison of the respective transpositional efficiencies of wild-type and sandwich transposon vectors. Transfections were done in human HeLa cells, and efficiency of transposition is expressed as a ratio of colony numbers in the presence versus in the absence of transposase. The graphs show that the sandwich transposon vector (black columns) has superior ability over wild-type transposons to integrate transgenes longer than 7 kb.

The efficiency of transposition of the sandwich

constructs was tested using the *in vivo* transposition assay and wild-type transposon constructs as controls for comparison. The sandwich transposon T/SA7.7 jumped about 3-fold more efficiently than the similar size, wild-type marker transposon T7.5, and about 2.2-fold more efficiently than T6.2, a wild-type transposon that contains the same transgene insert as T/SA7.7 (Fig. 59B). This result indicates that the sandwich vector is indeed more efficient in transposing relatively long DNA fragments than wild-type *SB*. Transposition of the sandwich element T/SA12.2 was still more efficient than that of a 10.3-kb-long wild-type transposon (Fig. 59B). However, the sandwich transposon apparently abides the same rule as wild-type *SB*, namely, that transposition rates are inversely proportional to the length of the transposon (Fig. 59B) [259]. Our results suggest that increasing the numbers of binding sites for the transposase can improve transposition of large size transposable elements, and establish the sandwich element as a useful transposon vector for stable integration of large transgenes. The structure of the sandwich transposon is somewhat similar to that of the bacterial transposons *Tn5* and *Tn10*. These elements might have been fortuitously generated by transposition of two insertion sequence (IS) elements on both sides of an immobile segment containing antibiotic resistance genes (Fig. 2E) [88, 418]. This situation can also arise, probably by chance, in other transposition systems, resulting in new, composite, mobile elements. Indeed, a pair of *Paris* elements that flank a nonrepetitive sequence of more than 10 kb in an inverted orientation was shown to be able to transpose in *Drosophila virilis* [416].

Why does the sandwich vector transpose long transgenes better than the wild-type *SB* transposon? We have shown earlier that 1) long elements tend to transpose less efficiently than short ones, likely because the ends of long elements cannot pair easily during synaptic complex formation [259]; and 2) the DNA-bending protein HMGB1 plays an important role in *SB* transposition likely by aiding the pairing of transposon ends (see section 3.1.1.1.4) [335]. Thus, we suggest that an increase in the number of transposase binding sites (from four to eight) in the sandwich transposon can partially rescue synaptic complex formation of long elements, presumably due to the more pronounced action of transposase-transposase interactions and HMGB1 at the transposon inverted repeats. Artificially made,

sandwich-like *Mos1 mariner* elements similar to the ones described here for *SB* have been found to have increased mobility in *Drosophila* [419], suggesting common underlying mechanisms in composite transposon mobilization in the *Tc1/mariner* family.

Transposons and their hosts have coevolved, and developed strategies that reduce the negative effects on the host but ensure proliferation of the element [31]. Thus, those elements that were apparently very successful in propagating themselves within a genome and in colonizing new genomes through horizontal transmission, such as *SB* [49, 220], are unlikely to represent their most active forms. This predicts that hyperactive versions of transposases can be generated by mutational analysis, which is the case for several transposases including *Tn5* [36, 420, 421], *Tn10* [422], *Himar1* [423] and *SB* [414]. There are several mechanisms of hyperactivity in transposases. For example, hyperactive phenotypes of the bacterial element *Tn5* are due to either the reduction of the self-inhibitory activity of intact *Tn5* transposase [421], a reduced affinity of an inhibitor protein to the transposase [36], or an increase in the binding affinity of the transposase to its binding sites within the transposon IRs [420]. The combination of these three hyperactive mutants yields a synergistic effect, leading to an extraordinarily active transposase [424]. Interestingly, amino acid replacements that change glutamic acid (E) residues to lysine (K) led to hyperactive transposase versions in three different transposon systems, *Tn5* [420, 421], *Tn10* [422] and *Himar1* [423]. Introducing a proline residue, a secondary structure breaker, at a defined site in the Tn5 transposase also resulted in a hyperactive mutant [36].

Mutations into genes can be introduced in either a random or site-directed fashion. We explored two different approaches of site-directed mutagenesis of the *SB* transposase: 1) replacement of acidic amino acids with basic amino acids; and 2) incorporation of naturally occurring sequence variants into the transposase. Based on findings that some of the *Tn5*, *Tn10* and *Himar1* hyperactive mutations are acidic to basic amino acid replacements, we hypothesized that similar mutations also have the potential to increase transpositional activity of the *SB* transposase. There are altogether 28 aspartic acid (D) and glutamic acid (E) residues in the *SB* transposase, from which 15 are not predicted to have crucial functions in the transposase; these were replaced by either lysine (K) or arginine (R) residues.

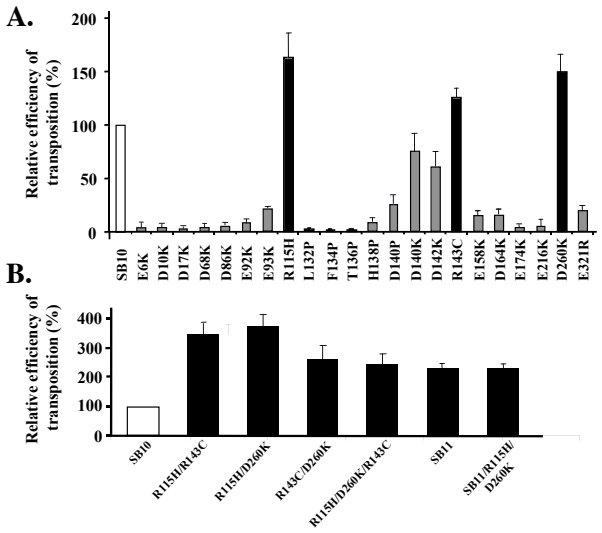


Figure 60. Effects of amino acid substitutions on the efficiency of *Sleeping Beauty* transposition. (A) Effects of single amino acid replacements. Shown are the transpositional efficiencies of 22 single-amino-acid mutants of the transposase relative to the wild-type (SB10) transposase (white column). From these, 15 are E to K/R mutations, 5 are proline mutations in the linker region between the DNA-binding- and catalytic domains and 2 are naturally occurring sequence variants. The three individual hyperactive mutants identified in this screen are shown as black columns. **(B)** Effects of hyperactive mutations in combinations. Transposition was assayed in human HeLa cells, and the activity of wild-type transposase (SB10, white column) is taken as a reference and set to 100%.

Mutations E6K, D10K, D17K, D68K, D86K, E92K, E93K, E158K, D164K, E174K, E216K and E321R reduced transposition frequency to barely measurable levels, whereas D140K and D142K reduced transposition to about 70% and 50%, respectively (Fig. 60A), indicating that these amino acids play critical roles in *SB* transposase activity. Importantly, however, transposition activity of D260K was about 40% higher than that of wild-type *SB* (Fig. 60A), demonstrating that an acidic-to-basic change in this position improves the function of the transposase.

Our second approach for mutagenesis of the *SB* transposase was to introduce amino aids that naturally occur in *SB* or related transposases. Such an approach has been shown to be useful for the generation of hyperactive versions of *SB* [414]. We evaluated the effects of two amino acid changes in the transposase: R115H and R143C. The R115H substitution was made based on a comparison between *SB* and the *Tdr1* transposase in zebrafish [80]. These two transposable elements represent closely related subfamilies of *Tc1*-like transposons in fish genomes, and show about 80% identity in transposase sequence. Therefore, these two sequences probably represent variants of a transposase that had been selected for activity in nature. The amino acid residue in position 115 in the *Tdr1* transposase is a histidine, which is expected to preserve the positive charge in this position of the transposase polypeptide. The second mutant that we tested, the R143C substitution, is a naturally occurring mutation in the *SB* transposase, possibly generated at a mutable CpG site in the transposase gene. The R143C mutant is also called *SB9*, and represents a particular intermediate version of the transposase that we obtained during the reconstruction process of the *SB* transposase gene [85]. Both the R115H and the R143C mutants showed

hyperactivity: R115H by about 60% and R143C by about 25% compared to the wild-type *SB* transposase (Fig. 60A).

Next, we asked the question whether combinations of the D260K, R115H, and R143C hyperactive mutations would result in an additive or a synergistic effect. Towards this end, the three possible double mutants and a triple mutant was engineered. R115H/D260K showed a 3.7-fold, R115H/R143C a 3.2-fold-, R143C/D260K a 2.6-fold-, and the R115H/D260K/R143C combination a 2.3-fold increase in transposition activity compared to the wild-type transposase (Fig. 60B). These results indicate that the R115H mutation acts synergistically with both D260K and R143C. We sought to determine whether incorporation of the previously described T136R/M243Q/VVA253HVR hyperactive mutations (collectively referred to as *SB11*) [414] would further increase transposition activity of our hypertransposing mutants. *SB11* showed a 2.3-fold higher activity than wild-type transposase, and this level of activity remained unchanged when *SB11* was combined with the R115H/D260K mutations (Fig. 60B). Altogether, our results demonstrate that a mutagenesis approach to the development of hypractive transposases is viable.

In summary, by using a limited site-directed mutagenesis screen, we identified hyperactive versions of the *SB* transposase. The rationale behind the change of all nonconserved acidic amino acid residues to basic amino acids is that in several transposition systems, including *Tn5* [420, 421], *Tn10* [422] and *Himar1* [423], some hyperactive mutations fall into this class. Acidic-to-basic amino acid changes might eliminate (or at least reduce) the unfavorable charge-charge interaction between the acidic amino acid residues and the negatively charged phosphate backbone of the transposon (or target) DNA [420], or might overcome the self-inhibiting properties of transposase [421]. Most of the mutations that we introduced into the *SB* transposase resulted in a decrease in the efficiency of transposition (Fig. 60A), suggesting very little functional redundancy in the transposase sequence. A marked sensitivity of transposase to mutations was previously noted for the *Mos1 mariner* element [74]. Nevertheless, one of the substitutions, the D260K mutation, produced a hyperactive phenotype. The aspartic acid in position 260 is either lysine or arginine in other *Tc1*-like transposases (Fig. 61), suggesting that lysine and arginine can better function in that

dc_67_10

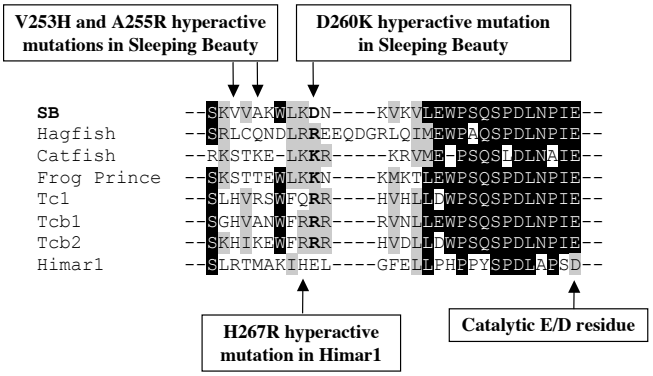


Figure 61. Locations of hyperactive mutations. Amino acid alignment of transposase segments of *Tc1*-like transposons and that of the *Himar1 mariner* element highlights a region where several hyperactive mutations are located. The alignment shows that D260 of *SB* is not conserved: several *Tc1*-like transposases contain lysine or arginine in this position. The E residue (D in *Himar1*) of the DDE catalytic triad is indicated.

sequence context. It is possible, that a particular version of fish *Tc1*-like transposases did contain K or R at position 260, but this amino acid got replaced at some point during transposase evolution, because it is functionally non-essential for the transposase. It is important to note that four hyperactive mutations reported from *Tc1*/*mariner* elements are located within the same eight-amino-acid segment in the catalytic domains of these transposases (Fig. 61). The D260K mutation (this work), and the V253H and A255R mutations [414] in *Sleeping Beauty*, and the H267R mutation in *Himar1* [423] all map to the same region just preceding the E/D residue of the DDE (DDD in *mariner* elements) catalytic triad (Fig. 61). It is therefore possible that these mutations result in a slight conformational change that is more favorable for catalysis. Because three out of these four mutations are K and R replacements (Fig. 61), the local shift to positive charge might enhance target DNA capture, a function likely encoded in the catalytic domain [112, 114]. Further biochemical work will be required to substantiate either of these hypotheses.

Using the apparently successful paradigm of incorporating phylogenetically conserved amino acids from related transposases into *SB*, 53 single amino acid variants were tested for activity in human HeLa cells. 25 of these substitutions resulted in hyperactivity as compared to the original *SB* transposase, underscoring the biological relevance of our phylogenetic approach. Subsequently, this collection of substitutions was supplemented with hyperactive mutations reported earlier [337, 414, 425, 426], resulting in a library of 41 clones, each containing a single hyperactive mutation. None of these individual mutations resulted in hyperactivity higher than 4-fold. However, as demonstrated above, combinations of individual mutations could potentially result in additive or synergistic effects. Due to the large number of possible combinations of 41 variants, it was necessary to develop a high-throughput, PCR-based, DNA-shuffling strategy and screening in mammalian cells. A

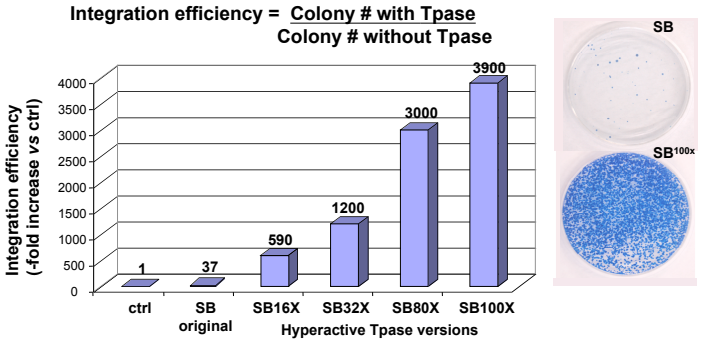


Figure 62. Comparison of different hyperactive versions of the *SB* transposase in transfected human HeLa cells. (***right***) Respective Petri dishes with stained, antibiotic-resistant cell colonies obtained with the original *SB* transposase as well as with the *SB100X* hyperactive variant.

library of mutant transposase genes was established with an average number of two mutations per gene in the hope of identifying pairwise, synergistic combinations. The best 38 clones isolated from 2,000 clones screened in total showed up to 25-fold higher activity as compared to the original *SB* transposase. On average, the best 8 of the 38 clones carried 3.6 mutations per gene, supporting our hypothesis that the key to achieve a higher degree of hyperactivity is finding the right combination of multiple hyperactive variants. We observed that 16 of the 41 hyperactive mutations were never recovered in combinations ("unfriendly" mutations), while 25 appeared repeatedly in the most hyperactive versions, indicating that only a fraction of hyperactive combinations are compatible with others ("friendly" mutations). Four clones were selected and used as a base for further rounds of manual combinations with "friendly" mutations identified by the screen. This strategy yielded a series of mutants displaying hyperactivity in a range of ~10-80-fold (Fig. 62). The most hyperactive version, hereafter referred to as *SB100X*, contained six combinatorial units that yield nearly 4,500,000 possible combinations, underscoring the necessity of combining "high throughput" and "analytical" strategies [427]. The availability of the novel hyperactive transposases may contribute to the development of efficient and safe non-viral vectors that would greatly facilitate clinical implementation of *ex vivo* and *in vivo* gene therapies and functional genomics studies.

3.2.2 *Frog Prince* transposon-based RNAi vectors mediate efficient gene knockdown in human cells (Paper XII)

The recent discovery and development of RNA interference (RNAi) to knock down the expression of a gene-of-interest has brought a widely applicable tool into the toolbox of the

molecular or developmental biologist. RNAi is a mechanism of post-transcriptional gene silencing mediated by double-stranded RNA (dsRNA) (reviewed in [428]). While introduction of long dsRNA works well in such organisms as *C. elegans* and *Drosophila*, it induces a strong cytotoxic effect in mammalian somatic cells [429]. Synthetic, 21-23-nucleotide short interfering RNAs (siRNAs) were shown to circumvent this response [430, 431]. However, despite their efficient and specific suppression of gene expression, siRNAs are costly to manufacture, the silencing effect is short-lived, and generation of stable knock-down cell lines (or organisms) is not possible. A refinement of this technique demonstrated that vector-based siRNAs allowed efficient gene silencing in transgenic cell derivatives due to stable chromosomal integration. These vectors express siRNAs through either convergent or divergent transcription [432, 433], or by expression from an Pol III promoter, such as U6 or H1, of a hairpin-containing inverted repeat, called short-hairpin RNA (shRNA) [434-436].

Initially, plasmid DNA was used as the vector, but this is inefficient in chromosomal integration. Stable cell lines containing and expressing shRNAs can be generated by cotransfection of a selectable marker or by placing the selectable marker on the shRNA expression plasmid [434]. However, expression of the selectable marker does not guarantee that the shRNA expression cassette is integrated and that it is expressed at a level high enough to elicit the desired silencing effect in either of these cases. This is because chromosomal integration of plasmid constructs is probably preceded by random breakage in the plasmid, which could prevent linkage of the marker and the shRNA expression cassette [437]. Viral vectors, such as retroviruses and lentiviruses, were developed to alleviate these problems [438, 439]. However, viral vectors impose problems such as a requirement for specialized laboratories, and complicated construction and preparation. Clearly, RNAi technology would benefit from the development of simple, plasmid-based vector systems that allow efficient chromosomal integration and stable expression of shRNA expression cassettes both in tissue culture and *in vivo*.

Transposons can be harnessed and developed into useful genetic tools (see section 1.6). In particular, the *SB* element has been developed as an efficient vector for gene transfer and insertional mutagenesis in vertebrates as well as for gene therapy in humans

[239]. Two recent reports have shown the utility of the *SB* system as an shRNA vector. Heggestad et al. (2004) demonstrated stable knockdown of both GFP and lamin A, whereas Chen and colleagues reported long-lasting knockdown of the human huntingtin transcript in human cells [280, 318]. The availability of multiple, transposon-based vector systems broadens the utility of these elements as genetic tools. Thus, another transposon of the *Tc1*/*mariner* superfamily, named *Frog Prince* (*FP*), was reconstructed from inactive elements found in the genome of the frog species *Rana pipiens* (see section 3.1.5) [382].

We assessed the utility of *FP*-based transposon vectors for efficient delivery and expression of shRNAs in cultured mammalian cells. The H1 Pol III promoter expression cassette from pSUPER [434] was subcloned into the *FP* transposon to drive expression of shRNAs. The H1 expression cassette was subcloned into the 3' end of the *FP* transposon, between the polyadenylation signal of the neomycin-resistance (*neo*) expression cassette and the right terminal inverted repeat, and the resulting construct was denoted as pFP/Neo-H1 (Fig. 63A). EGFP was chosen as the initial target for proof-of-principle tests. Four anti-EGFP shRNA oligonucleotides were designed using web-based parameters, and subcloned into both pSUPER and pFP/Neo-H1 for transient analysis. One of these (hereafter referred to as EGFP4) resulted in over 90% reduction in mean EGFP fluorescence in transient transfection assays using both pSUPER-based and transposon-based vectors with either orientation of the H1 shRNA expression cassette (not shown). The

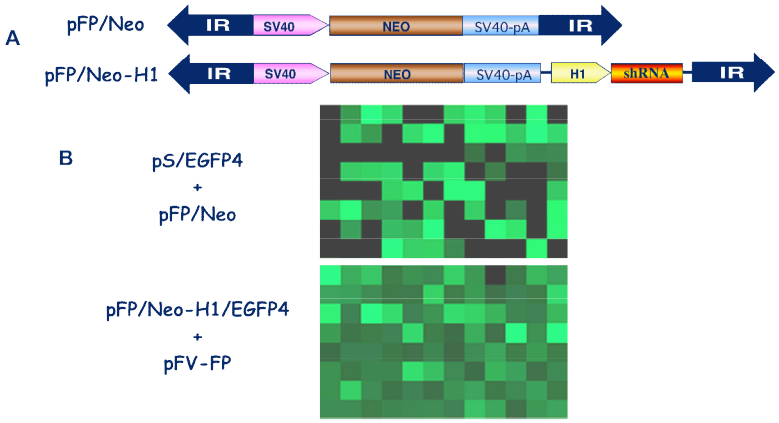


Figure 63. Stable knockdown of EGFP expression with *Frog Prince*-based shRNA expression vectors. (A) Schematic of *Frog Prince*-based shRNA vectors. The parental FP transposon contains the neomycin resistance gene (NEO) driven by the SV40 promoter (SV40) and followed by the SV40 polyadenlyation signal (SV40-pA) between the left and right flanking terminal inverted repeats (IRs). The H1 promoter expression cassette from pSUPER was subcloned between the polyadenylation signal and the right IR to generate pFP/Neo-H1. **(B)** Comparison of efficiencies of colony formation and EGFP knockdown with pSUPER- versus pFP-based shRNA vectors. A HeLa-derived cell line stably expressing EGFP was cotransfected with pSUPER/EGFP4 and pFP/Neo (upper panel), or pFP/Neo-H1/EGFP4 and pFV-FP (lower panel). Two days after transfection, the cells were diluted and plated into 96-well plates. After two weeks under G418 selection, EGFP expression was analyzed for resistant colonies in each well with a plate reader. Colonies were then fixed, stained and counted. The EGFP expression level per colony was calculated and graphically displayed using the Treeview program. Each square pixel represents a well and is color-coded. Black represents empty wells, while the shades of green reflect the level of EGFP expression per colony, with dark meaning a low level and bright meaning a high level.

EGFP4 shRNA oligo was subcloned into pFP/Neo-H1 to generate pFP/Neo-H1/EGFP4. EGFP knockdown was examined in a stably transgenic, HeLa-derived, EGFP-expressing clonal cell line. Cells were transfected with either pFP/Neo-H1-EGFP4 plus pFV-FP (transposition-mediated stable transgenesis) or with pS/EGFP4 plus pFP/Neo as a selectable marker ("no-transposition" control). The cells were plated to a 96-well plate such that pFP/Neo-H1-EGFP4-transfected cells gave at least one colony in almost every well. After two weeks under G418 selection, the level of EGFP expression per well was measured using a plate reader. The cells were then fixed, stained, and the colonies counted. The value of EGFP expression per colony was calculated and expressed in a graphical format. In Fig. 63B, each rectangle represents a 96-well plate, with each pixel being one well. The pixels are color-coded according to the level of EGFP expression per colony - with black meaning no colonies, and increasingly lighter shades of green meaning higher EGFP expression per colony. pFP/Neo-H1-EGFP4 was clearly more efficient at forming antibiotic resistant colonies, as 95 of the 96 wells contained colonies, whereas only 52 of 96 wells contained colonies for pS/EGFP4 (Fig. 63B). Furthermore, the average level of EGFP expression in wells with colonies was overall reduced in pFP/Neo-H1-EGFP4 colonies compared to pS/EGFP4. We conclude that FP transposon-based shRNA vectors are more efficient at colony formation and EGFP knockdown than non-transposon, plasmid-based vectors.

In summary, we showed that the *Frog Prince*-based shRNA system is capable of knocking down transcripts in human cells. In addition, it functions both transiently and in stable cell clones, either individually picked or pooled. Our results show that both the number and the knockdown effect of shRNA-expressing colonies are higher with the transposon- versus the plasmid-based approach. When the transposon vector system is used for establishing stable knockdown lines, the probability that an antibiotic-resistant clone is also a good knockdown clone was higher. One explanation is that transposition-mediated genomic integration maintains the shRNA cassette and the selectable marker physically linked within the transposable element. This can enhance the stability of the gene knockdown effect, compared to non-transposase-mediated integration, where the cleavage of the plasmid DNA

occurs randomly prior to chromosomal integration [437], which may lead to physical separation of the selectable maker and the shRNA expression cassette.

Together with previous studies demonstrating *SB* transposon-based stable knockdown of gene expression [280, 318], our data establish the advantages of transposon-based shRNA vectors for both stable and long-term knockdown of a target transcript. The availability of multiple transposon-based RNAi vectors increases the utility of these elements for regulating gene expression in vertebrates. The *FP*-shRNA system lends itself not only to generation of stable knockdown lines in cell culture, but also to generation of knockdown model organisms. The *FP* transposon system has shown activity in a wide host range of vertebrate cell types in culture, as well as in zebrafish embryos (see section 3.1.5) [382]. With the inclusion of either an inducible promoter or a tissue-specific promoter, the *FP*-shRNA system would allow spatial and temporal control over the silencing of developmentally important or essential genes, as well as generation of animal models of human disease through localized inhibition of specific targets.

RNA interference has also generated much interest as an alternate method for gene therapy, focusing on loss-of-function rather than gain-of-function. Current targets being tested include cancer, single gene disorders and viral infections (reviewed in [440]). The *FP*-shRNA system is perfectly suited as a vector for these applications. In addition to the benefits stated above over plasmid- and viral-based systems, the *FP* transposon is expected to be active in many, if not all, human cell types.

3.2.3 Comparative analysis of transposable element vector systems in human cells (Paper XIII)

The above sections provide important examples for the utilty of *SB* and *FP* transposon-based gene vectors for genetic applications. However, the transposon toolkit is expanding, as other transposons have been recently shown to catalyze efficient transposition in vertebrate model organisms. The *Tol2* transposon is a member of the *hAT* superfamily of TEs, and is endogenous in the medaka fish (*Orizyas latipes*) [242]. *Tol2* is the preferred transposon

system for transgenesis and insertional mutagenesis in zebrafish, and it is considered to be the current standard for the functionality and activity of the *hAT* transposon family in vertebrates including frog, chicken and mouse cells (reviewed in [441]). *PiggyBac* (*PB*) was isolated from the cabbage looper moth (*Trichoplusia ni*), and is a founder of the *piggyBac* superfamily of transposons [442]. It is active in mouse and human cells [256, 443, 444], and shows a great potential as a tool for transposon-based reprogramming of induced pluripotent stem (iPS) cells for regenerative medicine [445, 446] as well as for insertional mutagenesis in mice [256, 443, 447].

The *SB*, *Tol2* and *piggyBac* transposon systems not only differ in their phylogenetic origin, but might also differ in their biochemical properties affecting their activities under specific experimental conditions. Thus, an informed decision on which transposon vector system to use for a particular experimental goal should ideally be based on a comparative assessment of the properties of each system. However, a careful side-by-side characterization of these diverse transposon vector systems has been lacking.

Overall transpositional activity presents one of the main limiting factors for any transposon application. In order to enhance transposition efficiency, significant efforts have been put into modifications of the transposon systems, such as codon optimization for enhanced transposase expression, production of hyperactive transposase variants by mutagenesis and modification of the transposon IRs [337, 414, 425, 426]. Two recent outcomes of such efforts are a mouse codon-optimized *piggyBac* transposase gene (*mPB*) [448] and a hyperactive *SB* transposase called *SB100X* (see section 3.2.1) [427]; both represent the most effective transposases currently available.

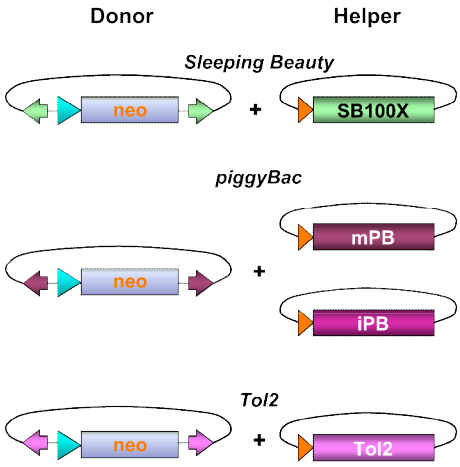


Figure 64. Transposase and transposon constructs for comparative studies. *Donor constructs*. Arrows represent transposon inverted terminal repeats, blue arrowheads: SV40 promoter; neo: neomycin resistance gene. *Helper constructs*. Orange arrowheads: CAGGS promoter driving expression of the different transposases.

In order to compare the activities of *SB100X*, *piggyBac* (including *mPB* and the native insect gene termed *iPB*) and *Tol2*, the respective

transposase coding regions were cloned into identical, CAGGS promoter-driven expression vectors (Fig. 64). This arrangement ensured expression of all four proteins from the same promoter, and excluded possible interference from different plasmid backbones. To be able to quantify transposition by colony forming assays, we inserted an *SV40neo* selection cassette between the IRs of the transposons previously shown to be minimally required for efficient transposition of each system [259, 449, 450]. To further minimize the difference between the systems, all transposons were inserted at the same site of their carrier plasmids. Thus, the only difference between the transposon vectors was the IR sequences, which are specifically required for interaction with their cognate transposases (Fig. 64).

Two-component transposition systems allow for the optimization of the transposition reaction by regulating the relative amounts of the transposon and transposase components in the transfected cells, which is of great importance for tuning transposon systems affected by overproduction inhibition (OPI), a phenomenon that results in inhibition of transpositional activity by excess transposase expression (reviewed in [322]). Thus, to characterize and compare transpositional activities of the three transposons in an unbiased fashion, and to find the optimal transposition (transfection) conditions for each system, we generated two independent transposase titration curves for two different transposon dosages in transfected HeLa cells. First we addressed OPI and enzyme activities by transfecting very “low” amounts of transposon donor plasmid (10 ng DNA per $2.5x10^7$ transfected cells) and varying amounts of transposase helper plasmids (Fig. 65).

SB100X reached its peak activity at as low as 5 ng of the transposase plasmid transfected, while *Tol2* required the highest amount (125 ng) of transfected helper plasmid to obtain its maximal activity (Fig. 65). *mPB* reached its peak activity at 50 ng of helper plasmid, whereas *iPB* needed as much as 250 ng of transfected transposase plasmid to produce its highest colony number (Fig. 65), consistent with a weaker expression of the insect transposase gene than the mouse codon-optimized gene in human cells. Neither of the three systems showed a plateau effect after having reached the peak (which would have indicated saturation of the transposition reaction). Instead, they all exhibited a distinct decrease in their

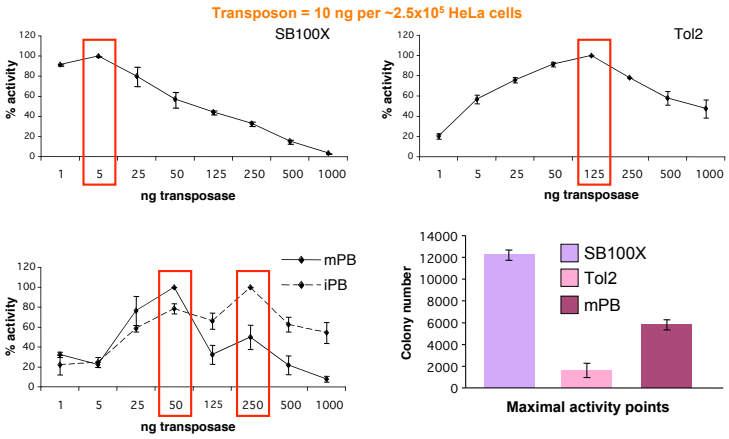


Figure 65. Transposition activity curves of *SB100X*, *Tol2* and *PB* transposons in low transposon DNA conditions. Transposition activity was measured under fixed amount of transposon plasmid (10 ng) cotransfected with increasing amounts of transposase expression plasmids into HeLa cells. Charts were derived from three independent transposition assays, error bars represent standard error of the mean (SEM). (***bottom right***) Transposition efficiency as measured by colony numbers obtained at the peak conditions normalized with transfection efficiencies for each system.

activities (Fig. 65), consistent with OPI. This phenomenon is well documented for *SB* [337, 414], but was not yet observed with the *Tol2* transposon system [257, 258], even though it is known that the *Ac* element, another member of the hAT superfamily, is regulated by OPI [451]. We next compared the numbers of antibiotic-resistant colonies produced by each system under conditions that supported their peak activities, normalized to transfection efficiencies (as measured by expression of a cotransfected *Venus* marker). At their peak activities under these conditions, the estimated transposition efficiencies per transfected cell population were about 10% for *SB100X*, 1% for *Tol2* and 3,5% for the *mPB* system (Fig. 65).

Next we addressed OPI and enzyme activities by transfecting “high” amounts of transposon donor plasmid (500 ng DNA per 2.5x10^7 transfected cells) and varying amounts of transposase helper plasmids. *SB100X* reached its peak activity at 50 ng of transfected helper plasmid, while *mPB* and *Tol2* both required 250 ng transposase plasmids to reach their maximal transpositional efficiencies (not shown). As seen before, all three systems showed a decrease in colony numbers beyond the peak conditions, consistent with OPI. At their peak activities under these conditions, the estimated transposition efficiencies per transfected cell population were about 31% for *SB100X*, 12% for *Tol2* and 27% for *mPB*. In sum, *Sleeping Beauty* was found to be the most efficient and *Tol2* the least efficient transposon in HeLa cells, with *piggyBac* showing an intermediate level of activity. Our data suggest that OPI has the capacity to downregulate all three systems, underscoring the importance of careful optimization of transposon system components for gene transfer experiments.

The development of efficient and safe non-viral vectors would greatly facilitate clinical implementation of cell- and gene-based therapies. It has been previously shown that *ex vivo* transfection of human CD34$^+$ cells with *SB100X* resulted in efficient gene marking with resulting robust and stable gene expression, as well as multi-lineage hematopoietic reconstitution after transplantation into immunodeficient mice [427, 452].

In order to compare the *SB100X* and *mPB* systems in clinically relevant human cells, CD34$^+$ cell populations enriched in hematopoietic stem/progenitor cells (HSCs) were co-transfected by nucleofection with varying amounts of a GFP-marked *SB* and *PB* transposons and CMV promoter-driven helper plasmid encoding the *SB100X* and *mPB* transposases. Stable gene transfer was assessed by GFP$^+$ colony forming units (CFUs) in clonogenic assays following *in vitro* differentiation into the erythroid (CFU-E) and granulocyte-monocyte-macrophage (CFU-GM) lineages (Fig. 66). *SB100X* produced increasingly higher numbers of CFU-E colonies as compared to *mPB* in a transposon component dosage-dependent manner (Fig. 66). An optimization of *piggyBac* transposition against that of *SB* under conditions previously determined to be optimal for the *SB* system in HSCs [427] revealed that *SB100X* was significantly more efficient in human CD34$^+$ cells than *mPB* in both CFU-E and CFU-GM clonogenic assays (% GFP$^+$ 33-35% for *SB100X* vs 7-9% for *mPB*, Fig. 66). Thus, the robust transposition of the *SB* system previously observed in HeLa cells can be translated to clinically relevant human HSCs.

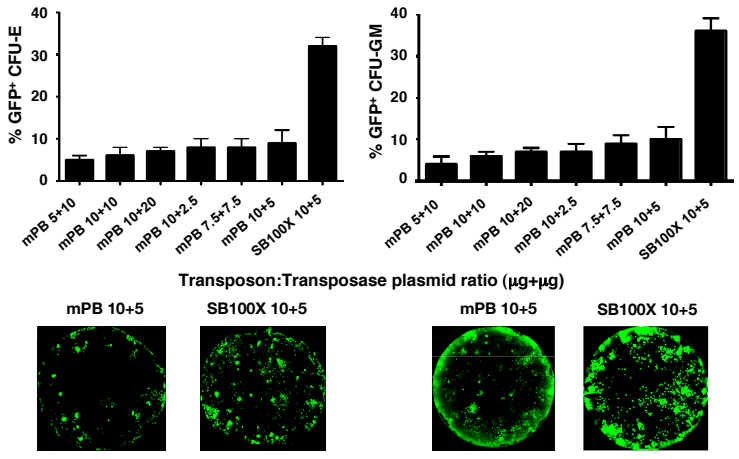


Figure 66. Transposition mediated by *Sleeping Beauty* and *piggyBac* in primary human CD34$^+$ hematopoietic stem cells. Cord blood-derived CD34+ cells were nucleofected with a *mPB* or *SB* transposon plasmid along with plasmids encoding the respective *mPB* or *SB100X* transposases. The amount of transposon and transposase plasmid is indicated (in μg). The total amount of DNA tranfected into the cells was kept constant (15 μg) by topping up with carrier DNA. The total numbers of GFP+ CFUs in the erythroid (CFU-E) and granulocytic/monocytic/macrophage (CFU-GM) lineages were compared. The mean values of GFP+ CFUs +/- standard deviation are shown. Representative mosaic images of the culture plate reflect the overall GFP+ colonies.

For many transposon applications it is necessary to titrate the

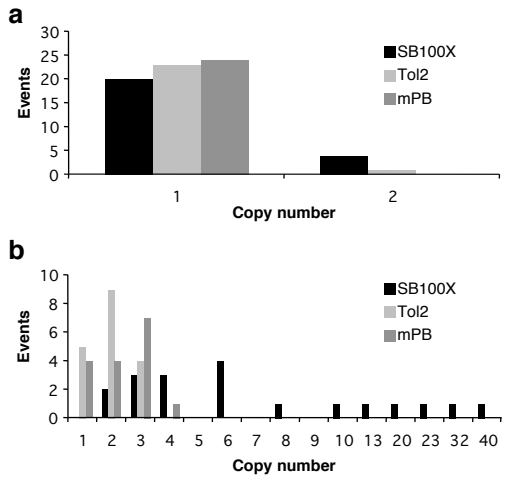


Figure 67. Transposon copy numbers generated by *SB100X*, *Tol2* and *mPB*. Transposon copy number for each system was detected by dot-blot analysis of genomic DNA obtained from HeLa clones resulting from transfections at activity peak point conditions. **(a)** Copy numbers in "low" transposon conditions. 24 clones for each system were analyzed. **(b)** Copy numbers in "high" transposon conditions. 19, 18 and 16 clones for *SB100X*, *Tol2* and *mPB*, respectively, were analyzed.

transposon components in order to control the numbers of genomic transposon insertions per transposed cell. For example, for loss-of-function mutagenesis and for therapeutic applications it is advantageous to keep copy numbers low, whereas somatic mutagenesis for cancer gene discovery often requires the cumulative effects of multiple insertions in the same cell. Genomic DNA prepared from HeLa cell clones obtained at the "low" and "high" transposon dosages at the peak activities of each of the transposon systems as described above was analyzed for integrated transposon copy number per clone by dot blotting. Altogether we analyzed 24 clones obtained at the "low" conditions for each system, and 19, 18 and 16 clones obtained at the "high" conditions for *SB100X*, *Tol2* and *mPB*, respectively. At the "low" conditions, the vast majority of clones had a single insertion for all of the *SB100X*, *Tol2* and *mPB* systems, and none of the clones had more than two integrated transposons (Fig. 67A). At the "high" conditions, *SB100X* generated insertions in a range of 2-40 copies (most of them falling in the range between 2 and 6 insertions), *Tol2* in a range of 1-3 copies and *mPB* in a range of 1-4 copies per cell clone, corresponding to an average copy number of ~10/clone for *SB100X*, and ~2/clone for both *Tol2* and *mPB* (Fig. 67B). The results suggest that the *Tol2* and *mPB* systems tend to produce insertions in the low copy range even when availability of the transposon is not limiting the transposition reaction, whereas *SB100X* transposition can occur within a relatively wide range of copy number that is tunable by dosing the transposon DNA. After correcting the transposition efficiencies as measured by the colony formation assays at the peak conditions for both the "low" as well as the "high" conditions with the average numbers of insertions per colony, relative transpositional activities of the three transposon systems were calculated: the *SB100X*

transposase is about 12-16-fold more potent than *Tol2*, and about 3-6-fold more potent than *mPB* in overall transpositional activity in HeLa cells.

Transposons can be succesfully harnessed as vechicles for introducing transgenes into the genomes of recipient cells. However, any exogenous transcription unit introduced into the host genome is a potential target for postintegrative epigenetic modifications that could result in some degree of expressional silencing. This attenuation of transgene expression might be induced by several factors such as high copy number, chromatin context at the insertion site, transgene expression units such as promoter sequences and particular features of the introduced transgenes including DNA sequence [453]. In order to compare stability of transgene expression by the *SB100X*, *Tol2* and *mPB* systems, a CAGGS/Venus/IRES/neo/SV40pA expression cassette was cloned between the IRs of each transposon, allowing for selection of chromosomal integration events by antibiotic selection and monitoring of transgene expression by *Venus* fluorescence. This arrangement ensured that possible differences in transgene expression between the three systems are not conferred by differences in transgene sequences, but either by intrinsic features of transposon IRs or by transposon target site selection. The transposon constructs were transfected in HeLa cells in conjunction with their cognate transposases under conditions that predominantly yield a single transposon insertion per cell (confirmed by dot-blotting, data not shown), and do not yield transposase-independent genomic insertions. The same expression cassette containing no transposon sequences and transfected in the absence of any transposase served as control for silencing of random plasmid integration events. After three weeks of selection in G418, cell clones were regularly replated and monitored for *Venus* fluorescence for 6 additional weeks in the absence of antibiotic selection.

Altogether we picked and monitored 294 *SB100X*, 186 *Tol2*, 178 *mPB* and 98 control clones, and detected 18 transposon-associated- and 26 plasmid-associated silencing events ranging from mosaicism to complete loss of transgene expression (Table 1). All of the three transposon systems showed low levels of transgene silencing in the range of 2-3% of the cell clones (Table 1). In contrast to the low silencing frequencies of transposon-delivered

	Monitored clones	Silencing events	%	Complete silencing	%	Mosaic colonies	%	Uniform silencing	%
SB100X	294	5	1.7	2	0.7	3	1	0	0
Tol2	186	7	3.8	4	2.2	3	1.6	0	0
mPB	178	6	3.4	0	0	6	3.4	0	0
control	98	26	26.5	8	8.16	10	10.2	8	8.16

Table 1. Postintegrative transgene expression in HeLa cells

transgenes, as many as 26% of the cell clones was affected by partial or complete loss of transgene expression in the control group (Table 1). Taken together, all three transposon vector systems generated low frequencies of silenced transgene integrations, underscoring the value of transgene delivery with transposon systems versus random plasmid integration.

The mutagenic potential of any given transposon system is one of the most important basic considerations for most applications ranging from forward genetic screens (where mutagenicity is desired) to gene therapy (where mutagenicity is undesired). Integration patterns of most transposable elements have been shown to be nonrandom. In previous studies, insertional patterns and preferences of *SB* and *PB* systems were quite extensively analyzed [367, 376, 443, 444, 454], in contrast to *Tol2*, whose insertional characteristics in mammalian genomes are poorly documented.

In order to gain insight into the target site selection properties of the *Tol2* system in human cells, we mapped 113 *Tol2* insertions onto human chromosomes. All of the HeLa cell chromosomes were targeted in this dataset, suggesting an overall random chromosomal distribution. With respect to transposition into genes, 48% of the insertions mapped to transcription units. This frequency of genic insertion is similar to that of the *PB* system (49-52%, depending on cell type), and considerably higher than that of *SB* (31-39%, depending on cell type) in mouse and human cells. Most of the intragenic insertions mapped to introns, which could be explained by the larger overall size of introns as compared to exon sequences. Interestingly, similar to *PB* and unlike *SB*, *Tol2* revealed a significant bias for inserting close to transcription start sites, suggesting that an open chromatin state around transcriptionally active chromosomal regions favors *Tol2* integration. Finally, *Tol2* insertion sites revealed significant underrepresentation within chromosomal regions with H3K27me3 histone marks (not shown) [455], typically associated with transcriptionally repressed heterochromatin. At the primary DNA sequence level, sequence logo analysis of the 113

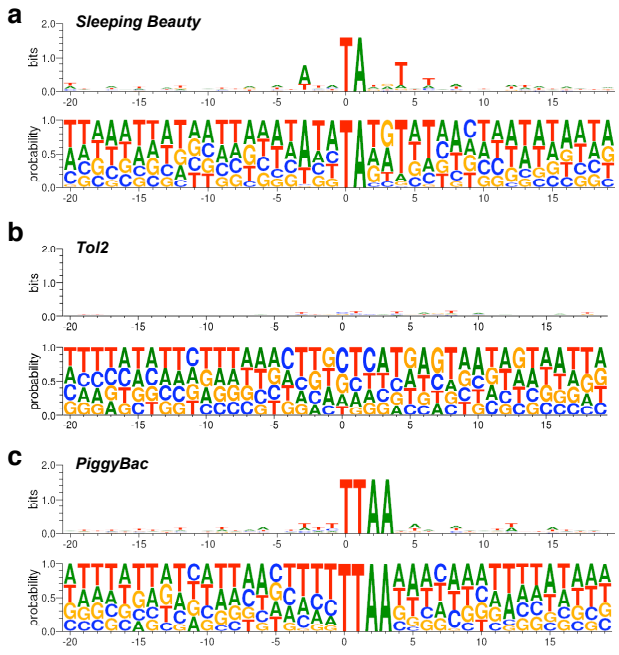


Figure 68. Sequence logo analysis of *SB100X*, *Tol2* and *PB* integration sites. Web logo analyses and nucleotide probability plots of 46 *SB100X* **(a)**, 113 *Tol2* **(b)** and 46 *PB* **(c)** integration sites in HeLa cells.

Tol2 insertions revealed no obvious consensus sequence and a lack of preference for a particular base composition of target DNA (Fig. 68), indicating that the *Tol2* element is promiscuous in its target site selection properties, at least on the level of primary DNA sequence. This is in sharp contrast to the *SB* and *PB* elements that use obligate TA and TTAA target sequences, respectively, in a generally AT-rich DNA context (Fig. 68) [376, 444]. In sum, out of the three transposon systems, the *Tol2* transposon displays the most random target site selection properties on the primary DNA sequence level, but a preference for integration close to the transcriptional regulatory regions of genes on the genomic level.

In summary, *SB100X* was found to be the most efficient system in terms of stable gene transfer under conditions where the availability of the transposon DNA is limiting the transposition reaction. This could make *SB100X* a powerful reagent for genomic mobilization of chromosomally resident transposons in insertional mutagenesis screens. Another particular application in which *SB* could be preferred to the other systems is gene transfer in hard-to-transfect cell types including stem cells. Indeed, the *SB* system was found superior to *PB* in its ability to give rise to robust and stable gene transfer and transgene expression in human $CD34^+$ cells, suggesting a potential advantage of the *SB* system under non-selective conditions in HSC-based gene therapy [427, 452]. Importantly, in a therapeutical setup (e.g., β-thalassemia, hemophilia), under non-selective conditions in HSCs, the measured difference in transpositional efficiencies is expected to be decisive in favor of *SB100X*.

All of the three transposon systems tested were found to be sensitive to OPI. OPI describes a phenomenon of decreasing transposition above a certain level of cellular transposase concentration [456]. The molecular basis of OPI is not known, but it is thought to

occur when transposase is present in excess concentration driving the formation of transposition-deficient transposase oligomers [322]. OPI has been described for a wide variety of transposons, including the *Tc1*/*mariner* ([322] and references therein) and *hAT* [451] superfamilies. Negative regulation of the *SB* system by OPI is well known in the literature (reviewed in [322]), available data for *PB* is contradictory [443, 444, 448, 457], whereas *Tol2* has been claimed to be immune to OPI [257, 258]. We found clear indications for OPI for both the *mPB* and *Tol2* transposases, but it appears that these two transposons are less affected by this mechanism than *SB*. Both the *mPB* and *Tol2* systems were found to be active in a wide range of transposase expression in our experiments. In contrast, it is advisable to titrate the *SB* transposase either by dosing expression constructs or by the choice of promoter driving transposase expression [458].

Chromosomal target site selection as well as transgene copy number are key issues to be considered for genetic applications of any transposon system. For example, single copy insertions away from endogenous genes is a clear advantage for safe gene transfer in human gene therapeutic applications. The *SB* system appears to satisfy these needs the best: the insertion pattern generated by *SB100X* is close to random at the genome level with no apparent bias for insertion near transcriptional regulatory regions of genes [427], and it is possible to generate predominantly single-copy insertions by carefully dosing the transposon components. The *PB* and *Tol2* systems appear to be less favorable for potential therapeutic applications because, even though they tend to generate low copy number insertions, they appear to prefer genes [443, 444, 447, 454] and their upstream regulatory regions for insertion, respectively. Unlike in therapeutic applications, hitting genes by insertional elements is the goal with forward mutagenesis screens, and all three transposons studied here are excellent tools for such purposes. For example, the propensity of *Tol2* to insert close to transcriptional start sites of genes is particularly advantageous for enhancer trapping [459, 460]. The propensity of *PB* to insert into transcription units [443, 444, 447, 454] supports genome-wide mutagenesis with gene trap cassettes [461]. In addition to these fundamental differences in global, genome-wide preferences for transposon insertion, significant biases in integration site selection at the primary DNA sequence as well as local

DNA structure levels are also evident. For example, protein-induced deformability characterized by an alternating *V*step pattern was shown to be associated with preferred *SB* insertion sites, whereas *piggyBac* and *Tol2* integration sites lack such consistent, clear-cut structural patterns [462, 463]. The insertional biases associated with vector systems represent the main limitation to full genome coverage with individual transposon-based vectors. Thus, in this respect, the utility of transposons for mutagenesis is greatly enhanced by the availability of multiple, alternative vector systems with distinct preferences for insertion, such as *Sleeping Beauty*, *Tol2* and *piggyBac*.

Any transgene vector system should provide long-term expression of transgenes. Transgenes delivered by non-viral approaches often form long, repeated arrays (concatemers) that are targets for transcriptional silencing by heterochromatin formation. In addition, long-term expression of transgenes delivered by retroviruses has been shown to be compromised by transcriptional silencing [464]. It was recently shown that the zinc finger protein ZFP809 bridges the integrated proviral DNA of the murine leukaemia virus and the TRIM28 transcriptional co-repressor in embryonic stem cells [465]. Thus, sequence elements in the vector itself can predispose the cargo for silencing. Here we carried out an analysis to address if similar mechanisms act to silence transgene insertions delivered by transposon vectors. The cut-and-paste mechanism of DNA transposition results in a single copy of the transgene per insertion locus; thus, concatemer-induced gene silencing is unlikely to be an issue with transposition-mediated gene transfer. Indeed, we found that transposon insertions delivered by the *Sleeping Beauty*, *Tol2* and *piggyBac* systems only rarely (<4% of all insertions) undergo silencing in HeLa cells. In contrast, we observed a high frequency (26%) of silencing in the control group transfected with constructs encoding an inactive transposase along with the transposon. This silencing was probably caused by high copy number and the likely formation of tandem array patterns of transgenes as a result of random plasmid integrations. However, putative silencing of transposon-carried genes may depend on the site of insertion (position effects). Our data suggest that the three transposon systems examined here rarely target heterochromatic chromosomal regions for insertion, consistent with stable transgene expression observed in hundreds of independent insertions. An

additional factor that may provoke transgene silencing is the cargo DNA, in particular the type of promoter used to drive expression of the gene of interest. Indeed, it was previously shown that transgene constructs delivered into mouse cells using *SB* transposition can be subject to epigenetic regulation by CpG methylation, and that a determinant of epigenetic modifications of the integrating transposon vector is the cargo transgene construct, with the promoter playing a major role [453]. Our data suggest that the CAGGS enhancer/promoter element is unlikely to trigger a silencing cascade, consistent with ubiquitous expression of this promoter in several tissues and cell types. Finally, could it be that, similar to retroviruses, certain sequence motifs in the transposon vectors are recognized by mediators of silencing in the cell? Our data suggest that this is unlikely for all three transposons investigated here, at least in HeLa cells. Furthermore, several studies have established that *SB*-mediated transposition provides long-term expression *in vivo*. For example, stable transgene expression from *SB* vectors was seen in mice after gene delivery in the liver [266, 313, 466], lung [314, 320], the brain [317] and in the blood following hamatopoietic reconstitution *in vivo* [427, 452]. Thus, although our understanding of all the factors that will ultimately determine the expressional fate of an integrated transposon is still rudimentary, it appears that transposon vectors have the capacity to provide long-term expression of transgenes both *in vitro* and *in vivo*. Our results suggest that *Sleeping Beauty*, *Tol2* and *piggyBac* are attractive complementary research tools for gene transfer in mammalian systems.

3.2.4 Towards safer vectors for gene therapy I: Transcriptional shielding of *Sleeping Beauty*'s genetic cargo with insulators (Paper III)

Transposons are emerging alternatives to retroviral vectors for use in gene therapy applications (see section 1.6.3). Our results on the transcriptional activities associated with *SB* vectors (section 3.1.1.1.1) suggest that *SB*-based vectors may have a safety advantage as compared to retrovirus-based vectors due to the lack of strong element-intrinsic promoter activities. However, trans-activation of host gene expression upon vector integration may

eventually arise from strong promoter/enhancer elements that are not components of the transposon vector itself, but instead are components of the cargo transgene cassette.

In order to simulate transposon insertion upstream of a gene, an *SB* transposon carrying an SV40-neo cassette (representing a model therapeutic gene) was cloned immediately upstream of the TATA-box minimal promoter-driven luciferase gene (representing a model host gene). This arrangement mimics the integration of an *SB* transposon carrying a therapeutic gene in close proximity to a host gene driven by its own weak promoter. As shown in Fig. 69, insertion of the SV40-neo cassette carried by the *SB* transposon leads to a 40- to 100-fold activation (depending on the orientation of insertion) of luciferase gene expression, consistent with transactivation of the TATA-box promoter by the SV40 enhancer. Indeed, the transactivating ability of an “empty” transposon is reduced as compared to the cargo-containing transposon.

We next tested whether the safety of the model transposon vector can be further improved by flanking the SV40 cassette by chicken beta-globin insulator (HS4) sequences. The HS4 sequence at the 5'-end of the chicken beta-globin locus has the two defining properties of an insulator: it prevents an enhancer from acting on a promoter when placed between them ("enhancer blocking"), and acts as a barrier to chromosomal position effect when it surrounds a stably integrated reporter. Further dissection of the core revealed that HS4 is a compound element in which the enhancer blocking and barrier activities can be separated [467]. A CTCF binding site in a 250-bp core element of HS4 is necessary and sufficient for enhancer blocking activity [467,

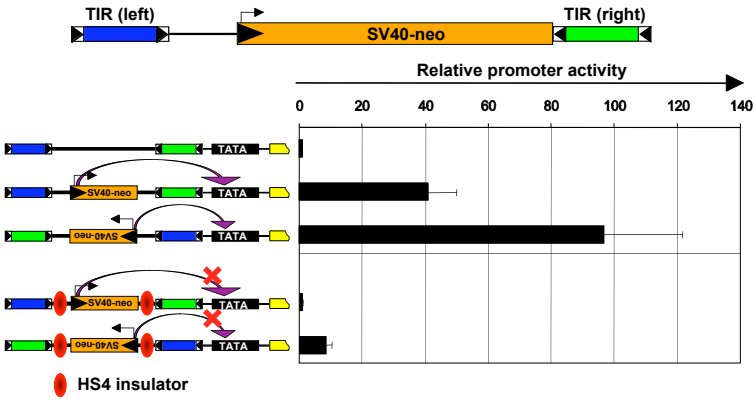


Figure 69. Upregulation of test promoters by vector-borne expression units, and its shielding with insulators. An *SB* transposon carrying a SV40-neo transgene cassette can activate transcription of a nearby minimal promoter (TATA-box). The capacity of the cargo gene to activate adjacent promoter elements can be efficiently reduced by flanking the transgene with HS4 insulator core elements. Transcriptional activity was determined by transient luciferase assays in HeLa cells, and activity of the TATA-box was arbitrarily set to value 1. Blue box: left IR/DR of *SB*; green box: right IR/DR of *SB*; small triangles in the boxes: transposase binding sites; arrowhead: SV40 enhancer/promoter element; orange box: neo marker; red ovals: HS4 insulator elements; arrows indicate the direction of transcription initiated at the SV40 promoter.

468]. The SV40-neo cassette was flanked by the core HS4 elements in the SB vector in two possible orientations with regard to the transgene cassette. The insulated transposons displayed a 7- to 51-fold reduction in luciferase transcativation as compared to uninsulated vectors (Fig. 69). Enhancer blocking was efficient in both orientations of transposon integration with regard to the luciferase gene, and in both orientations of the insulators within the transposon (Fig. 69).

Activities of promoter/enhancer elements may be subject of tissue/cell type-specific regulation. In order to substantiate our observations on the transcriptional activities of *SB* transposon-derived gene vectors, some of the luciferase reporter constructs described above were transfected into primary human T cells. The results largely confirmed the data obtained in HeLa cells in that an SV40-neo cassette-containing *SB* transposon can significantly upregulate the TATA-box promoter in T cells, which can efficiently be blocked by flanking the transgene cassette with HS4 insulators (not shown). Taken together, incorporation of HS4 insulator sequences in *SB*-based vectors reduces transactivation of promoters by transposon-borne enhancers, and thus may significantly increase the safety of these vectors due to a reduced risk of transcriptional activation of host genes situated close to a transposon insertion site.

In summary, *SB*-based technologies for nonviral gene transfer gained significant ground over the past couple of years [310]. Thus, the results presented in this study bear practical relevance to the use and safety of *SB* transposon vectors in clinically relevant applications. First and foremost, the 65-bp region within the 5'-UTR of the *SB* transposon that mediates HMG2L1's activity on transcriptional regulation is not included in *SB* vectors, because the ~160-bp DNA situated between the left IR and the transposase coding region is not required for transposition [337]. We further showed that transcription from the UTRs towards the outside of the *SB* transposon is negligible, and occurs at rates comparable to that by the eukaryotic core promoter TATA-box in both HeLa and primary human T cells (as discussed in detail in section 3.1.1.1.1).

We have also shown that the cargo transgene sequence carried by *SB* vectors can exert a profound effect on the activity of a transcription unit linked *in cis* to the transposon vector, due to the transcriptional enhancer element of the transgene cassette. This presents a safety issue, because therapeutic expression cassettes may inadvertently upregulate a proto-oncogene or other signalling factor that happens to be close to the transposon insertion site (as discussed in section 1.6.3.1). One strategy that is currently considered to lower the risk of insertional mutagenesis by integrating vector systems is transcriptional confinement of an expression unit upon genomic integration, which serves two purposes: allow position-independent expression of the transgene (for efficiency), and prevent trans-activation of a cellular gene (for safety). The HS4 chromatin insulator of the chicken beta-globin locus has both of these activities [467], and was shown to improve the expression performance of murine retrovirus [469, 470], lentivirus [471, 472] as well as AAV [473] vectors by protecting them from chromosomal position effects. In addition, a suppression of clonal dominance was found with HS4-insulated lentiviral vectors [474], suggesting reduced upregulation of host genes upon vector integration due to enhancer blocking by the insulator. In sum, *SB* transposon-based vectors have a favorable safety profile, because they are fairly inert in their transcriptional activities, and because insulator elements can successfully be incorporated in the next generation of transposon vectors. Thus, *SB* vectors are expected to have only a limited ability to upregulate a cellular gene located in the vicinity of a transposon insertion site.

3.2.5 Towards safer vectors for gene therapy II: Targeted *Sleeping Beauty* transposition in human cells (Paper XIV)

None of the vector systems currently used either in preclinical experiments or in clinical trials described above displays DNA sequence preferences specific enough for targeted insertion into a defined location in the human genome. However, with any vector that integrates into chromosomes in a nearly random manner (the *SB* transposon could theoretically insert into

any of the ~10^8 TA sites in the human genome) comes the potential risk of insertional activation or inactivation of cellular genes (as discussed in section 1.6.3.1) [325].

Integration into selected sites in the genome would simultaneously ensure appropriate expression of the transgene (lack of position effects), and prevent hazardous effects to the organism due to insertional mutagenesis of cellular genes (lack of genotoxicity). Targeted gene delivery can rely on distinct mechanisms, and we envision three distinct molecular strategies for targeted *SB* transposition, making use of heterologous DNA-binding domains (DBDs) that are either fused to: 1) the *SB* transposase itself, 2) a protein domain that binds the transposon DNA, or 3) a protein domain that interacts with the *SB* transposase (Fig. 70).

The premise of the first approach is that upon binding of the engineered transposase to a specific target site specified by the heterologous DBD, transposon insertion may occur in adjacent regions (Fig. 70A). However, altering sequence-specificity of most recombinases may prove difficult, since they do not have spatially separable catalytic and target DBDs that could be modularly replaced irrespectively of each other. In addition, some proteins display sensitivity to fusions with foreign peptides, domains or proteins, possibly due to altered folding of the resulting chimeric protein. Thus, fusions may result in abolished or limited enzymatic activity. Nevertheless, there is some evidence for the feasibility of using transposase fusions to target insertions to a certain extent to specific sites. Namely, fusions of the bacterial IS30 transposase with the λ repressor and with the DBD of the transcription factor Gli1 showed

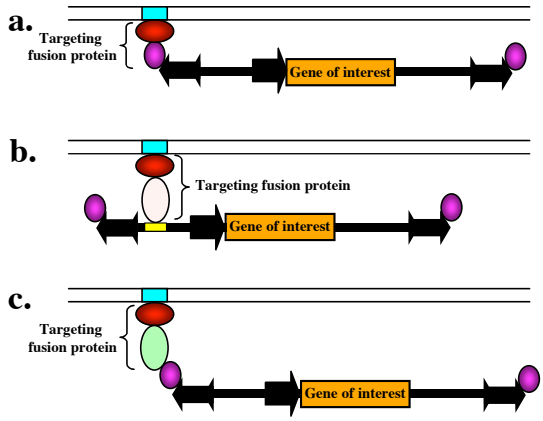


Figure 70. Experimental strategies for targeting *Sleeping Beauty* transposition. The common components of the targeting systems include a transposable element that contains the IRs (arrowheads) and a gene of interest equipped with a suitable promoter. The transposase (purple circle) binds to the IRs and catalyzes transposition. A DNA-binding protein domain (red oval) recognizes a specific sequence (turquoise box) in the target DNA (parallel lines). **(a)** Targeting with transposase fusion proteins. Targeting is achieved by fusing a specific DNA-binding protein domain to the transposase. **(b)** Targeting with fusion proteins that bind the transposon DNA. Targeting is achieved by fusing a specific DNA-binding protein domain to another protein (white oval) that binds to a specific DNA sequence within the transposable element (yellow box). In this strategy, the transposase is not modified. **(c)** Targeting with fusion proteins that interact with the transposase. Targeting is achieved by fusing a specific DNA-binding protein domain to another protein (light green oval) that interacts with the transposase. In this strategy, neither the transposase nor the transposon is modified.

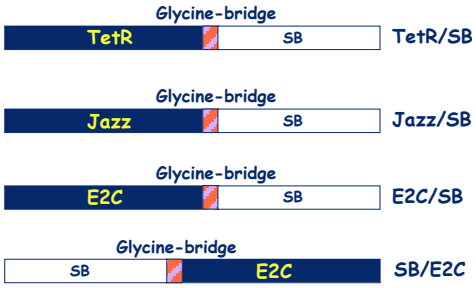


Figure 71. Design and transpositional activities of transposase fusions. Schematic representation of the fusion proteins that consist of the *SB* transposase fused to the tetracycline repressor (TetR), the Jazz or the E2C ZF proteins. All fusions contain a glycine-bridge consisting of ten consecutive glycine residues to provide a flexible linker between the fusion partners.

altered insertions profiles in *E. coli* and in zebrafish embryos, respectively, using plasmid targets [475]. Furthermore, direct fusions of the *Mos1* and *piggyBac* eukaryotic transposases with the GAL4 DBD were shown to retain transpositional activity, and to result in site-selective transposon insertion in a plasmid-to-plasmid experimental setup in mosquito embryos [476]. Transposition mediated by the chimeric *Mos1* transposase into the UAS-containing target plasmid occurred at a 96 % frequency at the same TA located 954 bp away from the targeted UAS sequence. Transposition by the GAL4-*PB* fusion protein into a plasmid containing the UAS target sequence occurred at a 67% frequency into a TTAA site located 1103 bp upstream of the UAS. Another group reported that only N-terminal fusions to the *SB* transposase retained transpositional activity, and that fusions of HSB5 (a third-generation improved *SB* transposase) with the GAL4 and E2C (a synthetic, zinc finger protein recognizing an 18-bp target site in the 5'-untranslated region of the human *erbB-2* gene) DBDs resulted in a drop in transposition efficiency to ~20-26% of unfused HSB5 [477]. Nevertheless, these fusion transposases showed targeted transposon integration in plasmid-based assays in cultured human cells. Targeted transposition events were enriched about 11-fold in a 443-bp window around a 5-mer UAS site and about 8-fold in a 443-bp window around a 5-mer repeat of the E2C binding site in the target plasmids, respectively, as compared with integration patterns mediated by unfused HSB5. However, cell-based assays failed to detect targeting of the E2C binding site in the genomic context.

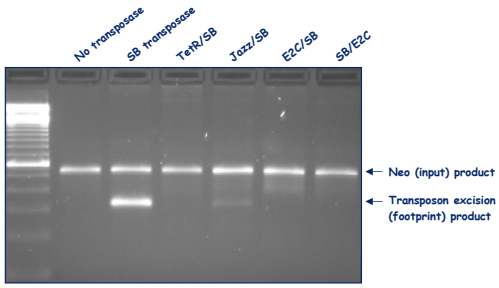


Figure 72. A fusion protein consisting of the *SB* transposase and the Jazz zinc finger protein retains transposon excision activity. HeLa cells were cotransfected with a *neo*-marked transposon plasmid and vectors expressing the proteins indicated. Transposon excision is assayed with PCR that amplifies a footprint product. PCR-amplification of the *neo* marker inside the transfected transposon donor serves as a loading control.

Fusion proteins containing the *SB* transposase and either the bacterial tetracycline

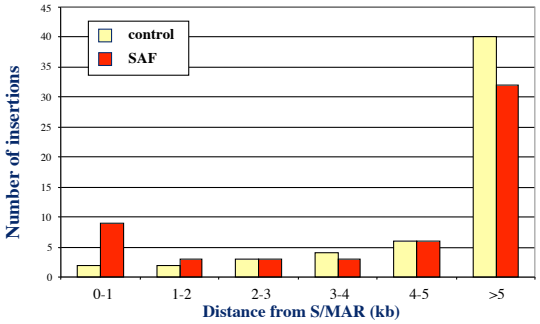


Figure 73. Preferential insertion into S/MARs by transposon targeting. Transposition events were recovered from transformed cells, and human chromosomal DNA flanking the insertion sites was analyzed with respect to proximity to chromosomal S/MARs. The MAR-Wiz program was used to predict the presence of an S/MAR in the vicinity of a transposon insertion. Distances were categorized, and the numbers of insertions obtained in the presence and in the absence of the targeting fusion protein in each category are shown.

repressor (TetR) that specifically binds the tetracycline operator sequence, or the Jazz and E2C ZF peptides were engineered (Fig. 71). Transposon excision activity of the fusion proteins was tested using a PCR-based excision assay [348]. Out of four constructs tested, only the Jazz/SB fusion showed detectable activity in human HeLa cells, although at a clearly reduced efficiency compared to unfused transposase (Fig. 72). In line with the excision data, the Jazz/SB fusion was found to retain transpositional activity at about 10-15% of the wild-type level. However, a PCR survey on genomic DNA isolated from transformant cells generated using Jazz/SB as transposase source revealed no indication of targeted transposition into the utrophin locus, and no occurrence of the 9-bp binding site of Jazz within a 1-kb window around the transposon insertion sites (data not shown). Taken together, the results establish that most direct fusions to the *SB* transposase have negative effects on transpositional activity, and suggest that ZFs with higher specificity in terms of DNA binding will be required for targeted transposition.

The second strategy is based on a fusion protein with dual DNA-binding activity that has the capacity to bind to two DNA molecules that contain binding sites of the respective fusion partners, thereby bringing them into close proximity (Fig. 70B). A similar mechanism of bridging of DNA molecules by proteins might act in targeting some P element transposon vectors in *Drosophila*. *P* element insertion is essentially random at the genome scale. However, *P* elements containing regulatory sequences from the *engrailed* gene show some insertional specificity by frequently inserting near the endogenous, parental gene [478, 479]. This

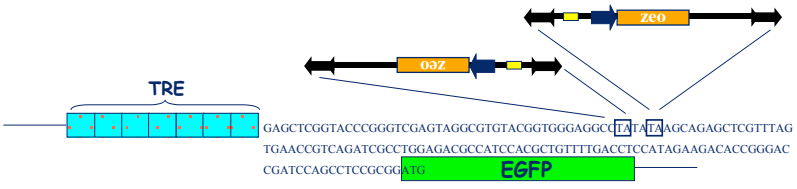


Figure 74. Targeted transposition close to the TRE. Two transposon insertions in close proximity of the targeted TRE region, in the two possible orientations, within two TA dinucleotides of the CMV promoter TATA-box are shown.

phenomenon, called transposon "homing", tends to be region-specific [479] with transposon integrations distributed over several kilobase pair regions near the targeted loci. Potential *SB* targeting by such mechanism was assessed by engineering a LexA operator into a benign site within an *SB* transposon vector. Targeted transposition events into endogenous chromosomal MAR sequences (Fig. 73) as well as a chromosomally integrated tetracycline response element (Fig. 74) were recovered by employing targeting fusion proteins containing LexA and either the SAF-box, a protein domain that binds to chromosomal MARs, or the tetracycline repressor (TetR). The targeted transposition events identified in these experiments were likely mediated by simultaneous binding of the targeting fusion protein to both transposon and target DNA. This strategy shows promise, because it does not measurably interfere with the transposition process.

The third strategy for targeted *SB* transposition is based on protein-protein interactions between a targeting protein and the SB transposase (Fig. 70C). Either naturally occurring or engineered transposase interactors may tether the transpositional machinery to specific DNA sites, potentially leading to integration into nearby regions. Importantly, several, naturally occurring transposable elements evolved strategies for targeted insertion into defined chromosomal sites or regions, and the mechanisms of targeted insertions often rely on protein-protein interactions between a transposon-encoded factor and a cellular, DNA-binding host factor. For example, based upon observations for a role of LEDGF/p75 in directing HIV integration into expressed transcription units, *in vitro* studies have shown increased integration near λ repressor binding sites by fusing either the full-length LEDGF/p75 or the LEDGF/p75 IN-binding domain to the DBD of phage λ repressor protein [480]. In an analogous fashion, Sir4p (which mediates targeted insertion of the yeast *Ty5* retrotransposon into heterochromatin) fused to the *E. coli* LexA DBD was shown to result in integration hot spots for *Ty5* near LexA operators [481]. Altogether, these observations suggest a general model wherein interactions between transposase/IN and DNA-bound proteins mediate insertional target choice.

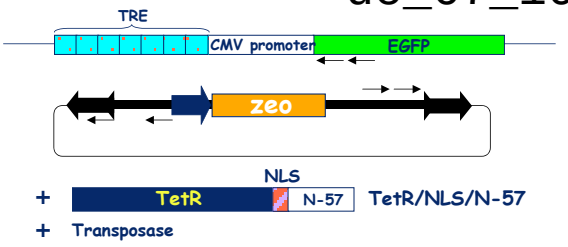


Figure 75. Transposon targeting using a strategy based on protein-protein interactions between a targeting fusion protein and the SB transposase. The targeting fusion protein consists of the tetracycline repressor (TetR) that binds to the TRE, a nuclear localization signal (NLS), a glycine-bridge and the N-terminal protein interaction domain of the *SB* transposase (N-57). The targeted chromosomal locus as described is a tetracyclin resoponse element (TRE)-driven EGFP gene stably integrated into HeLa cell chromosomes, the tranposon is an unmodified, antibiotic-marked *SB* element.

Such a strategy was successfully adapted for targeted *SB* transposition by coexpressing the SB transposase with a targeting fusion protein consisting of a specific DBD and a subdomain of the *SB* transposase that mediates protein-protein interactions between transposase subunits (Fig. 75) [45]. This domain spans the N-terminal helix-turn-helix domain (termed N57 for containing 57 amino acids) of the *SB* transposase (Fig. 75) [45]. Targeted transposition into a chromosomally integrated tetracycline response element using a TetR-N57 fusion was monitored in human cells [411]. By using this strategy, >10% of cells receiving transposon insertions contained at least one transposition event within the targeted chromosomal region. Insertions obtained by this strategy occurred at multiple sites within a 2.5-kb window, and featured some insertion hot spots (Fig. 76). A significant advantage of this technology as compared to direct transposase fusions is that the transposase polypeptide does not have to be modified; thus, potential negative effects on transposase activity are eliminated.

Technologies for site-directed transgene integration into safe regions in the human genome would reduce the potential genotoxic effects of transposon insertion, thereby contributing to an overall improvement of the safety profile of transposon-based gene vectors for human applications. There are several factors affecting site-selectivity of integrating vector systems. These include primary sequence and physical structure of the DNA at the targeted region, accessibility of specific chromosomal sites determined by chromatin components, expression of endogenous proteins that may compete for binding, and the specificity as well as capacity of chimeric proteins in DNA-

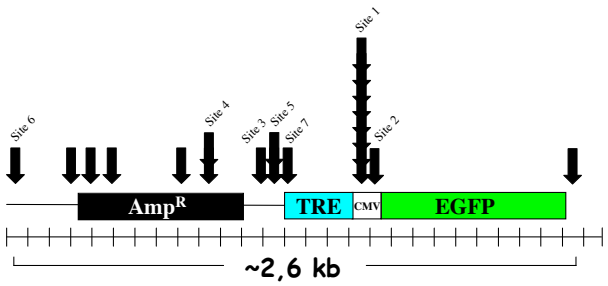


Figure 76. Mapping of targeted SB insertions. Mapping with respect to the TRE-EGFP target isolated from six independent experiments is shown. Multiple arrows represent independent insertions into the same site.

binding as well as in catalytic functions. The major challenges facing the development of this technology are around three major issues. 1) Future work will have to focus on the identification of applicable, endogenous chromosomal target sites that fulfill the criteria for a genomic “safe harbor”, and on the selection of DNA-binding proteins that can be exploited for efficiently targeting transposition into those sites *in vivo*. In this respect, naturally occurring DBDs have some limitations for use as gene targeting agents. Namely, those DBDs that have physiological binding sites in the human genome recognize short DNA sequences present in multiple copies throughout the human genome, making targeted insertion with these DBDs impractical. Recognition sites of 18 bp would be expected to be unique in the human genome. Artificial ZFs offer a potential solution. Their modular character in structure and function is the key advantage in engeneering of proteins that are able to recognize theoretically any sequence in the human genome [482]. Each individual zinc finger binds 3-4 bp DNA, thus a set of 64 domains would cover recognition of any desired DNA sequence. 2) Direct fusions of DBDs to transposase proteins appear to interfere with the biochemical activities of the transposase; thus, a systematic evaluation of protein spacer sequences linking the two fusion partners will be required in order to allow rational design of direct transposase fusions. 3) A major hurdle for targeting systems engineered from promiscuously integrating vectors (such as *Sleeping Beauty*) are the considerable off-target insertions in the context of the human genome. Despite the fact that targeted integrations can be generated, non-targeted insertions can still occur at high frequencies, because the natural DNA-binding capacities of the transposase competes with that of the foreign DBD used for targeting. Keeping such off-target insertions at a minimum remains a major challenge that could be at least in part be addressed by engineering of DNA-binding domains with high specificity as well as affinity towards targeted sites. Although these hurdles are yet to be overcome before technologies of targeted gene insertion can be considered for applications, recent evidence suggests that target-selected transgene insertion into desired regions in the human genome is a realistic goal.

4 SUMMARY: Major discoveries and conclusions

Transposons are discrete segments of DNA that have the distinctive ability to move and replicate within genomes. Transposons were discovered in the 1940's by Barbara McClintock (who later was awarded with the Nobel Prize for this discovery) in the maize genome, and have since been found ubiquitous in essentially all living organisms. The process of element movement is generally called transposition, and can contribute to insertional mutagenesis, altered gene expression and recombination. Transposons make up significant fractions of genomes; for example, about 45% of the human genome is composed of sequences of a variety of different elements. Transposons are best viewed as molecular parasites that propagate themselves using resources of the host cell. The transposition process is under regulation by both element- and host cell-encoded mechanisms and factors; these include transcriptional regulation of transposase expression, regulation of synaptic complex assemby, modulation of both transposon excision as well as integration by components of the chromatin and various factors that contribute to the target site selection properties of a given element. Transposition also has profound effects on the cell's life by modulating cellular pathways and gene functions by interacting with host proteins and by insertional mutagenesis. Despite their parasitic nature, there is increasing evidence that transposable elements are a powerful force in gene evolution. Indeed, about 50 human genes are derived from transposable elements, among them genes that are responsible for immunoglobulin gene recombination in all vertebrates.

Transposons are natural gene delivery vehicles, and have been revolutionizing genomic manipulations in diverse model systems. Transposons have been developed as useful tools for genomic manipulations, including transgenesis and insertional mutagenesis, in invertebrate animal systems as well as in plants, but similar technologies have been impossible in vertebrate systems for a long time, for the simple reason that the vast majority of DNA transposons are extinct in vertebrate genomes.

We made the following major discoveries and conclusions:

- Molecular phylogenetic data were used to construct a synthetic transposon, *Sleeping Beauty (SB)*, which could be identical or equivalent to an ancient element that dispersed in fish genomes in part by horizontal transmission between species. A consensus sequence of a transposase gene of the salmonid subfamily of elements was engineered by eliminating the inactivating mutations. *SB* transposase binds to two sites (the DRs) within the inverted repeats (IRs) of salmonid transposons in a substrate-specific manner, and mediates precise, cut-and-paste transposition in fish as well as in mouse and human cells.

- We investigated transcriptional activities of *SB* in order to assess its potential to alter host gene expression upon integration. The untranslated regions (UTRs) of the transposon direct convergent, inwards-directed transcription. Transcription from the 5'-UTR of *SB* is upregulated by the host-encoded factor HMG2L1, and requires a 65-bp region not present in commonly used *SB* vectors. The *SB* transposase antagonizes the effect of HMG2L1, suggesting that natural transposase expression is under a negative feedback regulation. *SB* transposon vectors lacking the 65-bp region associated with HMG2L1-dependent upregulation exhibit benign transcriptional activities, at a level up to 100-times lower than that of the MLV retrovirus long terminal repeat.

- We established that the DNA-bending, high mobility group protein, HMGB1 is a host-encoded cofactor of *SB* transposition. Transposition was severely reduced in mouse cells deficient in HMGB1. This effect was rescued by transient overexpression of HMGB1, and was partially complemented by HMGB2, but not with the HMGA1 protein. Overexpression of HMGB1 in wild-type mouse cells enhanced transposition, indicating that HMGB1 can be a limiting factor of transposition. *SB* transposase was found to interact with HMGB1 *in vivo*, suggesting that the transposase may recruit HMGB1 to transposon DNA. HMGB1 stimulated preferential binding of the transposase to the DR more distant from the

cleavage site, and promoted bending of DNA fragments containing the transposon IR. We propose that the role of HMGB1 is to ensure that transposase-transposon complexes are first formed at the internal DRs, and to subsequently promote juxtaposition of functional sites in transposon DNA, thereby assisting the formation of synaptic complexes.

- We used the *SB* element as a tool to probe transposon-host cell interactions in vertebrates. The Miz-1 transcription factor was identified as an interactor of the *SB* transposase in a yeast two-hybrid screen. Through its association with Miz-1, the *SB* transposase downregulates cyclin D1 expression in human cells, as evidenced by differential gene expression analysis using microarray hybridization. Downregulation of cyclin D1 results in a prolonged G1 phase of the cell-cycle and retarded growth of transposase-expressing cells. G1 slowdown is associated with a decrease of cyclin D1/cdk4-specific phosphorylation of the retinoblastoma protein. Both cyclin D1 downregulation and the G1 slowdown induced by the transposase require Miz-1. A temporary G1 arrest enhances transposition, suggesting that *SB* transposition is favored in the G1 phase of the cell-cycle, where the nonhomologous end joining (NHEJ) pathway of DNA repair is preferentially active. Because NHEJ is required for efficient *SB* transposition, the transposase-induced G1 slowdown is probably a selfish act on the transposon's part to maximize the chance for a successful transposition event.

- We have shown that DNA CpG methylation upregulates transposition of IR/DR elements in the *Tc1/mariner* superfamily. CpG methylation provokes the formation of a tight chromatin structure at the transposon DNA, likely aiding the formation of a catalytically active complex by facilitating synapsis of sites bound by the transposase.

- We established that the distribution of experimentally induced *SB* insertions in the human genome can be considered fairly random, because most chromosomes can serve as a target; no obvious hotspots with multiple insertions were found, and no preference for coding versus non-coding DNA was observed. We further showed that the *SB* element

displays a certain degree of specificity in target site utilization at the DNA sequence and structural level. A palindromic AT-repeat consensus sequence with bendability and a symmetrical pattern of hydrogen bonding sites in the major groove of the target DNA define preferred sites for integration.

- A novel open reading frame-trapping method was used to isolate uninterrupted transposase coding regions from the genome of the frog species *Rana pipiens*. The isolated clones were about 90% identical to a predicted transposase gene sequence from *Xenopus laevis*. None of these native genes was found to be active. Therefore, a consensus sequence of the transposase gene was derived. This engineered transposase and the transposon inverted repeats together constitute the components of a novel transposon system that we named *Frog Prince* (*FP*). *FP* has only about fifty percent sequence similarity to *SB*, and catalyzes efficient cut-and-paste transposition in fish, amphibian and mammalian cell lines. We demonstrate high-efficiency gene trapping in human cells using *FP* transposition. *FP* is the most efficient DNA-based transposon from vertebrates described to date, and shows about 70% higher activity in zebrafish cells than *SB*. *Frog Prince* can greatly extend our possibilities for genetic analyses in vertebrates.

- *Hsmar1*, one of the two subfamilies of *mariner* transposons in humans, is an ancient element that entered the primate genome lineage ~50 million years ago. Although *Hsmar1* elements are inactive due to mutational damage, one particular copy of the transposase gene has apparently been under selection. This transposase coding region is part of the SETMAR gene, in which a histone methylatransferase SET domain is fused to an *Hsmar1* transposase domain. A phylogenetic approach was taken to reconstruct the ancestral *Hsmar1* transposase gene that we named *Hsmar1-Ra*. The *Hsmar1*-Ra transposase efficiently mobilizes *Hsmar1* transposons by a cut-and-paste mechanism in human cells and zebrafish embryos. *Hsmar1*-Ra can also mobilize short inverted-repeat transposable elements (MITEs) related to *Hsmar1* (*MiHsmar1*), thereby establishing a functional relationship between an *Hsmar1* transposase source and these MITEs.

MiHsmar1 excision is two orders-of-magnitude more efficient than that of long elements, thus providing an explanation for their high copy number. We show that the SETMAR protein binds, and introduces single-strand nicks into *Hsmar1* inverted repeat sequences *in vitro*. Pathway choice for DNA break repair was found to be characteristically different in response to transposon cleavage mediated by *Hsmar1*-Ra and SETMAR *in vivo*. Whereas nonhomologous end-joining plays a dominant role in repairing excision sites generated by the *Hsmar1*-Ra transposase, DNA repair following cleavage by SETMAR predominantly follows a homology-dependent pathway. The novel transposon system can be a useful tool for genome manipulations in vertebrates, and for investigations into the transpositional dynamics and contribution of these elements to primate genome evolution.

- Ancient, inactive copies of transposable elements of the *PIF/Harbinger* superfamily have been described in vertebrates. Based on a predicted consensus sequence, we reconstructed the functional components of the *Harbinger3_DR* transposon in zebrafish, including a transposase and a second, transposon-encoded protein of unknown function that has a Myb-like trihelix domain. The reconstructed *Harbinger* transposon shows efficient cut-and-paste transposition in human cells, and preferentially inserts into a 15-bp consensus target sequence. The Myb-like protein is required for transposition, and physically interacts with the transposase. The Myb-like protein enables transposition in part by promoting nuclear import of the transposase, and by binding to the transposon ends. We investigated the functions of two, transposon-derived human proteins: HARBI1, a domesticated transposase-derived protein and NAIF1 that contains a trihelix motif similar to that described in the Myb-like protein. Physical interaction, subcellular localization and DNA-binding activities of HARBI1 and NAIF1 suggest strong functional homologies between the *Harbinger3_DR* system and their related, host-encoded counterparts.

- The *SB* transposon is a promising vector for transgenesis in vertebrates, and is being developed as a novel, nonviral system for gene therapeutic purposes. A mutagenesis approach was undertaken to improve various aspects of the transposon, including safety and overall efficiency of gene transfer in human cells. We constructed a “sandwich” transposon, in which the DNA to be mobilized is flanked by two complete *SB* elements arranged in an inverted orientation. The sandwich element has superior ability to transpose >10 kb transgenes, thereby extending the cloning capacity of *SB*-based vectors. We derived hyperactive versions of the *SB* transposase by single-amino-acid substitutions. These mutations act synergistically, and result in an almost 4-fold enhancement of activity when compared to the wild-type transposase. We also created a library of mutant *SB* transposase genes by using an *in vitro* evolution paradigm. One particular mutant, called *SB100X*, exhibited a ~100-fold enhancement of transposition as compared to the wild-type transposase, and showed robust gene transfer efficiencies in mouse embryos as well as human hematopoietic stem cells. The improved vector system should prove useful for efficient gene transfer in vertebrates.

- We have developed a stable RNA interference (RNAi) delivery system that is based on the *FP* transposon. This plasmid-based vector system combines the gene silencing capabilities of H1 polymerase III promoter-driven short hairpin RNAs (shRNA) with the advantages of stable and efficient genomic integration of the shRNA cassette mediated by transposition. We show that the *FP*-based shRNA expressing system can efficiently knock down the expression of genes in human cells. Transposon-mediated genomic integration ensures that the shRNA expression cassette and a selectable marker gene within the transposon remain intact and physically linked. We demonstrate that a major advantage of our vector system over plasmid-based shRNA delivery is both its enhanced frequency of intact genomic integration as well as higher target suppression in transgenic human cells. Due to its simplicity and effectiveness, transposon-based RNAi is an emerging tool to facilitate analysis of gene function through the establishment of stable loss-of-function cell lines.

- Transposon-based gene vectors have become indispensable tools in vertebrate genetics for applications ranging from insertional mutagenesis and transgenesis in model species to gene therapy in humans. The transposon toolkit is expanding, but a careful, side-by-side characterization of the diverse transposon systems has been lacking. Here we compared the *SB*, *piggyBac* and *Tol2* transposons with respect to overall activity, overproduction inhibition (OPI), target site selection and transgene copy number as well as long-term expression in human cells. *SB* was the most efficient system under conditions where the availability of the transposon DNA is limiting the transposition reaction including hard-to-transfect hematopoietic stem/progenitor cells, and the most sensitive to OPI, underpinning the need for careful optimization of the transposon components. *SB* and *piggyBac* were about equally active, and both more efficient than *Tol2*, under nonrestrictive conditions. All three systems provided long-term transgene expression in human cells with minimal signs of silencing. *SB*, *Tol2* and *piggyBac* constitute complementary research tools for gene transfer in mammalian cells with important implications for fundamental and translational research.

- Incorporation of chicken beta-globin HS4 insulator sequences in *SB*-based vectors reduces transactivation of model promoters by transposon-borne enhancers, and thus may lower the risk of transcriptional activation of host genes situated close to a transposon insertion site.

- Random chromosomal transposition is clearly undesired for human gene therapeutic applications due to potential genotoxic effects associated with transposon integration. We demonstrated targeted chromosomal insertion of the *SB* transposon in human cells. We established a successful strategy based on targeting proteins that can bind both transposon and target DNA to direct *SB* element transposition into the vicinity of a specific DNA sequence in the human genome. Furthermore, transposon targeting based on protein-protein interactions between the *SB* transposase and a targeting fusion containing the N-terminal protein interaction domain of *SB* is a successful strategy to direct SB integrations into a given locus in the human genome. This approach was found

to enable a ~10^7-fold enrichment of transgene insertion at a desired locus. Our results provide proof-of-principle for directing chromosomal insertion of an otherwise randomly integrating genetic element into preselected sites. Targeted transposition could be a powerful technology for safe transgene integration in human applications.

5 REFERENCES

1. Kidwell MG, Lisch DR (2001). Transposable elements, parasitic DNA, and genome evolution. *Int J Org Evol*; 55: 1-24.
2. Orgel LE, Crick FH (1980). Selfish DNA: the ultimate parasite. *Nature*; 284: 604-607.
3. Ohno S (1972). So much "junk" DNA in our genome. *Brookhaven Symp Biol*; 23: 366-370.
4. Makalowski W (2003). Genomics. Not junk after all. *Science*; 300: 1246-1247.
5. Brosius J (1991). Retroposons - seeds of evolution. *Science*; 251: 753.
6. Nowak R (1994). Mining treasures from 'junk DNA'. *Science*; 263: 608-610.
7. McClure MA (1991). Evolution of retroposons by acquisition or deletion of retrovirus-like genes. *Mol Biol Evol*; 8: 835-856.
8. Xiong Y, Eickbush TH (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J*; 9: 3353-3362.
9. Trono D (1995). HIV accessory proteins: leading roles for the supporting cast. *Cell*; 82: 189-192.
10. Kim A, Terzian C, Santamaria P, Pelisson A, Purd'homme N, Bucheton A (1994). Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. *Proc Natl Acad Sci U S A*; 91: 1285-1289.
11. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. (2001). Initial sequencing and analysis of the human genome. *Nature*; 409: 860-921.
12. Malik HS, Burke WD, Eickbush TH (1999). The age and evolution of non-LTR retrotransposable elements. *Mol Biol Evol*; 16: 793-805.
13. Fanning T, Singer M (1987). The LINE-1 DNA sequences in four mammalian orders predict proteins that conserve homologies to retrovirus proteins. *Nucleic Acids Res*; 15: 2251-2260.
14. Moran JV, Gilbert N (2002) Mammalian LINE-1 retrotransposons and related elements. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington DC: ASM Press. pp. 836-869.
15. Sassaman DM, Dombroski BA, Moran JV, Kimberland ML, Naas TP, DeBerardinis RJ, et al. (1997). Many human L1 elements are capable of retrotransposition. *Nat Genet*; 16: 37-43.
16. Dewannieux M, Esnault C, Heidmann T (2003). LINE-mediated retrotransposition of marked Alu sequences. *Nat Genet*; 35: 41-48.
17. Dewannieux M, Heidmann T (2005). LINEs, SINEs and processed pseudogenes: parasitic strategies for genome modeling. *Cytogenet Genome Res*; 110: 35-48.
18. Ullu E, Tschudi C (1984). Alu sequences are processed 7SL RNA genes. *Nature*; 312: 171-172.
19. Weichenrieder O, Wild K, Strub K, Cusack S (2000). Structure and assembly of the Alu domain of the mammalian signal recognition particle. *Nature*; 408: 167-173.
20. Plasterk RH, Izsvak Z, Ivics Z (1999). Resident aliens: the Tc1/mariner superfamily of transposable elements. *Trends Genet*; 15: 326-332.
21. Kapitonov VV, Jurka J (2001). Rolling-circle transposons in eukaryotes. *Proc Natl Acad Sci U S A*; 98: 8714-8719.
22. Kapitonov VV, Jurka J (2006). Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A*; 103: 4540-4545.
23. Hedges RW, Jacob AE (1974). Transposition of ampicillin resistance from RP4 to other replicons. *Mol Gen Genet*; 132: 31-40.
24. Kunze R (1996). The maize transposable element activator (Ac). *Curr Top Microbiol Immunol*; 204: 161-194.
25. Emmons SW, Yesner L, Ruan KS, Katzenberg D (1983). Evidence for a transposon in Caenorhabditis elegans. *Cell*; 32: 55-65.
26. Jacobson JW, Medhora MM, Hartl DL (1986). Molecular structure of a somatically unstable transposable element in Drosophila. *Proc Natl Acad Sci U S A*; 83: 8684-8688.
27. Smit AF, Riggs AD (1996). *Tiggers* and DNA transposon fossils in the human genome. *Proc Natl Acad Sci USA*; 93: 1443-1448.
28. Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K (1992). The *pogo* transposable element family of *Drosophila melanogaster*. *Mol Gen Genet*; 232: 126-134.
29. Plasterk RH (1996). The Tc1/*mariner* transposon family. *Curr Top Microbiol Immunol*; 204: 125-143.
30. Robertson HM (1995). The Tc1-*mariner* superfamily of transposons in animals. *J Insect Physiol*; 41: 99-105.
31. Hartl DL, Lohe AR, Lozovskaya ER (1997). Modern thoughts on an ancyent marinere: Function, evolution, regulation. *Annu Rev Genet*; 31: 337-358.
32. Capy P, Vitalis R, Langin T, Higuet D, Bazin C (1996). Relationships between transposable elements based upon the integrase-transposase domains: is there a common ancestor? *J Mol Evol*; 42: 359-368.

33. Doak TG, Doerder FP, Jahn CL, Herrick G (1994). A proposed superfamily of transposase genes: transposon-like elements in ciliated protozoa and a common "D35E" motif. *Proc Natl Acad Sci USA*; 91: 942-946.
34. Machida C, Machida Y (1989). Regulation of IS1 transposition by the insA gene product. *J Mol Biol*; 208: 567-574.
35. Lavoie BD, Chaconas G (1996). Transposition of phage Mu DNA. *Curr Top Microbiol Immunol*; 204: 83-102.
36. Weinreich MD, Mahnke-Braam L, Reznikoff WS (1994). A functional analysis of the Tn5 transposase. Identification of domains required for DNA binding and multimerization. *J Mol Biol*; 241: 166-177.
37. Jain C, Kleckner N (1993). Preferential cis action of IS10 transposase depends upon its mode of synthesis. *Mol Microbiol*; 9: 249-260.
38. Derbyshire KM, Hwang L, Grindley ND (1987). Genetic analysis of the interaction of the insertion sequence IS903 transposase with its terminal inverted repeats. *Proc Natl Acad Sci U S A*; 84: 8049-8053.
39. Ichikawa H, Ikeda K, Amemura J, Ohtsubo E (1990). Two domains in the terminal inverted-repeat sequence of transposon Tn3. *Gene*; 86: 11-17.
40. Craigie R (1996). Quality control in Mu DNA transposition. *Cell*; 85: 137-140.
41. Hauer B, Shapiro JA (1984). Control of Tn7 transposition. *Mol Gen Genet*; 194: 149-158.
42. Mahillon J, Chandler M (1998). Insertion sequences. *Microbiol Mol Biol Rev*; 62: 725-774.
43. Becker HA, Kunze R (1997). Maize Activator transposase has a bipartite DNA binding domain that recognizes subterminal sequences and the terminal inverted repeats. *Mol Gen Genet*; 254: 219-230.
44. Haren L, Ton-Hoang B, Chandler M (1999). Integrating DNA: transposases and retroviral integrases. *Annu Rev Microbiol*; 53: 245-281.
45. Izsvák Z, Khare D, Behlke J, Heinemann U, Plasterk RH, Ivics Z (2002). Involvement of a bifunctional, paired-like DNA-binding domain and a transpositional enhancer in *Sleeping Beauty* transposition. *J Biol Chem*; 277: 34581-34588.
46. Pietrokovski S, Henikoff S (1997). A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol Gen Genet*; 254: 689-695.
47. Vos JC, Plasterk RH (1994). Tc1 transposase of *Caenorhabditis elegans* is an endonuclease with a bipartite DNA binding domain. *EMBO J*; 13: 6125-6132.
48. Franz G, Loukeris TG, Dialektaki G, Thompson CR, Savakis C (1994). Mobile Minos elements from Drosophila hydei encode a two-exon transposase with similarity to the paired DNA-binding domain. *Proc Natl Acad Sci U S A*; 91: 4746-4750.
49. Ivics Z, Izsvák Z, Minter A, Hackett PB (1996). Identification of functional domains and evolution of Tc1-like transposable elements. *Proc Natl Acad Sci USA*; 93: 5008-5013.
50. Czerny T, Schaffner G, Busslinger M (1993). DNA sequence recognition by Pax proteins: bipartite structure of the paired domain and its binding site. *Genes Dev*; 7: 2048-2061.
51. Pellizzari L, Tell G, Damante G (1999). Co-operation between the PAI and RED subdomains of Pax-8 in the interaction with the thyroglobulin promoter. *Biochem J*; 337 (Pt 2): 253-262.
52. Breitling R, Gerber JK (2000). Origin of the paired domain. *Dev Genes Evol*; 210: 644-650.
53. van Pouderoyen G, Ketting RF, Perrakis A, Plasterk RH, Sixma TK (1997). Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C. elegans* in complex with transposon DNA. *EMBO J*; 16: 6044-6054.
54. Prère MF, Chandler M, Fayet O (1990). Transposition in *Shigella dysenteriae*: isolation and analysis of IS911, a new member of the IS3 group of insertion sequences. *J Bacteriol*; 172: 4090-4099.
55. Gehring WJ, Qian YQ, Billeter M, Furukubo-Tokunaga K, Schier AF, Resendez-Perez D, et al. (1994). Homeodomain-DNA recognition. *Cell*; 78: 211-223.
56. Feng JA, Johnson RC, Dickerson RE (1994). Hin recombinase bound to DNA: the origin of specificity in major and minor groove interactions. *Science*; 263: 348-355.
57. Kapitonov VV, Jurka J (2005). RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. *PLoS Biol*; 3: e181.
58. Agrawal A, Eastman QM, Schatz DG (1998). Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*; 394: 744-751.
59. Hiom K, Melek M, Gellert M (1998). DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. *Cell*; 94: 463-470.
60. Miller WJ, Hagemann S, Reiter E, Pinsker W (1992). P-element homologous sequences are tandemly repeated in the genome of Drosophila guanche. *Proc Natl Acad Sci USA*; 89: 4018-4022.
61. Jones JM, Gellert M (2004). The taming of a transposon: V(D)J recombination and the immune system. *Immunol Rev*; 200: 233-248.

62. Spanopoulou E, Zaitseva F, Wang FH, Santagata S, Baltimore D, Panayotou G (1996). The homeodomain region of Rag-1 reveals the parallel mechanisms of bacterial and V(D)J recombination. *Cell*; 87: 263-276.
63. Difilippantonio MJ, McMahan CJ, Eastman QM, Spanopoulou E, Schatz DG (1996). RAG1 medaiates signal sequence recognition and recruitment of RAG2 in V(D)J recombination. *Cell*; 87: 253-262.
64. Dreyfus DH (1992). Evidence suggesting an evolutionary relationship between transposable elements and immune system recombination sequences. *Mol Immunol*; 29: 807-819.
65. Bushman FD, Engelman A, Palmer I, Wingfield P, Craigie R (1993). Domains of the integrase protein of human immunodeficiency virus type 1 responsible for polynucleotidyl transfer and zinc binding. *Proc Natl Acad Sci USA*; 90: 3428-3432.
66. Eijkelenboom AP, van den Ent FM, Vos A, Doreleijers JF, HÂrd K, Tullius TD, et al. (1997). The solution structure of the amino-terminal HHCC domain of HIV-2 integrase: a three-helix bundle stabilized by zinc. *Curr Biol*; 7: 739-746.
67. Kulkosky J, Jones KS, Katz RA, Mack JP, Skalka AM (1992). Residues critical for retroviral integrative recombination in a region that is highly conserved among retroviral/retrotransposon integrases and bacterial insertion sequence transposases. *Mol Cell Biol*; 12: 2331-2338.
68. Kim DR, Dai Y, Mundy CL, Yang W, Oettinger MA (1999). Mutations of acidic residues in RAG1 define the active site of the V(D)J recombinase. *Genes Dev*; 13: 3070-3080.
69. Landree MA, Wibbenmeyer JA, Roth DB (1999). Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V(D)J recombination. *Genes Dev*; 13: 3059-3069.
70. Dyda F, Hickman AB, Jenkins TM, Engelman A, Craigie R, Davies DR (1994). Crystal structure of the catalytic domain of HIV-1 integrase: similarity to other polynucleotidyl transferases. *Science*; 266: 1981-1986.
71. Bujacz G, Alexandratos J, Wlodawer A, Merkel G, Andrake M, Katz RA, et al. (1997). Binding of different divalent cations to the active site of avian sarcoma virus integrase and their effects on enzymatic activity. *J Biol Chem*; 272: 18161-18168.
72. Goldgur Y, Dyda F, Hickman AB, Jenkins TM, Craigie R, Davies DR (1998). Three new structures of the core domain of HIV-1 integrase: an active site that binds magnesium. *Proc Natl Acad Sci U S A*; 95: 9150-9154.
73. van Luenen HG, Colloms SD, Plasterk RH (1994). The mechanism of transposition of Tc3 in C. elegans. *Cell*; 79: 293-301.
74. Lohe AR, De Aguiar D, Hartl DL (1997). Mutations in the mariner transposase: the D,D(35)E consensus sequence is nonfunctional. *Proc Natl Acad Sci U S A*; 94: 1293-1297.
75. Richardson JM, Colloms SD, Finnegan DJ, Walkinshaw MD (2009). Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. *Cell*; 138: 1096-1108.
76. Katayanagi K, Miyagawa M, Matsushima M, Ishikawa M, Kanaya S, Ikehara M, et al. (1990). Three-dimensional structure of ribonuclease H from *E. coli*. *Nature*; 347: 306-309.
77. Ariyoshi M, Vassylyev DG, Iwasaki H, Nakamura H, Shinagawa H, Morikawa K (1994). Atomic structure of the RuvC resolvase: a holliday junction-specific endonuclease from *E. coli*. *Cell*; 78: 1063-1072.
78. Lampe DJ, Churchill ME, Robertson HM (1996). A purified *mariner* transposase is sufficient to mediate transposition *in vitro*. *EMBO J*; 15: 5470-5479.
79. Fischer SE, van Luenen HG, Plasterk RH (1999). Cis requirements for transposition of Tc1-like transposons in C. elegans. *Mol Gen Genet*; 262: 268-274.
80. Izsvák Z, Ivics Z, Hackett PB (1995). Characterization of a Tc1-like transposable element in zebrafish (*Danio rerio*). *Mol Gen Genet*; 247: 312-322.
81. Franz G, Savakis C (1991). Minos, a new transposable element from Drosophila hydei, is a member of the Tc1-like family of transposons. *Nucleic Acids Res*; 19: 6646.
82. Merriman PJ, Grimes CD, Ambroziak J, Hackett DA, Skinner P, Simmons MJ (1995). *S* elements: a family of Tc1-like transposons in the genome of *Drosophila melanogaster*. *Genetics*; 141: 1425-1438.
83. Ke Z, Grossman GL, Cornel AJ, Collins FH (1996). *Quetzal*: A transposon of the Tc1 family in the mosquito *Anopheles albimanus*. *Genetica*; 98: 141-147.
84. Lam WL, Seo P, Robison K, Virk S, Gilbert W (1996). Discovery of amphibian Tc1-like transposon families. *J Mol Biol*; 257: 359-366.
85. Ivics Z, Hackett PB, Plasterk RH, Izsvak Z (1997). Molecular reconstruction of Sleeping Beauty, a Tc1-like transposon from fish, and its transposition in human cells. *Cell*; 91: 501-510.
86. Moschetti R, Caggese C, Barsanti P, Caizzi R (1998). Intra- and interspecies variation among *Bari*-1 elements of the *melanogaster* species group. *Genetics*; 150: 239-250.

87. Wang H, Hartswood E, Finnegan DJ (1999). *Pogo* transposase contains a putative helix-turn-helix DNA binding domain that recognises a 12 bp sequence within the terminal inverted repeats. *Nucleic Acids Res*; 27: 455-461.
88. Kleckner N, Chalmers RM, Kwon D, Sakai J, Bolland S (1996). Tn10 and IS10 transposition and chromosome rearrangements: mechanism and regulation in vivo and in vitro. *Curr Top Microbiol Immunol*; 204: 49-82.
89. Craig NL (1996). Transposon Tn7. *Curr Top Microbiol Immunol*; 204: 27-48.
90. Kaufman PD, Rio DC (1992). P element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. *Cell*; 69: 27-39.
91. Lodish H, Berk A, Matsudaira P, Kaiser CA, Krieger M, Scott MP, et al. (2004) Molecular Cell Biology. New York: Freeman. 417 p.
92. Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J (1990). High-frequency P element loss in *Drosophila* is homolog dependent. *Cell*; 62: 515-525.
93. Lohe AR, Timmons C, Beerman I, Lozovskaya ER, Hartl DL (2000). Self-inflicted wounds, template-directed gap repair and a recombination hotspot. Effects of the mariner transposase. *Genetics*; 154: 647-656.
94. Mizuuchi K (1992). Polynucleotidyl transfer reactions in transpositional DNA recombination. *J Biol Chem*; 267: 21273-21276.
95. Curcio MJ, Derbyshire KM (2003). The outs and ins of transposition: from mu to kangaroo. *Nat Rev Mol Cell Biol*; 4: 865-877.
96. Mizuuchi K (1997). Polynucleotidyl transfer reactions in site-specific DNA recombination. *Genes Cells*; 2: 1-12.
97. Craig NL (1995). Unity in transposition reactions. *Science*; 270: 253-254.
98. Turlan C, Chandler M (2000). Playing second fiddle: second-strand processing and liberation of transposable elements from donor DNA. *Trends Microbiol*; 8: 268-274.
99. Gellert M (2002). V(D)J recombination: RAG proteins, repair factors, and regulation. *Annu Rev Biochem*; 71: 101-132.
100. Kennedy AK, Guhathakurta A, Kleckner N, Haniford DB (1998). Tn10 transposition via a DNA hairpin intermediate. *Cell*; 95: 125-134.
101. Miskey C, Papp B, Mates L, Sinzelle L, Keller H, Izsvak Z, et al. (2007). 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*.
102. Luo G, Ivics Z, Izsvak Z, Bradley A (1998). Chromosomal transposition of a Tc1/mariner-like element in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*; 95: 10769-10773.
103. van Gent DC, Mizuuchi K, Gellert M (1996). Similarities between initiation of V(D)J recombination and retroviral integration. *Science*; 271: 1592-1594.
104. Bhasin A, Goryshin IY, Reznikoff WS (1999). Hairpin formation in Tn5 transposition. *J Biol Chem*; 274: 37021-37029.
105. Bainton R, Gamas P, Craig NL (1991). Tn7 transposition in vitro proceeds through an excised transposon intermediate generated by staggered breaks in DNA. *Cell*; 65: 805-816.
106. Craig NL (1997). Target site selection in transposition. *Annu Rev Biochem*; 66: 437-474.
107. Goryshin IY, Miller JA, Kil YV, Lanzov VA, Reznikoff WS (1998). Tn5/IS50 target recognition. *Proc Natl Acad Sci USA*; 95: 10716-10721.
108. Bainton RJ, Kubo KM, Feng JN, Craig NL (1993). Tn7 transposition: target DNA recognition is mediated by multiple Tn7-encoded proteins in a purified in vitro system. *Cell*; 72: 931-943.
109. Devine SE, Boeke JD (1996). Integration of the yeast retrotransposon Ty1 is targeted to regions upstream of genes transcribed by RNA polymerase III. *Genes Dev*; 10: 620-633.
110. Kirchner J, Connolly CM, Sandmeyer SB (1995). Requirement of RNA polymerase III transcription factors for in vitro position-specific integration of a retroviruslike element. *Science*; 267: 1488-1491.
111. Zou S, Ke N, Kim JM, Voytas DF (1996). The Saccharomyces retrotransposon Ty5 integrates preferentially into regions of silent chromatin at the telomeres and mating loci. *Genes Dev*; 10: 634-645.
112. Junop MS, Haniford DB (1997). Factors responsible for target site selection in Tn10 transposition: a role for the DDE motif in target DNA capture. *EMBO J*; 16: 2646-2655.
113. van Luenen HG, Plasterk RH (1994). Target site choice of the related transposable elements Tc1 and Tc3 of Caenorhabditis elegans. *Nucleic Acids Res*; 22: 262-269.
114. Katzman M, Sudol M (1995). Mapping domains of retroviral integrase responsible for viral DNA specificity and target site selection by analysis of chimeras between human immunodeficiency virus type 1 and visna virus integrases. *J Virol*; 69: 5687-5696.
115. Davies CJ, Hutchison CA (1995). Insertion site specificity of the transposon Tn3. *Nucleic Acids Res*; 23: 507-514.

116. Lodge JK, Berg DE (1990). Mutations that affect Tn5 insertion into pBR322: importance of local DNA supercoiling. *J Bacteriol*; 172: 5956-5960.
117. Davies DR, Goryshin IY, Reznikoff WS, Rayment I (2000). Three-dimensional structure of the Tn5 synaptic complex transposition intermediate. *Science*; 289: 77-85.
118. Kuduvalli PN, Rao JE, Craig NL (2001). Target DNA structure plays a critical role in Tn7 transposition. *EMBO J*; 20: 924-932.
119. Bender J, Kleckner N (1992). Tn10 insertion specificity is strongly dependent upon sequences immediately adjacent to the target-site consensus sequence. *Proc Natl Acad Sci USA*; 89: 7996-8000.
120. Pribil PA, Haniford DB (2000). Substrate recognition and induced DNA deformation by transposase at the target-capture stage of Tn10 transposition. *J Mol Biol*; 303: 145-159.
121. Liao GC, Rehm EJ, Rubin GM (2000). Insertion site preferences of the P transposable element in Drosophila melanogaster. *Proc Natl Acad Sci USA*; 97: 3347-3351.
122. Milot E, Belmaaza A, Rassart E, Chartrand P (1994). Association of a host DNA structure with retroviral integration sites in chromosomal DNA. *Virology*; 201: 408-412.
123. Muller HP, Varmus HE (1994). DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J*; 13: 4704-4714.
124. Katz RA, Gravuer K, Skalka AM (1998). A preferred target DNA structure for retroviral integrase in vitro. *J Biol Chem*; 273: 24190-24195.
125. Pruss D, Reeves R, Bushman FD, Wolffe AP (1994). The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem*; 269: 25031-25041.
126. Cost GJ, Boeke JD (1998). Targeting of human retrotransposon integration is directed by the specificity of the L1 endonuclease for regions of unusual DNA structure. *Biochemistry*; 37: 18081-18093.
127. Jurka J, Klonowski P, Trifonov EN (1998). Mammalian retroposons integrate at kinkable DNA sites. *J Biomol Struct Dyn*; 15: 717-721.
128. Lampe DJ, Grant TE, Robertson HM (1998). Factors affecting transposition of the *Himar*1 *mariner* transposon *in vitro*. *Genetics*; 149: 179-187.
129. Laski FA, Rio DC, Rubin GM (1986). Tissue specificity of Drosophila P element transposition is regulated at the level of mRNA splicing. *Cell*; 44: 7-19.
130. Seleme MC, Busseau I, Malinsky S, Bucheton A, Teninges D (1999). High-frequency retrotransposition of a marked I factor in Drosophila melanogaster correlates with a dynamic expression pattern of the ORF1 protein in the cytoplasm of oocytes. *Genetics*; 151: 761-771.
131. Dupressoir A, Heidmann T (1996). Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice. *Mol Cell Biol*; 16: 4495-4503.
132. Ergun S, Buschmann C, Heukeshoven J, Dammann K, Schnieders F, Lauke H, et al. (2004). Cell type-specific expression of LINE-1 open reading frames 1 and 2 in fetal and adult human tissues. *J Biol Chem*; 279: 27753-27763.
133. Trelogan SA, Martin SL (1995). Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis. *Proc Natl Acad Sci USA*; 92: 1520-1524.
134. Branciforte D, Martin SL (1994). Developmental and cell type specificity of LINE-1 expression in mouse testis: implications for transposition. *Mol Cell Biol*; 14: 2584-2592.
135. Eide D, Anderson P (1988). Insertion and excision of Caenorhabditis elegans transposable element Tc1. *Mol Cell Biol*; 8: 737-746.
136. Emmons SW, Yesner L (1984). High-frequency excision of transposable element Tc 1 in the nematode Caenorhabditis elegans is limited to somatic cells. *Cell*; 36: 599-605.
137. Sijen T, Plasterk RH (2003). Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature*; 426: 310-314.
138. Dalrymple B, Arber W (1985). Promotion of RNA transcription on the insertion element IS30 of E. coli K12. *EMBO J*; 4: 2687-2693.
139. Reimmann C, Moore R, Little S, Savioz A, Willetts NS, Haas D (1989). Genetic structure, function and regulation of the transposable element IS21. *Mol Gen Genet*; 215: 416-424.
140. Duval-Valentin G, Normand C, Khemici V, Marty B, Chandler M (2001). Transient promoter formation: a new feedback mechanism for regulation of IS911 transposition. *EMBO J*; 20: 5802-5811.
141. Raina R, Cook D, Fedoroff N (1993). Maize Spm transposable element has an enhancer-insensitive promoter. *Proc Natl Acad Sci USA*; 90: 6355-6359.
142. Kunze R, Stochaj U, Laufs J, Starlinger P (1987). Transcription of transposable element Activator (Ac) of Zea mays L. *EMBO J*; 6: 1555-1563.
143. Kaufman PD, Rio DC (1991). Drosophila P-element transposase is a transcriptional repressor in vitro. *Proc Natl Acad Sci USA*; 88: 2613-2617.

144. Yoder JA, Walsh CP, Bestor TH (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet*; 13: 335-340.
145. Roberts D, Hoopes BC, McClure WR, Kleckner N (1985). IS10 transposition is regulated by DNA adenine methylation. *Cell*; 43: 117-130.
146. Yin JC, Krebs MP, Reznikoff WS (1988). Effect of dam methylation on Tn5 transposition. *J Mol Biol*; 199: 35-45.
147. Kunze R, Starlinger P, Schwartz D (1988). DNA methylation of the maize transposable element Ac interferes with its transcription. *Mol Gen Genet*; 214: 325-327.
148. Hata K, Sakaki Y (1997). Identification of critical CpG sites for repression of L1 transcription by DNA methylation. *Gene*; 189: 227-234.
149. Thayer RE, Singer MF, Fanning TG (1993). Undermethylation of specific LINE-1 sequences in human cells producing a LINE-1-encoded protein. *Gene*; 133: 273-277.
150. Yu F, Zingler N, Schumann G, Stratling WH (2001). Methyl-CpG-binding protein 2 represses LINE-1 expression and retrotransposition but not Alu transcription. *Nucleic Acids Res*; 29: 4493-4501.
151. Walsh CP, Chaillet JR, Bestor TH (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet*; 20: 116-117.
152. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T (2001). Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature*; 411: 212-214.
153. Hirochika H, Okamoto H, Kakutani T (2000). Silencing of retrotransposons in arabidopsis and reactivation by the ddm1 mutation. *Plant Cell*; 12: 357-369.
154. Lavie L, Maldener E, Brouha B, Meese EU, Mayer J (2004). The human L1 promoter: variable transcription initiation sites and a major impact of upstream flanking sequence on promoter activity. *Genome Res*; 14: 2253-2260.
155. Tchenio T, Casella JF, Heidmann T (2000). Members of the SRY family regulate the human LINE retrotransposons. *Nucleic Acids Res*; 28: 411-415.
156. Muotri AR, Chu VT, Marchetto MC, Deng W, Moran JV, Gage FH (2005). Somatic mosaicism in neuronal precursor cells mediated by L1 retrotransposition. *Nature*; 435: 903-910.
157. Minakami R, Kurose K, Etoh K, Furuhata Y, Hattori M, Sakaki Y (1992). Identification of an internal cis-element essential for the human L1 transcription and a nuclear factor(s) binding to the element. *Nucleic Acids Res*; 20: 3139-3145.
158. Becker KG, Swergold GD, Ozato K, Thayer RE (1993). Binding of the ubiquitous nuclear transcription factor YY1 to a cis regulatory sequence in the human LINE-1 transposable element. *Hum Mol Genet*; 2: 1697-1702.
159. Yang N, Zhang L, Zhang Y, Kazazian HH, Jr. (2003). An important role for RUNX3 in human L1 transcription and retrotransposition. *Nucleic Acids Res*; 31: 4929-4940.
160. Hartl DL, Lozovskaya ER, Nurminsky DI, Lohe AR (1997). What restricts the activity of *mariner*-like transposable elements? *Trends Genet*; 13: 197-201.
161. Surette MG, Buch SJ, Chaconas G (1987). Transpososomes: stable protein-DNA complexes involved in the in vitro transposition of bacteriophage Mu DNA. *Cell*; 49: 253-262.
162. Craigie R, Mizuuchi K (1987). Transposition of Mu DNA: joining of Mu to target DNA can be uncoupled from cleavage at the ends of Mu. *Cell*; 51: 493-501.
163. Lavoie BD, Chaconas G (1993). Site-specific HU binding in the Mu transpososome: conversion of a sequence-independent DNA-binding protein into a chemical nuclease. *Genes Dev*; 7: 2510-2519.
164. Robertson CA, Nash HA (1988). Bending of the bacteriophage lambda attachment site by Escherichia coli integration host factor. *J Biol Chem*; 263: 3554-3557.
165. Surette MG, Lavoie BD, Chaconas G (1989). Action at a distance in Mu DNA transposition: an enhancer-like element is the site of action of supercoiling relief activity by integration host factor (IHF). *EMBO J*; 8: 3483-3489.
166. Chalmers R, Guhathakurta A, Benjamin H, Kleckner N (1998). IHF modulation of Tn10 transposition: sensory transduction of supercoiling status via a proposed protein/DNA molecular spring. *Cell*; 93: 897-908.
167. Signon L, Kleckner N (1995). Negative and positive regulation of Tn10/IS10-promoted recombination by IHF: two distinguishable processes inhibit transposition off of multicopy plasmid replicons and activate chromosomal events that favor evolution of new transposons. *Genes Dev*; 9: 1123-1136.
168. Crellin P, Chalmers R (2001). Protein-DNA contacts and conformational changes in the Tn10 transpososome during assembly and activation for cleavage. *EMBO J*; 20: 3882-3891.
169. van Gent DC, Hiom K, Paull TT, Gellert M (1997). Stimulation of V(D)J cleavage by high mobility group proteins. *EMBO J*; 16: 2665-2670.

170. Aidinis V, Bonaldi T, Beltrame M, Santagata S, Bianchi ME, Spanopoulou E (1999). The RAG1 homeodomain recruits HMG1 and HMG2 to facilitate recombination signal sequence binding and to enhance the intrinsic DNA-bending activity of RAG1-RAG2. *Mol Cell Biol*; 19: 6532-6542.
171. Aiyar A, Hindmarsh P, Skalka AM, Leis J (1996). Concerted integration of linear retroviral DNA by the avian sarcoma virus integrase in vitro: dependence on both long terminal repeat termini. *J Virol*; 70: 3571-3580.
172. Farnet CM, Bushman FD (1997). HIV-1 cDNA integration: requirement of HMG I(Y) protein for function of preintegration complexes in vitro. *Cell*; 88: 483-492.
173. Li L, Yoder K, Hansen MS, Olvera J, Miller MD, Bushman FD (2000). Retroviral cDNA integration: stimulation by HMG I family proteins. *J Virol*; 74: 10965-10974.
174. Hartl D (2001). Discovery of the transposable element mariner. *Genetics*; 157: 471-476.
175. Zhang L, Dawson A, Finnegan DJ (2001). DNA-binding activity and subunit interaction of the mariner transposase. *Nucleic Acids Res*; 29: 3566-3575.
176. Dawson A, Finnegan DJ (2003). Excision of the Drosophila mariner transposon Mos1. Comparison with bacterial transposition and V(D)J recombination. *Mol Cell*; 11: 225-235.
177. Richardson JM, Dawson A, O'Hagan N, Taylor P, Finnegan DJ, Walkinshaw MD (2006). Mechanism of Mos1 transposition: insights from structural analysis. *Embo J*; 25: 1324-1334.
178. Richardson JM, Finnegan DJ, Walkinshaw MD (2007). Crystallization of a Mos1 transposase-inverted-repeat DNA complex: biochemical and preliminary crystallographic analyses. *Acta Crystallogr Sect F Struct Biol Cryst Commun*; 63: 434-437.
179. Auge-Gouillou C, Brillet B, Hamelin MH, Bigot Y (2005). Assembly of the mariner Mos1 synaptic complex. *Mol Cell Biol*; 25: 2861-2870.
180. Auge-Gouillou C, Brillet B, Germon S, Hamelin MH, Bigot Y (2005). Mariner Mos1 transposase dimerizes prior to ITR binding. *J Mol Biol*; 351: 117-130.
181. Grewal SI, Jia S (2007). Heterochromatin revisited. *Nat Rev Genet*; 8: 35-46.
182. Bender J (2004). Chromatin-based silencing mechanisms. *Curr Opin Plant Biol*; 7: 521-526.
183. Lahue E, Heckathorn J, Meyer Z, Smith J, Wolfe C (2005). The Saccharomyces cerevisiae Sub2 protein suppresses heterochromatic silencing at telomeres and subtelomeric genes. *Yeast*; 22: 537-551.
184. Allshire RC, Javerzat JP, Redhead NJ, Cranston G (1994). Position effect variegation at fission yeast centromeres. *Cell*; 76: 157-169.
185. Thon G, Klar AJ (1992). The clr1 locus regulates the expression of the cryptic mating-type loci of fission yeast. *Genetics*; 131: 287-296.
186. Ayoub N, Goldshmidt I, Cohen A (1999). Position effect variegation at the mating-type locus of fission yeast: a cis-acting element inhibits covariegated expression of genes in the silent and expressed domains. *Genetics*; 152: 495-508.
187. Clarke L, Amstutz H, Fishel B, Carbon J (1986). Analysis of centromeric DNA in the fission yeast Schizosaccharomyces pombe. *Proc Natl Acad Sci U S A*; 83: 8253-8257.
188. Nakaseko Y, Adachi Y, Funahashi S, Niwa O, Yanagida M (1986). Chromosome walking shows a highly homologous repetitive sequence present in all the centromere regions of fission yeast. *EMBO J*; 5: 1011-1021.
189. Lee SE, Mitchell RA, Cheng A, Hendrickson EA (1997). Evidence for DNA-PK-dependent and -independent DNA double-strand break repair pathways in mammalian cells as a function of the cell cycle. *Mol Cell Biol*; 17: 1425-1433.
190. Takata M, Sasaki MS, Sonoda E, Morrison C, Hashimoto M, Utsumi H, et al. (1998). Homologous recombination and non-homologous end-joining pathways of DNA double-strand break repair have overlapping roles in the maintenance of chromosomal integrity in vertebrate cells. *EMBO J*; 17: 5497-5508.
191. Jackson SP, Jeggo PA (1995). DNA double-strand break repair and V(D)J recombination: involvement of DNA-PK. *Trends Biochem Sci*; 20: 412-415.
192. Lin WC, Desiderio S (1994). Cell cycle regulation of V(D)J recombination-activating protein RAG-2. *Proc Natl Acad Sci USA*; 91: 2733-2737.
193. Lee J, Desiderio S (1999). Cyclin A/CDK2 regulates V(D)J recombination by coordinating RAG-2 accumulation and DNA repair. *Immunity*; 11: 771-781.
194. Jowett JB, Planelles V, Poon B, Shah NP, Chen ML, Chen IS (1995). The human immunodeficiency virus type 1 vpr gene arrests infected T cells in the G2 + M phase of the cell cycle. *J Virol*; 69: 6304-6313.
195. He J, Choe S, Walker R, Di Marzio P, Morgan DO, Landau NR (1995). Human immunodeficiency virus type 1 viral protein R (Vpr) arrests cells in the G2 phase of the cell cycle by inhibiting p34cdc2 activity. *J Virol*; 69: 6705-6711.
196. Re F, Braaten D, Franke EK, Luban J (1995). Human immunodeficiency virus type 1 Vpr arrests the cell cycle in G2 by inhibiting the activation of p34cdc2-cyclin B. *J Virol*; 69: 6859-6864.

197. Zhao Y, Cao J, O'Gorman MR, Yu M, Yogev R (1996). Effect of human immunodeficiency virus type 1 protein R (vpr) gene expression on basic cellular function of fission yeast Schizosaccharomyces pombe. *J Virol*; 70: 5821-5826.
198. de Noronha CM, Sherman MP, Lin HW, Cavrois MV, Moir RD, Goldman RD, et al. (2001). Dynamic disruptions in nuclear envelope architecture and integrity induced by HIV-1 Vpr. *Science*; 294: 1105-1108.
199. Goh WC, Rogel ME, Kinsey CM, Michael SF, Fultz PN, Nowak MA, et al. (1998). HIV-1 Vpr increases viral expression by manipulation of the cell cycle: a mechanism for selection of Vpr in vivo. *Nat Med*; 4: 65-71.
200. Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998). Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete Saccharomyces cerevisiae genome sequence. *Genome Res*; 8: 464-478.
201. Bryk M, Banerjee M, Murphy M, Knudsen KE, Garfinkel DJ, Curcio MJ (1997). Transcriptional silencing of Ty1 elements in the RDN1 locus of yeast. *Genes Dev*; 11: 255-269.
202. Bachman N, Gelbart ME, Tsukiyama T, Boeke JD (2005). TFIIIB subunit Bdp1p is required for periodic integration of the Ty1 retrotransposon and targeting of Isw2p to S. cerevisiae tDNAs. *Genes Dev*; 19: 955-964.
203. Yieh L, Hatzis H, Kassavetis G, Sandmeyer SB (2002). Mutational analysis of the transcription factor IIIB-DNA target of Ty3 retroelement integration. *J Biol Chem*; 277: 25920-25928.
204. Aye M, Dildine SL, Claypool JA, Jourdain S, Sandmeyer SB (2001). A truncation mutant of the 95-kilodalton subunit of transcription factor IIIC reveals asymmetry in Ty3 integration. *Mol Cell Biol*; 21: 7839-7851.
205. Xie W, Gai X, Zhu Y, Zappulla DC, Sternglanz R, Voytas DF (2001). Targeting of the yeast Ty5 retrotransposon to silent chromatin is mediated by interactions between integrase and Sir4p. *Mol Cell Biol*; 21: 6606-6614.
206. Peters JE, Craig NL (2001). Tn7: smarter than we thought. *Nat Rev Mol Cell Biol*; 2: 806-814.
207. Kuduvalli PN, Mitra R, Craig NL (2005). Site-specific Tn7 transposition into the human genome. *Nucleic Acids Res*; 33: 857-863.
208. Loomis WF, Welker D, Hughes J, Maghakian D, Kuspa A (1995). Integrated maps of the chromosomes in Dictyostelium discoideum. *Genetics*; 141: 147-157.
209. Winckler T, Dingermann T, Glockner G (2002). Dictyostelium mobile elements: strategies to amplify in a compact genome. *Cell Mol Life Sci*; 59: 2097-2111.
210. Winckler T, Szafranski K, Glockner G (2005). Transfer RNA gene-targeted integration: an adaptation of retrotransposable elements to survive in the compact Dictyostelium discoideum genome. *Cytogenet Genome Res*; 110: 288-298.
211. Chung T, Siol O, Dingermann T, Winckler T (2007). Protein interactions involved in tRNA gene-specific integration of Dictyostelium non-long terminal repeat retrotransposon TRE5-A. *Mol Cell Biol*.
212. Robertson HM (1993). The *mariner* transposable element is widespread in insects. *Nature*; 362: 241-245.
213. Robertson HM, Zumpano KL (1997). Molecular evolution of an ancient *mariner* transposon, *Hsmar*1, in the human genome. *Gene*; 205: 203-217.
214. Lampe DJ, Walden KK, Robertson HM (2001). Loss of transposase-DNA interaction may underlie the divergence of mariner family transposable elements and the ability of more than one mariner to occupy the same genome. *Mol Biol Evol*; 18: 954-961.
215. Garcia-Fernandez J, Bayascas-Ramirez JR, Marfany G, Munoz-Marmol AM, Casali A, Baguna J, et al. (1995). High copy number of highly similar mariner-like transposons in planarian (Platyhelminthe): evidence for a trans-phyla horizontal transfer. *Mol Biol Evol*; 12: 421-431.
216. Kidwell MG (1992). Horizontal transfer of P elements and other short inverted repeat transposons. *Genetica*; 86: 275-286.
217. Kidwell MG (1992). Horizontal transfer. *Curr Opin Genet Dev*; 2: 868-873.
218. Houck MA, Clark JB, Peterson KR, Kidwell MG (1991). Possible horizontal transfer of Drosophila genes by the mite Proctolaelaps regalis. *Science*; 253: 1125-1128.
219. Engels W (1989) P elements in Drosophila melanogaster. In: M BD, E HM, editors. Mobile DNA I. Washington D.C.: American Society of Microbiology Washington D.C. pp. 437-483.
220. Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995). Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. *Mol Biol Evol*; 12: 62-72.
221. Silva JC, Loreto EL, Clark JB (2004). Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol*; 6: 57-71.
222. Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993). Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. *Cell*; 72: 595-605.

223. Collins J, Forbes E, Anderson P (1989). The Tc3 family of transposable genetic elements in Caenorhabditis elegans. *Genetics*; 121: 47-55.
224. Medhora M, Maruyama K, Hartl DL (1991). Molecular and functional analysis of the mariner mutator element Mos1 in Drosophila. *Genetics*; 128: 311-318.
225. Barry EG, Witherspoon DJ, Lampe DJ (2004). A bacterial genetic screen identifies functional coding sequences of the insect mariner transposable element Famar1 amplified from the genome of the earwig, Forficula auricularia. *Genetics*; 166: 823-833.
226. Kidwell MG, Lisch DR (2002) In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington DC: ASM Press. pp. 60.
227. Kazazian HHJ (1999). An estimated frequency of endogenous insertional mutations in humans. *Nat Genet*; 22: 130.
228. Rubin GM, Kidwell MG, Bingham PM (1982). The molecular basis of P-M hybrid dysgenesis: the nature of induced mutations. *Cell*; 29: 987-994.
229. Bradley D, Carpenter R, Sommer H, Hartley N, Coen E (1993). Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of Antirrhinum. *Cell*; 72: 85-95.
230. Selinger DA, Chandler VL (1999). Major recent and independent changes in levels and patterns of expression have occurred at the b gene, a regulatory locus in maize. *Proc Natl Acad Sci U S A*; 96: 15007-15012.
231. Colot V, Haedens V, Rossignol JL (1998). Extensive, nonrandom diversity of excision footprints generated by Ds-like transposon Ascot-1 suggests new parallels with V(D)J recombination. *Mol Cell Biol*; 18: 4337-4346.
232. Deininger PL, Batzer MA (1999). Alu repeats and human disease. *Mol Genet Metab*; 67: 183-193.
233. Casa AM, Brouwer C, Nagel A, Wang L, Zhang Q, Kresovich S, et al. (2000). Inaugural article: the MITE family heartbreaker (Hbr): molecular markers in maize. *Proc Natl Acad Sci U S A*; 97: 10083-10089.
234. Klobutcher LA, Herrick G (1995). Consensus inverted terminal repeat sequence of Paramecium IESs: resemblance to termini of Tc1-related and Euplotes Tec transposons. *Nucleic Acids Res*; 23: 2006-2013.
235. Hartl DL (2000). Molecular melodies in high and low C. *Nat Rev Genet*; 1: 145-149.
236. Wessler SR (1996). Turned on by stress. Plant retrotransposons. *Curr Biol*; 6: 959-961.
237. Pickeral OK, Makalowski W, Boguski MS, Boeke JD (2000). Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. *Genome Res*; 10: 411-415.
238. Ivics Z, Izsvák Z (2004). Transposable elements for transgenesis and insertional mutagenesis in vertebrates: a contemporary review of experimental strategies. *Methods Mol Biol*; 260: 255-276.
239. Ivics Z, Li MA, Mates L, Boeke JD, Nagy A, Bradley A, et al. (2009). Transposon-mediated genome manipulation in vertebrates. *Nat Methods*; 6: 415-422.
240. Cooley L, Kelley R, Spradling A (1988). Insertional mutagenesis of the Drosophila genome with single P elements. *Science*; 239: 1121-1128.
241. Zwaal RR, Broeks A, van Meurs J, Groenen JT, Plasterk RH (1993). Target-selected gene inactivation in Caenorhabditis elegans by using a frozen transposon insertion mutant bank. *Proc Natl Acad Sci U S A*; 90: 7431-7435.
242. Koga A, Suzuki M, Inagaki H, Bessho Y, Hori H (1996). Transposable element in fish. *Nature*; 383: 30.
243. Fischer SE, Wienholds E, Plasterk RH (2001). Regulated transposition of a fish transposon in the mouse germ line. *Proc Natl Acad Sci U S A*; 98: 6759-6764.
244. Rio DC, Barnes G, Laski FA, Rine J, Rubin GM (1988). Evidence for *Drosophila* P element transposase activity in mammalian cells and yeast. *J Mol Biol*; 200: 411-415.
245. Russell WL, Kelly EM, Hunsicker PR, Bangham JW, Maddux SC, Phipps EL (1979). Specific-locus test shows ethylnitrosourea to be the most potent mutagen in the mouse. *Proc Natl Acad Sci USA*; 76: 5818-5819.
246. Driever W, Solnica-Krezel L, Schier AF, Neuhauss SC, Malicki J, Stemple DL, et al. (1996). A genetic screen for mutations affecting embryogenesis in zebrafish. *Development*; 123: 37-46.
247. Haffter P, Granato M, Brand M, Mullins MC, Hammerschmidt M, Kane DA, et al. (1996). The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio*. *Development*; 123: 1-36.
248. Kile BT, Hentges KE, Clark AT, Nakamura H, Salinger AP, Liu B, et al. (2003). Functional genetic analysis of mouse chromosome 11. *Nature*; 425: 81-86.
249. Babinet C, Morello D, Renard JP (1989). Transgenic mice. *Genome*; 31: 938-949.
250. Garrick D, Fiering S, Martin DI, Whitelaw E (1998). Repeat-induced gene silencing in mammals. *Nat Genet*; 18: 56-59.

251. Bushman FD (2003). Targeting survival: integration site selection by retroviruses and LTR-retrotransposons. *Cell*; 115: 135-138.
252. Amsterdam A, Burgess S, Golling G, Chen W, Sun Z, Townsend K, et al. (1999). A large-scale insertional mutagenesis screen in zebrafish. *Genes Dev*; 13: 2713-2724.
253. Zambrowicz BP, Friedrich GA (1998). Comprehensive mammalian genetics: history and future prospects of gene trapping in the mouse. *Int J Dev Biol*; 42: 1025-1036.
254. Drabek D, Zagoraiou L, deWit T, Langeveld A, Roumpaki C, Mamalaki C, et al. (2003). Transposition of the Drosophila hydei Minos transposon in the mouse germ line. *Genomics*; 81: 108-111.
255. Zagoraiou L, Drabek D, Alexaki S, Guy JA, Klinakis AG, Langeveld A, et al. (2001). In vivo transposition of Minos, a Drosophila mobile element, in mammalian tissues. *Proc Natl Acad Sci U S A*; 98: 11474-11478.
256. 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: 473-483.
257. Kawakami K, Noda T (2004). Transposition of the Tol2 element, an Ac-like element from the Japanese medaka fish Oryzias latipes, in mouse embryonic stem cells. *Genetics*; 166: 895-899.
258. Balciunas D, Wangensteen KJ, Wilber A, Bell J, Geurts A, Sivasubbu S, et al. (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet*; 2: e169.
259. Izsvák Z, Ivics Z, Plasterk RH (2000). *Sleeping Beauty*, a wide host-range transposon vector for genetic transformation in vertebrates. *J Mol Biol*; 302: 93-102.
260. Huang X, Wilber AC, Bao L, Tuong D, Tolar J, Orchard PJ, et al. (2006). Stable gene transfer and expression in human primary T cells by the Sleeping Beauty transposon system. *Blood*; 107: 483-491.
261. Davidson AE, Balciunas D, Mohn D, Shaffer J, Hermanson S, Sivasubbu S, et al. (2003). Efficient gene delivery and gene expression in zebrafish using the Sleeping Beauty transposon. *Dev Biol*; 263: 191-202.
262. Grabher C, Henrich T, Sasado T, Arenz A, Wittbrodt J, Furutani-Seiki M (2003). Transposon-mediated enhancer trapping in medaka. *Gene*; 322: 57-66.
263. Balciunas D, Davidson AE, Sivasubbu S, Hermanson SB, Welle Z, Ekker SC (2004). Enhancer trapping in zebrafish using the Sleeping Beauty transposon. *BMC Genomics*; 5: 62.
264. Sinzelle L, Vallin J, Coen L, Chesneau A, Pasquier DD, Pollet N, et al. (2006). Generation of trangenic Xenopus laevis using the Sleeping Beauty transposon system. *Transgenic Res*.
265. Yergeau DA, Mead PE (2007). Manipulating the Xenopus genome with transposable elements. *Genome Biol*; 8 Suppl 1: S11.
266. Yant SR, Meuse L, Chiu W, Ivics Z, Izsvak Z, Kay MA (2000). Somatic integration and long-term transgene expression in normal and haemophilic mice using a DNA transposon system. *Nat Genet*; 25: 35-41.
267. Dupuy AJ, Fritz S, Largaespada DA (2001). Transposition and gene disruption in the male germline of the mouse. *Genesis*; 30: 82-88.
268. Horie K, Yusa K, Yae K, Odajima J, Fischer SE, Keng VW, et al. (2003). Characterization of Sleeping Beauty transposition and its application to genetic screening in mice. *Mol Cell Biol*; 23: 9189-9207.
269. Carlson CM, Dupuy AJ, Fritz S, Roberg-Perez KJ, Fletcher CF, Largaespada DA (2003). Transposon mutagenesis of the mouse germline. *Genetics*; 165: 243-256.
270. Geurts AM, Collier LS, Geurts JL, Oseth LL, Bell ML, Mu D, et al. (2006). Gene mutations and genomic rearrangements in the mouse as a result of transposon mobilization from chromosomal concatemers. *PLoS Genet*; 2: e156.
271. Kitada K, Ishishita S, Tosaka K, Takahashi RI, Ueda M, Keng VW, et al. (2007). Transposon-tagged mutagenesis in the rat. *Nat Methods*.
272. Lu B, Geurts AM, Poirier C, Petit DC, Harrison W, Overbeek PA, et al. (2007). Generation of rat mutants using a coat color-tagged Sleeping Beauty transposon system. *Mamm Genome*; 18: 338-346.
273. Miskey C, Izsvak Z, Kawakami K, Ivics Z (2005). DNA transposons in vertebrate functional genomics. *Cell Mol Life Sci*; 62: 629-641.
274. Mates L, Izsvak Z, Ivics Z (2007). Technology transfer from worms and flies to vertebrates: transposition-based genome manipulations and their future perspectives. *Genome Biol*; 8 Suppl 1: S1.
275. Horie K, Kuroiwa A, Ikawa M, Okabe M, Kondoh G, Matsuda Y, et al. (2001). Efficient chromosomal transposition of a Tc1/mariner- like transposon Sleeping Beauty in mice. *Proc Natl Acad Sci U S A*; 98: 9191-9196.

276. Geurts AM, Wilber A, Carlson CM, Lobitz PD, Clark KJ, Hackett PB, et al. (2006). Conditional gene expression in the mouse using a Sleeping Beauty gene-trap transposon. *BMC Biotechnol*; 6: 30.
277. Dupuy AJ, Akagi K, Largaespada DA, Copeland NG, Jenkins NA (2005). Mammalian mutagenesis using a highly mobile somatic Sleeping Beauty transposon system. *Nature*; 436: 221-226.
278. Collier LS, Carlson CM, Ravimohan S, Dupuy AJ, Largaespada DA (2005). Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature*; 436: 272-276.
279. Dupuy AJ, Jenkins NA, Copeland NG (2006). Sleeping beauty: a novel cancer gene discovery tool. *Hum Mol Genet*; 15 Spec No 1: R75-79.
280. Heggestad AD, Notterpek L, Fletcher BS (2004). Transposon-based RNAi delivery system for generating knockdown cell lines. *Biochem Biophys Res Commun*; 316: 643-650.
281. Kaufman CD, Izsvak Z, Katzer A, Ivics Z (2005). *Frog Prince* transposon-based RNAi vectors mediate efficient gene knockdown in human cells. *J RNAi Gene Silenc*; 1: 97-104.
282. Bushman F, Lewinski M, Ciuffi A, Barr S, Leipzig J, Hannenhalli S, et al. (2005). Genome-wide analysis of retroviral DNA integration. *Nat Rev Microbiol*; 3: 848-858.
283. Raz E, van Luenen HG, Schaerringer B, Plasterk RH, Driever W (1998). Transposition of the nematode Caenorhabditis elegans Tc3 element in the zebrafish Danio rerio. *Curr Biol*; 8: 82-88.
284. Fadool JM, Hartl DL, Dowling JE (1998). Transposition of the mariner element from Drosophila mauritiana in zebrafish. *Proc Natl Acad Sci U S A*; 95: 5182-5186.
285. Kawakami K, Shima A, Kawakami N (2000). 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 U S A*; 97: 11403-11408.
286. Nasevicius A, Ekker SC (2000). Effective targeted gene 'knockdown' in zebrafish. *Nat Genet*; 26: 216-220.
287. Hamlet MR, Yergeau DA, Kuliyev E, Takeda M, Taira M, Kawakami K, et al. (2006). Tol2 transposon-mediated transgenesis in Xenopus tropicalis. *Genesis*; 44: 438-445.
288. Dupuy AJ, Clark K, Carlson CM, Fritz S, Davidson AE, Markley KM, et al. (2002). Mammalian germ-line transgenesis by transposition. *Proc Natl Acad Sci U S A*; 99: 4495-4499.
289. Wilber A, Frandsen JL, Geurts JL, Largaespada DA, Hackett PB, McIvor RS (2006). RNA as a source of transposase for sleeping beauty-mediated gene insertion and expression in somatic cells and tissues. *Mol Ther*; 13: 625-630.
290. Carlson CM, Frandsen JL, Kirchhof N, McIvor RS, Largaespada DA (2005). Somatic integration of an oncogene-harboring Sleeping Beauty transposon models liver tumor development in the mouse. *Proc Natl Acad Sci U S A*; 102: 17059-17064.
291. Verma IM, Somia N (1997). Gene therapy - Promises, problems and prospects. *Nature*; 389: 239-242.
292. Dobbelstein M (2003). Viruses in therapy--royal road or dead end? *Virus Res*; 92: 219-221.
293. Thomas CE, Ehrhardt A, Kay MA (2003). Progress and problems with the use of viral vectors for gene therapy. *Nat Rev Genet*; 4: 346-358.
294. Sinn PL, Sauter SL, McCray PB, Jr. (2005). Gene therapy progress and prospects: development of improved lentiviral and retroviral vectors--design, biosafety, and production. *Gene Ther*; 12: 1089-1098.
295. VandenDriessche T, Collen D, Chuah MK (2003). Biosafety of onco-retroviral vectors. *Curr Gene Ther*; 3: 501-515.
296. Scherdin U, Rhodes K, Breindl M (1990). Transcriptionally active genome regions are preferred targets for retrovirus integration. *J Virol*; 64: 907-912.
297. Schiedner G, Morral N, Parks RJ, Wu Y, Koopmans SC, Langston C, et al. (1998). Genomic DNA transfer with a high-capacity adenovirus vector results in improved in vivo gene expression and decreased toxicity. *Nat Genet*; 18: 180-183.
298. Ehrhardt A, Kay MA (2002). A new adenoviral helper-dependent vector results in long-term therapeutic levels of human coagulation factor IX at low doses in vivo. *Blood*; 99: 3923-3930.
299. Thorrez L, VandenDriessche T, Collen D, Chuah MK (2004). Preclinical gene therapy studies for hemophilia using adenoviral vectors. *Semin Thromb Hemost*; 30: 173-183.
300. Chuah MK, Collen D, VandenDriessche T (2003). Biosafety of adenoviral vectors. *Curr Gene Ther*; 3: 527-543.
301. Kafri T, Morgan D, Krahl T, Sarvetnick N, Sherman L, Verma I (1998). Cellular immune response to adenoviral vector infected cells does not require de novo viral gene expression: implications for gene therapy. *Proc Natl Acad Sci U S A*; 95: 11377-11382.
302. Miller DG, Rutledge EA, Russell DW (2002). Chromosomal effects of adeno-associated virus vector integration. *Nat Genet*; 30: 147-148.

303. Zaiss AK, Muruve DA (2005). Immune responses to adeno-associated virus vectors. *Curr Gene Ther*; 5: 323-331.
304. Chuah MK, Collen D, Vandendriessche T (2004). Preclinical and clinical gene therapy for haemophilia. *Haemophilia*; 10 Suppl 4: 119-125.
305. Manno CS, Arruda VR, Pierce GF, Glader B, Ragni M, Rasko J, et al. (2006). Successful transduction of liver in hemophilia by AAV-Factor IX and limitations imposed by the host immune response. *Nat Med*; 12: 342-347.
306. Glover DJ, Lipps HJ, Jans DA (2005). Towards safe, non-viral therapeutic gene expression in humans. *Nat Rev Genet*; 6: 299-310.
307. Li S, Ma Z (2001). Nonviral gene therapy. *Curr Gene Ther*; 1: 201-226.
308. Abdallah B, Sachs L, Demeneix BA (1995). Non-viral gene transfer: applications in developmental biology and gene therapy. *Biol Cell*; 85: 1-7.
309. Niidome T, Huang L (2002). Gene therapy progress and prospects: nonviral vectors. *Gene Ther*; 9: 1647-1652.
310. Ivics Z, Izsvak Z (2006). Transposons for gene therapy! *Curr Gene Ther*; 6: 593-607.
311. Yant SR, Ehrhardt A, Mikkelsen JG, Meuse L, Pham T, Kay MA (2002). Transposition from a gutless adeno-transposon vector stabilizes transgene expression in vivo. *Nat Biotechnol*; 20: 999-1005.
312. Bowers WJ, Mastrangelo MA, Howard DF, Southerland HA, Maguire-Zeiss KA, Federoff HJ (2006). Neuronal precursor-restricted transduction via in utero CNS gene delivery of a novel bipartite HSV amplicon/transposase hybrid vector. *Mol Ther*; 13: 580-588.
313. Ohlfest JR, Frandsen JL, Fritz S, Lobitz PD, Perkinson SG, Clark KJ, et al. (2005). Phenotypic correction and long-term expression of factor VIII in hemophilic mice by immunotolerization and nonviral gene transfer using the Sleeping Beauty transposon system. *Blood*; 105: 2691-2698.
314. Liu L, Mah C, Fletcher BS (2006). Sustained FVIII Expression and Phenotypic Correction of Hemophilia A in Neonatal Mice Using an Endothelial-Targeted Sleeping Beauty Transposon. *Mol Ther*; 13: 1006-1015.
315. Ortiz-Urda S, Thyagarajan B, Keene DR, Lin Q, Fang M, Calos MP, et al. (2002). Stable nonviral genetic correction of inherited human skin disease. *Nat Med*; 8: 1166-1170.
316. Montini E, Held PK, Noll M, Morcinek N, Al-Dhalimy M, Finegold M, et al. (2002). In vivo correction of murine tyrosinemia type I by DNA-mediated transposition. *Mol Ther*; 6: 759-769.
317. Ohlfest JR, Demorest ZL, Motooka Y, Vengco I, Oh S, Chen E, et al. (2005). Combinatorial antiangiogenic gene therapy by nonviral gene transfer using the sleeping beauty transposon causes tumor regression and improves survival in mice bearing intracranial human glioblastoma. *Mol Ther*; 12: 778-788.
318. Chen ZJ, Kren BT, Wong PY, Low WC, Steer CJ (2005). Sleeping Beauty-mediated down-regulation of huntingtin expression by RNA interference. *Biochem Biophys Res Commun*; 329: 646-652.
319. He CX, Shi D, Wu WJ, Ding YF, Feng DM, Lu B, et al. (2004). Insulin expression in livers of diabetic mice mediated by hydrodynamics-based administration. *World J Gastroenterol*; 10: 567-572.
320. Belur LR, Frandsen JL, Dupuy AJ, Ingbar DH, Largaespada DA, Hackett PB, et al. (2003). Gene insertion and long-term expression in lung mediated by the Sleeping Beauty transposon system. *Mol Ther*; 8: 501-507.
321. Liu L, Sanz S, Heggestad AD, Antharam V, Notterpek L, Fletcher BS (2004). Endothelial targeting of the Sleeping Beauty transposon within lung. *Mol Ther*; 10: 97-105.
322. Izsvák Z, Ivics Z (2004). Sleeping beauty transposition: biology and applications for molecular therapy. *Mol Ther*; 9: 147-156.
323. Hackett PB, Ekker SC, Largaespada DA, McIvor RS (2005). Sleeping beauty transposon-mediated gene therapy for prolonged expression. *Adv Genet*; 54: 189-232.
324. Essner JJ, McIvor RS, Hackett PB (2005). Awakening gene therapy with Sleeping Beauty transposons. *Curr Opin Pharmacol*; 5: 513-519.
325. Baum C, von Kalle C, Staal FJ, Li Z, Fehse B, Schmidt M, et al. (2004). Chance or necessity? Insertional mutagenesis in gene therapy and its consequences. *Mol Ther*; 9: 5-13.
326. Wu X, Li Y, Crise B, Burgess SM (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science*; 300: 1749-1751.
327. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F (2002). HIV-1 integration in the human genome favors active genes and local hotspots. *Cell*; 110: 521-529.
328. Mitchell RS, Beitzel BF, Schroder AR, Shinn P, Chen H, Berry CC, et al. (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol*; 2: E234.

329. Hacein-Bey-Abina S, Le Deist F, Carlier F, Bouneaud C, Hue C, De Villartay JP, et al. (2002). Sustained correction of X-linked severe combined immunodeficiency by ex vivo gene therapy. *N Engl J Med*; 346: 1185-1193.
330. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, et al. (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science*; 302: 415-419.
331. Baum C (2007). What are the consequences of the fourth case? *Mol Ther*; 15: 1401-1402.
332. Thrasher AJ, Gaspar HB (2007). Severe Adverse Event in Clinical Trial of Gene Therapy for X-SCID. *ASGT press release*.
333. Sinzelle L, Pollet N, Bigot Y, Mazabraud A (2005). Characterization of multiple lineages of Tc1-like elements within the genome of the amphibian Xenopus tropicalis. *Gene*; 349: 187-196.
334. Cui Z, Geurts AM, Liu G, Kaufman CD, Hackett PB (2002). Structure-function analysis of the inverted terminal repeats of the *Sleeping Beauty* transposon. *J Mol Biol*; 318: 1221-1235.
335. Zayed H, Izsvak Z, Khare D, Heinemann U, Ivics Z (2003). The DNA-bending protein HMGB1 is a cellular cofactor of *Sleeping Beauty* transposition. *Nucleic Acids Res*; 31: 2313-2322.
336. Leaver MJ (2001). A family of Tc1-like transposons from the genomes of fishes and frogs: evidence for horizontal transmission. *Gene*; 271: 203-214.
337. Zayed H, Izsvak Z, Walisko O, Ivics Z (2004). Development of hyperactive sleeping beauty transposon vectors by mutational analysis. *Mol Ther*; 9: 292-304.
338. Bewley CA, Gronenborn AM, Clore GM (1998). Minor groove-binding architectural proteins: structure, function, and DNA recognition. *Annu Rev Biophys Biomol Struct*; 27: 105-131.
339. Walisko O, Schorn A, Rolfs F, Devaraj A, Miskey C, Izsvak Z, et al. (2008). Transcriptional activities of the Sleeping Beauty transposon and shielding its genetic cargo with insulators. *Mol Ther*; 16: 359-369.
340. Yamada M, Ohkawara B, Ichimura N, Hyodo-Miura J, Urushiyama S, Shirakabe K, et al. (2003). Negative regulation of Wnt signalling by HMG2L1, a novel NLK-binding protein. *Genes Cells*; 8: 677-684.
341. Leung PC, Teplow DB, Harshey RM (1989). Interaction of distinct domains in Mu transposase with Mu DNA ends and an internal transpositional enhancer. *Nature*; 338: 656-658.
342. Bustin M (1999). Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins. *Mol Cell Biol*; 19: 5237-5246.
343. Richet E, Abcarian P, Nash H (1986). Synapsis of attachment sites during lambda integrative recombination involves capture of a naked DNA by a protein-DNA complex. *Cell*; 46: 1011–1021.
344. Sherr CJ (1995). Mammalian G1 cyclins and cell cycle progression. *Proc Assoc Am Physicians*; 107: 181-186.
345. Baldin V, Lukas J, Marcote MJ, Pagano M, Draetta G (1993). Cyclin D1 is a nuclear protein required for cell cycle progression in G1. *Genes Dev*; 7: 812-821.
346. Peukert K, Staller P, Schneider A, Carmichael G, Hanel F, Eilers M (1997). An alternative pathway for gene regulation by Myc. *EMBO J*; 16: 5672-5686.
347. Walisko O, Izsvak Z, Szabo K, Kaufman CD, Herold S, Ivics Z (2006). Sleeping Beauty transposase modulates cell-cycle progression through interaction with Miz-1. *Proc Natl Acad Sci U S A*; 103: 4062-4067.
348. Izsvák Z, Stuwe EE, Fiedler D, Katzer A, Jeggo PA, 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.
349. Yant SR, Kay MA (2003). Nonhomologous-end-joining factors regulate DNA repair fidelity during *Sleeping Beauty* element transposition in mammalian cells. *Mol Cell Biol*; 23: 8505-8518.
350. Emerman M (1996). HIV-1, Vpr and the cell cycle. *Curr Biol*; 6: 1096-1103.
351. Lomonte P, Everett RD (1999). Herpes Simplex Virus type 1 immediate-early protein Vmw110 inhibits progression of cells through mitosis and from G1 into S phase of the cell cycle. *J Virol*; 73: 9456-9467.
352. Lu M, Shenk T (1999). Human cytomegalovirus UL69 protein induces cells to accumulate in G1 phase of the cell cycle. *J Virol*; 73: 676-683.
353. Cayrol C, Flemington EK (1996). The Epstein-Barr virus bZIP transcription factor Zta causes G0/G1 cell cycle arrest through induction of cyclin-dependent kinase inhibitors. *EMBO J*; 15: 2748-2759.
354. Izumiya Y, Lin S-F, Ellison TJ, Levy AM, Mayeur GL, Izumiya C, et al. (2003). Cell cycle regulation by Kaposi's sarcoma-associated herpesvirus K-bZIP: direct interaction with Cyclin-CDK2 and induction of G1 growth arrest. *J Virol*; 77: 9652-9661.
355. Chen CJ, Makino S (2004). Murine coronavirus replication induces cell cycle arrest in G0/G1 phase. *J Virol*; 78: 5658-5669.

356. Yusa K, Takeda J, Horie K (2004). Enhancement of Sleeping Beauty transposition by CpG methylation: possible role of heterochromatin formation. *Mol Cell Biol*; 24: 4004-4018.
357. Ikeda R, Kokubu C, Yusa K, Keng VW, Horie K, Takeda J (2007). Sleeping beauty transposase has an affinity for heterochromatin conformation. *Mol Cell Biol*; 27: 1665-1676.
358. Spradling AC, Stern DM, Kiss I, Roote J, Laverty T, Rubin GM (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. *Proc Natl Acad Sci U S A*; 92: 10824-10830.
359. Gorin AA, Zhurkin VB, Olson WK (1995). B-DNA twisting correlates with base-pair morphology. *J Mol Biol*; 247: 34-48.
360. Ivanov VI, Minchenkova LE (1995). The A-form of DNA: in search of the biological role. *Mol Biol*; 28: 1258-1271.
361. Brukner I, Sanchez R, Suck D, Pongor S (1995). Sequence-dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *EMBO J*; 14: 1812-1818.
362. Olson WK, Gorin AA, Lu XJ, Hock LM, Zhurkin VB (1998). DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc Natl Acad Sci USA*; 95: 11163-11168.
363. Haapa-Paananen S, Rita H, Savilahti H (2002). DNA transposition of bacteriophage Mu. A quantitative analysis of target site selection in vitro. *J Biol Chem*; 277: 2843-2851.
364. Hallet B, Rezsohazy R, Mahillon J, Delcour J (1994). IS231A insertion specificity: consensus sequence and DNA bending at the target site. *Mol Microbiol*; 14: 131-139.
365. Lee GS, Neiditch MB, Sinden RR, Roth DB (2002). Targeted transposition by the V(D)J recombinase. *Mol Cell Biol*; 22: 2068-2077.
366. Nakai H, Montini E, Fuess S, Storm TA, Grompe M, Kay MA (2003). AAV serotype 2 vectors preferentially integrate into active genes in mice. *Nat Genet*; 34: 297-302.
367. Yant SR, Wu X, Huang Y, Garrison B, Burgess SM, Kay MA (2005). High-resolution genome-wide mapping of transposon integration in mammals. *Mol Cell Biol*; 25: 2085-2094.
368. Vinogradov AE (1998). Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry*; 31: 100-109.
369. Lam WL, Lee TS, Gilbert W (1996). Active transposition in zebrafish. *Proc Natl Acad Sci USA*; 93: 10870-10875.
370. Sundararajan P, Atkinson PW, O'Brochta DA (1999). Transposable element interactions in insects: crossmobilization of hobo and Hermes. *Insect Mol Biol*; 8: 359-368.
371. Arca B, Zabalou S, Loukeris TG, Savakis C (1997). Mobilization of a Minos transposon in Drosophila melanogaster chromosomes and chromatid repair by heteroduplex formation. *Genetics*; 145: 267-279.
372. Maruyama IN, Rakow TL, Maruyama HI (1995). cRACE: a simple method for identification of the 5' end of mRNAs. *Nucleic Acids Res*; 23: 3796-3797.
373. Bronchain OJ, Hartley KO, Amaya E (1999). A gene trap approach in Xenopus. *Curr Biol*; 9: 1195-1198.
374. Simmons MJ, Bucholz LM (1985). Transposase titration in Drosophila melanogaster: a model of cytotype in the P-M system of hybrid disgenesis. *Proc Natl Acad Sci USA*; 82: 8119-8123.
375. Klinakis AG, Zagoraiou L, Vassilatis DK, Savakis C (2000). Genome-wide insertional mutagenesis in human cells by the Drosophila mobile element Minos. *EMBO Rep*; 1: 416-421.
376. Vigdal TJ, Kaufman CD, Izsvak Z, Voytas DF, Ivics Z (2002). Common physical properties of DNA affecting target site selection of sleeping beauty and other Tc1/mariner transposable elements. *J Mol Biol*; 323: 441-452.
377. Hacker U, Nystedt S, Barmchi MP, Horn C, Wimmer EA (2003). piggyBac-based insertional mutagenesis in the presence of stably integrated P elements in Drosophila. *Proc Natl Acad Sci U S A*; 100: 7720-7725.
378. Lohe AR, Hartl DL (2002). Efficient mobilization of mariner in vivo requires multiple internal sequences. *Genetics*; 160: 519-526.
379. Dematteï MV, Auge-Gouillou C, Pollet N, Hamelin MH, Meunier-Rotival M, Bigot Y (2000). Features of the mammal mar1 transposons in the human, sheep, cow, and mouse genomes and implications for their evolution. *Mamm Genome*; 11: 1111-1116.
380. Robertson HM, Martos R (1997). Molecular evolution of the second ancient human mariner transposon, Hsmar2, illustrates patterns of neutral evolution in the human genome lineage. *Gene*; 205: 219-228.
381. Feschotte C, Jiang N, Wessler SR (2002). Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*; 3: 329-341.
382. Miskey C, Izsvak Z, Plasterk RH, Ivics Z (2003). The Frog Prince: a reconstructed transposon from Rana pipiens with high transpositional activity in vertebrate cells. *Nucleic Acids Res*; 31: 6873-6881.

383. Thornton JW (2004). Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*; 5: 366-375.
384. Lampe DJ, Walden KK, Sherwood JM, Roberstson HM (2000) Genetic engineering of insects with mariner transposons. In: Handler AM, James AA, editors. Insect Transgenesis: methods and applications: CRC Press. pp. 237-248.
385. Bryan G, Garza D, Hartl D (1990). Insertion and excision of the transposable element *mariner* in *Drosophila*. *Genetics*; 125: 103-114.
386. Dufresne M, Hua-Van A, El Wahab HA, Ben M'Barek S, Vasnier C, Teysset L, et al. (2007). Transposition of a fungal miniature inverted-repeat transposable element through the action of a Tc1-like transposase. *Genetics*; 175: 441-452.
387. Feschotte C, Mouches C (2000). Evidence that a family of miniature inverted-repeat transposable elements (MITEs) from the Arabidopsis thaliana genome has arisen from a pogo-like DNA transposon. *Mol Biol Evol*; 17: 730-737.
388. Izsvak Z, Ivics Z, Shimoda N, Mohn D, Okamoto H, Hackett PB (1999). Short inverted-repeat transposable elements in teleost fish and implications for a mechanism of their amplification. *J Mol Evol*; 48: 13-21.
389. Lee SH, Oshige M, Durant ST, Rasila KK, Williamson EA, Ramsey H, et al. (2005). The SET domain protein Metnase mediates foreign DNA integration and links integration to nonhomologous end-joining repair. *Proc Natl Acad Sci U S A*; 102: 18075-18080.
390. Jenuwein T, Laible G, Dorn R, Reuter G (1998). SET domain proteins modulate chromatin domains in eu- and heterochromatin. *Cell Mol Life Sci*; 54: 80-93.
391. Cordaux R, Udit S, Batzer MA, Feschotte C (2006). Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci U S A*; 103: 8101-8106.
392. Liu D, Bischerour J, Siddique A, Buisine N, Bigot Y, Chalmers R (2007). The human SETMAR protein preserves most of the activities of the ancestral Hsmar1 transposase. *Mol Cell Biol*; 27: 1125-1132.
393. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. (2003). An active DNA transposon family in rice. *Nature*; 421: 163-167.
394. Feschotte C, Osterlund MT, Peeler R, Wessler SR (2005). DNA-binding specificity of rice mariner-like transposases and interactions with Stowaway MITEs. *Nucleic Acids Res*; 33: 2153-2165.
395. Zhang X, Feschotte C, Zhang Q, Jiang N, Eggleston WB, Wessler SR (2001). P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci U S A*; 98: 12572-12577.
396. Loot C, Santiago N, Sanz A, Casacuberta JM (2006). The proteins encoded by the pogo-like Lemi1 element bind the TIRs and subterminal repeated motifs of the Arabidopsis Emigrant MITE: consequences for the transposition mechanism of MITEs. *Nucleic Acids Res*; 34: 5238-5246.
397. Gloor GB, Moretti J, Mouyal J, Keeler KJ (2000). Distinct P-element excision products in somatic and germline cells of Drosophila melanogaster. *Genetics*; 155: 1821-1830.
398. Plasterk RH (1991). The origin of footprints of the Tc1 transposon of *Caenorhabditis elegans*. *EMBO J*; 10: 1919-1925.
399. Rubin E, Levy AA (1997). Abortive gap repair: underlying mechanism for Ds element formation. *Mol Cell Biol*; 17: 6294-6302.
400. Lee GS, Neiditch MB, Salus SS, Roth DB (2004). RAG proteins shepherd double-strand breaks to a specific pathway, suppressing error-prone repair, but RAG nicking initiates homologous recombination. *Cell*; 117: 171-184.
401. Kapitonov VV, Jurka J (1999). Molecular paleontology of transposable elements from Arabidopsis thaliana. *Genetica*; 107: 27-37.
402. Jurka J, Kapitonov VV (2001). PIFs meet Tourists and Harbingers: a superfamily reunion. *Proc Natl Acad Sci U S A*; 98: 12315-12316.
403. Zhang X, Jiang N, Feschotte C, Wessler SR (2004). PIF- and Pong-like transposable elements: distribution, evolution and relationship with Tourist-like miniature inverted-repeat transposable elements. *Genetics*; 166: 971-986.
404. Grzebelus D, Yau YY, Simon PW (2006). Master: a novel family of PIF/Harbinger-like transposable elements identified in carrot (Daucus carota L.). *Mol Genet Genomics*; 275: 450-459.
405. Casola C, Lawing AM, Betran E, Feschotte C (2007). PIF-like Transposons Are Common in Drosophila and Have Been Repeatedly Domesticated to Generate New Host Genes. *Mol Biol Evol*.
406. Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, et al. (2003). Mobilization of a transposon in the rice genome. *Nature*; 421: 170-172.

407. Kapitonov VV, Jurka J (2004). Harbinger transposons and an ancient HARBI1 gene derived from a transposase. *DNA Cell Biol*; 23: 311-324.
408. Boyer LA, Latek RR, Peterson CL (2004). The SANT domain: a unique histone-tail-binding module? *Nat Rev Mol Cell Biol*; 5: 158-163.
409. Yang G, Zhang F, Hancock CN, Wessler SR (2007). Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*; 104: 10962-10967.
410. Lv B, Shi T, Wang X, Song Q, Zhang Y, Shen Y, et al. (2006). Overexpression of the novel human gene, nuclear apoptosis-inducing factor 1, induces apoptosis. *Int J Biochem Cell Biol*; 38: 671-683.
411. Ivics Z, Katzer A, Stuwe EE, Fiedler D, Knespel S, Izsvak Z (2007). Targeted Sleeping Beauty transposition in human cells. *Mol Ther*; 15: 1137-1144.
412. Wistuba A, Kern A, Weger S, Grimm D, Kleinschmidt JA (1997). Subcellular compartmentalization of adeno-associated virus type 2 assembly. *J Virol*; 71: 1341-1352.
413. Frey M, Reinecke J, Grant S, Saedler H, Gierl A (1990). Excision of the En/Spm transposable element of Zea mays requires two element-encoded proteins. *Embo J*; 9: 4037-4044.
414. Geurts AM, Yang Y, Clark KJ, Liu G, Cui Z, Dupuy AJ, et al. (2003). Gene transfer into genomes of human cells by the sleeping beauty transposon system. *Mol Ther*; 8: 108-117.
415. Karsi A, Moav B, Hackett P, Liu Z (2001). Effects of insert size on transposition efficiency of the *Sleeping Beauty* transposon in mouse cells. *Mar Biotechnol*; 3: 241-245.
416. Petrov DA, Schutzman JL, Hartl DL, Lozovskaya ER (1995). Diverse transposable elements are mobilized in hybrid dysgenesis in *Drosophila virilis*. *Proc Natl Acad Sci USA*; 92: 8050-8054.
417. Vos JC, De Baere I, Plasterk RH (1996). Transposase is the only nematode protein required for *in vitro* transposition of Tc1. *Genes Dev*; 10: 755-761.
418. Berg DE, Berg CM, Sasakawa C, Zhou M, Reznikoff WS (1984). Bacterial transposon Tn5: evolutionary inferences. *Mol Biol Evol*; 1: 411-422.
419. Lozovsky ER, Nurminsky D, Wimmer EA, Hartl DL (2002). Unexpected stability of mariner transgenes in Drosophila. *Genetics*; 160: 527-535.
420. Zhou M, Reznikoff WS (1997). Tn5 transposase mutants that alter DNA binding specificity. *J Mol Biol*; 271: 362-373.
421. Wiegand TW, Reznikoff WS (1992). Characterization of two hypertransposing Tn5 mutants. *J Bacteriol*; 174: 1229-1239.
422. Sakai J, Kleckner N (1996). Two classes of Tn10 transposase mutants that suppress mutations in the Tn10 terminal inverted repeat. *Genetics*; 144: 861-870.
423. Lampe DJ, Akerley BJ, Rubin EJ, Mekalanos JJ, Robertson HM (1999). Hyperactive transposase mutants of the Himar1 mariner transposon. *Proc Natl Acad Sci USA*; 96: 11428-11433.
424. Goryshin IY, Reznikoff WS (1998). Tn5 *in vitro* transposition. *J Biol Chem*; 273: 7367-7374.
425. Baus J, Liu L, Heggestad AD, Sanz S, Fletcher BS (2005). Hyperactive transposase mutants of the Sleeping Beauty transposon. *Mol Ther*; 12: 1148-1156.
426. Yant SR, Park J, Huang Y, Mikkelsen JG, Kay MA (2004). Mutational analysis of the N-terminal DNA-binding domain of sleeping beauty transposase: critical residues for DNA binding and hyperactivity in mammalian cells. *Mol Cell Biol*; 24: 9239-9247.
427. Mates L, Chuah MK, Belay E, Jerchow B, Manoj N, Acosta-Sanchez A, et al. (2009). Molecular evolution of a novel hyperactive Sleeping Beauty transposase enables robust stable gene transfer in vertebrates. *Nat Genet*; 41: 753-761.
428. Fire A (1999). RNA-triggered gene silencing. *Trends Genet*; 15: 358-363.
429. Gil J, Esteban M (2000). Induction of apoptosis by the dsRNA-dependent protein kinase (PKR): mechanism of action. *Apoptosis*; 5: 107-114.
430. Caplen NJ, Parrish S, Imani F, Fire A, Morgan RA (2001). Specific inhibition of gene expression by small double-stranded RNAs in invertebrate and vertebrate systems. *Proc Natl Acad Sci U S A*; 98: 9742-9747.
431. Elbashir SM, Harborth J, Lendeckel W, Yalcin A, Weber K, Tuschl T (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*; 411: 494-498.
432. Miyagishi M, Taira K (2002). Development and application of siRNA expression vector. *Nucleic Acids Res Suppl*: 113-114.
433. Tran N, Cairns MJ, Dawes IW, Arndt GM (2003). Expressing functional siRNAs in mammalian cells using convergent transcription. *BMC Biotechnol*; 3: 21.
434. Brummelkamp TR, Bernards R, Agami R (2002). A system for stable expression of short interfering RNAs in mammalian cells. *Science*; 296: 550-553.
435. Paul CP, Good PD, Winer I, Engelke DR (2002). Effective expression of small interfering RNA in human cells. *Nat Biotechnol*; 20: 505-508.

436. Yu JY, DeRuiter SL, Turner DL (2002). RNA interference by expression of short-interfering RNAs and hairpin RNAs in mammalian cells. *Proc Natl Acad Sci U S A*; 99: 6047-6052.
437. Bishop JO (1996). Chromosomal insertion of foreign DNA. *Reprod Nutr Dev*; 36: 607-618.
438. Abbas-Terki T, Blanco-Bose W, Deglon N, Pralong W, Aebischer P (2002). Lentiviral-mediated RNA interference. *Hum Gene Ther*; 13: 2197-2201.
439. Devroe E, Silver PA (2002). Retrovirus-delivered siRNA. *BMC Biotechnol*; 2: 15.
440. Caplen NJ (2004). Gene therapy progress and prospects. Downregulating gene expression: the impact of RNA interference. *Gene Ther*; 11: 1241-1248.
441. Kawakami K (2007). Tol2: a versatile gene transfer vector in vertebrates. *Genome Biol*; 8 Suppl 1: S7.
442. Fraser MJ, Brusca JS, Smith GE, Summers MD (1985). Transposon-mediated mutagenesis of a baculovirus. *Virology*; 145: 356-361.
443. Wang W, Lin C, Lu D, Ning Z, Cox T, Melvin D, et al. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*; 105: 9290-9295.
444. Wilson MH, Coates CJ, George AL, Jr. (2007). PiggyBac Transposon-mediated Gene Transfer in Human Cells. *Mol Ther*; 15: 139-145.
445. Woltjen K, Michael IP, Mohseni P, Desai R, Mileikovsky M, Hamalainen R, et al. (2009). piggyBac transposition reprograms fibroblasts to induced pluripotent stem cells. *Nature*; 458: 766-770.
446. Yusa K, Rad R, Takeda J, Bradley A (2009). Generation of transgene-free induced pluripotent mouse stem cells by the piggyBac transposon. *Nat Methods*; 6: 363-369.
447. Liang Q, Kong J, Stalker J, Bradley A (2009). Chromosomal mobilization and reintegration of Sleeping Beauty and PiggyBac transposons. *Genesis*; 47: 404-408.
448. Cadinanos J, Bradley A (2007). Generation of an inducible and optimized piggyBac transposon system. *Nucleic Acids Res*; 35: e87.
449. Urasaki A, Morvan G, Kawakami K (2006). Functional dissection of the Tol2 transposable element identified the minimal cis-sequence and a highly repetitive sequence in the subterminal region essential for transposition. *Genetics*; 174: 639-649.
450. Li X, Lobo N, Bauser CA, Fraser MJ, Jr. (2001). The minimum internal and external sequence requirements for transposition of the eukaryotic transformation vector piggyBac. *Mol Genet Genomics*; 266: 190-198.
451. Kunze R, Behrens U, Courage-Franzkowiak U, Feldmar S, Kuhn S, Lutticke R (1993). Dominant transposition-deficient mutants of maize Activator (Ac) transposase. *Proc Natl Acad Sci U S A*; 90: 7094-7098.
452. Xue X, Huang X, Nodland SE, Mates L, Ma L, Izsvak Z, et al. (2009). Stable gene transfer and expression in cord blood-derived CD34+ hematopoietic stem and progenitor cells by a hyperactive Sleeping Beauty transposon system. *Blood*; 114: 1319-1330.
453. Garrison BS, Yant SR, Mikkelsen JG, Kay MA (2007). Postintegrative gene silencing within the Sleeping Beauty transposition system. *Mol Cell Biol*; 27: 8824-8833.
454. Galvan DL, Nakazawa Y, Kaja A, Kettlun C, Cooper LJ, Rooney CM, et al. (2009). Genome-wide mapping of PiggyBac transposon integrations in primary human T cells. *J Immunother*; 32: 837-844.
455. Cuddapah S, Jothi R, Schones DE, Roh TY, Cui K, Zhao K (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res*; 19: 24-32.
456. Lohe AR, Hartl DL (1996). Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. *Mol Biol Evol*; 13: 549-555.
457. Wu SC, Meir YJ, Coates CJ, Handler AM, Pelczar P, Moisyadi S, et al. (2006). piggyBac is a flexible and highly active transposon as compared to Sleeping Beauty, Tol2, and Mos1 in mammalian cells. *Proc Natl Acad Sci U S A*; 103: 15008-15013.
458. Mikkelsen JG, Yant SR, Meuse L, Huang Z, Xu H, Kay MA (2003). Helper-Independent Sleeping Beauty transposon-transposase vectors for efficient nonviral gene delivery and persistent gene expression in vivo. *Mol Ther*; 8: 654-665.
459. Parinov S, Kondrichin I, Korzh V, Emelyanov A (2004). Tol2 transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes in vivo. *Dev Dyn*; 231: 449-459.
460. Kondrychyn I, Garcia-Lecea M, Emelyanov A, Parinov S, Korzh V (2009). Genome-wide analysis of Tol2 transposon reintegration in zebrafish. *BMC Genomics*; 10: 418.
461. Wang W, Bradley A, Huang Y (2009). A piggyBac transposon-based genome-wide library of insertionally mutated Blm-deficient murine ES cells. *Genome Res*; 19: 667-673.
462. Geurts AM, Hackett CS, Bell JB, Bergemann TL, Collier LS, Carlson CM, et al. (2006). Structure-based prediction of insertion-site preferences of transposons into chromosomes. *Nucleic Acids Res*; 34: 2803-2811.
463. Hackett CS, Geurts AM, Hackett PB (2007). Predicting preferential DNA vector insertion sites: implications for functional genomics and gene therapy. *Genome Biol*; 8 Suppl 1: S12.

464. Jahner D, Stuhlmann H, Stewart CL, Harbers K, Lohler J, Simon I, et al. (1982). De novo methylation and expression of retroviral genomes during mouse embryogenesis. *Nature*; 298: 623-628.
465. Wolf D, Goff SP (2009). Embryonic stem cells use ZFP809 to silence retroviral DNAs. *Nature*; 458: 1201-1204.
466. Aronovich EL, Bell JB, Khan SA, Belur LR, Gunther R, Koniar B, et al. (2009). Systemic correction of storage disease in MPS I NOD/SCID mice using the sleeping beauty transposon system. *Mol Ther*; 17: 1136-1144.
467. Recillas-Targa F, Pikaart MJ, Burgess-Beusse B, Bell AC, Litt MD, West AG, et al. (2002). Position-effect protection and enhancer blocking by the chicken beta-globin insulator are separable activities. *Proc Natl Acad Sci U S A*; 99: 6883-6888.
468. Bell AC, West AG, Felsenfeld G (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell*; 98: 387-396.
469. Emery DW, Yannaki E, Tubb J, Stamatoyannopoulos G (2000). A chromatin insulator protects retrovirus vectors from chromosomal position effects. *Proc Natl Acad Sci U S A*; 97: 9150-9155.
470. Emery DW, Yannaki E, Tubb J, Nishino T, Li Q, Stamatoyannopoulos G (2002). Development of virus vectors for gene therapy of beta chain hemoglobinopathies: flanking with a chromatin insulator reduces gamma-globin gene silencing in vivo. *Blood*; 100: 2012-2019.
471. Malik P, Arumugam PI, Yee JK, Puthenveetil G (2005). Successful correction of the human Cooley's anemia beta-thalassemia major phenotype using a lentiviral vector flanked by the chicken hypersensitive site 4 chromatin insulator. *Ann N Y Acad Sci*; 1054: 238-249.
472. Ramezani A, Hawley TS, Hawley RG (2003). Performance- and safety-enhanced lentiviral vectors containing the human interferon-beta scaffold attachment region and the chicken beta-globin insulator. *Blood*; 101: 4717-4724.
473. Inoue T, Yamaza H, Sakai Y, Mizuno S, Ohno M, Hamasaki N, et al. (1999). Position-independent human beta-globin gene expression mediated by a recombinant adeno-associated virus vector carrying the chicken beta-globin insulator. *J Hum Genet*; 44: 152-162.
474. Evans-Galea MV, Wielgosz MM, Hanawa H, Srivastava DK, Nienhuis AW (2007). Suppression of clonal dominance in cultured human lymphoid cells by addition of the cHS4 insulator to a lentiviral vector. *Mol Ther*; 15: 801-809.
475. Szabo M, Muller F, Kiss J, Balduf C, Strahle U, Olasz F (2003). Transposition and targeting of the prokaryotic mobile element IS30 in zebrafish. *FEBS Lett*; 550: 46-50.
476. Maragathavally KJ, Kaminski JM, Coates CJ (2006). Chimeric Mos1 and piggyBac transposases result in site-directed integration. *Faseb J*; 20: 1880-1882.
477. Yant SR, Huang Y, Akache B, Kay MA (2007). Site-directed transposon integration in human cells. *Nucleic Acids Res*; 35: e50.
478. Hama C, Ali Z, Kornberg TB (1990). Region-specific recombination and expression are directed by portions of the Drosophila engrailed promoter. *Genes Dev*; 4: 1079-1093.
479. Kassis JA, Noll E, VanSickle EP, Odenwald WF, Perrimon N (1992). Altering the insertional specificity of a Drosophila transposable element. *Proc Natl Acad Sci U S A*; 89: 1919-1923.
480. Ciuffi A, Diamond TL, Hwang Y, Marshall HM, Bushman FD (2006). Modulating Target Site Selection During Human Immunodeficiency Virus DNA Integration In Vitro with an Engineered Tethering Factor. *Hum Gene Ther*; 17: 960-967.
481. Zhu Y, Dai J, Fuerst PG, Voytas DF (2003). Controlling integration specificity of a yeast retrotransposon. *Proc Natl Acad Sci U S A*; 100: 5891-5895.
482. Mandell JG, Barbas CF, 3rd (2006). Zinc Finger Tools: custom DNA-binding domains for transcription factors and nucleases. *Nucleic Acids Res*; 34: W516-523.