dc_1908_21

# Evolutionary genetics in the era of genome-scale modelling

Balázs Papp

Synthetic and Systems Biology Unit

Institute of Biochemistry

Biological Research Center, Eötvös Loránd Kutatási Hálózat

Szeged

2021

## Table of Contents

## I. Introduction

Many outstanding questions in evolutionary biology depend on the mutational effects that govern the complex relationship between genotype and phenotype. For example, why are most inactivating mutations have little phenotypic effects? And how do multiple mutations interact with each other to produce a novel phenotype? Resolving such issues requires an understanding of how genotypes map onto phenotypes. While genotype-phenotype maps have long been investigated at the level of individual proteins (Dean and Thornton 2007), analyses of larger gene networks lag behind. Developing quantitative frameworks to interrogate mutational effects in large cellular networks would be important for at least two fundamental reasons. First, such a framework would provide mechanistic insights into complex evolutionary phenomena, from the emergence of evolutionary novelties that hinge on multiple mutations to the evolution of minimized genomes. Second, it would transform evolutionary biology into a more predictive discipline by allowing specific predictions on the outcome of evolution. A predictive framework would give us a clue which genes are likely to be lost, to mutate, or change expression during evolution. Beyond catalyzing basic research, such a framework would also have practical relevance. Among others, it would allow forecasting evolutionary changes in pathogenic microbes (Sommer, et al. 2017) and inform the engineering of novel biosynthetic pathways (Johannes and Zhao 2006; Notebaart, et al. 2018).

Recent advances in systems biology provide an unprecedented opportunity to build computational models that map from mutations and environmental changes to phenotypes. These computational approaches rely on mathematical models of specific molecular systems and come in different flavors. These models range from detailed kinetic models of smaller metabolic systems (Teusink,

et al. 1998) and regulatory circuits (Chen, et al. 2004) to constraint-based models of genome-scale metabolic networks (Price, et al. 2004). Constraint-based metabolic models are especially appealing for studying the relationship between genotypes and phenotypes in large networks and have already provided valuable insights into the evolution of metabolic gene contents and phenotypes of microbial species (Feist and Palsson 2008). These models start from high-quality metabolic network reconstructions (Price, et al. 2004). These reconstructions are typically built through integrating genome annotation data, information from enzyme databases, such as KEGG (Kanehisa and Goto 2000) and BRENDA (Schomburg, et al. 2002)) and the primary literature and involve extensive manual curation. The network of biochemical reactions is then converted into a mathematical representation and analyzed using constrained-based methods (Box 1). In particular, a widely used method termed flux balance analysis (FBA) calculates the optimal flow of metabolites through the network as a function of available nutrients in the environment. These predictions have been extensively tested and showed high agreement with empirical data (Edwards, et al. 2001; Snitkin, et al. 2008; Oberhardt, et al. 2009).

Constraint-based models have important conceptual advantages over both small-scale biochemical models and graph-theoretical approaches that make them especially well-suited to interrogate the genotype-phenotype map. First, they can be applied on a genomic scale. As these models require only few empirical parameters beyond the structure of the metabolic network, they can capture the behavior of large metabolic systems that encompass all enzyme encoding genes of an organisms, that is, hundreds to few thousands of genes. As a consequence, these models allow comparison with results of high-throughput omics data and also allow incorporation of multiple modalities of omics data (Yizhak, et al. 2010; Lloyd, et al. 2018). Second, unlike graph-theoretical approaches,

constraint-based models are based on sound biochemical principles and can compute functional states of the network while explicitly taking into account the nutrient environment (Box 1). However, I wish to emphasize that the constraint-based analyses framework also suffers from several important limitations owing to the lack of enzyme kinetic information. For example, investigating the phenotypic effect of minor changes in enzyme activity, as opposed to complete loss or gain of enzymes, and predicting metabolite concentrations remain a formidable challenge.

This thesis focuses on four major research topics, each of which employs genome-scale metabolic networks to address a long-standing issue in evolutionary genetics. First, how do mutations modulate each other's phenotypic effects, that is, how do genetic interactions arise at the mechanistic level? And how accurately can we computationally predict which gene pairs show a genetic interaction based on a detailed knowledge of the metabolic network? Second, can we predict the gene content of endosymbiotic bacteria that have highly reduced genomes? That is, can we predict which genes are lost and which are kept during millions of years of reductive genome evolution? Third, can we computationally predict the outcome and genetic basis of adaptation to new environments? More specifically, how does the ability to utilize new nutrients arise from existing low-level enzymatic side activities? Finally, how do evolutionary novelties arise that demand the simultaneous acquisition of multiple mutations?

**Box 1. Constraint-based analysis of metabolic networks**

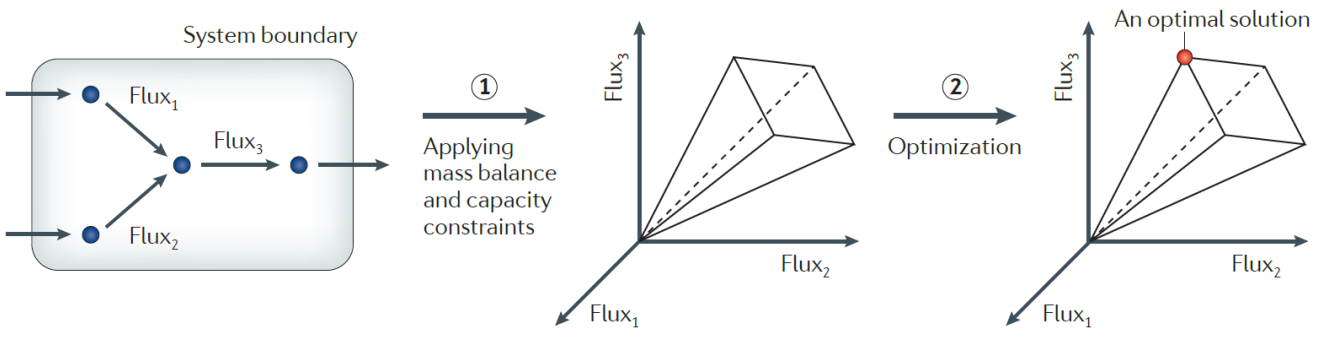



*Figure reproduced from (Papp, et al. 2011).*

The functional properties of genome-scale metabolic networks are generally studied using constraint-based methods (Price, et al. 2004). Such methods apply physicochemical and biological constraints to define the range of achievable functional states (flux states) of the network, without relying on enzyme kinetic information. There are two fundamental types of constraints: (i) balance constraints, such as the conservation of mass, that is, at steady-state there is no accumulation or depletion of internal metabolites, and (ii) capacity constraints, that is, bounds that constrain the values of individual fluxes. For example, the rates of irreversible reactions must have a minimum value of zero. The nutrient environment is set up by allowing certain metabolites to enter the system through applying capacity constraints. Together, the applied constraints limit the allowable functional states of the network and define a solution space, which is typically a polytope in a high-dimensional space (step 1 in the figure). A widely used strategy, termed flux balance analysis (FBA), can then be employed to identify steady state flux states of the network that maximize a particular network function (step 2). This optimization step is generally achieved using linear programming and serve three main purposes (Price, et al. 2004): (i) exploration of the biochemical potential of the network, such as the maximum yield of producing metabolites; (ii) identification of likely physiological states based on the assumption that microbial cells have evolved towards maximum growth efficiency. This is done by representing growth as a pseudo-reaction in which all biomass compounds required for growth are drained from the network; (iii) rational design of networks that improve the production of desired compounds.

## II. Genetic interactions in metabolic networks

*Key papers: (Szappanos, et al. 2011), (Harrison, et al. 2007) (see Appendix)*

The phenotypic effect of a mutation often depends on the presence of other mutations in the genome, a phenomenon termed genetic interaction or epistatic interaction. Genetic interactions are the key to understand the functional relationships between genes, the extent to which organisms tolerate deleterious mutations, as well as the underpinnings of complex genetic diseases. In the past decade, high-throughput studies have generated comprehensive maps of genetic interactions between genes in several organisms, including budding yeast (*Saccharomyces cerevisiae*) (Costanzo, et al. 2010; Costanzo, et al. 2016), *E. coli* (Babu, et al. 2014) and human cell lines (Horlbeck, et al. 2018). These works focused on loss-of-function mutations and revealed two main forms of genetic interactions: (i) *negative genetic interactions* (synthetic sick or lethal / aggravating) when two mutations enhance each other's harmful effects, potentially indicating functional compensation between them, and (ii) *positive genetic interactions* (antagonistic / diminishing) when a mutation has a smaller than expected deleterious effect in the presence of another deleterious mutation. However, despite the rapid accumulation of experimental data on genetic interactions, several questions remain open about the organization and mechanistic underpinnings of epistasis. In the past years, I contributed to three outstanding issues (Harrison, et al. 2007; Szappanos, et al. 2011):

1) The first high-throughput genetic interaction screens in yeast provided the first glimpse into the overall organizational principles of genetic interaction networks (Costanzo, et al. 2010). A major finding of these studies was that the vast majority of genes show few genetic interactions, while a small number of 'hub' genes are highly connected in the

genetic interaction network. Note that this pattern is consistent with those in other biological networks (such as protein-protein interaction networks), yet the underlying mechanisms might be entirely different (Barabasi and Oltvai 2004). Importantly, hub genes in genetic interaction network tend to display severe fitness defect when deleted and are highly pleiotropic (i.e. affect multiple cellular processes)(Costanzo, et al. 2010), but the underlying mechanisms has remained unknown.

2) Genetic interactions between genes are highly specific, with only ~3% of tested gene pairs showing an experimentally detectable interaction (Costanzo, et al. 2010). Is it possible to computationally predict which specific gene pair would show a genetic interaction based on our knowledge of the biochemical circuits in which they participate? This would be useful not only to accurately predict genetic interactions on a large scale, but also to understand the links between genetic and molecular interaction networks. Crucially, reconciling discrepancies between empiricial and predicted genetic interactions would allow us to refine the metabolic model and generate new biological hypotheses.
3) Several lines of evidences indicate that genetic interactions themselves might often be environment dependent (You and Yin 2002; Remold and Lenski 2004; Bandyopadhyay, et al. 2011). However, we poorly understand the mechanisms of this phenomenon and how it contributes to the apparent phenotypic silence of many gene deletions.

Addressing these issues requires computational systems biology models, which allow studying how these genetic phenomena emerge from the molecular interactions of hundreds of proteins. Genome-scale flux balance analysis models of cellular metabolism allow researchers to calculate the phenotypic effect of gene deletions and provides mechanistic insights into why most genes

appear to be phenotypically silent (Papp, et al. 2004). Therefore, to tackle the above questions, we applied an integrated systems biology approach by constructing a large-scale empirical genetic interaction map of yeast metabolism and integrating the data with a genome-scale metabolic network model. Our analyses yielded four major insights, discussed in turn below (Harrison, et al. 2007; Szappanos, et al. 2011).

*Organization principles of genetic interaction networks*

First, the computational model successfully captured the high genetic interaction connectivity, for both positive and negative interactions, of genes that have a large contribution to fitness (Figure 1). Importantly, the genome-scale model incorporates information only on the stoichiometry of biochemical reactions and growth requirements of the cell without explicitly accounting for gene regulation and enzyme kinetic details. Therefore, this result suggests that genetic interaction hubs emerge from the structure of metabolic networks. Modelling also offered a mechanistic explanation for the existence of hubs in the genetic interaction network. Hubs are driven by pleiotropic enzymes that participate in multiple biological processes by contributing to the biosynthesis of multiple key biomass precursors. As a result, the phenotypic impact of their loss can potentially be shaped by several other enzymes in the network, yielding numerous epistatic interactions. Clearly, further empirical works on enzyme pleiotropy are needed to test how well this hypothesis explains experimentally observed genetic interaction hubs.

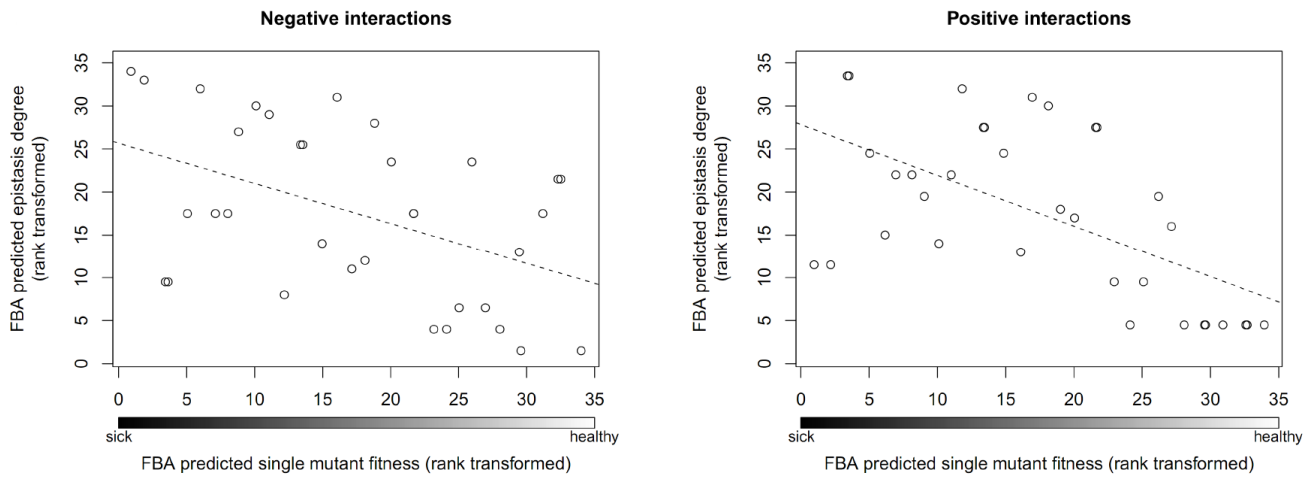



***Figure 1. Genes with large fitness contribution show many genetic interactions in the computational model.*** *Predicted single mutant fitness correlates well with both the number of predicted negative and positive genetic interactions (Spearman's rho=-0.59, P<$10^{-3}$ and rho=-0.47, P=0.005, respectively). Only genes with nonzero predicted fitness defects are shown. Variables are rank transformed. Figure reproduced from (Szappanos, et al. 2011).*

*Computational prediction of individual genetic interactions*

Second, by comparing genetic interaction data with a genome-scale model of metabolism, we provided the first large-scale assessment of our ability to predict individual genetic interactions using genome-scale metabolic models. Analysis of high-confidence experimental data across ~67,500 metabolic gene pairs uncovered a strong enrichment of *in vivo* interactions among computationally predicted ones (100-fold and 60-fold enrichment for negative and positive genetic interactions, respectively, corresponding to precision values of 0.5 and 0.11, respectively; see Figure 2). Thus, gene pairs that show a strong epistasis in the model are highly likely to also show an interaction in the experiment. This is rather remarkable as the modelling framework is simple and does not rely on detailed enzyme kinetic or regulatory information. However, the metabolic

model fails to capture the majority of *in vivo* detected genetic interactions (97% and 89% of the negative and positive interactions, respectively; see Figure 2). Overall, it appears that this low success rate comes from overestimating the fitness of double mutants, possibly because many *in vivo* observed genetic interactions arise from regulatory effects that are not captured by the structure of the metabolic network. We anticipate that more sophisticated models that take into account allosteric regulations and account for widespread regulatory / signaling interactions with non-metabolic genes (Mulleder, et al. 2016) will be needed to more accurately capture the metabolic behavior of mutant cells.

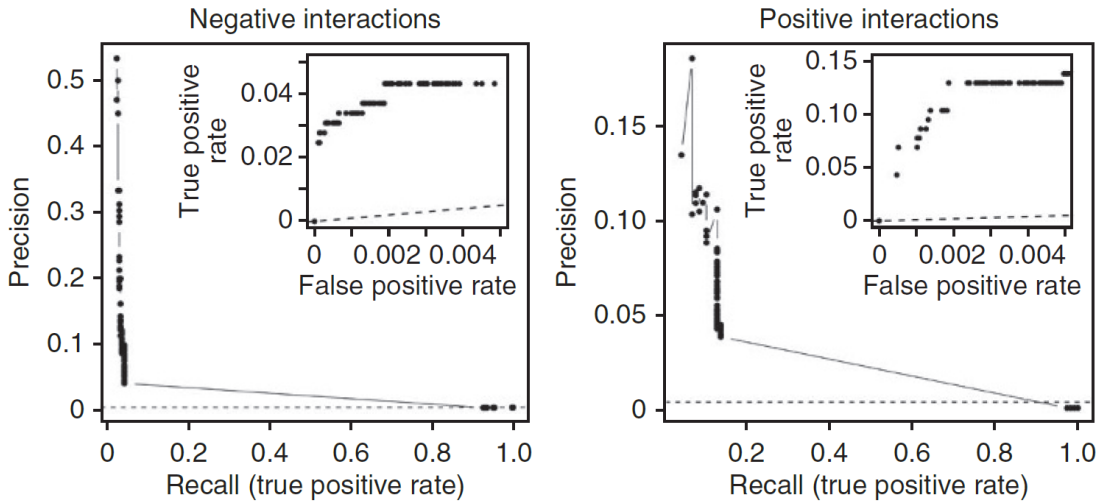



***Figure 2. Accuracy of predicting genetic interactions using a genome-scale metabolic network model in yeast.*** *We visualized prediction accuracy by plotting precision (fraction of predicted interactions that can be verified by experimental data) against recall (fraction of experimentally observed interactions that are successfully identified by the model) at different predicted genetic interaction score cutoffs. In addition, figure insets show ROC curves. Analysis is based on 325 in vivo negative and 116 in vivo positive interactions among 67,517 gene pairs for which high-confidence experimental data was available. Figure reproduced from (Szappanos, et al. 2011).*

*Automated refinement of the metabolic model based on genetic interaction data*

Can we make use of the discrepancies between empirical data and model predictions to refine the metabolic model itself? In principle, the large number of experimentally observed genetic interactions offers a rich source of information to modify the model in a data-driven way. To this end, we developed a machine learning method that automatically suggests modifications to the model that improve its ability to predict negative genetic interactions (Figure 3A). Allowed modifications included changes in reaction reversibility, removing reactions and modifying the set of biomass compounds deemed essential for growth. The method employs a genetic algorithm to minimize false predictions in a two-stage process (Figure 3A). Importantly, we minimized model false predictions globally (i.e. not on a mutant by mutant basis) and using experimental growth data on both single and double mutants. Overall, the method proposed several modifications that together improved the fit of the model to the data (i.e. 100–267% increase in recall and 44–59% increase in precision; see Figure 3B). Among the suggested modifications, we found the removal of the *de novo* NAD biosynthesis pathway starting from aspartate. This pathway is present in *E. coli* (Flachmann, et al. 1988), but was probably erroneously included in the yeast network. Indeed, a follow-up experiment confirmed that removing this pathway specifically allows the correct prediction of nicotinic acid auxotrophy of mutants affecting the kynurenine pathway. We anticipate that similar machine learning methods have the potential to facilitate the development of more accurate metabolic network models for metabolic engineering and systems biology and will also contribute to the growing field of automated scientific discovery (King, et al. 2009).

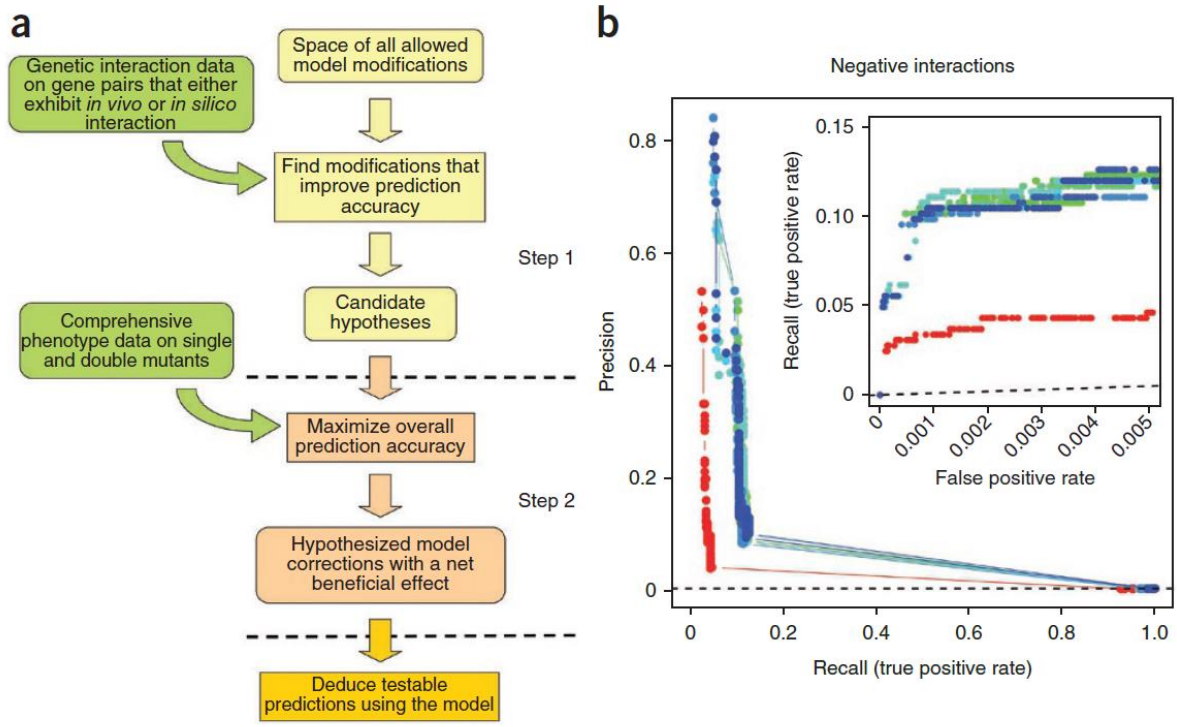



***Figure 3. Automated refinement of the metabolic model.*** *(A) Workflow of the refinement algorithm. Because evaluating each model is computationally intensive (i.e. a large number of gene deletions should be simulated for each individual model), we employed a two-step procedure to make use of all available phenotypic data while maintaining computational feasibility. In the first step, we searched for models by evaluating a model on only those gene pairs that display either in vivo interaction or in silico interaction according to the original model. Because genetic interactions are very rare both in vivo and in silico, most gene pairs examined in this study show no interaction and omitting them significantly speeds up the exploration of the hypothesis space. In the second step, we defined a new, very restricted hypothesis space based on the most successful models from the first step, but searched for models that improve overall prediction accuracy as assessed by a comprehensive evaluation of each model in the population. (B) Impact of model refinement on prediction accuracy using 8 independent runs of the algorithm. Figure shows the congruency of the modified (blue to green) and original (red) models to the empirical genetic interaction data by both precision recall and partial ROC curves (inset). Dashed lines represent prediction accuracy expected by chance. Note that while the same dataset was used for both model refinement and evaluation in this plot, a cross-validation procedure confirmed significant model improvement (Szappanos, et al. 2011). Figure reproduced from (Szappanos, et al. 2011).*

*Environmental-dependence of genetic interactions*

Finally, by computationally analyzing how genetic interactions change across dozens of nutrient conditions, we found that genetic interactions often depend on the prevailing environments. The study focused on synthetic lethal interactions, which is an extreme form of negative genetic interactions where the double gene deletant shows a no-growth phenotype that is not displayed be either single deletion mutant. In particular, out of 98 gene pairs that show synthetic lethality in at least one condition, only ~15% display this interaction under all nutrient conditions (Harrison, et al. 2007). Our work offered two scenarios for such environmental dependency, both of which was experimentally confirmed (Figure 4): (i) one or both genes of the synthetically interacting pairs become essential upon environmental change, or (ii) the double mutant becomes viable in a different environment. The first scenario has important implications for our understanding of genetic redundancy. Synthetic lethal interactions are often thought to indicate functional redundancy between gene pairs, e.g. through alternative pathways or gene duplicates (Hartman, et al. 2001). Crucially, our work shows that many genes involved in synthetic interactions in one environment become essential in another environment, indicating that their redundancy is more apparent than real. More generally, the robustness of metabolic networks against genetic perturbations is likely to be a by-product of adaptation to survive in a large variety of nutrient conditions.

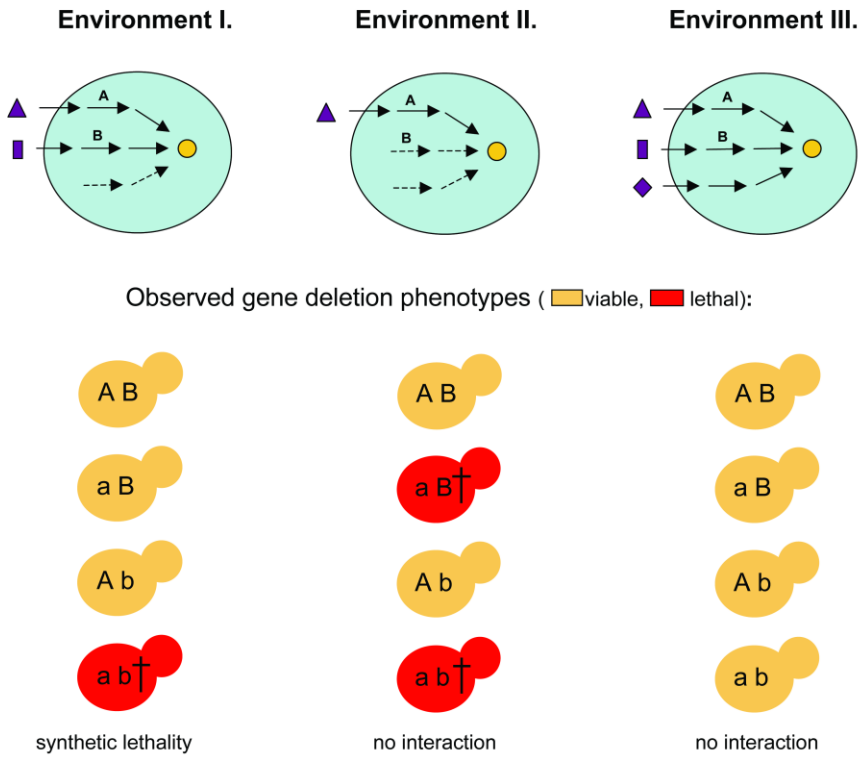



***Figure 4. Conceptual model to explain the environment-dependency of synthetic lethal interactions.*** *An essential metabolic intermediate (yellow circle) can be synthesized via three independent pathways. Enzyme encoding genes A and B are in synthetic lethal interaction in Environment I, where precursor nutrients of both pathways are present in the environment. However, the presence of gene B is unable to compensate inactivation of gene A in Environment II, and the double mutant ab is rescued by a third pathway in Environment III. Figure reproduced from (Harrison, et al. 2007).*

## III. Predicting genome reduction

*Key paper: (Pál, et al. 2006) (see Appendix)*

*Further related papers: (Fehér, et al. 2007), (Yizhak, et al. 2011)*

One of the central questions in the post genomic era is understanding which organisms have which genes. Typically, such inferences are drawn *a posteriori*, that is having discovered that an organism has a given gene we then construct hypothesizes about its ecology or biology. For example, we infer that because mice have abundant olfactory receptors they need to detect many chemicals in their natural environment. But is it possible to do the inverse and hence have an *a priori* and predictive theory for a genome? That is, can we take an organism's ecology and predict which genes it should have with any accuracy? At first sight this seems an almost impossible task due to the large diversity of genes and gene combinations that may perform similar functions. However, it might be possible to predict changes in gene content during reductive evolution of genomes, that is, when genes are lost on a massive scale, as seen in endosymbotic bacteria. In such situations, knowing the initial genomic composition, relevant selection pressures and functional constraints may allow us to predict the outcome of evolution.

*Computational prediction of genome reduction*

As first attempt to probe the feasibility of predicting long-term genomic evolution, we asked whether, given the genome of *E. coli*, we can predict the metabolism of *Buchnera*, an intracellular symbiont with a heavily reduced genome, that was derived from *E. coli* (Figure 5). *Buchnera* have

evolved from its free-living ancestors approximately 200 million years ago and lost 75% of their genes, reaching nearly minimal gene sets (~600 genes) needed to sustain life. We used a genome-scale model of *E. coli* metabolism, and setup the model to mimic the lifestyle of the endosymbiont based on available physiological evidence. Specifically, *Buchnera* consume glucose and glutamate while supply their aphid hosts with essential amino acids and riboflavin that are in shortage in the hosts' diets. Using a series of flux balance analysis simulations, we considered sequentially the fate of randomly selected gene deletions and asked, given the ecology of *Buchnera*, whether these would be effectively neutral or not. Repeatedly simulating successive gene loss events until no further genes could be deleted without impairing *in silico* growth, we obtained a set of minimal networks (Figure 5). Remarkably, comparison of the gene complements of these in silico minimal networks with three then available *B. aphidicola* genomes revealed that gene presence / absence can be predicted with high accuracy (area under the ROC curve =0.794 – 0.802, depending on the endosymbiont genome). Thus, 200 million years of reductive genome evolution is surprisingly well predictable based on knowledge of the organism's distant ancestor and its current lifestyle. Moreover, in a follow-up work, we reconstructed the intermediate steps of genome reduction in the *Buchnera* lineage and demonstrated that not only the outcome, but also the order of gene loss events are predictable (Yizhak, et al. 2011).

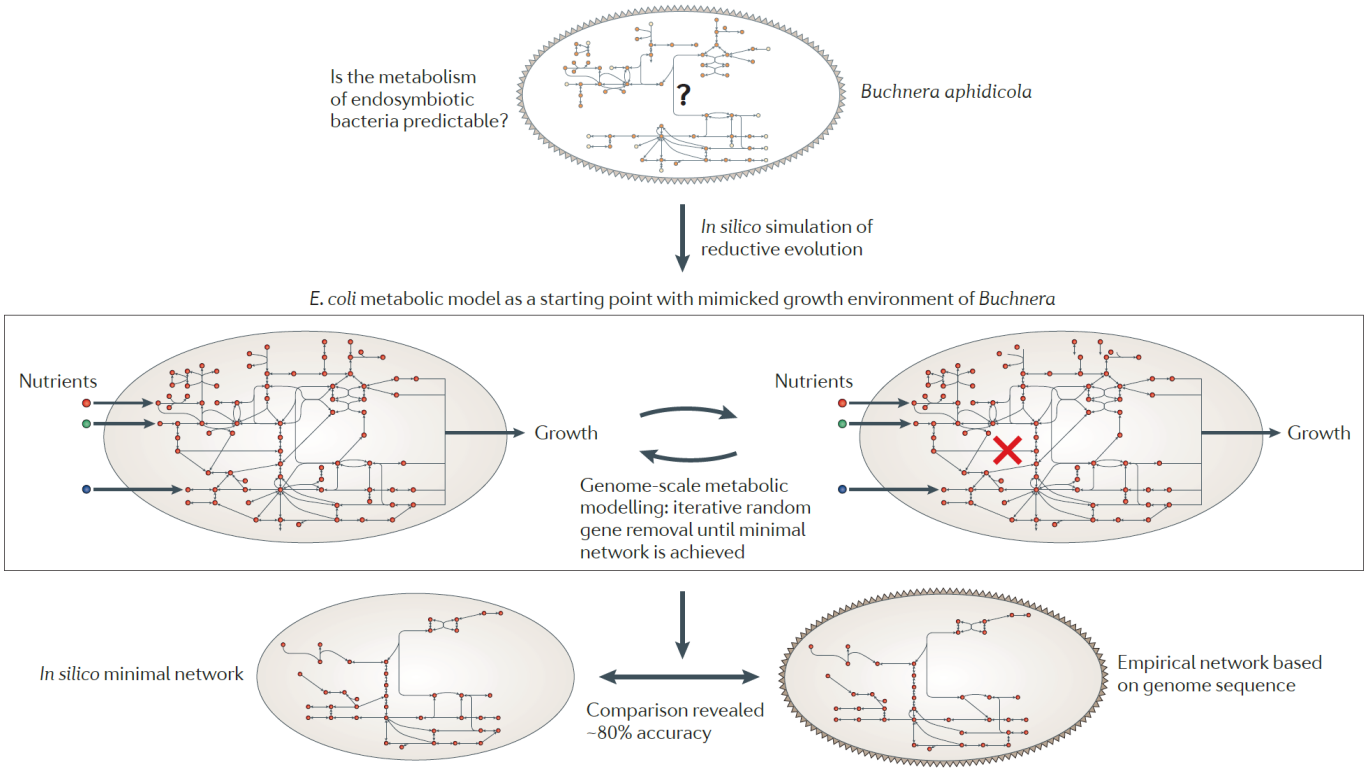



***Figure 5. A genome-scale metabolic modelling approach to predict genome reduction in endosymbiotic bacteria.*** *The computational approach uses the genome-scale metabolic model of present day E. coli as a proxy for the free-living ancestor of the endosymbiont B. aphidicola and mimicks the lifestyle of the present day endosymbiont to predict the impact of individual gene loss events along an evolutionary trajectory. During the evolutionary simulations, minimal metabolic networks were generated by repeatedly simulating gene loss events (red cross) until no further genes could be removed without impairing in silico growth. The computationally predicted minimal networks showed high overlap with the metabolic gene contents of real Buchnera (lower panel). Figure reproduced from (Papp, et al. 2011).*

### *Chance and necessity in the evolution of minimal networks*

Even closely related *Buchnera* strains vary in their gene complements. In principle, such variation in reductive genome evolution may reflect both differences in selective forces (ecology) and chance events, yielding differences in the order of gene deletions and hence a choice between

alternative cellular pathways. The stochastic nature of the deletions in our simulations introduces an historical accident component that ends up predicting that such variety should exist. Comparing the minimal networks from repeated simulations, we indeed found support for partially different evolutionary outcomes that arise from chance events. Simulated minimal reaction sets differ, on average, by 12% of their reactions. This variability represents phenotypically nearly equivalent alternative gene loss trajectories owing to the presence of parallel metabolic pathways in the ancestral bacterium. For example, *E. coli* can convert acetate to acetyl-CoA through two parallel pathways (Kumari, et al. 1995). In line with this, we found that the simulated minimal networks always contain only one of the two pathways and *Buchnera* strains have also retained only one of them.

Remarkably, these analyses successfully predicted core genes that all *Buchnera* genomes should share as well as genes that should be present in some, but not all, *Buchnera* genomes due to historical chance events (Figure 6). This is an important result because it shows that variation between taxa in metabolic capabilities need not simply reflect ecological differences, as typically presumed, but may just be historical contingency. We expect that the role of chance events will be more prominent when horizontal gene transfer plays an important role in the evolution of metabolic gene contents.

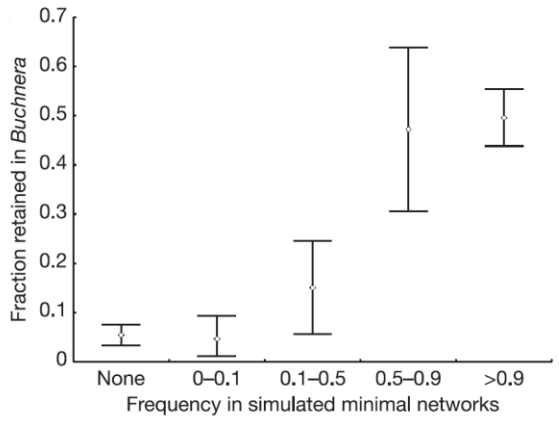



***Figure 6. Chance and necessity in genome reduction trajectories.*** *Repeated evolutionary simulations predicted core genes that should be present in all minimal metabolic networks and genes that should be present in some but not all networks. Indeed, genes in the former group are much more likely to be retained in the genome of B. aphidicola Bp. Error bars indicate 95% confidence intervals. Chi-square test: n=874, Chi-square=222.6, d.f.=4, P<$10^{-46}$. Figure reproduced from (ref).*

*Implications for synthesizing minimal genomes*

These results also have implications for the synthesis of minimal genomes in the laboratory. An important strategy to identify the set of genes essential for cellular life is to inactivate genes individually (Fehér, et al. 2007). However, due the presence of parallel pathways and functional compensation between different genes, the set of essential genes of any organism must be only a subset of the minimal genome as non-essential gene can easily become essential in some genomic contexts. Our simulations quantify this discrepancy. Specifically, we found that the list of essential genes in a free-living bacterium underestimates the minimal gene set by 45% in the metabolic network. Indeed, more recent practical computational tools to prioritize genomic regions for removal integrate genome-scale metabolic modelling with gene essentiality data (Wang and Maranas 2018).

## IV. Underground metabolism and the predictability of adaptive evolution

*Key paper: (Notebaart, et al. 2014) (see Appendix)*

*Further related papers: (Notebaart, et al. 2018), (Guzman, et al. 2019)*

Understanding how new molecular pathways emerge during adaptation is one of the central issues in evolutionary and systems biology. In the most well-understood networks, small-molecule metabolism, the prevailing paradigm is that evolution capitalizes on the weak side activities of pre-existing enzymes (Jensen 1976). This paradigm rests on several key empirical observations. First, most enzymes are catalytically promiscuous, that is they have limited substrate specificities and show measurable, albeit weak, catalytic activity for alternative substrates (Khersonsky and Tawfik 2010). These so-called underground enzyme activities appear to be widespread (Kuznetsova, et al. 2006; Huang, et al. 2012). Second, underground activities can be enhanced by few mutational steps and hence serve as starting points for new enzyme functions in directed evolution experiments in the lab (Aharoni, et al. 2005). Weak ancestral activities towards non-preferred substrates have also been demonstrated to contribute to the functional diversification of enzyme families in the wild (Huang, et al. 2012). Third, comparative genomic studies have established that new metabolic pathways are typically patched together from homologs of other enzymes that function in different parts of the network (Rison, et al. 2002; Schmidt, et al. 2003).

While much has been learned about the biochemical mechanisms of these activities in a few well-studied enzymes, the extent to which underground reactions provide novelties in the context of the entire cellular system remains completely unexplored. This gap of knowledge is far from trivial as

many underground reactions, while providing a new catalytic function, might be isolated from the rest of the network, might only contribute to new pathways that are functionally redundant with existing ones, or might even be harmful.

*Reconstructing the underground metabolic network of E. coli*

To address the above gap, we applied a systems-level approach that provides insights into the architecture of underground metabolism and also enables the prediction of the role of underground activities in adaptation to nutrient conditions (Notebaart, et al. 2014). In brief, we focused on *Escherichia coli*, which is the most comprehensively characterized organism in terms of enzymatic activities. For instance, enzyme databases, such as BRENDA (Scheer, et al. 2011) contains hundreds of publications on non-native side activities of enzymes that are not included in the canonical metabolic network reconstruction (i.e. native network). Based on such databases and literature survey, we therefore built an *in silico* underground metabolic network reconstruction of *Escherichia coli* and integrated it with the native genome-scale metabolic model of this organism (Feist, et al. 2007) (Figure 7). Overall, we included 262 underground reactions and 277 metabolites that are not present in the native network. Two lines of evidence indicate that these reactions occur at very low rates and are physiologically irrelevant in the wild-type cell: (i) The measured catalytic activities of these reactions are ~200-fold lower than the native activities of the same enzymes, and (ii) metabolites introduced into the network via underground reactions are rarely observed experimentally. The resulting reconstruction is the first comprehensive computational model of the underground metabolism of any organism.

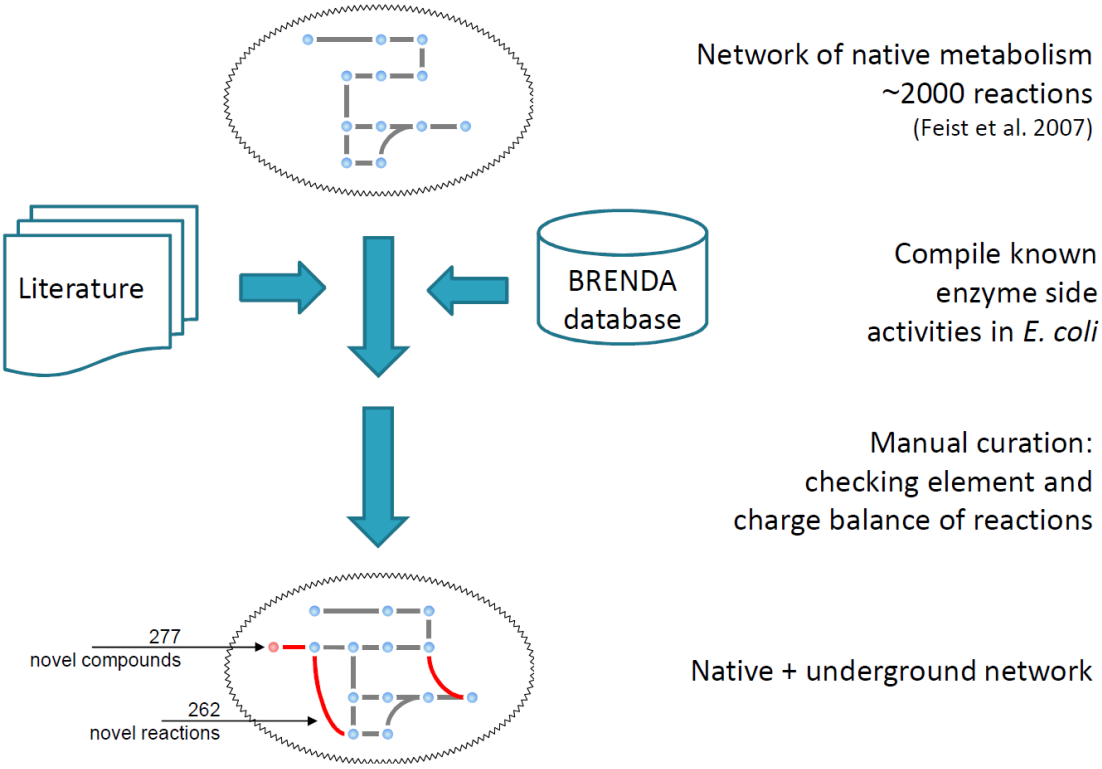



***Figure 7. Workflow of the reconstruction process.*** *Grey and red lines indicate native and underground reactions, respectively.*

*The architecture of underground metabolism*

Introducing novel biochemical reactions through enzyme side activities into to the native network may create cross-wiring between existing metabolites, introduce dead-ends or result in isolated reactions that are not connected to the rest of the network (Figure 8). Analysis of our underground network reconstruction suggests that a large fraction (45%) of these reactions can be fully wired into the native network, while only a minority is completely isolated from the rest of metabolism (Figure 8). But can the fully connected underground reactions potentially contribute to the

formation of key biomass precursors and hence be useful for the cell? To test this, we decomposed the network into biochemically relevant pathways, so called elementary flux modes (EFMs) that can maintain steady-state flux from nutrient uptake to biomass production, and analyzed the properties of such pathways. Taking a large sampling of EFMs shows that all underground reactions that can be fully wired into the network can carry flux from glucose to biomass precursors and at similar efficiencies (i.e. yield) as the native reactions of corresponding enzymes. Together, these analyses showed that a substantial proportion of underground reactions can form new pathways with high potential biological relevance.

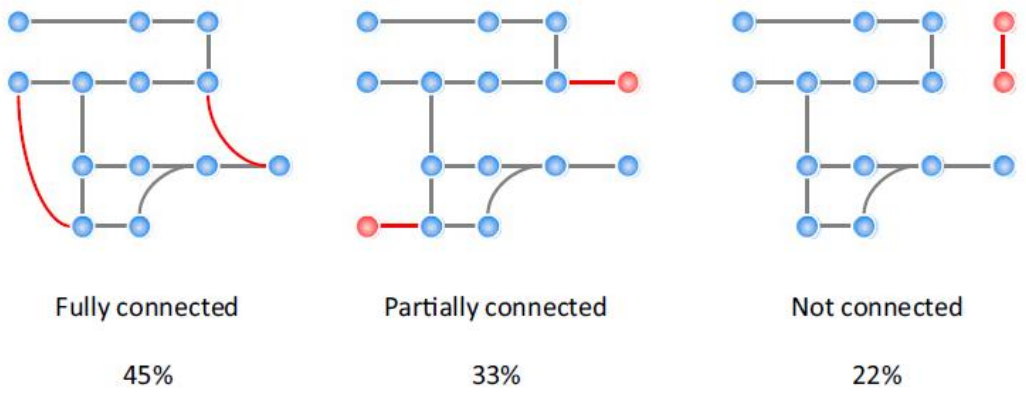



***Figure 8. Connectedness of underground reactions in the native network.*** *The connectedness of each underground reaction was assessed individually, hence some unconnected or partially connected metabolites might become fully connected in the presence of other underground reactions. Nodes denote metabolites (blue and red correspond to native and novel metabolites, respectively) and edges denote biochemical reactions (gray and red correspond to native and underground, respectively). Figure reproduced from (Notebaart, et al. 2014).*

*Characterizing the evolutionary potential of underground metabolism*

How frequently do underground enzyme activities serve as raw materials for evolutionary adaptation to new environments? To address this, we conducted an integrated *in silico* and an *in vivo* survey to characterize the evolutionary potential of *E. coli* to adapt to hundreds of novel nutrient conditions. First, we computationally predicted the impact of adding underground reactions to the native network on maximum growth across a variety of environments. Because we were interested in the potential to evolve towards new nutrients, we assumed that all underground reactions can be utilized (i.e. there are no enzyme kinetic or regulatory constraints). We simulated growth using FBA in ~2700 nutrient conditions that encompass the full range of carbon, nitrogen, sulfur and phosphorous sources that can be transported to the network. The analysis revealed dozens of cases where underground reactions allow or improve growth in previously uncharacterized growth environments when their activity is increased (Figure 9). Specifically, underground reactions enabled growth in 19 new conditions and improved growth in 31 conditions. To put these figures into context, we note that the native network shows growth in 645 environments. Thus, increasing the total reaction content of the *E. coli* network by ~11% expanded its range of utilizable carbon sources by ~3%. Most of the growth improvements were conferred by single underground reactions (shown in Figure 8), with only a minority requiring multiple reactions simultaneously (see (Notebaart, et al. 2014)). Overall, we estimate that ~15% of the underground reactions that can be fully wired into the network confer an advantage when added individually, while an additional 5% are beneficial in combination with other side activities (Figure 9).

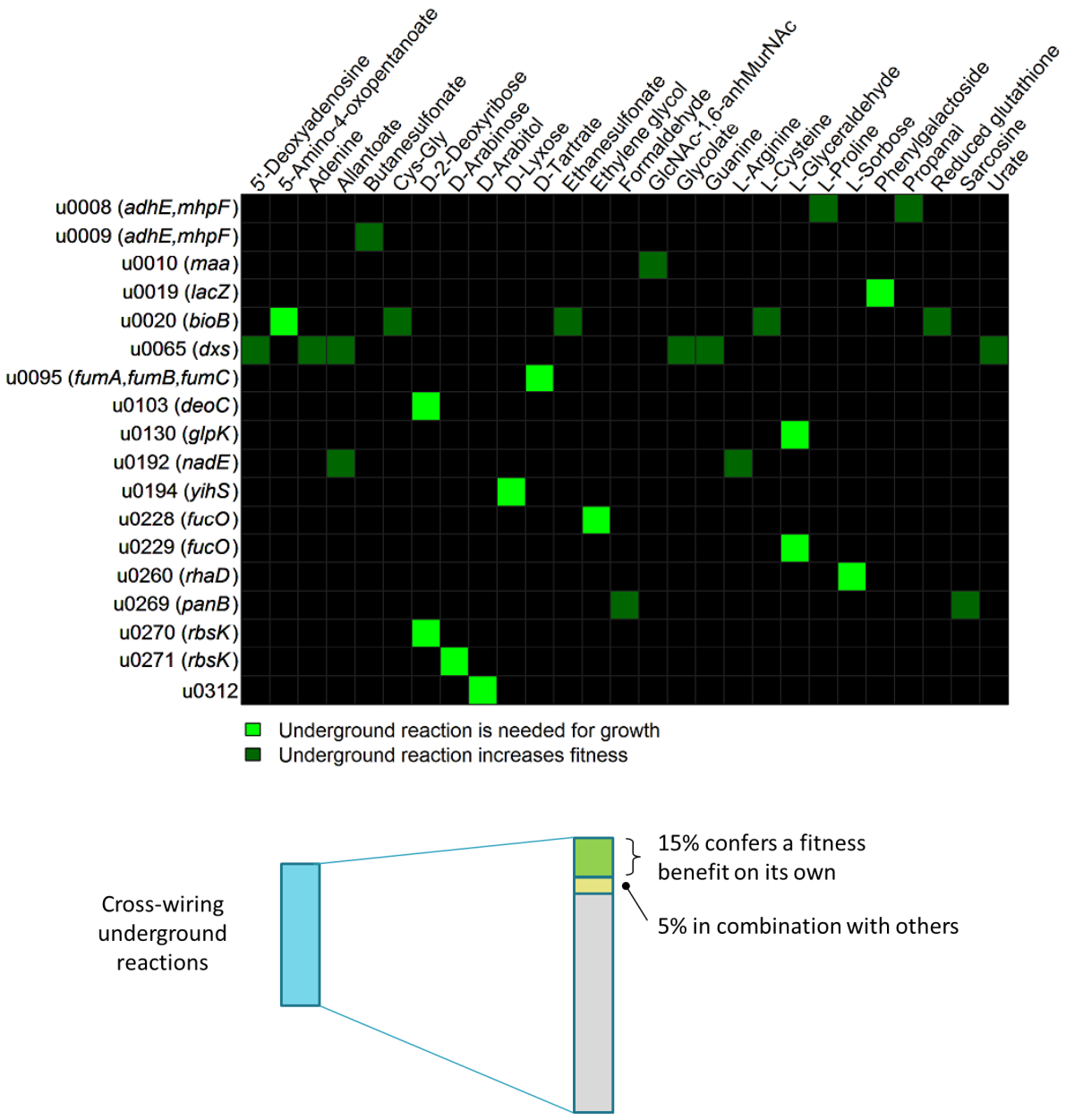



***Figure 9. In silico growth improvements conferred by underground reactions.*** *Heatmap showing the computationally predicted growth advantages conferred by adding single underground activities to the native network across different nutrient conditions (upper panel). Only aerobic carbon sources are shown here. Bright green squares indicate utilization of a nutrient on which the native network does not grow. Lower panel depicts the estimated adaptive potential of underground reactions. Upper panel reproduced from (Notebaart, et al. 2014).*

Second, we experimentally estimated the potential of underground activities in adaptation to new nutrient sources. We carried out a genome-wide gene overexpression screen in *E. coli* and measured growth under 194 carbon sources (Patrick, et al. 2007; Kim, et al. 2010). Overall, we identified 17 genes that improved growth upon overexpression in at least one of 17 specific carbon sources. Out of the 17 genes, 11 encoded enzymes and 9 of these had known underground activities. Notably, 6 of these enzymes conferred growth on a carbon source where the wild-type was unable to growth. Our screen offers an estimate of the *in vivo* evolutionary potential of individual underground reactions: strong overexpression of single genes expands the range of utilizable nutrients by 6% (from 85 to 90 of the tested nutrient conditions). Note that this must be an underestimate as our experimental assay is unable to detect evolutionary novelties that require the simultaneous amplification of multiple underground activities or those that provide only a small fitness benefit. Together, these analyses strongly support the notion that evolution can capitalize on underground reactions both to enhance growth in existing environments and to exploit completely new nutrient sources.

*Predicting the genetic basis of adaptation to novel environments*

Is it possible to predict which enzyme confers a growth benefit in which new environment through amplifying its side activity? In other words, can we predict the genetic basis of evolutionary adaptation to new environments? Comparison of the computational predictions with the genome-wide overexpression experiment showed a remarkable agreement (Figure 10). Specifically, the computational model successfully predicted 44% of the carbon sources on which amplification of

an enzyme conferred or improved growth, an overlap that is statistically highly significant ($P<10^{-13}$). For instance, the metabolic network model predicted that amplification of the side activity of YihS enables growth on D-lyxose. In line with this prediction, wild-type *E. coli* is unable to grow on D-lyxose, but becomes capable of utilizing it when the gene encoding YihS is overexpressed. Notably, this is a highly specific prediction as none of the other ~4000 overexpressed proteins conferred growth on this carbon source. These results demonstrate that it is possible to predict the genetic basis of evolution towards new nutrient environments based on a detailed knowledge of an organism's underground metabolism.

*Toward predicting evolution through spontaneous mutations*

The above analyses employ gene overexpression experiments to validate the computational predictions. As such, they leave it unclear whether the underlying genetics of adaptation can be also predicted in a population of bacteria that evolve through spontaneous mutations. Overexpression experiments might be poor representation of real evolutionary processes for at least two reasons: (i) strong artificial overexpression might induce phenotypes that are not readily accessible through single mutations arising spontaneously and (ii) the same phenotype might be reached by mutations in several distinct genes, making it challenging to predict which of these genes are actually mutated during evolution. We addressed these issues in a more recent work in collaboration with Adam Feist's lab by conducting a series of automated laboratory evolution experiments to adapt *E. coli* to novel carbon sources (Guzman, et al. 2019). By focusing on 5 non-native carbon sources that are predicted to be reachable through specific underground reactions, we showed that E. coli repeatedly acquired the ability to utilize them. More strikingly, in 4 out of 5 carbon sources, the genes underlying the phenotypic innovations were accurately predicted by

computational model simulations incorporating underground reactions. Eventually, this work demonstrates that computational systems biology models can be employed to predict the trajectory and outcome of adaptive evolution under certain circumstances.

More broadly, as our work offers a system-level framework to predict evolution based on the knowledge of 'underground' phenotypic potentials, we anticipate that it will have far-reaching potential for various application areas from bioengineering to medical genetics. Specifically, the role of gain-of-function enzyme mutations in tumor evolution is becoming increasingly recognized (e.g. (Dang, et al. 2009)) and our work provides a new framework to systematically study the phenotypic potential of such mutations.

| Carbon source | Enzyme | Known or proposed underground reaction | Phenotype | *In silico* |
|---|---|---|---|---|
| D-Lyxose | Mannose isomerase (*yihS*) | D-Lyxose ↔ D-Xylulose | ++ | ● |
| D-2-Deoxyribose | Ribokinase (*rbsK*) | D-2-Deoxyribose + ATP → D-2-Deoxyribose 5-P + ADP + $H^+$ | ++ | ● |
| D-Tartrate | Fumarase A (*fumA*), Fumarase B (*fumB*) | D-Tartrate ↔ Oxaloacetate + $H_2O$ | ++ | ● |
| Phenylgalactoside | β-galactosidase (*lacZ*) | Phenylgalactoside + $H_2O$ → D-Galactose + Phenol | + | ● |
| m-Tartrate | D-malate dehydrogenase (*dmlA*) * | m-Tartrate ↔ D-Glycerate + $CO_2$ | ++ | ○ |
| Putrescine | Acetaldehyde dehydrogenase (*mhpF*) * | 4-Aminobutanal + H-R + NAD ↔ 4-Aminobutanoyl-R + NADH + $H^+$ (R = $CoA_{,,}$ OH) | ++ | ○ |
| D-Malate | 3-isopropylmalate dehydrogenase (*leuB*) * | D-Malate + NAD ↔ Pyruvate + $CO_2$ + NADH + $H^+$ | + | ○ |
| β-Methyl-D-Glucoside | 6-phospho-beta-glucosidase (*bglB*) | β-Methyl-D-Glucoside 6-P + $H_2O$ ↔ β-D-Glucose 6-P + Methanol | + | ○ |
| Monomethyl Succinate | Carboxylic ester hydrolase (*ybfF*) * | Monomethyl Succinate + $H_2O$ ↔ Succinate + Methanol | + | ○ |

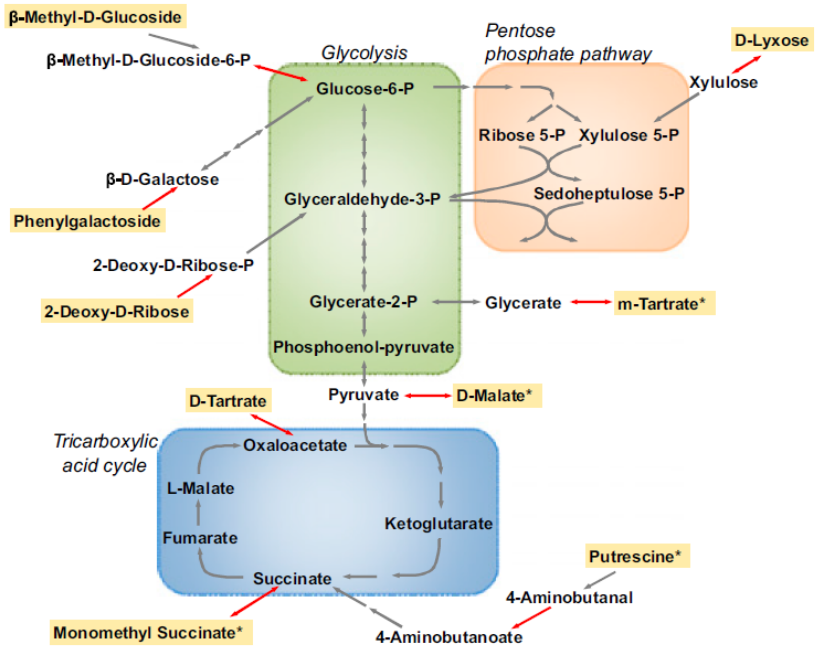



***Figure 10. Prediction of growth advantages conferred by amplified underground activities.*** *List of nine enzymes that enable (++) or improve (+) growth on specific carbon sources when overexpressed experimentally. All of these enzymes have known or presumed side activities (underground reactions supported by indirect evidence are denoted by asterisks). Forty-four percent of the experimentally confirmed phenotypes were also predicted by the computational model (●). Lower panel depicts a schematic map of central carbon metabolism with underground reactions (red arrows) that confer a growth advantage on specific carbon sources (highlighted in yellow) when overexpressed.*

## V. Simple paths to complex adaptations

*Key publication: (Szappanos, et al. 2016) (see Appendix)*

*Further related paper: (Pal and Papp 2017)*

Explaining the origin of evolutionary novelties remains a central challenge in evolutionary biology. Traits that require the simultaneous emergence of multiple mutations, none of which seemingly confer a benefit individually, pose an especially daunting challenge for evolutionists. Such traits are often referred to as complex adaptations and might be difficult to evolve, not because of physical or chemical constraints, but because of the dynamics of how mutations spread in the population. Darwin himself was well aware of this challenge: “if it could be demonstrated that any complex organ existed, which could not possibly have been formed by numerous, successive, slight modifications, my theory would absolutely break down” (Darwin 1859).

Proponents of intelligent design generally refer to the pseudoscientific theory of irreducible complexity, claiming that such complex traits cannot be explained by Darwinian evolution: they are considered too complex to evolve from simpler predecessors through natural selection acting upon a series of advantageous naturally occurring mutations. However, complex adaptations are omnipresent at all levels of biological organization. For example, they frequently occur in molecular systems, including the establishment of disulfide bonds in protein molecules, the origin of multi-step metabolic pathways and regulatory – DNA interactions (**Hiba! A hivatkozási forrás nem található.**1). Overall, the widespread occurrence of complex adaptations indicate that they can readily evolve in nature. Thus, a theory is needed that accounts for their rapid evolution.

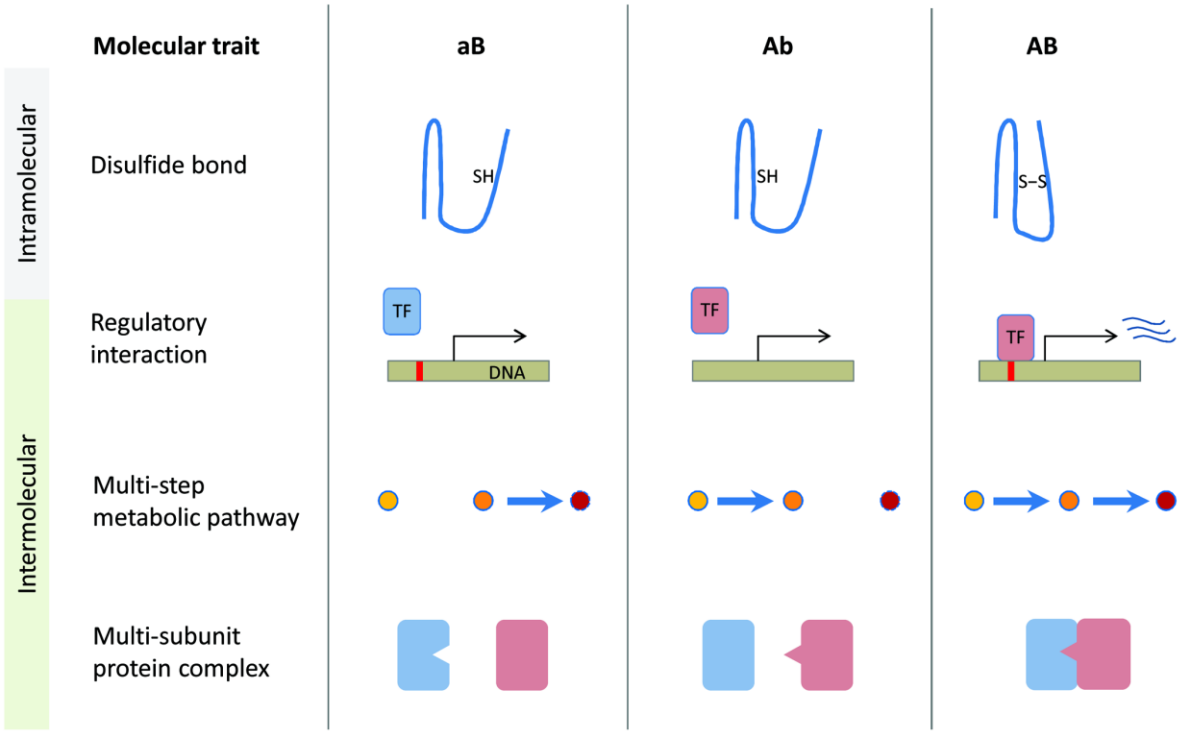



***Figure 11. Major forms and examples of complex adaptations in molecular traits.*** *The origin of a new disulfide bond (S-S) from two nearby sulfhydryl groups (–SH) within the same protein represents an example of intramolecular complex adaptation. The evolutionary establishment of new transcription factor – DNA binding site interactions, metabolic pathways involving multiple steps and multi-subunit protein complexes can be considered as intermolecular complex adaptations demanding specific mutations in multiple genes. Note that two mutations (a → A and b → B) have to occur simultaneously to confer a fitness benefit (adaptation) in all these traits. Yellow, orange and red circles represent the substrate, intermediate metabolite and end product, respectively, of a schematic metabolic pathway. Figure reproduced from (Pal and Papp 2017).*

One influential theoretical model for complex adaptations invokes the accumulation of neutral mutations which prepare the ground for later beneficial mutations that eventually lead to

innovations (Wagner 2008). A related population genetic theory argues that large populations harbor a reservoir of non-adaptive mutations in which a second mutation can be beneficial and go to fixation (Weissman, et al. 2009; Lynch and Abegg 2010). However, these processes are expected to be very slow compared to those where the intermediate steps are facilitated by adaptive bypasses (see below for one such bypass). Furthermore, these model lacks direct empirical support in molecular networks.

*The varying environment scenario of complex adaptations*

Here I discuss a conceptually simple model to resolve the paradox of complex adaptation. This scenario is closely related to the notion of pre-adaptation and purely relies on the successive accumulation of beneficial mutations. In brief, temporally varying environmental conditions select for single adaptive mutations that, as a by-product, serve as stepping stones towards the establishment of more complex phenotypes. Thus, complex adaptations can be accelerated in dynamically changing environments. The core of this idea has been proposed by Horowitz in his seminal paper on the early origin of metabolic pathways (Horowitz 1945) and is conceptually related to *in silico* studies of the evolution of RNA molecules and genetic circuits in varying environments (Kashtan, et al. 2007).

Specifically, we asked how novel nutrient utilization phenotypes (traits) can be established in a bacterial metabolic network by adding new enzymatic reactions to it. We hypothesized that varying environments promote the piecewise assembly of enzymatic reactions into novel multi-step pathways in an organism that already harbors a complex metabolic network. This

phenomenon is expected to arise if some single enzyme acquisitions confer a fitness advantage in some other specific environments in addition to contributing to the multi-step pathway. The acquisition of such enzymes act as molecular 'springboards' to facilitate further adaptive evolution.

*Testing the varying environment scenario – a computational approach*

To test this scenario, we first studied *in silico* the expansion of the *E. coli* metabolic network to utilize novel nutrients. It has been established that bacterial metabolic networks expand typically by horizontally acquiring new enzymatic and transporter genes involved in the utilization of nutrient sources (Pál, et al. 2005). Therefore, we examined how biochemical reactions that are absent in *E. coli* but present in other organisms (the 'universal reaction set') confer a benefit in a large panel of environments when added to the E. coli network. Our dataset comprised more than 2,500 reactions not present in E. coli and ~1,700 possible nutrient sources. Flux balance analysis showed that expanding the E. coli network enabled growth in 321 of these novel nutrient environments. Importantly, acquiring the capacity to grow in a new environment typically demanded the addition of only one to three new enzymatic or transport reactions to the native *E. coli* network (Figure 12). Thus, most metabolic novelties can be reached by few mutational steps.

Our computational analyses revealed that new complex pathways can evolve via the successive acquisition of single biochemical reactions that each confer a benefit under specific environmental conditions. For example, there are 118 environments that require the simultaneous acquisition of two reactions and 8.5% of them can be accessed through an adaptive intermediate step (i.e. one

reaction confers a benefit in another environment). This is well illustrated by the ability to grow on chorismate, which can be reached through purely adaptive walks by first acquiring L-phenylalanine utilization (Figure 13).

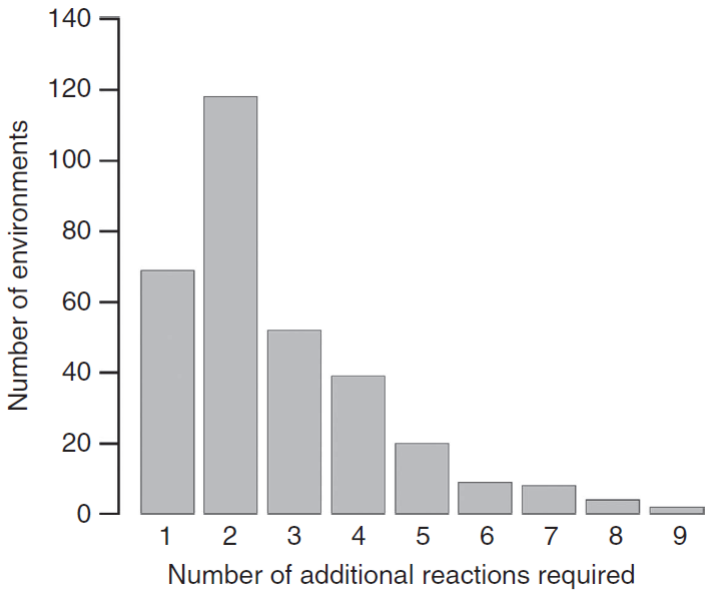



***Figure 12.*** *Metabolic novelties are only few mutational steps afar from the E. coli network. The plot shows the distribution of the number of minimum extra reactions needed for growth in 321 novel nutrient conditions.*

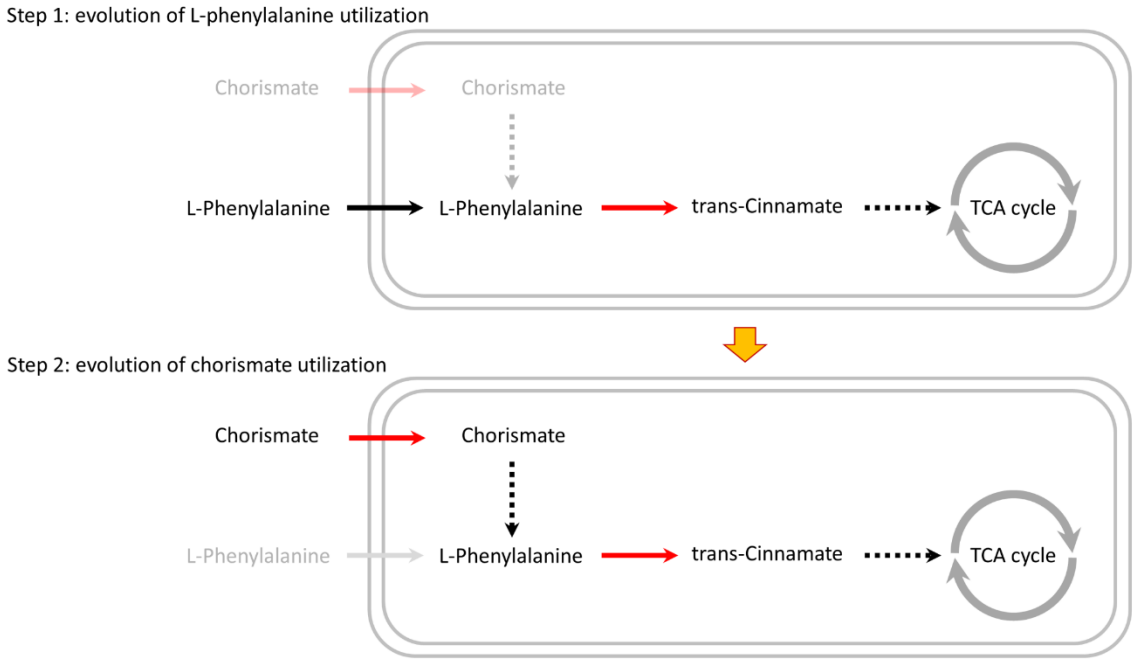



***Figure 13.*** *Example of a two-step pathway that becomes accessible for evolution through purely adaptive enzyme acquisitions. Chorismate and L-phenylalanine are carbon sources that cannot be utilized by E. coli K-12. Wheras chorismate utilization requires the joint acquisition of two reaction steps, one of these steps (catalysed by phenylalanine ammonia lyase) also enables the utilization of L-phenylalanine when acquired individually. Reproduced from (Szappanos, et al. 2016).*

Second, the varying environment scenario predicts that gain of new metabolic genes should occur in a defined order. In particular, if two enzymes are jointly required to support growth in a novel environment, and one enzyme confers a benefit on its own in another environment, then the latter enzyme should be gained earlier. We tested this by reconstructing the evolutionary history of gene gain events in 943 bacteria using phylogenetic methods. In line with the expectation, we found

that two-step pathways tend to be established the same order as predicted by the computational analysis.

*Experimental test of the varying environment scenario*

Last, we carried out a laboratory evolution study to adapt *E. coli* to two novel carbon sources and showed that evolving the ability to grow on one of them facilitated subsequent adaptation to the other. Specifically, we focused on two related carbon sources, ethylene glycol and propylene glycol, on which wild-type *E. coli* is unable to grow. We found that adaptation of an *E. coli* strain background with an elevated mutation rate occurred readily to propylene glycol, but not to ethylene glycol. Remarkably, a genotype that first adapted to propylene glycol showed at least ~100-fold increased frequency to adapt to ethylene glycol, indicating that the first adaptation served as a stepping stone to the second one. Further analysis showed that upregulation of the *fucO* gene alone confers growth on propylene glycol and increases the rate of adaptation to ethylene glycol. Notably, ethylene glycol utilization was achieved by amplification of the gene of the AldA enzyme, which acts in the same pathway as FucO (Figure 14). Indeed, we found that simultaneous overexpression of both *fucO* and *aldA* enabled growth on ethylene glycol, a phenotype that was not conferred by either *fucO* or *aldA* overexpression alone.

Taken together, the above results demonstrate that complex metabolic adaptations can evolve through adaptive intermediate mutations by stepwise expansion of nutrient utilization capabilities. This conclusion represents an important conceptual advance as there is no need to invoke the slow process of accumulating neutral intermediate mutations.

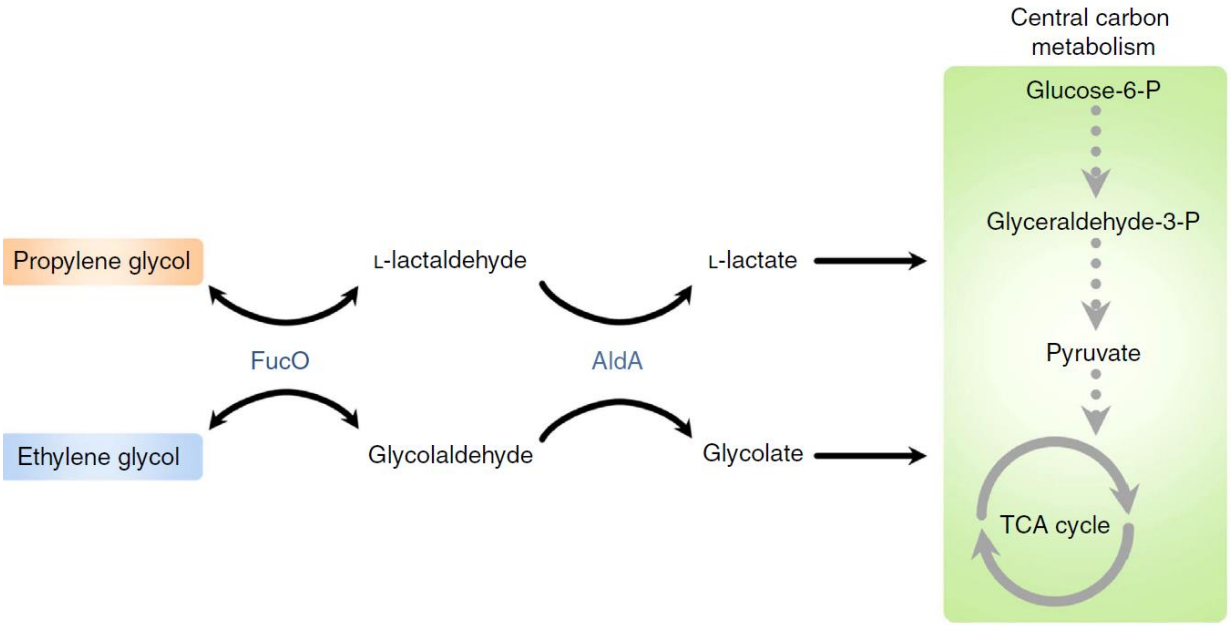



***Figure 14.*** *Gaining the ability to utilize propylene glycol enhances adaptation to ethylene glycol utilization. Pathway map of propylene glycol (PG) and ethylene glycol (EG) utilization. In the first step, FucO catalyses the oxidation of PG and EG to glycolaldehyde and L-lactaldehyde, respectively. In the second step, AldA oxidizes the products of FucO to hydroxycarboxylic acids which are channeled into the central carbon metabolism. We note that the affinity of AldA for L-lactaldehyde is higher than for glycolaldehyde, potentially explaining why growth on EG demands the amplification of aldA.*

*Implications of the varying environment scenario*

Our work has important ramifications for those studying the design principles of complex molecular pathways as well as for those aiming to create industrially useful microbes. First, our results suggest that deciphering the adaptive value of molecular pathways might often require studying their operation under multiple environmental conditions. Second, we anticipate that

evolutionary engineering of microbes to obtain desired phenotypes could be facilitated by temporally varying the traits under selection.

## VI. Summary of key results

This thesis is based on a series of publications that utilize computational systems biology modelling of metabolic networks to address several outstanding issues in evolutionary genetics. We reached the following major results, as published in the highlighted papers:

1) We probed the limits of predicting genetic interactions using genome-scale metabolic networks and showed that genetic interaction hubs are highly predictable, but individual genetic interactions are often missed by these computational models. Building on these discrepancies, we developed a machine learning method that refines the metabolic network model based on large-scale genetic interaction data.

Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C., **Papp, B.** (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics* **43:** 656

2) We showed that synthetic lethal genetic interactions often depend on the prevailing environments. Importantly, this is often caused by one or both genes of the synthetically interacting pairs becoming essential upon environmental change, indicating that the two genes are only partly redundant.

*Harrison, R., ***Papp, B.**, Pál, C., Oliver, S.G., Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A.* **104**: 2307-12.

3) By simulating the repeated loss of non-essential genes in a genome-scale metabolic model of *E. coli*, we showed that it is possible to predict the highly reduced gene content of closely related endosymbiotic bacteria that diverged ~200 million years ago.

*Pál, C., ***Papp, B.**, Lercher, M.J., Csermely, P., Oliver, S.G. and Hurst, L.D. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667-70.

4) We reconstructed a comprehensive network of known enzyme side activities (i.e. underground reactions) in *E. coli*, which is the first such reconstruction in any organism. By combining computational simulations and a high-throughput experimental survey across hundreds of nutrient environments, we predicted and confirmed new environments where enhanced activity of underground reactions confer growth. Our results demonstrate that the genetic basis of evolutionary adaptations via underground metabolism is computationally predictable.

Notebaart, R.A.*, Szappanos, B., Kintses, B., Pál, F., Györkei, A., Bogos, B., Lázár, V., Spohn, R., Csörgő, B., Wagner, A., Ruppin, E., Pál, C.*, **Papp, B.*** (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A.* **111**: 11762-11767.

5) We proposed a new model to resolve the paradox of complex adaptations, i.e. new traits that require the simultaneous emergence of multiple mutations, none of which seemingly confer a benefit individually. By studying the evolution of new nutrient utilization capabilities in metabolic networks, we showed that phenotypes accessible through the addition of a single reaction serve as stepping stones towards the later establishment of complex metabolic

phenotypes in another environment. Thus, temporally varying environmental conditions enable the step-by-step expansion of nutrient utilization capacities without the need to invoke non-adaptive processes.

Szappanos, B., Fritzemeier, J.C., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R.A., Lercher, M.J., Pál, C.*, **Papp, B.*** (2016) Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat Commun.* **7**:11607

Note that the highlighted papers are included in the *Appendix* at the end of the thesis.

## VII. Outlook: genome-scale modelling meets machine learning

Genome-scale metabolic models have proven successful in addressing research questions that can be formulated as a mapping from gene presence - absence to growth phenotypes. However, the basic constraint-based modelling framework has at least two major limitations that prevents its more extensive use in the field of evolutionary genetics. First, it remains a formidable challenge to investigate the phenotypic impact of genetic variants beyond the resolution of gene presence – absence (e.g. most nucleotide-level variants) or those that alter non-enzymatic genes (e.g. in regulatory genes). Second, they assume optimal behavior for the wild-type network and therefore no mutations can increase fitness without expanding the network itself. Clearly, as genetic variants that do not completely abolish enzyme functions are abound, these limitations must be overcome to better represent evolutionary processes.

One possible way to overcome these limitations is to integrate the constraint-based metabolic model with additional layers of mechanistic models to capture (i) quantitative details of enzyme kinetics, at least in some parts of the network (Fleming, et al. 2010; Cotten and Reed 2013), (ii) gene expression and gene regulatory circuits (Shlomi, et al. 2007; O'Brien, et al. 2013), and (iii) enzyme structures (Mih, et al. 2016). In theory, such efforts may eventually lead to whole-cell models that can predict the phenotypic impact of single nucleotide variants in most protein coding genes. However, building whole-cell models demand a more comprehensive knowledge of biochemical mechanisms and are highly prone to overfitting due to their sheer complexity (Oberhardt and Ruppin 2013). An alternative strategy is to infer genotype – phenotype associations using machine learning of large datasets and use genome-scale metabolic models as tools to provide a mechanistic structure to these inferences (i.e. white-box machine learning) (Yang, et al.

2019). In chapter 2, I have already presented such a machine learning approach that generates biological hypotheses from large-scale genetic interaction data and simultaneously improves the metabolic model (Szappanos, et al. 2011). Below, I briefly discuss the prospect of integrating biochemical network modelling with machine learning to build interpretable models that better capture genetic variants that occur in natural populations.

The advent of population genomics yielded thousands of fully sequenced genomes of individuals from the same species. Genetic variation that exist in natural populations has the potential to illuminate the genetic basis of phenotypic traits using association studies, such as GWAS (Visscher, et al. 2017). A recent method combines mechanistic and machine learning genotype – phenotype models in an innovative way to identify the genetic basis of antimicrobial phenotypes from sequence data (Kavvas, et al. 2020). The method makes use of the fact that antimicrobial resistance phenotypes to certain drugs are causally linked to metabolic alterations. In brief, the method represents the allelic variants of each bacterial strain as a set of allele-specific flux constraints (i.e. upper / lower flux bounds) and optimally separates drug resistant and sensitive strains in the flux space using a variant of flux balance analysis. Thus, the framework is a machine learning classifier that infers allele-specific flux effects underlying resistance. As such it also provides causal biochemical network explanation of the classification. The method classifies resistant and sensitive *Mycobacterium tuberculosis* strains with high accuracy and recapitulates known resistance mechanisms. Similar approaches, including methods that integrate various types of omics data (Culley, et al. 2020), hold great promise towards building genome-scale models that predict the phenotypic effects of natural variants.

**Acknowledgements**

I'm most indebted to Csaba Pál, Laurence D. Hurst, Steven G. Oliver and Bálint Kintses, from whom I've learned the most. I'm grateful to members of the Papp and Pál labs, especially Balázs Szappanos, Károly Kovács, Viktória Lázár and Bálint Csörgő, who have heavily contributed to the results presented here. I'm lucky to having worked with Martin Lercher and Richard Notebaart. I thank György Pósfai and Péter Horváth for fostering an outstanding and supportive community to do systems biology research in. My lab has been enjoying generous financial supports from the Hungarian Academy of Sciences, the Wellcome Trust, the BRC and the National Research, Development and Innovation Office. Last but not least, I feel extremely lucky to have a loving and supporting family.

**References**

Aharoni A, Gaidukov L, Khersonsky O, Mc QGS, Roodveldt C, Tawfik DS. 2005. The 'evolvability' of promiscuous protein functions. Nat Genet 37:73-76.

Babu M, Arnold R, Bundalovic-Torma C, Gagarinova A, Wong KS, Kumar A, Stewart G, Samanfar B, Aoki H, Wagih O, et al. 2014. Quantitative genome-wide genetic interaction screens reveal global epistatic relationships of protein complexes in Escherichia coli. PLoS Genet 10:e1004120.

Bandyopadhyay S, Mehta M, Kuo D, Sung MK, Chuang R, Jaehnig EJ, Bodenmiller B, Licon K, Copeland W, Shales M, et al. 2011. Rewiring of genetic networks in response to DNA damage. Science 330:1385-1389.

Barabasi AL, Oltvai ZN. 2004. Network biology: understanding the cell's functional organization. Nat Rev Genet 5:101-113.

Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ. 2004. Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15:3841-3862.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S, et al. 2010. The genetic landscape of a cell. Science 327:425-431.

Costanzo M, VanderSluis B, Koch EN, Baryshnikova A, Pons C, Tan G, Wang W, Usaj M, Hanchard J, Lee SD, et al. 2016. A global genetic interaction network maps a wiring diagram of cellular function. Science 353.

Cotten C, Reed JL. 2013. Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. BMC Bioinformatics 14:32.

Culley C, Vijayakumar S, Zampieri G, Angione C. 2020. A mechanism-aware and multiomic machine-learning pipeline characterizes yeast cell growth. Proceedings of the National Academy of Sciences 117:18869-18879.

Dang L, White DW, Gross S, Bennett BD, Bittinger MA, Driggers EM, Fantin VR, Jang HG, Jin S, Keenan MC, et al. 2009. Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature 462:739-744.

Darwin C. 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life: London : John Murray, 1859.

Dean AM, Thornton JW. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet 8:675-688.

Edwards JS, Ibarra RU, Palsson BO. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol 19:125-130.

Fehér T, Papp B, Pál C, Pósfai G. 2007. Systematic genome reductions: theoretical and experimental approaches. Chem Rev 107:3498-3513.

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121.

Feist AM, Palsson BO. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat Biotechnol 26:659-667.

Flachmann R, Kunz N, Seifert J, Gutlich M, Wientjes FJ, Laufer A, Gassen HG. 1988. Molecular biology of pyridine nucleotide biosynthesis in Escherichia coli. Cloning and characterization of quinolinate synthesis genes nadA and nadB. Eur J Biochem 175:221-228.

Fleming RMT, Thiele I, Provan G, Nasheuer HP. 2010. Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. Journal of Theoretical Biology 264:683-692.

Guzman GI, Sandberg TE, LaCroix RA, Nyerges A, Papp H, de Raad M, King ZA, Hefner Y, Northen TR, Notebaart RA, et al. 2019. Enzyme promiscuity shapes adaptation to novel growth substrates. Mol Syst Biol 15:e8462.

Harrison R, Papp B, Pal C, Oliver SG, Delneri D. 2007. Plasticity of genetic interactions in metabolic networks of yeast. Proc Natl Acad Sci U S A 104:2307-2312.

Hartman JLt, Garvik B, Hartwell L. 2001. Principles for the buffering of genetic variation. Science 291:1001-1004.

Horlbeck MA, Xu A, Wang M, Bennett NK, Park CY, Bogdanoff D, Adamson B, Chow ED, Kampmann M, Peterson TR, et al. 2018. Mapping the Genetic Landscape of Human Cells. Cell 174:953-967 e922.

Horowitz NH. 1945. On the Evolution of Biochemical Syntheses. Proc Natl Acad Sci U S A 31:153-157.

Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, Sorrelle N, Piechulla B, Barkman TJ. 2012. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. Proc Natl Acad Sci U S A 109:2966-2971.

Jensen RA. 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425.

Johannes TW, Zhao H. 2006. Directed evolution of enzymes and biosynthetic pathways. Curr Opin Microbiol 9:261-267.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27-30.

Kashtan N, Noor E, Alon U. 2007. Varying environments can speed up evolution. Proceedings of the National Academy of Sciences 104:13711-13716.

Kavvas ES, Yang L, Monk JM, Heckmann D, Palsson BO. 2020. A biochemically-interpretable machine learning classifier for microbial GWAS. Nature Communications 11:2580.

Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79:471-505.

Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD. 2010. Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5'-phosphate synthesis. Mol Syst Biol 6:436.

King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, et al. 2009. The automation of science. Science 324:85-89.

Kumari S, Tishel R, Eisenbach M, Wolfe AJ. 1995. Cloning, characterization, and functional expression of acs, the gene which encodes acetyl coenzyme A synthetase in Escherichia coli. J Bacteriol 177:2878-2886.

Kuznetsova E, Proudfoot M, Gonzalez CF, Brown G, Omelchenko MV, Borozan I, Carmel L, Wolf YI, Mori H, Savchenko AV, et al. 2006. Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. J Biol Chem 281:36149-36161.

Lloyd CJ, Ebrahim A, Yang L, King ZA, Catoiu E, O'Brien EJ, Liu JK, Palsson BO. 2018. COBRAme: A computational framework for genome-scale models of metabolism and gene expression. PLoS Comput Biol 14:e1006302.

Lynch M, Abegg A. 2010. The rate of establishment of complex adaptations. Mol Biol Evol 27:1404-1414.

Mih N, Brunk E, Bordbar A, Palsson BO. 2016. A Multi-scale Computational Platform to Mechanistically Assess the Effect of Genetic Variation on Drug Responses in Human Erythrocyte Metabolism. PLoS Comput Biol 12:e1005039.

Mulleder M, Calvani E, Alam MT, Wang RK, Eckerstorfer F, Zelezniak A, Ralser M. 2016. Functional Metabolomics Describes the Yeast Biosynthetic Regulome. Cell 167:553-565 e512.

Notebaart RA, Kintses B, Feist AM, Papp B. 2018. Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol 49:108-114.

Notebaart RA, Szappanos B, Kintses B, Pal F, Gyorkei A, Bogos B, Lazar V, Spohn R, Csorgo B, Wagner A, et al. 2014. Network-level architecture and the evolutionary potential of underground metabolism. Proc Natl Acad Sci U S A 111:11762-11767.

O'Brien EJ, Lerman JA, Chang RL, Hyduke DR, Palsson BØ. 2013. Genome-scale models of metabolism and gene expression extend and refine growth phenotype prediction. Molecular Systems Biology 9:693.

Oberhardt M, Ruppin E. 2013. Taming the complexity of large models. EMBO reports 14:848-848.

Oberhardt MA, Palsson BO, Papin JA. 2009. Applications of genome-scale metabolic reconstructions. Mol Syst Biol 5:320.

Pal C, Papp B. 2017. Evolution of complex adaptations in molecular systems. Nat Ecol Evol 1:1084-1092.

Pál C, Papp B, Lercher MJ. 2005. Horizontal gene transfer depends on gene content of the host. Bioinformatics 21 Suppl 2:ii222-ii223.

Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. 2006. Chance and necessity in the evolution of minimal metabolic networks. Nature 440:667-670.

Papp B, Notebaart RA, Pal C. 2011. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet 12:591-602.

Papp B, Pál C, Hurst LD. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature 429:661-664.

Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. 2007. Multicopy suppression underpins metabolic evolvability. Mol Biol Evol 24:2716-2722.

Price ND, Reed JL, Palsson BO. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2:886-897.

Remold SK, Lenski RE. 2004. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. Nat Genet 36:423-426.

Rison SC, Teichmann SA, Thornton JM. 2002. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. J Mol Biol 318:911-932.

Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, Sohngen C, Stelzer M, Thiele J, Schomburg D. 2011. BRENDA, the enzyme information system in 2011. Nucleic Acids Res 39:D670-676.

Schmidt S, Sunyaev S, Bork P, Dandekar T. 2003. Metabolites: a helping hand for pathway evolution? Trends Biochem Sci 28:336-341.

Schomburg I, Chang A, Schomburg D. 2002. BRENDA, enzyme data and metabolic information. Nucleic Acids Res 30:47-49.

Shlomi T, Eisenberg Y, Sharan R, Ruppin E. 2007. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. Mol Syst Biol 3:101.

Snitkin ES, Dudley AM, Janse DM, Wong K, Church GM, Segre D. 2008. Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. Genome Biol 9:R140.

Sommer MOA, Munck C, Toft-Kehler RV, Andersson DI. 2017. Prediction of antibiotic resistance: time for a new preclinical paradigm? Nat Rev Microbiol 15:689-696.

Szappanos B, Fritzemeier J, Csorgo B, Lazar V, Lu X, Fekete G, Balint B, Herczeg R, Nagy I, Notebaart RA, et al. 2016. Adaptive evolution of complex innovations through stepwise metabolic niche expansion. Nat Commun 7:11607.

Szappanos B, Kovacs K, Szamecz B, Honti F, Costanzo M, Baryshnikova A, Gelius-Dietrich G, Lercher MJ, Jelasity M, Myers CL, et al. 2011. An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 43:656-662.

Teusink B, Walsh MC, van Dam K, Westerhoff HV. 1998. The danger of metabolic pathways with turbo design. Trends Biochem Sci 23:162-169.

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 2017. 10 Years of GWAS Discovery: Biology, Function, and Translation. The American Journal of Human Genetics 101:5-22.

Wagner A. 2008. Neutralism and selectionism: a network-based reconciliation. Nat Rev Genet 9:965-974.

Wang L, Maranas CD. 2018. MinGenome: An In Silico Top-Down Approach for the Synthesis of Minimized Genomes. ACS Synth Biol 7:462-473.

Weissman DB, Desai MM, Fisher DS, Feldman MW. 2009. The rate at which asexual populations cross fitness valleys. Theoretical Population Biology 75:286-300.

Yang JH, Wright SN, Hamblin M, McCloskey D, Alcantar MA, Schrübbers L, Lopatkin AJ, Satish S, Nili A, Palsson BO, et al. 2019. A White-Box Machine Learning Approach for Revealing Antibiotic Mechanisms of Action. Cell 177:1649-1661.e1649.

Yizhak K, Benyamini T, Liebermeister W, Ruppin E, Shlomi T. 2010. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics 26:i255-260.

Yizhak K, Tuller T, Papp B, Ruppin E. 2011. Metabolic modeling of endosymbiont genome reduction on a temporal scale. Mol Syst Biol 7:479.

You L, Yin J. 2002. Dependence of epistasis on environment and mutation severity as revealed by in silico mutagenesis of phage t7. Genetics 160:1273-1281.

## Appendix


1) Szappanos, B., Kovács, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pál, C., **Papp, B.** (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genetics* **43:** 656

2) *Harrison, R., ***Papp, B.**, Pál, C., Oliver, S.G., Delneri, D. (2007) Plasticity of genetic interactions in metabolic networks of yeast. *Proc Natl Acad Sci U S A.* **104**: 2307-12.

3) *Pál, C., ***Papp, B.**, Lercher, M.J., Csermely, P., Oliver, S.G. and Hurst, L.D. (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**: 667-70.

4) Notebaart, R.A.*, Szappanos, B., Kintses, B., Pál, F., Györkei, A., Bogos, B., Lázár, V., Spohn, R., Csörgő, B., Wagner, A., Ruppin, E., Pál, C.*, **Papp, B.*** (2014) Network-level architecture and the evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A.* **111**: 11762-11767.

5) Szappanos, B., Fritzemeier, J.C., Csörgő, B., Lázár, V., Lu, X., Fekete, G., Bálint, B., Herczeg, R., Nagy, I., Notebaart, R.A., Lercher, M.J., Pál, C.*, **Papp, B.*** (2016) Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat Commun.* **7**:11607

nature genetics

# An integrated approach to characterize genetic interaction networks in yeast metabolism

Balázs Szappanos[1,10], Károly Kovács[1,10], Béla Szamecz[1], Frantisek Honti[1,2], Michael Costanzo[3,4], Anastasia Baryshnikova[3,4], Gabriel Gelius-Dietrich[5], Martin J Lercher[5], Márk Jelasity[6], Chad L Myers[7], Brenda J Andrews[3,4], Charles Boone[3,4], Stephen G Oliver[8], Csaba Pál[1] & Balázs Papp[1,9,10]

© 2011 Nature America, Inc. All rights reserved.

**Although experimental and theoretical efforts have been applied to globally map genetic interactions, we still do not understand how gene-gene interactions arise from the operation of biomolecular networks. To bridge the gap between empirical and computational studies, we i, quantitatively measured genetic interactions between ~185,000 metabolic gene pairs in *Saccharomyces cerevisiae,* ii, superposed the data on a detailed systems biology model of metabolism and iii, introduced a machine-learning method to reconcile empirical interaction data with model predictions. We systematically investigated the relative impacts of functional modularity and metabolic flux coupling on the distribution of negative and positive genetic interactions. We also provide a mechanistic explanation for the link between the degree of genetic interaction, pleiotropy and gene dispensability. Last, we show the feasibility of automated metabolic model refinement by correcting misannotations in NAD biosynthesis and confirming them by *in vivo* experiments.**

Recent large-scale genetic analyses of yeast have enabled the systematic screening of pairwise genetic interactions and provided valuable insights into the functional organization of a eukaryotic cell[1] as well as genetic networks underlying specific biological processes[2,3]. Despite the rapid growth in quantitative data on genetic interactions, we still have only a limited understanding of the molecular mechanisms through which one mutation modifies the phenotypic effect of another. Furthermore, although the general properties of genetic interaction networks have been explored phenomenologically[1,4], we often lack a mechanistic understanding of these patterns. For example, a recent large-scale study reported that single mutants with severe fitness defects tend to have numerous genetic interactions[1], a phenomenon that still awaits explanation. Finally, the systematic generation of biological hypotheses from the welter of phenotypic data produced by interaction screens remains a major challenge. By examining how cellular phenotypes arise from the operation of molecular networks, systems biology offers great promise for meeting these challenges.

Metabolism is one of the best characterized cellular subsystems and is especially suited for system-level studies of the genotype-phenotype relationship and, hence, genetic interactions. This is because first, high-quality metabolic network reconstructions are available that specify the chemical reactions catalyzed by hundreds of enzymes and cover the molecular function for a substantial fraction of the genome (for example, 15% in yeast)[5]. Second, these reconstructions can be converted into computational models to calculate the phenotype of both wild-type and mutant cells using constraint-based analysis tools[6] such as flux balance analysis (FBA). This imposes mass balance and capacity constraints to define the space of feasible steady-state flux distributions of the network and then identifies optimal network states that maximize biomass yield, a proxy for growth. Despite its simplicity and low data requirements, this modeling framework has shown great predictive power and has been successfully applied to various research problems[7], including predicting the viability of single-gene deletants[8] and model-driven analysis of high-throughput data[8–10]. Although some properties of genetic interaction networks have also been addressed using FBA, these earlier studies were exclusively[11,12] or mainly[13,14] theoretical because of the lack of large-scale genetic interaction data for metabolic genes.

To bridge the gap between theory and experiment, we have systematically measured genetic interactions between pairs of metabolic genes in yeast and combined these data with a detailed metabolic network reconstruction. Quantitative measurement of the fitness of single and double mutants has enabled us to detect both negative (aggravating) and positive (alleviating) interactions (that is, the double mutant has a lower or higher fitness, respectively, than would be expected from the product of the single-mutant fitnesses).

[1]Institute of Biochemistry, Biological Research Centre, Szeged, Hungary. [2]Department of Biology and Biochemistry, University of Bath, Bath, UK. [3]Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [4]Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada. [5]Department of Computer Science, Heinrich-Heine-University, Düsseldorf, Germany. [6]Research Group on Artificial Intelligence, University of Szeged and HAS, Szeged, Hungary. [7]Department of Computer Science & Engineering, University of Minnesota, Minneapolis, Minnesota, USA. [8]Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, UK. [9]Cambridge Systems Biology Centre and Department of Genetics, University of Cambridge, Cambridge, UK. [10]These authors contributed equally to this work. Correspondence should be addressed to B.P. (pappb@brc.hu).

Received 30 December 2010; accepted 5 May 2011; published online 29 May 2011; doi:10.1038/ng.846

Our integrated approach had three major goals. First, we investigated the distribution of genetic interactions within and across functional modules as defined by classical annotation groups and network-based mathematical methods. Second, we performed constraint-based analysis of the network to simulate mutational effects and predict interactions *in silico*. We then employed our *in vivo* interaction data to test the model's ability to capture the general properties of genetic interaction networks and to assess the validity of its specific predictions. Third, we automated the reconciliation of empirical interaction data with model predictions and used discrepancies to update the metabolic network and direct biological discovery.

## RESULTS

### Constructing a genetic interaction map of yeast metabolism

We selected genes for our genetic interaction map based on an updated reconstruction of the *S. cerevisiae* metabolic network, which consists of 1,412 reactions and accounts for 904 genes[10]. Genetic interaction data has been generated by large-scale synthetic genetic array (SGA) technology[15]. First, we performed new screens to construct a map that covers all major metabolic subsystems, except for transfer RNA aminoacylation. The screens involved construction of high-density arrays of double mutants by crossing 613 query mutants, including 78 hypomorphic alleles of essential genes, against an array of 470 null mutants, producing double mutants for 184,624 unique gene pairs. The fitness of single and double mutants was assessed quantitatively by measuring colony size[16]. We calculated interaction scores ($\varepsilon$) based on the deviation of the double-mutant fitness ($f_{12}$) from the product of the corresponding single-mutant fitnesses ($\varepsilon = f_{12} - f_1 \cdot f_2$)[17]. Second, we supplemented our measurements with data from our recent large-scale genetic interaction screen[1], which employed the same experimental procedure as the present study but represented genes in all functional categories, including metabolism.

Overall, our combined dataset covers more than 80% of metabolic network genes, including 82 essential genes, and provides interaction scores for 215,907 pairs, 57% of which have been independently screened more than once. Applying a previously defined confidence threshold that proved informative in functional analyses[1], we detected 3,572 negative and 1,901 positive interactions (Online Methods). We focused on interactions between null mutations of non-essential genes (176,821 pairs) because of their better coverage and easier interpretation; data on essential genes has only been used for specific analyses. Additionally, we also defined a high-confidence interaction set based on the reproducibility of replicate experiments and used it when very low false-positive rates were required.

### Genetic interactions are frequent between functional modules

We took advantage of our quantitative genetic interaction map to empirically test earlier predictions about the distribution of interactions within and between metabolic functional modules. Specifically, a computational study based on FBA suggested that i, genetic interactions are enriched within metabolic annotation groups, and ii, interactions between different functional groups tend to be either exclusively negative or exclusively positive, a property termed 'monochromaticity'[11].

First, we report a modest, but significant, enrichment of both negative (1.6-fold, $P < 10^{-3}$) and positive (2.5-fold, $P < 10^{-15}$) interactions within classically defined functional modules. For example, lipid metabolism is especially enriched in genetic interactions, with sterol metabolism and fatty acid biosynthesis being primarily enriched in positive interactions, and both forms of interactions are overrepresented in sphingolipid metabolism (**Fig. 1**). Notably, the enrichments remain after controlling for potential confounding variables, such as paralogy[18], physical interaction[3] or single-mutant fitness[1] (Online Methods), and become more pronounced when using the high-confidence interaction set (3.8-fold and 8.7-fold enrichment of negative and positive interactions, respectively). However, as **Figure 1** shows, the majority of genetic interactions occur between genes assigned to different metabolic functions (93% of negative and 90% of positive, or 86% and 73%, respectively, when using high-confidence interactions). The fact that even strongly enriched functional groups, such as fatty acid biosynthesis, have numerous interactions with other groups indicates widespread pleiotropy across metabolic subsystems.

Next, we asked whether interactions between different functional groups tend to

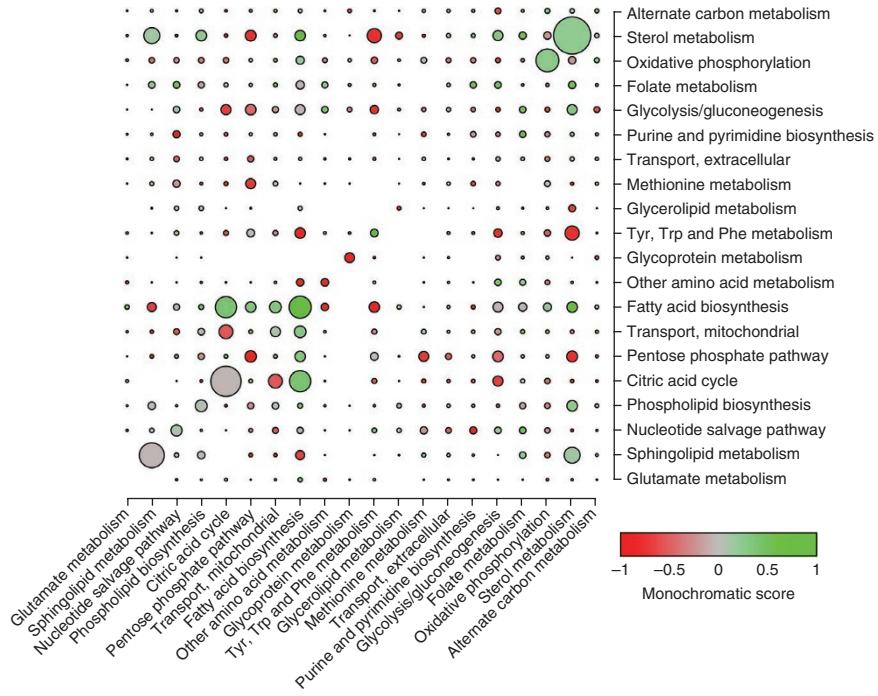



**Figure 1** Distribution and monochromaticity of genetic interactions between functional groups. The radii of the circles represent the fraction of screened gene pairs that show genetic interaction within and between functional annotation groups (for example, sterol metabolism has the highest prevalence of interactions with a value of 0.225). Enrichment of genetic interactions within functional groups is visually apparent and corresponds to the larger circles on the diagonal. The colors of the circles reflect the monochromatic score defined as the normalized ratio of positive to all interacting pairs (Online Methods). Functional groups displaying only positive genetic interactions between each other have a monochromatic score of +1 (green), whereas those interacting purely negatively have a score of −1 (red). The background ratio of positive to all interactions (0.348) corresponds to a score of 0 (gray). Only the top 20 functional groups with the largest number of screened gene pairs and those genes assigned to only one functional group are included in the plot.

© 2011 Nature America, Inc. All rights reserved.

be either exclusively negative or positive. In agreement with theoretical predictions[11], we found a statistically significant excess of monochromaticity among pairs of functional groups in the real data compared to randomized interaction maps ($P < 10^{-4}$). For example, whereas sterol metabolism displays almost purely negative interactions with tyrosine, tryptophan and phenylalanine metabolism, it predominantly interacts positively with fatty acid biosynthesis (**Fig. 1**). Nevertheless, monochromaticity in our genetic interaction map is modest: only ~24–34% more monochromatic pairs were found than expected by chance, a conclusion that remained qualitatively the same when using high-confidence interactions (**Supplementary Table 1**).

As an alternative to functional groups defined based on classical biochemical pathways, unbiased mathematical methods have been developed to measure functional relatedness based on coherent usage of reactions in the metabolic network[6,19]. In particular, flux coupling[20] provides a biochemically sound definition of functional relatedness and has strong physiological and evolutionary relevance[21–23]. To further investigate the distribution of genetic interactions within and between functional modules, we identified flux-coupled gene pairs computationally (that is, pairs of reactions where the activity of one reaction implies the activity of the other, either reciprocally or in one direction; Online Methods). In agreement with results obtained using annotation groups, although we find that both negative (twofold) and positive (2.7-fold) interactions are enriched in flux-coupled pairs ($P < 10^{-6}$ and $P < 10^{-8}$, respectively), the overwhelming majority (> 97%) of both forms of interactions occur between uncoupled genes, even when only high-confidence interactions are investigated (> 93%).

In conclusion, both definitions of functional relatedness reveal that most genetic interactions connect across distinct functional modules, extending an earlier estimate that synthetic lethal interactions are 3.5 times more likely to span pairs of protein-protein interaction pathways than to occur within such pathways[24]. Furthermore, our finding that both negative and positive interactions tend to occur between metabolic modules is consistent with recent observations that both forms of interactions primarily connect genes belonging to different protein complexes[1,16].

## A systems model explains genetic interaction connectivity

To further explore the organizational principles of the genetic interaction network, we next investigated its degree distribution using a computational model of metabolism. A prominent attribute of genetic interaction networks, also shared by other biological networks[25], is that the majority of genes show few interactions, and a minority of 'hub' genes are highly connected[1,4]. Furthermore, a recent study uncovered a strong correlation between the number of genetic interactions a gene shows and the fitness defect associated with its deletion (dispensability)[1], a pattern also confirmed by our empirical metabolic interaction map (**Supplementary Fig. 1**). Nevertheless, the tendency of 'sick' single mutants to engage in an especially high number of both negative and positive interactions remains unexplained. Intuitively, one expects that a strongly deleterious single mutation can mask a large number of mildly deleterious mutations in other genes and, hence, show numerous positive interactions. However, a similar logic would imply a paucity of negative interactions for sick mutants (meaning a sick deletant is less likely to be made worse by other mutations), an expectation that is inconsistent with observations[1].

To probe whether a simple structural model of metabolism is able to capture the above properties of genetic interaction networks, we computed *in silico* interaction degrees and single-mutant fitness using FBA. Similar to the empirical data, *in silico* genetic interaction degree is also unevenly distributed, with only ~12% of genes accounting for the majority (~85%) of interactions. Most remarkably, the model predicted a strong negative correlation between single-mutant fitness and genetic interaction degree for both positive and negative interactions, confirming the trend observed in the experimentally derived genetic interaction network (Spearman's $\rho = -0.89$ and $\rho = -0.66$, respectively). Notably, these trends remained when genes without any *in silico* fitness contribution were excluded from the analysis ($\rho = -0.59$, $P < 10^{-3}$ for positive interactions and $\rho = -0.47$, $P = 0.005$ for negative interactions; **Fig. 2a**), showing that the associations are not simply caused by the presence of silent reactions in the metabolic model.

Having established its ability to capture the high genetic interaction connectivity of sick mutants, we asked the metabolic model to provide mechanistic explanations. One reason why a gene might have numerous genetic interactions is that it contributes to multiple biological processes (that is, it is highly pleiotropic), and hence, the phenotypic effect of its deletion may be modulated by a large number of other genes, each of them negatively or positively affecting a different aspect of its functionality. Indeed, it has been reported that genetic interaction hubs often display multifunctionality[1]. If highly

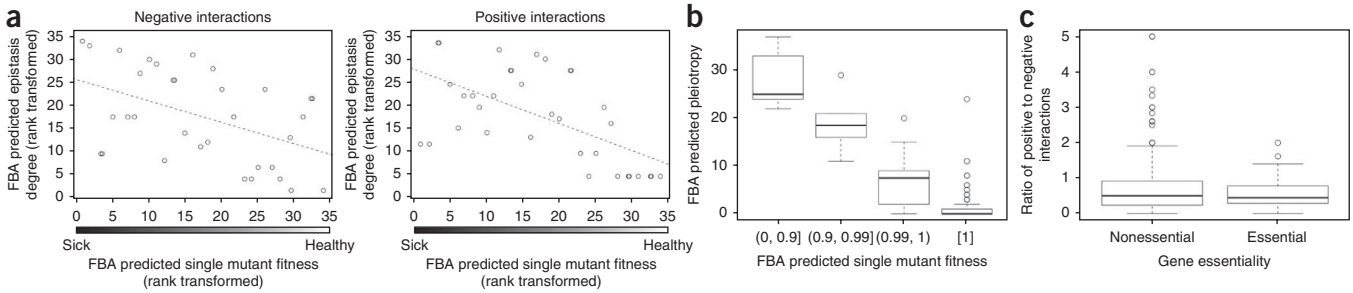



**Figure 2** Degree distribution of genetic interaction networks and gene dispensability. (**a**) Both negative and positive genetic interaction degrees predicted by FBA showed negative correlations with predicted single-mutant fitness. Only genes with nonzero *in silico* fitness defects are shown, and variables are rank transformed. See the Online Methods for details on selecting independent data points (genes) for the statistical analysis. To improve the visual representation of coincident data points, we added a small amount of noise over the *x* axis for plotting. (**b**) The FBA-predicted single-gene deletion effect is strongly associated with predicted system-level pleiotropy degree (that is, the number of biosynthetic processes to which a gene contributes). See the Online Methods for details on the gene selection procedure. (**c**) Comparison of the empirically determined positive-to-negative genetic interaction ratio between null mutants of non-essential genes and hypomorphic alleles of essential genes revealed no significant difference. Horizontal lines of the boxplots correspond to the medians, and the bottoms and the tops of the boxes show the twenty-fifth and seventy-fifth percentiles, respectively. Whiskers show either the maximum (minimum) value or 1.5 times the interquartile range of the data, whichever is smaller (larger). Points more than 1.5 times the interquartile range above the third quartile or below the first quartile are plotted individually as outliers.

© 2011 Nature America, Inc. All rights reserved.

pleiotropic genes also have (on average) a large fitness contribution, then we would expect a negative correlation between single-mutant fitness and interaction degree. Although pleiotropy is difficult to define empirically, the FBA framework offers a rigorous approach to compute pleiotropy and test this idea. To do this, we determined the number of key metabolites (so called biomass components, including amino acids, nucleotides, and so on) whose maximal production is affected by the absence of each gene (Online Methods)[26]. In accordance with our hypothesis, we found a strong association between the number of biosynthetic processes to which a gene contributes and the predicted fitness of its deletant ($\rho = -0.83$, $P < 10^{-9}$ on raw data for genes with a nonzero deletion effect; see also **Fig. 2b**). Moreover, pleiotropy correlates with both *in silico* and *in vivo* genetic interaction degrees (negative degree: $\rho = 0.55$ and $\rho = 0.24$; positive degree: $\rho = 0.62$ and $\rho = 0.25$, respectively; $P < 10^{-8}$ in all cases). Given the close association between computationally derived single-mutant fitness and pleiotropy, we next performed partial correlation analyses to disentangle the effects of these factors on *in silico* interaction degrees. Our multivariate analyses revealed that, although positive interaction degree is determined by single-mutant fitness (a finding consistent with the idea that severe mutations can mask numerous milder mutations), negative interaction degree is driven by pleiotropy (**Supplementary Table 2**).

Taken together, these computational results suggest that the structure of the metabolic network dictates both the fitness contribution (and hence positive interaction degree) and the functional pleiotropy (and hence negative interaction degree) of genes. Future empirical studies of pleiotropy will help to clarify whether these mechanisms also adequately explain *in vivo* genetic interaction degrees.

### No prevalent positive interactions in essential genes

A recent FBA study suggested that non-lethal mutations in essential metabolic genes have strikingly different interaction patterns compared to null mutations of non-essential genes[14]. Specifically, it was predicted that essential metabolic genes frequently show positive interactions with other metabolic genes regardless of their function or the latter's essentiality, strongly skewing the ratio of positive to negative interactions. Although a small-scale empirical analysis was consistent with this prediction[14], it remained to be seen whether it was supported by large-scale experiments. Accordingly, we mapped genetic interactions between hypomorphic alleles[2] of a set of essential genes and null mutants of non-essential genes, screening 39,086 pairs. If positive interactions were indeed highly abundant between gene pairs involving an essential reaction, then we should observe a strong bias toward positive interactions for essential genes. Although we found that essential genes have an increased number of positive interactions, they also show more negative interactions, and therefore their ratio of positive to negative interactions is virtually identical to those of non-essential genes (Wilcoxon test $P = 0.89$; **Fig. 2c**). In sum, we failed to find empirical evidence for the predicted high prevalence of positive genetic interactions for essential metabolic genes. Given that the only experimental study reporting abundant positive interactions investigated only a handful of non-metabolic essential genes[14], we speculate that the discrepancy between the small-scale study[14] and our results could partly be because of sampling bias in the former.

### Fine-scale evaluation of predicted genetic interactions

Our comprehensive genetic interaction map provides an unprecedented opportunity to assess the FBA framework's ability to predict individual interactions. To rigorously estimate the fraction of true predicted interactions (precision) and the fraction of experimentally observed interactions that are captured by the model (recall or true-positive rate), we selected a set of high-confidence empirical interactions between non-essential genes (Online Methods) and excluded genes that are associated with poorly characterized network parts (blocked reactions[20]). This resulted in 325 negative and 116 positive interactions among 67,517 non-essential gene pairs. We found that experimentally identified interactions are highly overrepresented among predicted strong interactions, with up to 100-fold and 60-fold enrichment for negative and positive interactions, respectively (with precision values of 50% and 11%, respectively; **Fig. 3**). Although this confirms that the highest predicted interaction scores have high physiological relevance[13], we find that only a minority of empirical interactions are captured by the model at the same cutoff points (the recall values were 2.8% and 12.9% for negative and positive interactions, respectively), a conclusion that remained unchanged when an alternative algorithm[27], an alternative interaction score[11] or a less compartmentalized metabolic model[28] was employed to compute interactions (**Supplementary Figs. 2a–c**). Notably, only a minority of gene pairs that show negative (7.6%) or positive (3%) interactions *in vivo* display nonzero interaction scores of the opposite sign *in silico*, indicating that the low recall of the model stems from missed genetic interactions, not from misclassification of the two forms of interactions.

Why are so many genetic interactions missed by the model? First, as single-mutant fitness predictions are far from perfect[8,10], one might expect that interaction between two non-essential genes could be missed simply because one or the other gene is essential in the model. Indeed, ~24% of negative and ~22% of positive interactions are missed because of misprediction of single-mutant viability. Although the true-positive rate of genetic interaction predictions slightly improves when genes falsely predicted to be essential are excluded, the majority of empirical interactions are still not captured by the model. In particular, FBA predicts strong negative interaction scores for only 3.7% of *in vivo* negative interactions, indicating that it overpredicts double-mutant fitness in the majority of these gene pairs. Second, weak *in vivo* genetic interactions might be inherently less reproducible by the metabolic model. Although this idea is supported by an improved true-positive rate for strong *in vivo* interactions (~17% for $\varepsilon \leq -0.5$ and 25% for $\varepsilon \geq 0.15$), we conclude that even the strongest interactions are frequently missed by the model. Third, FBA predicts optimal

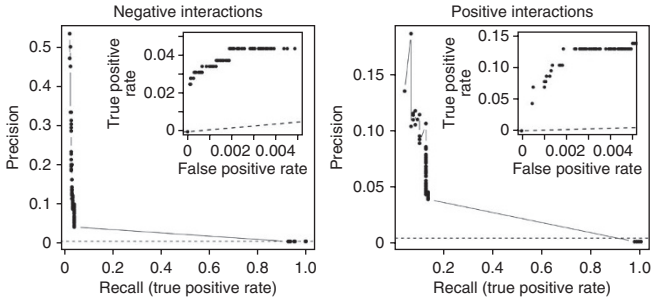



**Figure 3** Comparison of computationally predicted and empirically determined genetic interactions. We evaluated prediction accuracy by visualizing the trade-off between precision (fraction of predicted interactions that are supported by empirical data) and recall (fraction of empirical interactions that are successfully identified by the model), and true-positive and false-positive rates (partial receiver operating characteristic (ROC) curves, inset) at different *in silico* genetic interaction score cutoffs. Dashed lines represent the levels of discrimination expected by chance. Note the different scale of the *y* axes for the negative and positive interactions.

© 2011 Nature America, Inc. All rights reserved.

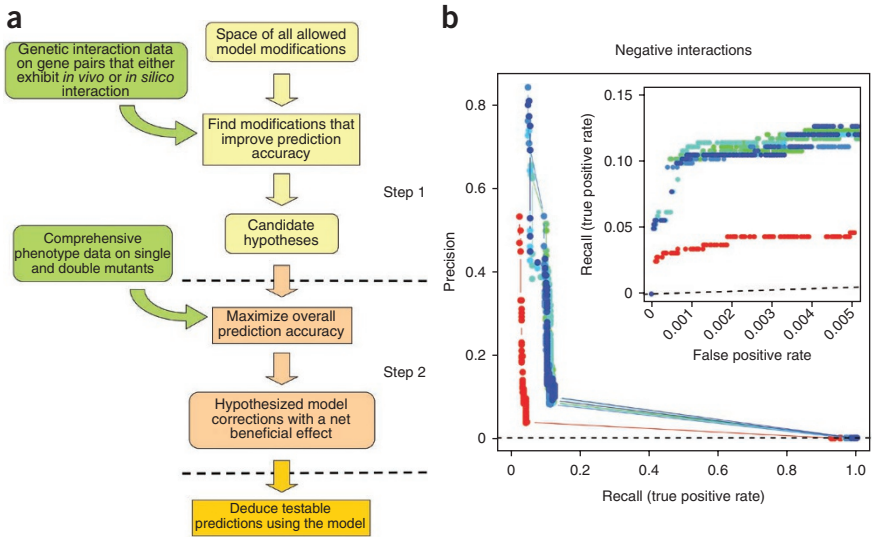



**Figure 4** Automated model refinement procedure. (**a**) Workflow of the two-stage model refinement method. In the first stage, a coarse-grained search is executed in which candidate models are evaluated only for those gene pairs that show interaction either *in vivo* or *in silico* according to the original model. In the second stage, the best models are refined in a restricted search space that is based on the results of the first stage but using all available data to evaluate the models. This two-stage approach made it feasible to explore a large space of candidate hypotheses while also making use of all available phenotypic data. (**b**) Results of eight independent runs of the model refinement algorithm. Fits of the modified (blue to green) and unmodified original (red) models to our empirical genetic interaction data are visualized by both precision recall and partial ROC curves (inset). Dashed lines represent the levels of discrimination expected by chance. Note that the same empirical dataset was used for both model refinement and model evaluation, meaning no unseen test data was used to generate these plots. For a cross-validation estimate of model improvement, see the main text and the **Supplementary Note**.

© 2011 Nature America, Inc. All rights reserved.

metabolic behavior without incorporating regulatory mechanisms. Consequently, reactions that are downregulated *in vivo* could nevertheless compensate deletions in other parts of the network *in silico*, and therefore the model likely underestimates mutational effects. To address this possibility, we used published quantitative transcriptome data[29] to identify non-expressed metabolic genes and constrained the corresponding reaction activities to zero in the simulations[30]. Imposing transcriptional constraints did not noticeably improve predictions (**Supplementary Fig. 2d**), suggesting that detailed information on other layers of regulation[31] (for example, metabolic regulation[32]), data on toxic intermediates and more sophisticated modeling frameworks (for example, regulatory FBA[33]) are needed to probe the performance limits of genome-scale models. Finally, aside from the limitations of FBA, some false predictions likely indicate incomplete knowledge or annotation errors in the metabolic network.

Numerous statistical methods have been proposed to predict genetic interactions by combining heterogeneous sources of genomic and functional data (for example, sequence homology, physical interaction, co-expression and so on)[34,35]. These statistical approaches serve complementary roles to FBA. Whereas biochemical modeling has the advantage of easy interpretability and offers direct mechanistic insights, statistical models may illuminate the amount of information available in large-scale datasets to predict genetic interactions. Thus, we asked whether such methods may substantially improve our knowledge of genetic interactions in the metabolic network.

To assess the performance of statistical modeling, we first compiled a dataset of gene-pair characteristics (following earlier studies[34,35] and based on metabolic network features but omitting any information on genetic interactions; **Supplementary Note**) and used data-mining methods (random forest[36] and logistic regression) to classify genetic interactions based on these features. Although an increased fraction of *in vivo* interactions could be retrieved, ~70% of negative and ~75% of positive interactions were still predicted with very low (<10%) precision (**Supplementary Fig. 3**). Thus, we conclude that the majority of genetic interactions are not well understood either in terms of biochemical processes or statistical associations. Notably, incorporating FBA-derived fitness and genetic interaction scores into statistical models boosts the precision of negative interaction predictions (**Supplementary Fig. 3**), indicating that biochemical modeling provides unique information that is not captured by purely statistical data integration.

## Automated model refinement using genetic interaction data

To reconcile discrepancies between empirical and computational genetic interaction maps, we developed a machine-learning method that automatically generates hypotheses to explain *in vivo* compensation (negative interaction) between genes. In contrast with a previously proposed approach[37] that reconciled experimental and computational growth data mutant by mutant, we sought to minimize model mispredictions globally (that is, using all available data) by using a two-stage genetic algorithm (**Fig. 4a** and **Supplementary Note**). The following types of changes to the model were allowed[37]: i, modifying reaction reversibility, ii, removing reactions and iii, altering the list of biomass compounds required for growth (**Supplementary Note**).

Our automated method suggested several modifications (**Supplementary Table 3**) that, together, considerably improved the fit of the model to our genetic interaction map (100–267% increase in recall and 44–59% increase in precision; **Fig. 4b**). Notably, cross-validation confirmed that our method also significantly ($P < 0.002$) improves the model's ability to predict genetic interactions that were not used in model refinement (with recall increased by ~87% on average; **Supplementary Note**).

As an example of a modification suggested by our method, it showed that omitting glycogen from the set of essential biomass components corrects two falsely predicted genetic interactions. This is congruent with glycogen's role as a reserve carbohydrate, which becomes important in nutrient-depleted or stress conditions[38]. Remarkably, our algorithm also revealed that removal of only one or two reactions from the network corrects the prediction of four negative interactions between alternative NAD biosynthesis pathways. In particular, the published network reconstruction[10] contains three biosynthetic routes for NAD, and removing the two-step path from aspartate to quinolinate uncovers pairwise compensation between the other two pathways (**Fig. 5a**). Notably, although *de novo* NAD synthesis from aspartate is present in *Escherichia coli*[39], it has no genes annotated in the yeast network, and bioinformatics analyses failed to find yeast homologs of the *E. coli*

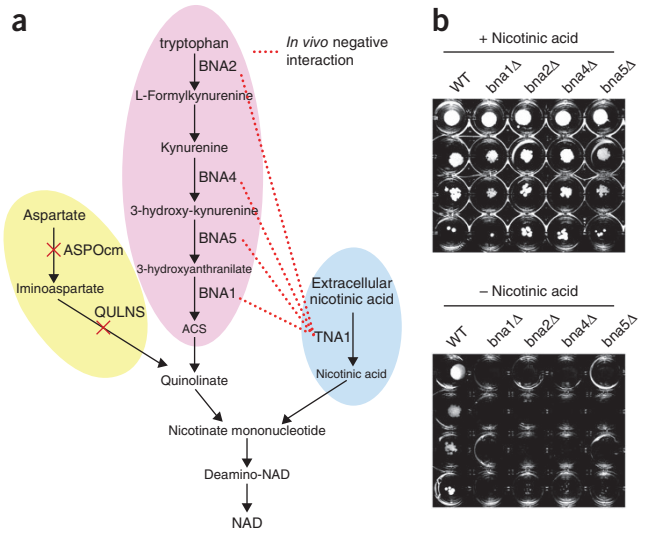



**Figure 5** Automated model refinement suggests modifications in NAD biosynthesis. (**a**) Biosynthetic routes to nicotinate mononucleotide in the yeast metabolic network reconstruction. Genes involved in the *de novo* pathway from tryptophan show negative genetic interactions with the nicotinic acid transporter gene *in vivo* but not *in silico* because of the presence of a two-step biosynthetic route from aspartate to quinolinate in the reconstruction (ASPOcm, aspartate oxidase; QULNS, quinolinate synthase). (**b**) Experimental verification of suggested model modifications. Deletion of genes for kynurenine pathway enzymes causes nicotinic acid auxotrophy. We spotted strains deleted for the genes of the kynurenine pathway (*bna1Δ*, *bna2Δ*, *bna4Δ* and *bna5Δ*) along with wild type (WT) in four serial dilutions on solid SC medium lacking histidine, arginine and lysine and incubated at 30 °C for 48 h in the presence and absence of nicotinic acid as indicated. To prevent diffusion of any substances that would complement nicotinic acid auxotrophy, the strains were grown separately from each other in a 24-well plate. Repeating the experiment using liquid media confirmed the nicotinic acid auxotrophy of the mutants (data not shown). Yeast strains used in the auxotrophy study are derivatives of the BY4741 yeast deletion collection[47,48].

enzymes (**Supplementary Note**). To further investigate whether quinolinate formation from aspartate might be wrongly included in the yeast reconstruction, we interrogated the metabolic model to deduce specific predictions for experimental testing. We found that only the refined model predicts the essentiality of genes in the kynurenine pathway (*BNA1*, *BNA2*, *BNA4* and *BNA5*) when nicotinic acid is absent from the medium. Next, we tested these predictions experimentally and confirmed that deletants of all four genes were nicotinic acid auxotrophs (**Fig. 5b**). Together, these results strongly suggest that the aspartate to NAD pathway is not present in yeast[40].

Our automated procedure identified additional erroneous predictions between NAD pathway genes and suggested further modifications (**Supplementary Table 3**), prompting us to thoroughly revise NAD biosynthesis in the published reconstruction. Based on inspection of interaction data, single-mutant phenotypes and literature information, we propose a number of changes including modifications of gene-reaction associations and reaction reversibilities (**Supplementary Fig. 4**). The revised model is not only consistent with literature data but also improves both interaction (12 corrections) and gene essentiality (1 correction) predictions.

## DISCUSSION

A system-level understanding of genetic interactions requires the integration of experimental and theoretical approaches. To progress toward this goal, we experimentally mapped interactions in yeast metabolism and systematically compared empirical data with predictions from a biochemical model. Our approach provides the first glimpse of genetic interactions in small-molecule metabolism and establishes the performance limits of a genome-scale metabolic model. We show that a simple structural model of metabolism captures several organizational properties of genetic interaction networks and suggests mechanistic hypotheses.

Notably, the computational model sheds new light on the relationship between the severity of mutational effects and genetic interactions. The FBA model not only captures the previously unexplained relationship between fitness effect and genetic interaction degree but also suggests a new mechanistic link between negative interaction degree and functional pleiotropy; the effect of mutations in pleiotropic genes may be modulated by mutations in a large number of other genes, each of them compensating a different aspect of the first gene's functionality.

Although we reported a coarse-grained consistency between model predictions and experiments, evaluation of individual interaction predictions revealed abundant discrepancies. In particular, FBA fails to capture the majority of experimentally determined genetic interactions, an attribute shared with statistical models built with data integration. Furthermore, interaction patterns of hypomorphic alleles of essential genes are grossly mispredicted, resulting in a discrepancy between our empirical data and a previous theoretical expectation about the high prevalence of positive interactions[14].

We can draw several conclusions from these inconsistencies. First, the quality and completeness of the metabolic reconstruction should be improved. Second, although null mutations can easily be represented in the FBA framework, simulation of hypomorphic alleles is inherently problematic as it hinges on assumptions about the relationship of enzyme activity to flux[41]. Third, the fact that a large number of *in vivo* instances of genetic interactions are not explained by the structure of the metabolic network suggests that regulation at both the gene expression and metabolite-enzyme levels should be taken into account in future attempts to realistically model metabolic behavior in genetically perturbed cells[42].

Most importantly, the comprehensive interaction map can be used to refine the metabolic model. Indeed, reconciling discrepancies between predicted and observed phenotypes is of central importance in developing systems biology models[43,44]. We showed the feasibility of an automated method to refine the metabolic model. We anticipate that similar approaches, coupled with high-throughput experimentation, have the potential to close the iterative cycles of generating and testing new hypotheses, leading to at least partial automation of biological discoveries[45,46].

**URLs.** Interaction data and modified metabolic reconstruction are available at http://www.utoronto.ca/boonelab/data/szappanos/; GLPK (GNU Linear Programming Kit); http://www.gnu.org/software/glpk/; CPLEX Optimizer, http://www-01.ibm.com/software/websphere/ilog/.

## METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

*Note: Supplementary information is available on the Nature Genetics website.*

ACKNOWLEDGMENTS

This work was supported by grants from The International Human Frontier Science Program Organization, the Hungarian Scientific Research Fund (OTKA PD 75261) and the 'Lendület Program' of the Hungarian Academy of Sciences

© 2011 Nature America, Inc. All rights reserved.

(B.P.), European Research Council (202591), Wellcome Trust and Hungarian Scientific Research Fund (C.P.), FEBS Long-Term Fellowship (B. Szamecz), Biotechnology & Biological Sciences Research Council (Grant BB/C505140/1) and the UNICELLSYS Collaborative Project (No. 201142) of the European Commission (S.G.O.), the US National Institutes of Health (1R01HG005084-01A1) and a seed grant from the University of Minnesota Biomedical Informatics and Computational Biology program (C.L.M.), the Canadian Institutes of Health Research (MOP-102629) (C.B. and B.J.A.) and the US National Institutes of Health (1R01HG005853-01) (C.B., B.J.A. and C.L.M.).

## AUTHOR CONTRIBUTIONS

M.C., C.L.M., B.J.A. and C.B. designed genetic interaction screens; A.B., M.C. and C.L.M. collected and analyzed raw data; B.P., C.P., M.J. and S.G.O. designed the computational study; B. Szappanos, K.K., F.H. and B.P. performed computational and statistical analyses; B. Szamecz performed auxotrophy experiments; G.G.-D. and M.J.L. developed software tools; and B.P., C.P., B. Szappanos, K.K., M.J. and S.G.O. wrote the paper.

## COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

© 2011 Nature America, Inc. All rights reserved.
npg

1. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
2. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**, 507–519 (2005).
3. Collins, S.R. *et al.* Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**, 806–810 (2007).
4. Tong, A.H. *et al.* Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813 (2004).
5. Feist, A.M., Herrgard, M.J., Thiele, I., Reed, J.L. & Palsson, B.O. Reconstruction of biochemical networks in microorganisms. *Nat. Rev. Microbiol.* **7**, 129–143 (2009).
6. Price, N.D., Reed, J.L. & Palsson, B.O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897 (2004).
7. Oberhardt, M.A., Palsson, B.O. & Papin, J.A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009).
8. Snitkin, E.S. *et al.* Model-driven analysis of experimentally determined growth phenotypes for 465 yeast gene deletion mutants under 16 different conditions. *Genome Biol.* **9**, R140 (2008).
9. Shlomi, T., Cabili, M.N., Herrgard, M.J., Palsson, B.O. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.* **26**, 1003–1010 (2008).
10. Mo, M.L., Palsson, B.O. & Herrgard, M.J. Connecting extracellular metabolomic measurements to intracellular flux states in yeast. *BMC Syst. Biol.* **3**, 37 (2009).
11. Segrè, D., Deluna, A., Church, G.M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83 (2005).
12. Deutscher, D., Meilijson, I., Kupiec, M. & Ruppin, E. Multiple knockout analysis of genetic robustness in the yeast metabolic network. *Nat. Genet.* **38**, 993–998 (2006).
13. Harrison, R., Papp, B., Pal, C., Oliver, S.G. & Delneri, D. Plasticity of genetic interactions in metabolic networks of yeast. *Proc. Natl. Acad. Sci. USA* **104**, 2307–2312 (2007).
14. He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nat. Genet.* **42**, 272–276 (2010).
15. Tong, A.H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
16. Baryshnikova, A. *et al.* Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat. Methods* **7**, 1017–1024 (2010).
17. Mani, R., St Onge, R.P., Hartman, J.L.t., Giaever, G. & Roth, F.P. Defining genetic interaction. *Proc. Natl. Acad. Sci. USA* **105**, 3461–3466 (2008).
18. DeLuna, A. *et al.* Exposing the fitness contribution of duplicated genes. *Nat. Genet.* **40**, 676–681 (2008).
19. Papin, J.A., Reed, J.L. & Palsson, B.O. Hierarchical thinking in network biology: the unbiased modularization of biochemical networks. *Trends Biochem. Sci.* **29**, 641–647 (2004).
20. Burgard, A.P., Nikolaev, E.V., Schilling, C.H. & Maranas, C.D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14**, 301–312 (2004).
21. Pál, C., Papp, B. & Lercher, M.J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37**, 1372–1375 (2005).
22. Bundy, J.G. *et al.* Evaluation of predicted network modules in yeast metabolism using NMR-based metabolite profiling. *Genome Res.* **17**, 510–519 (2007).
23. Notebaart, R.A., Teusink, B., Siezen, R.J. & Papp, B. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLOS Comput. Biol.* **4**, e26 (2008).
24. Kelley, R. & Ideker, T. Systematic interpretation of genetic interactions using protein networks. *Nat. Biotechnol.* **23**, 561–566 (2005).
25. Barabási, A.L. & Oltvai, Z.N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
26. Shlomi, T. *et al.* Systematic condition-dependent annotation of metabolic genes. *Genome Res.* **17**, 1626–1633 (2007).
27. Segrè, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
28. Kuepfer, L., Sauer, U. & Blank, L.M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430 (2005).
29. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344–1349 (2008).
30. Akesson, M., Forster, J. & Nielsen, J. Integration of gene expression data into genome-scale metabolic models. *Metab. Eng.* **6**, 285–293 (2004).
31. Daran-Lapujade, P. *et al.* The fluxes through glycolytic enzymes in *Saccharomyces cerevisiae* are predominantly regulated at posttranscriptional levels. *Proc. Natl. Acad. Sci. USA* **104**, 15753–15758 (2007).
32. Bouwman, J. *et al.* Metabolic regulation rather than de novo enzyme synthesis dominates the osmo-adaptation of yeast. *Yeast* **28**, 43–53 (2011).
33. Shlomi, T., Eisenberg, Y., Sharan, R. & Ruppin, E. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Mol. Syst. Biol.* **3**, 101 (2007).
34. Wong, S.L. *et al.* Combining biological networks to predict genetic interactions. *Proc. Natl. Acad. Sci. USA* **101**, 15682–15687 (2004).
35. Ulitsky, I., Krogan, N.J. & Shamir, R. Towards accurate imputation of quantitative genetic interactions. *Genome Biol.* **10**, R140 (2009).
36. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
37. Kumar, V.S. & Maranas, C.D. GrowMatch: an automated method for reconciling *in silico*/*in vivo* growth predictions. *PLOS Comput. Biol.* **5**, e1000308 (2009).
38. François, J. & Parrou, J.L. Reserve carbohydrates metabolism in the yeast *Saccharomyces cerevisiae*. *FEMS Microbiol. Rev.* **25**, 125–145 (2001).
39. Flachmann, R. *et al.* Molecular biology of pyridine nucleotide biosynthesis in *Escherichia coli*. Cloning and characterization of quinolinate synthesis genes nadA and nadB. *Eur. J. Biochem.* **175**, 221–228 (1988).
40. Panozzo, C. *et al.* Aerobic and anaerobic NAD+ metabolism in *Saccharomyces cerevisiae*. *FEBS Lett.* **517**, 97–102 (2002).
41. Kacser, H. & Burns, J.A. The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104 (1973).
42. Heinemann, M. & Sauer, U. Systems biology of microbial metabolism. *Curr. Opin. Microbiol.* **13**, 337–343 (2010).
43. Ideker, T., Galitski, T. & Hood, L. A new approach to decoding life: systems biology. *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
44. Kell, D.B. & Oliver, S.G. Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays* **26**, 99–105 (2004).
45. Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. & Palsson, B.O. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96 (2004).
46. King, R.D. *et al.* The automation of science. *Science* **324**, 85–89 (2009).
47. Winzeler, E.A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906 (1999).
48. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).

## ONLINE METHODS

**Experimental mapping of genetic interactions.** We used SGA methodology, an automated form of genetic analysis, to construct high-density arrays of double mutants (for details, see refs. 4,15.). Quantitative assessment of genetic interactions requires measurements of single- and double-mutant fitness and an estimate of the double-mutant fitness that would be expected based on the single-mutant phenotypes. Mutant fitnesses were derived from colony sizes after correcting systematic experimental biases (including positional effects, spatial effects, nutrient competition and screen batch effects)[16]. Single-mutant fitness was estimated using a set of control SGA screens, in which the queries carried a mutation in a neutral genomic locus[1]. Double-mutant fitness was estimated by employing the regular SGA protocol. We used the obtained single- ($f_i$ and $f_j$) and double-mutant fitnesses ($f_{ij}$) to derive genetic interaction measures as $\varepsilon = f_{ij} - f_i f_j$. A statistical confidence measure ($P$ value) was assigned to each interaction based on a combination of the observed variation of each double mutant across four experimental replicates and estimates of the background log-normal error distributions for the corresponding query and array mutants[1,16].

To explore the general properties of the metabolic genetic interaction map, we applied a previously suggested[1] confidence threshold of $|\varepsilon| > 0.08$ and $P < 0.05$ to define significantly interacting gene pairs. This threshold has been previously shown[1] to yield a good balance between coverage and precision and defines genetic interactions that cover at least ~35% of negative and ~18% of positive interactions deposited in BioGrid[49] with estimated precisions of ~63% and ~59%, respectively. In the case of replicate screens (for example, both AB and BA pairs were screened), we applied the following procedure: if replicate screens showed opposite interaction signs and at least one of them was significant, both pairs were removed; and if they showed the same interaction sign (both positive or both negative), the interaction with the lowest $P$ value was retained and both pairs were reported with that interaction. Comparison of interactions from screens performed in the present study with those from a full-genome study[1] showed a good correlation ($r = 0.76$) between interaction scores that were identified as significant by both studies. The high cross-study correlation allowed us to merge interaction data from the present study with interaction data on metabolic gene pairs from the genome-scale screens[1].

Additionally, we also defined a smaller high-confidence dataset in which all gene pairs were independently screened at least twice to minimize false interactions. Here, two genes were considered as interacting if at least one screen showed $|\varepsilon| > 0.08$ and $P < 0.05$ and another screen shows $P < 0.05$ with the same interaction sign, whereas non-interacting pairs are defined as those not showing $|\varepsilon| > 0.08$ and $P < 0.05$ in any of the screens. Any other gene pairs were removed from the high-confidence set. This resulted in 529 negative and 194 positive interactions between 122,875 gene pairs.

Interaction data can be downloaded from our website (see URLs).

**Analysis of the effect of functional relatedness, paralogy and protein-protein interactions on genetic interactions.** We used logistic regression analysis to test the association between genetic interaction and various categorical and continuous features (for example, paralogy, co-functionality, single mutant fitness, and so on). Functional annotation groups were as defined in the published metabolic reconstruction[10], and information on physical interactions between proteins was extracted from the BioGrid 2.0.58 database[49]. Paralog gene pairs were identified by performing all-against-all BLASTP similarity searches[50] of yeast open reading frames. We defined two genes as paralogs if i, the BLAST score had an expected value $E < 10^{-8}$, ii, alignment length exceeded 100 residues, iii, sequence similarity was > 30% and iv, they were not parts of transposons.

**Monochromaticity analysis.** To examine the monochromaticity of genetic interactions between pairs of functional annotation groups, we defined a monochromatic score (MC) as follows. Let $pr_{ij}$ denote the ratio of positive to all genetic interactions between group $i$ and $j$, and let $bpr$ denote the background ratio of positive to all interactions:

$$\text{if } pr_{ij} > bpr, \text{MC}_{ij} = (pr_{ij} - bpr)/(1 - bpr)$$

$$\text{if } pr_{ij} = bpr, \text{MC}_{ij} = 0$$

$$\text{if } pr_{ij} < bpr, \text{MC}_{ij} = (pr_{ij} - bpr)/bpr$$

A pair of groups showing purely positive (or purely negative) genetic interactions between each other has an MC score equal to +1 (or −1), whereas those reflecting the background ratio ($bpr$) have MC scores of 0. We computed MC scores based on those genes that are assigned to one functional group only. A pair of functional groups was considered monochromatic if $|\text{MC}_{ij}| > 0.5$.

To assess the significance of monochromaticity, we compared the monochromatic score of the experimentally determined genetic interaction network to those of 10,000 interaction maps that were constructed by randomizing the sign of each genetic interaction while keeping constant the total number of negative and positive interactions and conserving the annotation groups (see ref. 11). We restricted our analysis to those functional group pairs that showed at least two or three interactions between each other (**Supplementary Table 1**).

**Computing the impact of mutations and genetic interactions by flux balance analysis.** The recently reconstructed metabolic network (iMM904)[10] of *S. cerevisiae* was employed to simulate gene deletions. The reconstruction included 904 genes and 1,412 reactions and gave information on the stoichiometry and direction of biochemical reactions, their assignment to subcellular compartments and their associations to protein coding genes (including information on isoenzymes and enzyme complexes). Details of flux balance analysis (FBA) have been described elsewhere[6]. The simulated growth medium was set up to mimic the one used in the experiments (see the **Supplementary Note** for more details). *CAN1*, *LYP1*, *URA3*, *LEU2* and *MET17* were removed from the iMM904 reconstruction to mimic the strain background used in the experiments.

We employed linear programming to identify the maximum biomass yield of the wild-type network. The impact of gene deletions (null mutations) were calculated by constraining the corresponding reaction fluxes to zero and using either FBA or a linearized version of MOMA[27] to compute biomass yields of the mutant networks. Mutant fitness was defined as the biomass yield relative to wild type, and interaction between two mutations was calculated as follows: $\varepsilon = f_{12} - f_1 f_2$ (where $f_1$, $f_2$ and $f_{12}$ refer to the single and double mutant fitnesses, respectively). To compute the effect of a partial (non-null) mutation in a gene, we constrained the flux of its corresponding reaction to ≤50% of its wild-type level[14].

All calculations were carried out in the custom software package Sybil (G.G.-D. and M.J.L., unpublished data), developed in the R statistical environment[51] and using solvers GLPK and CPLEX (see URLs).

**Exploring the general properties of the FBA-derived genetic interaction map.** To generate an *in silico* genetic interaction map based on FBA, we computed interaction scores between all non-essential metabolic gene pairs and considered two genes as interacting if they had a predicted $|\varepsilon| > 10^{-4}$ (using a more stringent cutoff did not qualitatively affect our results). To investigate the relationship between *in silico* single-deletion fitness and other computed network properties (*in silico* genetic interaction degree and pleiotropy), we focused only on those genes i, whose reactions are not blocked (meaning they can attain a flux in some steady states of the network) and ii, whose removal affect the reaction content of the network (they do not have isoenzymes), thereby excluding genes that cannot have any single deletion effect in the model. Furthermore, some sets of genes would always produce identical phenotypes in the model simulations and cannot be treated as independent data points in statistical analyses (for example, genes encoding flux coupled reactions or subunits of the same protein complex). To avoid such a bias in our analysis, we represented each correlated gene set with one randomly chosen gene. These filtering procedures resulted in 193 genes.

**Computing system-level functional pleiotropy.** We used the metabolic model to derive a measure of functional pleiotropy for each metabolic gene. The model specifies a list of 54 metabolites that are essential for biomass formation and therefore for *in silico* growth (for example, amino acids, carbohydrates, fatty acids, and so on). We computed the maximum production yield of each biomass compound in the wild-type network by maximizing the flux through a pseudo reaction representing its secretion[26]. Next, we deleted each gene and

© 2011 Nature America, Inc. All rights reserved.

examined whether the knockout showed a reduction in the maximum production of a given compound (a flux reduction of $\geq 10^{-4}$ was considered significant). Finally, for each gene, we counted the number of biomass compounds whose maximal production is affected by its deletion. This number reflects the network-level multifunctionality, and hence, the pleiotropy of a gene.

**Identifying flux coupled genes in the network.** Coupled genes were identified by applying the flux coupling finder algorithm[20] on the metabolic network. We distinguished between coupled and uncoupled relationships between reaction pairs: i, coupled (fully and directionally coupled) meant that activity of one reaction fixed the activity of the other and vice versa, or activity of one reaction implies the activity of the other, but not the reverse; and ii, uncoupled: activity of one reaction does not imply the activity of the other and vice versa, indicating that the reactions are not required to operate together. Coupling relationships were calculated without assuming a fixed biomass composition to avoid coupling of a large set of fluxes to the biomass reaction[20].


49. Breitkreutz, B.J. *et al.* The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.* **36**, D637–D640 (2008).
50. Altschul, S.F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
51. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2007).


© 2011 Nature America, Inc. All rights reserved.


doi:10.1038/ng.846

# Plasticity of genetic interactions in metabolic networks of yeast

**Richard Harrison*, Balázs Papp*, Csaba Pál†, Stephen G. Oliver*‡, and Daniela Delneri*‡**

*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, United Kingdom; and †Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved December 12, 2006 (received for review August 17, 2006)

**Why are most genes dispensable? The impact of gene deletions may depend on the environment (plasticity), the presence of compensatory mechanisms (mutational robustness), or both. Here, we analyze the interaction between these two forces by exploring the condition-dependence of synthetic genetic interactions that define redundant functions and alternative pathways. We performed systems-level flux balance analysis of the yeast (*Saccharomyces cerevisiae*) metabolic network to identify genetic interactions and then tested the model's predictions with *in vivo* gene-deletion studies. We found that the majority of synthetic genetic interactions are restricted to certain environmental conditions, partly because of the lack of compensation under some (but not all) nutrient conditions. Moreover, the phylogenetic cooccurrence of synthetically interacting pairs is not significantly different from random expectation. These findings suggest that these gene pairs have at least partially independent functions, and, hence, compensation is only a byproduct of their evolutionary history. Experimental analyses that used multiple gene deletion strains not only confirmed predictions of the model but also showed that investigation of false predictions may both improve functional annotation within the model and also lead to the discovery of higher-order genetic interactions. Our work supports the view that functional redundancy may be more apparent than real, and it offers a unified framework for the evolution of environmental adaptation and mutational robustness.**

epistasis | genetic robustness | *Saccharomyces cerevisiae* | environmental dependence | flux balance analysis

One of the most striking discoveries of molecular genetics is that a large fraction of the protein-coding genes have negligible effects on growth rates under standard laboratory conditions. Recent systematic single-gene-deletion studies suggest that nearly 80% of yeast genes appear not to be essential for growth (1). Comparable large-scale experiments in free-living bacteria, worm, and mouse showed that the fraction of essential genes is generally low, typically in the range of 6–19% (2, 3).

Although much investigated, the causes and evolution of gene dispensability remain controversial (4–7). The high fraction of dispensable genes might reflect the capacity of organisms to compensate for null mutations by using either redundant gene duplicates or alternative metabolic pathways (mutational robustness) (4). Others have suggested that many of the seemingly dispensable genes have important fitness contributions only under special environmental conditions (environmental adaptation) (5). However, the potential links between adaptation to new environmental conditions and robustness against harmful mutations have remained largely unexplored. It may well be that these theories on gene dispensability are not mutually exclusive. Differences in the availability of external nutrients and/or intracellular metabolites across environmental conditions can have a large effect on the number of active metabolic pathways that can produce a given key cellular component (Fig. 1). Hence, the capacity to compensate null mutations may vary substantially between different nutritional environments. One clear prediction of this idea is that the impact of both single- and double-gene deletions should change across environmental conditions.

Several lines of evidence are compatible with this idea. First, data compiled from available large-scale phenotypic screens in yeast [see supporting information (SI) Table 2] suggest that at least 20% of the ≈5,000 apparently nonessential genes in *Saccharomyces cerevisiae* make a large contribution to fitness under at least 1 of the 31 investigated conditions. Moreover, most of these conditionally essential genes make a contribution in only one or a few environments (Fig. 2), suggesting that conditional growth defects for numerous other gene deletions remain to be discovered. Second, a gene-deletion phenotype frequently does not reflect simply the absence of a given gene but also the response of the cell to its absence. Such responses may involve the redistribution of enzymatic fluxes in the network and up-regulation of previously inactive genes (8, 9). Third, mutagenesis studies on *Escherichia coli* and viruses have shown a joint influence of environmental plasticity and epistatic genetic interactions on the effect of deleterious mutations (10, 11).

Using a combination of computational flux-balance analysis (FBA) and *in vivo* gene-deletion experiments, we have explored the link between epistatic genetic interactions and plasticity. FBA provides a rigorous computational framework for studying the impact of gene deletions (12). Based on steady-state assumptions and optimality criteria, this constraint-based method has been successfully applied for calculating the phenotypic behavior of the metabolic network (13) and the viability of single-gene-deletion strains in yeast (14). We restricted our attention to the most extreme form of genetic interaction [synthetic lethality (SL)], where a double deletant shows a no-growth phenotype that is not exhibited by either single deletant. The computational analyses suggest a strong dependence of genetic interactions on the prevailing environmental conditions, and this finding is supported by the experimental data presented below and by evidence from the literature.

Our study supports the view (15, 16) that mutational robustness is not a directly selected trait, but rather a byproduct of the evolution of biological networks toward survival under a wide range of environmental conditions (environmental robustness).

## Results

**FBA Reveals a High Frequency of Condition-Dependent Genetic Interactions.** We have extended previous studies (17) by applying FBA to a genome-scale metabolic network model of yeast (*S. cerevisiae*) to calculate genetic interactions. The previously recon-

Author contributions: R.H. and B.P. contributed equally to this work; B.P., S.G.O., and D.D. designed research; R.H., B.P., and D.D. performed research; R.H., B.P., C.P., and S.G.O. analyzed data; and B.P., C.P., S.G.O., and D.D. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS direct submission.

Abbreviations: FBA, flux-balance analysis; SD, synthetic defined; SL, synthetic lethal; SS, synthetic sickness; SSL, synthetic sick/lethal; YPD, yeast–peptone–dextrose.

‡To whom correspondence may be addressed. E-mail: steve.oliver@manchester.ac.uk or d.delneri@manchester.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0607153104/DC1.

© 2007 by The National Academy of Sciences of the USA



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

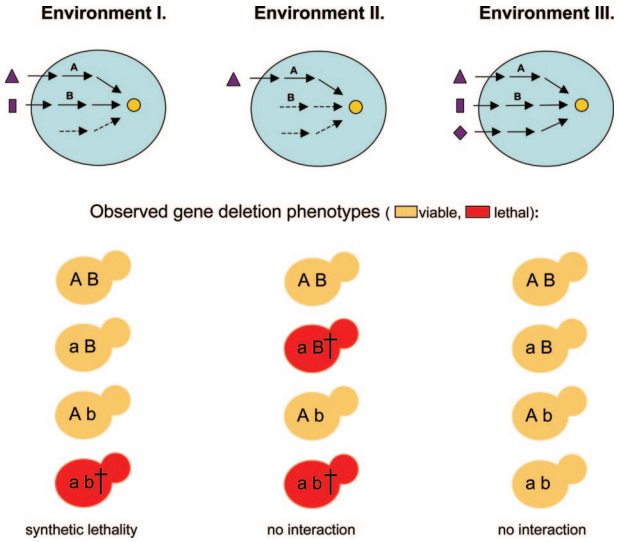



**Fig. 1.** Model to explain conditional synthetic lethality. A key metabolite (yellow circle) can be synthesized via three independent pathways. Metabolic genes A and B show synthetic lethality in Environment I, where starting nutrients of both pathways are present in the medium. However, B is unable to compensate deletion of A in Environment II, and the double mutant is rescued by the third pathway in Environment III.

structed metabolic network (18) consists of 672 genes and 745 unique biochemical reactions and incorporates external nutrients and the corresponding transport processes. The impacts of all possible single- and double-gene deletions were calculated for 53 nutritional environments, including various carbon sources (see *SI Materials and Methods*). The analysis identified 98 gene pairs that were predicted to be involved in a SL relationship under at least one of the conditions investigated (SI Tables 3 and 4). Only 14.3% of these SL relationships were displayed under all nutrient conditions investigated, and 50% of them are restricted to only one or two nutritional environments (Fig. 3).

The condition-specificity of interactions does not appear to be randomly distributed in the metabolic network. SL interactions between gene pairs annotated to different metabolic subsystems

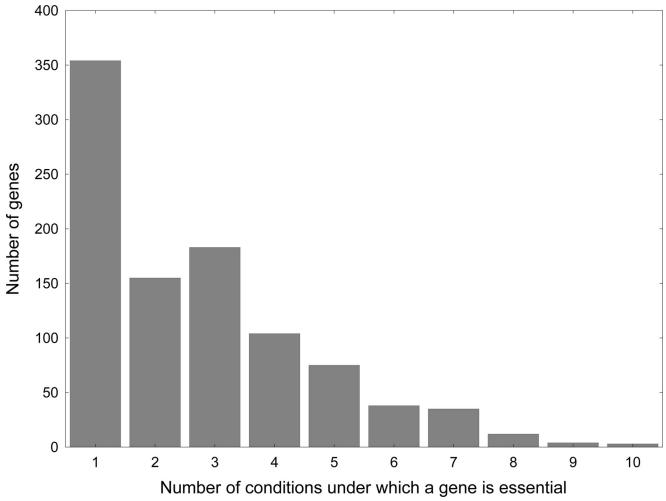



**Fig. 2.** Distribution of environmental specificity of single-gene deletion phenotypes. Gene deletions showing conditional growth phenotypes were compiled from published large-scale screens (see SI Table 2). Of 4,823 genes not essential for growth on YPD, 963 exhibited lethality or a strong growth defect under at least 1 of the 31 conditions investigated.

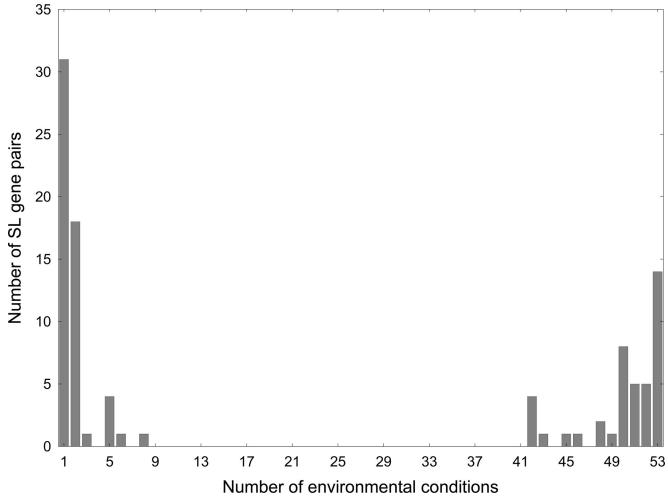



**Fig. 3.** Distribution of environmental specificity of predicted synthetic lethal interactions. The histogram shows the distribution of the number of simulated environments where each of the 98 gene pairs exhibits synthetic lethality (only gene pairs interacting in at least 1 of the 53 conditions investigated are included).

are present in a significantly smaller number of environments than those that are annotated to the same subsystem (Mann–Whitney $U$ test, $P < 0.02$; because enzymes catalyzing the same reaction, by definition, have the same functional annotation, they were excluded from this analysis). Moreover, more than half ($56.3 \pm 2.7\%$) of all genetic interactions remain undetected when only a single environment is investigated. These results not only provide a link between compensation of null mutations and the environment but also suggest that systematic genetic interaction screens (which are generally restricted to a single condition) may miss many of the extant interactions.

**Experimental Tests on the Reliability of the Model.** The predictions of the *in silico* model were tested by *in vivo* double-gene-deletion experiments (17 cases, Table 1, *Materials and Methods*) and by extracting published experimental data from the literature (32 cases; see *SI Materials and Methods*). This procedure enabled us to validate ≈60% of the total number of genetic interactions that we predicted to be present on either minimal or rich media (SI Table 5). Double deletants were constructed by sporulating and dissecting heterozygous diploids from crosses between two single-gene haploid deletants (see *SI Materials and Methods*). Next, we assessed the viability of double deletants by inspecting growth on plates. In 12 of the 17 cases investigated, we observed a clear synthetic sick or lethal (SSL) phenotype under the predicted growth condition (Table 1). The model also accurately captures changes in the presence of synthetic genetic interactions between media (see below). However, in five cases, the double mutant formed colonies qualitatively indistinguishable from the single-gene deletants.

These apparently false predictions may indicate that the model has only limited resolution. It may be, for example, that FBA accurately predicts the direction, but not the strength, of genetic interaction between genes. To explore whether weak genetic interactions, which are undetectable by a simple plate-growth assay, could be responsible for some of these false predictions, we measured the growth rates of all viable double deletants, and those of the corresponding single deletants, using an established protocol (19) (see *Materials and Methods*). In two of the five investigated cases, we found evidence for weak (but statistically significant) negative epistasis between the predicted gene pairs (Table 1).



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

**Table 1. Validation of *in silico* predictions by constructing double-mutant strains**

| Gene 1 | Gene 2 | Environment | Prediction success | Measured epistasis |
|---|---|---|---|---|
| *ASN1* | *ASN2* | SD | + | SL |
| *CHO2* | *PCT1* | YPD | + | SS, −0.372* |
| *CKI1* | *CHO2* | YPD | − | −0.089* |
| *CPT1* | *CHO2* | YPD | − | 0.020 |
| *ECM31* | *FEN2* | YPD | + | SL |
| *ECM31* | *FEN2* | SD | + | SL |
| *OPI3* | *PCT1* | YPD | − | 0.012 |
| *OPI3* | *CPT1* | YPD | − | 0.004 |
| *OPI3* | *CKI1* | YPD | − | −0.067* |
| *RPE1* | *ZWF1* | YPD | + | SL |
| *RPE1* | *ZWF1* | SD | + | SL |
| *SAM2* | *SAM1* | YPD | + | SL |
| *SAM2* | *SAM1* | SD | + | SL |
| *SPE1* | *FEN2* | SD | + | SL |
| *SPE2* | *FEN2* | SD | + | SL |
| *URA8* | *URA7* | YPD | + | SL |
| *URA8* | *URA7* | SD | + | SL |

A set of SL interactions predicted for nutrient-rich (YPD) and/or glucose minimal (SD) media were validated by measuring the epistasis between pairs of gene deletions (see *Materials and Methods*). Lack of growth of a double mutant is denoted by "SL" and synthetic sickness by "SSA." A prediction was considered successful if the double mutant had a visually apparent growth defect compared with single mutants in a plate growth assay (i.e. strong negative epistasis, SSL). *, $P < 10^{-5}$.

Although only a limited number of interactions were tested experimentally, the results suggest that FBA can reliably detect genetic interactions in the metabolic network of yeast. Future large-scale experimental screens are required to get a precise estimate of the fraction of false-positive and false-negative predictions. As a preliminary to such a larger study, we augmented our experimental results with literature data available on single- and double-deletant strains (see *SI Materials and Methods*). Overall, we were able to test 49 predicted interactions (SI Table 5) and estimate that ≈49% (24 of 49) of them were correct and that, in 53% of the cases, at least the sign of epistasis was consistent with the predictions. In a similar vein, FBA can identify ≈24% of a curated list (20) of previously described SL interactions between metabolic genes. Both values are at least two orders of magnitude higher than expected by chance ($P < 10^{-287}$, see *SI Materials and Methods*).

**Gene Duplicates Can Explain Many of the False Predictions.** Lack of an observable growth defect in three of the experimentally observed cases could be due to the presence of gene duplicates with redundant functions that are not represented in the current metabolic reconstruction. We investigated this possibility by determining whether one or the other member of the gene pairs investigated had a gene duplicate that might provide compensation for one missing function in the double deletant. One member of the gene pair had a paralog in all three cases. Construction of triple-deletion strains (*SI Materials and Methods*) revealed strong negative epistasis in all three cases (SI Fig. 5).

For example, *CHO2* and *CPT1* are erroneously predicted to show a synthetic genetic interaction on rich medium. We hypothesized that this interaction might be masked by *EPT1*, a duplicate of *CPT1*. The two encoded proteins show 56% amino acid sequence similarity to each other and have different primary catalytic activities. However, some studies suggest that although Cpt1p accounts for 95% of phosphatidylcholine synthesis *in vivo*, the remaining 5% is likely to be catalyzed by Ept1p (21). Remarkably, deletion of all three genes simultaneously resulted in a much stronger growth defect than observed for any of the pair-wise deletions (SI Fig. 5).

In addition, our data suggest that interaction between *OPI3* and *PCT1* is masked by *MUQ1*, a distant paralog of *PCT1* (the products of the two genes share 36% amino acid sequence similarity). The triple *opi3*Δ/*pct1*Δ/*muq1*Δ has a more severe phenotype than either double mutant (SI Fig. 5). Although Pct1p and Muq1p catalyze related reactions, they are generally believed to have different substrate specificities. Further biochemical studies will be needed to confirm whether Muq1p has the catalytic activity necessary to mitigate the effect of the *pct1*Δ/*opi3*Δ double deletion.

Detailed investigation of the false predictions can thus be used to generate novel biochemical hypotheses and refine the *in silico* model. Moreover, these results suggest that even duplicates with low sequence similarities and partly altered functions can compensate null mutations in each other. The importance of appropriate modeling of paralogs/gene duplicates is further underscored by inspection of false-negative predictions (i.e., true pair-wise interactions not predicted by the model). Many of these previously reported interacting gene pairs are predicted to participate in higher-order genetic interactions because of the presence of a gene duplicate with overlapping functions (i.e., an isoenzyme) (see SI Table 6). For example, *TDH2* and *TDH3* show synthetic lethality *in vivo* (22); however, our simulations show that *tdh2*Δ *tdh3*Δ double mutant is compensated by *TDH1*, a gene encoding an additional glyceraldehyde-3-phosphate dehydrogenase isoenzyme in the model. Lack of *in vivo* compensatory capacity of Tdh1p might be explained by its relatively low expression level compared with Tdh2p and Tdh3p (23). Thus, in addition to correctly assigning reactions to paralogous genes, incorporation of regulatory constraints (12) and information on maximum enzyme capacities would also be needed to more accurately model the behavior of isoenzymes. Transcriptional reprogramming upon gene deletion (8) may also have an influence on predicting deletant phenotypes.

**Two Explanations for the Condition-Specificity of Genetic Interactions.** We found empirical evidence of condition-specific epistasis for 14 validated SSL gene pairs, of which 11 were correctly predicted (SI Table 7), suggesting that our modeling framework is able to capture variation in the incidence of genetic interactions across a range of environmental conditions. There could be at least two explanations for the condition-dependence of these genetic interactions (see Fig. 1). First, members of the synthetically interacting gene pairs make important individual contributions to growth under different nutritional conditions. Alternatively, the double-deletant strain becomes viable under different conditions. There is experimental evidence for both explanations (Fig. 4, SI Table 7).

*CHO2* and *PCT1* are genes that encode two enzymes that each catalyze a step in two different pathways responsible for phosphatidylcholine synthesis (the phosphatidylethanolamine methylation pathway and the Kennedy pathway, respectively; see Fig. 4*A*). In agreement with our first explanation of condition-dependence, we find that these two genes can compensate null mutations in each other under nutrient-rich [yeast–peptone–dextrose (YPD)] conditions, but the *cho2*Δ deletant shows slow growth on glucose minimal medium (Fig. 4*B*). This indicates that, in the absence of exogenous choline, the Kennedy pathway (and hence *PCT1*), on its own, cannot support net phosphatidylcholine synthesis.

The second explanation for condition-dependence can be exemplified by the *SAM1*/*SAM2* duplicate gene pair, which encodes two distinct forms of *S*-adenosylmethionine (AdoMet) synthetase. Although differentially regulated (24), the two genes can compensate null mutations in one another, and the double mutants are inviable under nutrient-rich (YPD) conditions (see



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

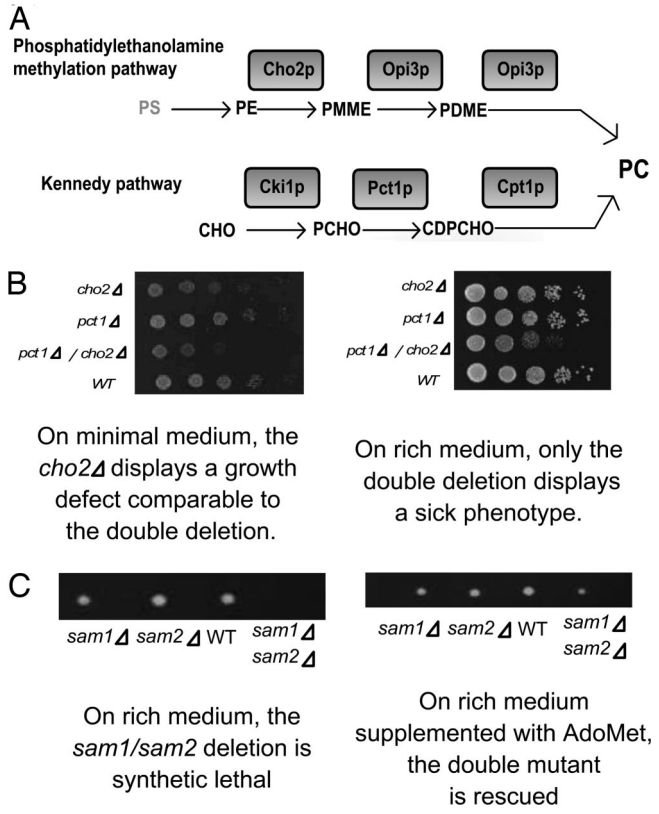



**Fig. 4.** Examples of environment-specific synthetic genetic interactions. (*A*) Alternative routes to phosphatidylcholine biosynthesis in yeast. Cho2p, phosphatidylethanolamine methyltransferase; Opi3p, phospholipid methyltransferase; Cki1p, choline kinase; Pct1p, cholinephosphate cytidylyltransferase; Cpt1p, sn-1,2-diacylglycerol cholinephosphotransferase; PS, phosphatidylserine; PE, phosphatidylethanolamine; PME, phosphatidyl-*N*-methylethanolamine; PDME, phosphatidyl-*N*-dimethylethanolamine; CHO, choline; PCHO, choline phosphate; CDPCHO, CDP-choline; PC, phosphatidylcholine. (*B*) One member of the SSL pair makes an important individual contribution to growth under a different condition. *CHO2* and *PCT1* can compensate null mutations in one another under nutrient-rich (YPD) conditions, but the *cho2*Δ mutant is slow growing on minimal medium. (*C*) The double deletant becomes viable under a different condition. The *SAM1/SAM2* duplicate gene pair, which encodes two distinct forms of *S*-adenosylmethionine (AdoMet) synthetase, can compensate null mutations in one another, and the double mutants are inviable under nutrient-rich (YPD) conditions. However, addition of AdoMet to the medium yields viable double mutants.

Table 1). However, addition of AdoMet (the enzymatic product of Sam1p/Sam2p) into the medium yields viable double mutants (Fig. 4*C*).

**Frequent Plasticity of Genetic Interactions Among Nonmetabolic Genes.** Having established the widespread occurrence of environment dependency of synthetic genetic interactions for metabolic genes, we asked whether condition dependency could be a general property of SSL interactions. First, we compiled a list of publicly available SSL interactions (25) discovered by a global genetic-interaction mapping approach (26) using a chemically defined glucose medium (those strains showing growth defects on minimal medium were excluded to ensure that the investigated interactions were not between unconditionally slow growing mutants, see *Materials and Methods*). Next, we collected viability data from published screens for single-gene-deletion phenotypes performed under 31 growth conditions (SI Table 2). In 57.4% of the investigated 2,666 SSL gene pairs, there is evidence that one or both members of the pairs make an essential contribution to growth under at least one of the 31 conditions investigated. This figure is likely to be an underestimate for two reasons: First, only a limited number of environments have been studied experimentally so far. Second, this estimate ignores cases where the double-deletant strain becomes viable under some other environmental condition. Moreover, there is some further support for a link between the extent of compensatory mechanisms and environmental specificity: genes for which evidence exists for conditional phenotypes have significantly more SL interactions than the rest of yeast's genes (Mann–Whitney *U* test, $P = 0.002$; see SI Fig. 6).

**Random Phylogenetic Cooccurence of SSL Pairs.** Comparative genomics studies indicate that members of functional modules (i.e., genes that contribute jointly to a given cellular function) evolve nonindependently and show a similar phylogenetic distribution across species. For example, genes encoding members of protein complexes or metabolic modules are frequently gained and lost together during evolutionary history (27, 28). Indeed, we could confirm that gene pairs encoding subunits of the same literature-curated protein complexes (29) have higher phylogenetic cooccurrence than random gene pairs. We calculated a score (30) for the cooccurrence of these gene pairs across 16 eukaryotic genomes and used randomization protocols to get an estimate of statistical significance (see *SI Materials and Methods*). As expected, the score for subunits of protein complexes is significantly higher than expected by chance ($P < 10^{-5}$, $n = 7{,}186$ pairs). Next, we asked if a similar result holds for experimentally determined SSL gene pairs. Using the same protocol as above, we found a strikingly different result. In contrast to members of protein complexes, gene pairs showing synthetic genetic interactions show no evidence for shared evolutionary history across species ($P = 0.107$, $n = 1{,}850$ pairs). Moreover, this finding cannot be explained by the likelihood of a low frequency of retention of redundant duplicates in all genomes; our result remains unchanged when all gene pairs showing even low sequence similarity to one another are excluded from the analysis ($P = 0.113$, $n = 1{,}780$ pairs, see *SI Materials and Methods*).

## Discussion

Systematic screens on SL genetic interactions in yeast (25) and worm (31) are providing invaluable insights into the organisms' compensatory capacity. However, because of the enormous number of possible gene combinations, a complete mapping of SL interactions is still some way off. For this and other reasons, there is a need to find systems-biology models that are able to provide efficient and reliable tools for predicting (higher-order) genetic interactions. FBA offers a rigorous theoretical framework for studying the impact of multiple gene deletions on yeast metabolism. It also has a major advantage over other suggested computational approaches (32, 33) in that it can investigate epistasis under various environments.

Previous theoretical studies relied exclusively on the biochemical consistency of FBA to calculate epistasis (17, 34, 35). This study attempts to experimentally validate synthetic genetic interactions predicted by a genome-scale metabolic model. Although the accuracy of FBA at predicting genetic interactions is comparable with previous approaches (32, 33), the method is far from perfect. Our work suggests that many of the apparently false predictions are not due to major conceptual problems with FBA but, rather, are due to incomplete annotation and incorrect modeling of isozymes. First, a duplicate of a given enzyme-encoding gene could be present in the genome, which, although not annotated as an isozyme and diverged in both its amino acid sequence and biological function, could retain the ability to compensate for the absence of the other gene. Second, redundancy of certain isozymes annotated in the model might be more



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

apparent than real because of incomplete compensatory capacity or regulatory differences between the gene copies (36). Thus, annotation of new enzymatic functions and incorporation of information on enzyme capacities and gene regulation (12) should lead to a refined model with more predictive power. Our study confirms the view that model building in systems biology is an iterative process (37) that proceeds by testing the predictions of the model against experimental data and then by using any discrepancies to revise and improve the model.

We used FBA to study the interplay between mutational robustness and the environment. Synthetic genetic interactions provide good examples of mutational robustness: members of these pairs are likely to be independent genes participating in alternative metabolic pathways or redundant gene duplicates. By integrating computational data with *in vivo* studies on double-gene deletants, we could show that synthetic genetic interactions are frequently restricted to particular environmental conditions, partly because genes involved in SL interactions under one condition frequently make an essential contribution to growth in another environment. The idea that compensating gene pairs bear distinct functional roles and are not redundant under all conditions is further supported by the observation that their phylogenetic cooccurrence is not different from those of functionally unrelated random gene pairs.

What could be the selective forces behind the evolutionary emergence of condition-specific compensation mechanisms? In principle, there are at least two possible routes. First, novel compensatory pathways might evolve to enhance robustness against spontaneously arising deleterious mutations and may later provide raw material for adaptation to new environments (38). Alternatively, adaptation toward new nutritional conditions may drive the evolution of novel metabolic pathways and, as a correlated response, some of these new pathways may also enhance the organism's ability to withstand harmful mutations under certain conditions. For example, we speculate that the ancestor of the choline transporter gene (*HNM1*) might have evolved to enable the cell to use exogenous choline and, as a side effect, provides robustness against null mutations in genes of the phosphatidylethanolamine methylation pathway when choline is present in the medium.

Several lines of theoretical reasoning and observation are consistent with the view that mutational robustness is a byproduct of other evolved properties of metabolic networks. First, the presence of compensating metabolic gene duplicates can be explained by gene dosage effects (5), differential regulation (39), or the capacity to filter nonheritable noise (40), without the need to invoke direct selection to favor mutational resilience. In a similar vein, computer simulations suggest that the evolution of several structural properties of metabolic networks can be explained by selection for enhanced growth rates (41). Second, population-genetics models have clearly shown that the selection pressure for enhanced mutational robustness is generally weak, of the order of mutation rates (42). Similar objections were raised to Fisher's selectionist theory of dominance (43). In contrast, evolution of environmental robustness is unproblematic from a population genetics point of view (42, 44), and mutational robustness might simply arise as a correlated response to selection for environmental robustness (15, 16). The finding that the extent of epistatic interactions is not independent of environmental specificity (SI Fig. 6 and ref. 10) provides evidence for a correlation between mutational and environmental robustness. Finally, the scenario of direct selection for mutational robustness would leave unexplained our observation that different genes can be compensated in different environments. Therefore, based on the above arguments, we conclude that mutational robustness of metabolic networks is unlikely to be a directly selected trait. Rather, it is a side effect of adaptation to survive in a large variety of nutrient conditions.

## Materials and Methods

**Analysis of Genetic Interactions in the Metabolic Network of Yeast.** We examined a recently updated (iLL672) metabolic network of *S. cerevisiae*, which contains 672 genes and 745 unique biochemical reactions including transport processes (18). The reconstruction also provides information on the association of genes with different metabolic subsystems (e.g., purine metabolism, phospholipid biosynthesis, etc.). One dubious reaction, corresponding to choline biosynthesis, was removed from the reconstruction because yeasts are unable to synthesize choline *de novo* (45). FBA of the metabolic network was used to calculate the impact of gene deletions on maximum biomass production rate (a proxy for fitness). Details of the FBA protocol have been described in ref. 12. SL interactions were identified by simulating all possible single- and double-gene deletions and screening for gene pairs where the single deletions had a <10% fitness effect, but the double mutant was unable to produce biomass (the use of different cut-offs led to very similar results). All deletion simulations were carried out in the *ura3*Δ *leu2*Δ *his3*Δ *met17*Δ *lys2*Δ genetic background to most closely mimic the strains used in the *in vivo* studies (see *SI Materials and Methods*).

To explore the condition dependency of SL interactions, we defined a large set of nutrient environments. First, we tested all external nutrients for their ability to support aerobic growth in minimal medium. This resulted in 50 minimal media containing different principal carbon sources, including glucose. Additionally, we defined a medium mimicking YPD, a medium where all possible external nutrients were allowed for uptake, and a minimal vitamin medium [lacking pantothenate because yeast is capable of *de novo* pantothenate biosynthesis (46)], resulting in 53 environmental conditions (for details see *SI Materials and Methods* and SI Table 8). All simulated growth media were supplemented with uracil, leucine, histidine, methionine, and lysine to complement the nutritional markers and also with vitamins (with or without pantothenate, see above) to further mimic the experimental conditions.

**Experimental Procedures.** The simulations identified 59 gene pairs showing SL on either nutrient-rich (YPD) or glucose minimal [synthetic defined (SD)] medium. Published data (1, 18) on single-deletion phenotypes for these two conditions enabled us, in a comprehensive manner, to identify gene pairs for which the viability of single deletants was correctly predicted.

To carry out *in vivo* validation, we considered initially those gene pairs that have, at most, one paralog that is not annotated as an isozyme in the model. This choice enabled us to test higher-order genetic interactions by constructing triple-deletion strains in cases where the double mutant was viable (see below). Because, by using this criterion, all isozymes were excluded from validation, we additionally incorporated three randomly selected isozyme pairs in our experimental set. Moreover, among the group of gene pairs containing several paralogs, we decided to test gene pairs involved in pantothenate and polyamine biosynthesis for which we had the highest number of predicted SL interactions but no literature support available. Finally, some of the selected gene pairs could not be verified because one or the other mutant strain was missing from the deletion collection or contained a second-site mutation (18). This selection procedure left us with a set of 13 gene pairs to validate *in vivo*. These pairs corresponded to 17 cases of synthetic lethality: 6 on YPD, 3 on glucose minimal medium (SD), and 4 on both media.

We constructed the predicted double mutants by crossing haploid yeast strains containing single-gene deletions in the BY4742 and BY4741 backgrounds following standard protocols (see *SI Materials and Methods*).

In cases where we failed to detect any growth defect by visual inspection of plates (no overt synthetic sick or lethal phenotype), we



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

performed accurate measurements of maximum growth rates of single and double mutants to estimate epistasis. Optical densities were measured by a Bioscreen C analyzer (Thermic Labsystems, Oy, Finland), and maximum growth rates were calculated by using an established protocol (19). Five cultures were grown for each strain in both YPD and SD. Maximum growth rates were averaged over the five replicates and divided by the wild-type value to yield a relative growth rate for each strain. Because additivity of the growth rates is equivalent to multiplicity of nonlogarithmic measures of fitness (47), we defined epistasis ($\varepsilon$) as the degree of departure from additivity of the relative growth rates ($\mu$), thus $\varepsilon = \mu_{AB} + \mu_{ab} - \mu_{Ab} - \mu_{aB}$.

See *SI Materials and Methods* for details on the construction of triple-gene deletants.

**Analysis of Global Genetic Interaction and Mutant Phenotype Data Sets.** We compiled a list of publicly available SSL interactions (25), discovered by the synthetic genetic array approach (26), on SD medium complemented with amino acids. Single-gene deletion strains that exhibit a pronounced growth defect (<80% of wild-type growth rate) on SD medium (48) were excluded from further analyses to ensure that the interactions were not between unconditionally slow growing mutants. This resulted in a list of 2,666 synthetic genetic interactions (1,230 of them being SL). Information on environment-specific phenotypes of single-gene deletions in nonessential genes was collected from published large-scale phenotypic screens (see SI Table 2). Our list of 31 experimental conditions included various nutrient and stress conditions, sporulation, and stationary phase, but excluded drug treatments. Only the strongest growth defects and phenotypes were considered as evidence for conditional fitness contributions.

We thank Lars Blank and Uwe Sauer for providing early access to the iLL672 metabolic reconstruction and to data from phenotypic screening of single mutants. We acknowledge valuable comments from Jonas Warringer on the growth-rate measurement protocol. R.H. was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) studentship. C.P. and B.P. are supported by the Hungarian Scientific Research Fund. B.P. is a Long-Term Fellow of The International Human Frontier Science Program Organization, C.P. is a Long-Term Fellow of the European Molecular Biology Organization, and D.D. is a Natural Environment Research Council (U.K.) Advanced Fellow. Work on systems biology in S.G.O.'s laboratory was supported by BBSRC. We acknowledge support from a Royal Society research grant (to D.D.).

1. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, *et al.* (2002) *Nature* 418:387–391.
2. Koonin EV (2003) *Nat Rev Microbiol* 1:127–136.
3. Wilson L, Ching YH, Farias M, Hartford SA, Howell G, Shao H, Bucan M, Schimenti JC (2005) *Genome Res* 15:1095–1105.
4. Wagner A (2000) *Nat Genet* 24:355–361.
5. Papp B, Pál C, Hurst LD (2004) *Nature* 429:661–664.
6. Blank LM, Kuepfer L, Sauer U (2005) *Genome Biol* 6:R49.
7. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li WH (2003) *Nature* 421:63–66.
8. Kafri R, Bar-Even A, Pilpel Y (2005) *Nat Genet* 37:295–299.
9. Fischer E, Sauer U (2003) *Eur J Biochem* 270:880–891.
10. Remold SK, Lenski RE (2004) *Nat Genet* 36:423–426.
11. You L, Yin J (2002) *Genetics* 160:1273–1281.
12. Price ND, Reed JL, Palsson BO (2004) *Nat Rev Microbiol* 2:886–897.
13. Famili I, Forster J, Nielsen J, Palsson BO (2003) *Proc Natl Acad Sci USA* 100:13134–13139.
14. Forster J, Famili I, Palsson BO, Nielsen J (2003) *Omics* 7:193–202.
15. Wagner A (2005) *Robustness and Evolvability of Living Systems* (Princeton Univ Press, Princeton).
16. Gibson G, Wagner G (2000) *BioEssays* 22:372–380.
17. Segrè D, Deluna A, Church GM, Kishony R (2005) *Nat Genet* 37:77–83.
18. Kuepfer L, Sauer U, Blank LM (2005) *Genome Res* 15:1421–1430.
19. Warringer J, Blomberg A (2003) *Yeast* 20:53–67.
20. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, *et al.* (2006) *J Biol* 5:11.
21. McMaster CR, Bell RM (1994) *J Biol Chem* 269:28010–28016.
22. McAlister L, Holland MJ (1985) *J Biol Chem* 260:15013–15018.
23. McAlister L, Holland MJ (1985) *J Biol Chem* 260:15019–15027.
24. Thomas D, Rothstein R, Rosenberg N, Surdin-Kerjan Y (1988) *Mol Cell Biol* 8:5132–5139.
25. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, *et al.* (2004) *Science* 303:808–813.
26. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, *et al.* (2001) *Science* 294:2364–2368.
27. Pál C, Papp B, Lercher MJ (2005) *Nat Genet* 37:1372–1375.
28. Campillos M, von Mering C, Jensen LJ, Bork P (2006) *Genome Res* 16:374–382.
29. Güldener U, Münsterkötter M, Kastenmüller G, Strack, N, van Helden J, Lemer C, Richelles J, Wodak SJ, García-Martínez J, Pérez-Ortín JE, *et al.* (2005) *Nucleic Acids Res* 33:D364–D368.
30. Huynen M, Snel B, Lathe W, III, Bork P (2000) *Genome Res* 10:1204–1210.
31. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) *Nat Genet* 38:896–903.
32. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, *et al.* (2004) *Proc Natl Acad Sci USA* 101:15682–15687.
33. Zhong W, Sternberg PW (2006) *Science* 311:1481–1484.
34. Thiele I, Vo TD, Price ND, Palsson BO (2005) *J Bacteriol* 187:5818–5830.
35. Ghim CM, Goh KI, Kahng B (2005) *J Theor Biol* 237:401–411.
36. Delneri D, Gardner DCJ, Oliver SG (1999) *Genetics* 153:1591–1600.
37. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) *Nature* 429:92–96.
38. Deutscher D, Meilijson I, Kupiec M, Ruppin E (2006) *Nat Genet* 38:993–998.
39. Ihmels J, Levy R, Barkai N (2004) *Nat Biotechnol* 22:86–92.
40. Kafri R, Levy M, Pilpel Y (2006) *Proc Natl Acad Sci USA* 103:11653–11658.
41. Pfeiffer T, Soyer OS, Bonhoeffer S (2005) *PLoS Biol* 3:e228.
42. Proulx SR (2005) *Am Nat* 165:147–162.
43. Wright S (1929) *Am Nat* 63:274–279.
44. Wagner GP, Booth G, Bagheri-Chaichian H (1997) *Evol Int J Org Evol* 51:329–347.
45. Howe AG, Zaremberg V, McMaster CR (2002) *J Biol Chem* 277:44100–44107.
46. White WH, Gunyuzlu PL, Toyn JH (2001) *J Biol Chem* 276:10794–10800.
47. Szafraniec K, Wloch DM, Sliwa P, Borts RH, Korona R (2003) *Genet Res* 82:19–31.
48. Warringer J, Ericson E, Fernandez L, Nerman O, Blomberg A (2003) *Proc Natl Acad Sci USA* 100:15724–15729.



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

de_1908_21

LETTERS

# Chance and necessity in the evolution of minimal metabolic networks

Csaba Pál[1,2]*, Balázs Papp[3]*, Martin J. Lercher[1,4], Péter Csermely[5], Stephen G. Oliver[3] & Laurence D. Hurst[4]

**It is possible to infer aspects of an organism's lifestyle from its gene content[1]. Can the reverse also be done? Here we consider this issue by modelling evolution of the reduced genomes of endosymbiotic bacteria. The diversity of gene content in these bacteria may reflect both variation in selective forces and contingency-dependent loss of alternative pathways. Using an *in silico* representation of the metabolic network of *Escherichia coli*, we examine the role of contingency by repeatedly simulating the successive loss of genes while controlling for the environment. The minimal networks that result are variable in both gene content and number. Partially different metabolisms can thus evolve owing to contingency alone. The simulation outcomes do preserve a core metabolism, however, which is over-represented in strict intracellular bacteria. Moreover, differences between minimal networks based on lifestyle are predictable: by simulating their respective environmental conditions, we can model evolution of the gene content in *Buchnera aphidicola* and *Wigglesworthia glossinidia* with over 80% accuracy. We conclude that, at least for the particular cases considered here, gene content of an organism can be predicted with knowledge of its distant ancestors and its current lifestyle.**

Naturally evolved, nearly minimal gene sets in closely related intracellular symbionts contain substantial differences[2]. The diversity of these evolved minimal gene sets may be the product of three fundamental processes: differences in initial genetic makeup; variation in selective forces within host cells; and differences in the order of gene deletions, resulting in a choice between alternative cellular pathways[2]. By modelling the reductive evolution of a detailed metabolic network, we first explore the evolutionary significance of the last of these alternatives.

Using the metabolic network of *Escherichia coli* K12 (ref. 3) as our model system has several advantages. First, the best evidence for the presence of alternative pathways within and across species comes from studies of metabolic networks[4]. Second, flux balance analysis provides a rigorous modelling framework for studying the impact of gene deletions[4,5]; the method relies on optimizing the steady-state use of the metabolic network to produce biomass components. Third, not only is the metabolic network of *E. coli* K12 one of the best studied cellular subsystems, but this organism is also a close relative of several endosymbiotic organisms[6], including *Buchnera aphidicola* and *Wigglesworthia glossinidia*. Cellular domestication has resulted in the elimination of 70–75% of the ancestral genome in these latter organisms[7].

The previously reconstructed metabolic network of *E. coli*[3] consists of 904 genes and 931 unique biochemical reactions, and incorporates external nutrients and the corresponding transport processes. The composition of a 'minimal reaction set' has been previously shown to depend strongly on the given environmental conditions[8]. Gradual evolution towards minimal genomes and the role of chance in this process, however, have remained unexplored. The smallest sets of genes that are compatible with cellular life will relate to the most favourable conditions, in which most nutrients are available from the environment. This situation is approximated by organisms with a strict intracellular lifestyle, where the host provides most of their nutrients[2]. Accordingly, we first characterized the simulated evolution of the network under nutrient-rich conditions (Supplementary Tables 1–3).

To explore systematically the combinatorial set of minimal metabolic reaction sets, we elaborated a simple algorithm for simulating gradual loss of metabolic enzymes. We remove a randomly chosen gene from the network and calculate the impact of this deletion on the production rate of biomass components (a proxy for fitness). If this rate is nearly unaffected, the deletion is assumed to be viable and the enzyme is considered to be permanently lost; otherwise, the gene is restored to the network. This procedure is repeated until no further enzymes can be deleted; that is, all remaining genes are essential for survival of the cell. This simulation was repeated 500 times, with each run providing an independent evolutionary outcome.

The resulting networks share on average 77% of their reactions, whereas only 25% would be shared by randomly deleting the same number of genes (Fig. 1a). This suggests that both selective constraints and historical contingencies influence the reductive evolution of metabolic networks. Owing to alternative metabolic pathways in the original *E. coli* network, numerous functionally equivalent minimal networks are possible, even under identical selective conditions. For the same reason, only 55% of the reactions are recoverable by single-gene deletion studies (Fig. 1b). The number of genes in the minimal networks is also variable (Fig. 1b), suggesting that there are differences in the number of enzymatic steps between alternative pathways. Deletions at the early stages of genome reduction may affect large genomic regions rather than single genes[9]. However, additional simulations showed that, although allowing such block deletions reduces the number of independent gene-loss events, it has no effect on the size and average similarity of the networks evolved (Supplementary Methods and Supplementary Table 4).

To compare our predictions against real evolutionary outcomes, we divided the *E. coli* enzymes into two mutually exclusive groups: enzymes ubiquitously present in the simulated minimal reaction sets (group A), and enzymes absent in some or all of the simulated sets (group B). If our analysis can approximate reductive evolution in other bacteria, we expect systematic differences in the relative frequencies of these enzymes between species with different lifestyles. As expected, the fraction of enzymes with ubiquitous presence in the simulated minimal reaction sets (group A) is especially high in

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany. [2]Department of Zoology, University of Oxford, Oxford OX1 3PS, UK. [3]Faculty of Life Sciences, The University of Manchester, Michael Smith Building, Oxford Road, Manchester M13 9PT, UK. [4]Department of Biology & Biochemistry, University of Bath, Claverton Down, Bath BA2 7AY, UK. [5]Department of Medical Chemistry, Semmelweis University, PO Box 260, H-1444 Budapest, Hungary.
*These authors contributed equally to this work.



© 2006 Nature Publishing Group

intracellular parasites and endosymbionts as compared with free-living microbes (Fig. 1c).

To investigate further how accurately the model describes reductive evolution in nature, we focused our simulations on three fully sequenced genomes of *B. aphidicola* strains[10–12] and *W. glossinidia*[13]. These are close relatives of *E. coli* with an evolved intracellular endosymbiotic lifestyle. Gene acquisition must have been a negligible factor in the evolution of these lineages (Supplementary Methods), providing a unique opportunity to study reductive evolution. Setting boundary conditions that mimic the relevant nutrient conditions and selective forces (Supplementary Tables 2 and 3), we performed simulations as described above.

Detailed physiological studies have shown that *Buchnera* supply their aphid hosts with riboflavin[14] and essential amino acids[15] that are lacking in their hosts' diets. To quantify the agreement between our predictions and the observed reductive evolution in *Buchnera*, while considering gene-content variation in simulated minimal genomes, we used a combined measure of sensitivity and specificity[16]. For each possible cutoff (that is, the minimal fraction of simulated genomes in which a gene must be present to predict its presence in *Buchnera*), Fig. 2a shows the fraction of true-positive predictions (sensitivity) plotted against the fraction of false-positive predictions (1–specificity). The area under the resulting curve gives a cutoff-independent measure of predictive accuracy[16]. For each of the *Buchnera* strains, the accuracy of the model is ~80% as compared with the 50% expected by chance (Fig. 2a). The above results remain valid when genes putatively transferred horizontally into *E. coli* since its split

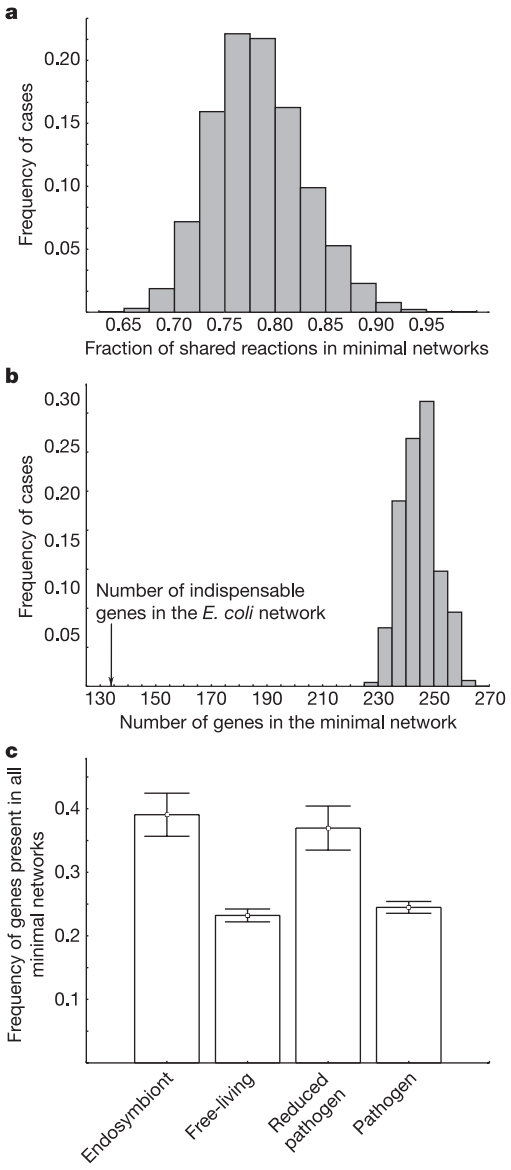



**Figure 1 | General properties of evolved minimal networks.** **a**, Distribution of the fraction of shared metabolic reactions between all possible pairs among 500 simulated minimal networks. Only reactions with annotated enzyme-encoding genes are shown. The resulting networks share 77 ± 4.4% (mean ± s.d.) of their reactions. The 500 networks were generated with random reaction content and the same distribution reaction numbers as the simulants. The average similarity across networks is 25 ± 2.7%. **b**, Distribution of the number of contributing genes in simulated minimal networks. Minimal reaction networks contain, on average, 245 ± 6.48 reactions (mean ± s.d.); however, only 134 of these genes (~55%) have a predicted fitness effect in the full original *E. coli* network (arrow). **c**, Distribution of genes consistently present in minimal networks in organisms with different lifestyles (Supplementary Table 11). Putative orthologues of *E. coli* enzymes were identified in 140 bacterial species. Shown is the fraction of these that are retained in all simulated minimal networks, summarized across species for each of four different lifestyles (values are the mean ± 2 s.e.m.). Analysis of variance: $n = 140$, $F = 62.9$, d.f. = 3, $P < 10^{-6}$.

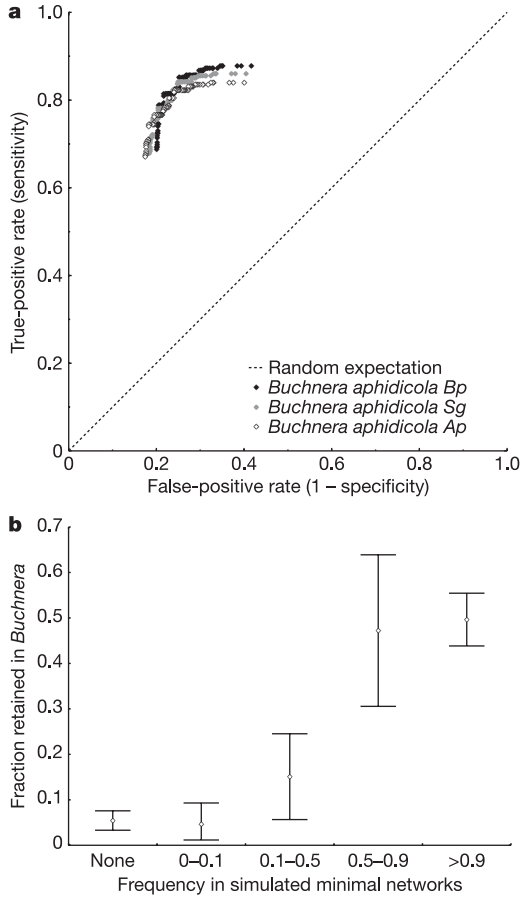



**Figure 2 | Comparison of reaction content of simulated and *Buchnera* metabolic networks.** **a**, Predictive accuracy for all possible cutoffs (receiver operating characteristic curve)[16]. *Bp*: *B. aphidicola*, endosymbiont of *Baizongia pistaciae*; *Sg*: *B. aphidicola*, endosymbiont of *Schizaphis graminum*; *Ap*: *B. aphidicola*, endosymbiont of *Acyrthosiphon pisum*. Overall accuracy (area under curve): $Bp = 0.802$, $Ap = 0.794$, $Sg = 0.800$. All results are highly significant, $P < 10^{-25}$ (see Supplementary Information). **b**, Presence or absence of reactions in *Buchnera aphidicola Bp*, averaged over genes within defined ranges of presence or absence in the simulated minimal reaction sets. Error bars indicate 95% confidence intervals. $\chi^2$-test: $n = 874$, $\chi^2 = 222.6$, d.f. = 4, $P < 10^{-46}$. For results on *Wigglesworthia glossinidia*, see Supplementary Fig. 2.



© 2006 Nature Publishing Group

from the *Buchnera* lineage are excluded from the analysis (Supplementary Methods and Supplementary Table 5). The model also accurately predicts several non-obvious features of *Buchnera* genomes: for example, the retention of particular reactions involved in oxidative phosphorylation and in pyruvate metabolism (Supplementary Table 6).

Consistent with the notion that genes vary widely in their propensity to be lost during reductive evolution, we find a strong correlation between the frequency of a reaction's presence in the simulated reduced networks and its retention in *Buchnera* (Fig. 2b). Metabolic pathways differ widely in their variability across simulated minimal sets (Supplementary Table 7). For example, it seems that there is only one way of producing some key cellular (biomass) components, including compounds for cell wall synthesis and some essential amino acids. By contrast, reactions involved in pyruvate metabolism, nucleotide salvage pathways or transport processes vary in their retention across simulations. For example, there are two distinct pathways by which *E. coli* can activate acetate to acetyl-coenzyme A (ref. 17). These two pathways have been shown experimentally to compensate for deletions in each other in *E. coli*[17], at least under some nutritional conditions. Consistent with this observation, the simulated minimal reaction sets always contain only one of the two pathways; accordingly, *Buchnera* strains have retained only one of the two pathways (Supplementary Table 8).

The above analysis relied on detailed knowledge of the lifestyle of *Buchnera*. Is it possible to predict gene content of an organism with much less information on lifestyle? *Wigglesworthia*, another endosymbiont and close relative of *E. coli*, is an obvious choice. *Wigglesworthia* provides some cofactors and vitamins for its host, the tsetse fly[18]. On the basis of the available physiological information[19], it is possible to model the evolution of the metabolic network of this organism with nearly 76% accuracy for the reaction content (Supplementary Fig. 2 and Table 3). It is likely that the available experiments underestimate the number of cofactors produced by the endosymbiont. We thus elaborated a systematic protocol to find the most likely set of cofactors synthesized by *Wigglesworthia* (Supplementary Methods). Based on the idea of greedy algorithms[20], the protocol iteratively adds biosynthetic components that must be produced for the host and calculates the impact on the accuracy of predicting the real reaction content of *Wigglesworthia*. In each round, the cofactor resulting in the best prediction is kept and a new round of simulations is started, adding again each of the remaining compounds one at a time (Supplementary Methods). The method substantially increases model accuracy up to 84% (Supplementary Table 5). It also results in a series of non-trivial predictions on the metabolic capability of *Wigglesworthia*. For example, it suggests that this organism retained the ability to synthesize not only protohaem, but also another related cofactor, haem O (Supplementary Methods).

Under a given selection pressure, simulated minimal reactions sets share 82% (*Wigglesworthia*) and 88% (*Buchnera*) of their reactions, respectively. This value drops to 65% when minimal gene sets across different models are compared. This suggests that variability in gene content among species reflects both variation in selection pressures and chance events in the evolutionary history of the endosymbionts (Supplementary Table 9).

Each loss of a reaction reduces the space available for further reductive evolution. This is most obvious for physiologically fully coupled reactions (such as those in linear pathways), which can only fulfil their metabolic function together[21]. As predicted, members of pairs are either lost or retained together in the investigated endosymbionts in 74–84% of cases, whereas only ~50–55% would be expected by chance (Supplementary Table 10).

Deviations between the model predictions and gene content of endosymbionts might be due to incomplete biochemical knowledge or inaccuracies in modelling the types and relative amounts of nutrient conditions and biosynthetic components required by the endosymbiont or the host cell. Finally, hosts and endosymbionts interact in ways that are not completely understood, and biomass production may be only a rough proxy for endosymbiont fitness. These caveats aside, our approach might be considered a step towards a predictive theory of gene-content evolution. Complementary to traditional approaches, in which lifestyle is inferred from genomic data, it seems possible to take an organism's ecology and to predict which genes it should have by *in silico* network analysis. Moreover, we find that evolutionary paths are contingent on prior gene deletion events, resulting in networks that generally do not represent the most economical solution in terms of the number of genes retained. Thus, history and chance seem to have significant roles not only in adaptive[22] but also in reductive evolution of genomes.

These results also have implications for the search for a minimal genome. By using comparative genomics[23,24] and systematic gene knock-out studies[25–27], traditional analyses of minimal gene sets aim to define a repertoire of genes that is necessary and sufficient to support cellular life[2]. The theoretical foundations of the minimal genome concept have remained, however, largely unexplored. We have established that the catalogue of essential genes in free-living species identified by single-gene deletion studies will underestimate the minimal gene set for metabolic system by about 45% (Fig. 1b). Such considerations, and the simulation techniques used to reach these conclusions, should inform attempts by experimentalists to construct minimal genomes by gradual evolution in the laboratory[28,29].

## METHODS

For full details on orthologue detection and statistical analyses, see Supplementary Methods.

**Flux balance analysis of the *E. coli* network.** A reconstructed metabolic network (*i*JR904 GSM/GPR)[3] of *E. coli* K12 was used in this study. The model consists of 931 unique biochemical reactions (including transport processes) and 904 genes. The metabolic reconstruction gives accurate information on the stoichiometry and direction of enzymatic reactions, on the presence of isoenzymes, and on enzymatic complexes. Details of flux balance analysis of the *E. coli* metabolic network have been described elsewhere[4,5]. In brief, it involves two fundamental steps: first, specification of mass balance constraints around intracellular metabolites; and second, maximization of the production of biomass components. The assumption of a steady state of metabolite concentrations specifies a series of linear equations of individual reaction fluxes, which is written in the form $S\mathbf{v} = 0$, where $S$ is the $mn$ stoichiometric matrix ($m$ being the number of metabolites and $n$ being the number of reactions) and $\mathbf{v}$ is the vector of individual fluxes through the network. An individual element $S_{ij}$ gives the contribution of the $j$-th reaction to metabolite $i$. A biomass reaction describes the relative contribution of metabolites to the cellular biomass. Availability of nutrients and directions of individual reactions were included as boundary conditions (Supplementary Tables 1–3). Using the linear programming package CPLEX 9.0.0, we identified the flux distribution that maximizes the rate of biomass production.

**Simulations on reductive evolution.** Following previously elaborated protocols[5], we start by investigating the behaviour of the *E. coli* metabolic network model under a given environmental condition (Supplementary Tables 1–3). Next, we remove a randomly chosen enzyme from the network and calculate the impact of this deletion on the production of biomass components (for a list, see Supplementary Tables 1–3). Enzyme deletions were simulated by constraining the flux of the corresponding reactions to zero and calculating the corresponding knockout flux configuration by established protocols[4,5]. A gene was classified as having no fitness effect if the biomass production rate of the knockout strain was reduced by less than a given cutoff; different cutoffs led to very similar results (Supplementary Table 5). Deletions of isoenzymes were considered to have no impact on fitness as long as at least one member remained. By contrast, deletion of any of the subunits of a protein complex was considered to result in zero flux through the corresponding reactions. Reactions with no annotated encoding genes were retained throughout the simulations. If the fitness effect of a simulated gene deletion was below the cutoff, the deletion was assumed to be viable and the enzyme was considered to be permanently lost. Otherwise, the gene was restored to the network. The procedure was repeated until no further enzymes could be deleted. This simulation was repeated 500 times; each run provided an independent evolutionary outcome.

The simulations that mimic the evolution of the *Buchnera* metabolic network relied on available biochemical evidence suggesting that glucose and glutamate are the principal carbon sources from which essential amino acids and riboflavin



© 2006 Nature Publishing Group

must be produced for the host (Supplementary Table 2). Besides amino acids, mononucleotides and fatty acids, among others, the biomass components that must be synthesized also include riboflavin. A previous study[30] estimated the population size of *Buchnera* as $N_e \approx 10^2$–$10^3$. Gene deletions are effectively neutral and can thus spread through a population if $|N_e s| < 1$, where $s$ is the selective effect of the gene deletion. Accordingly, the cutoff for the fitness effect of simulated gene deletions was set to $10^{-2}$. A less stringent cutoff (0.1) gave very similar results (Supplementary Table 6). For details of *Wigglesworthia* uptake and selective conditions, see Supplementary Table 3.

**Received 7 November; accepted 27 December 2005.**


1. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428,** 37–43 (2004).
2. Koonin, E. V. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nature Rev. Microbiol.* **1,** 127–136 (2003).
3. Reed, J. L., Vo, T. D., Schilling, C. H. & Palsson, B. O. An expanded genome-scale model of *Escherichia coli K-12* (iJR904 GSM/GPR). *Genome Biol.* **4,** R54 (2003).
4. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2,** 886–897 (2004).
5. Edwards, J. S. & Palsson, B. O. The *Escherichia coli MG1655* in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97,** 5528–5533 (2000).
6. Gil, R., Latorre, A. & Moya, A. Bacterial endosymbionts of insects: insights from comparative genomics. *Environ. Microbiol.* **6,** 1109–1122 (2004).
7. Klasson, L. & Andersson, S. G. Evolution of minimal-gene-sets in host-dependent bacteria. *Trends Microbiol.* **12,** 37–43 (2004).
8. Burgard, A. P., Vaidyaraman, S. & Maranas, C. D. Minimal reaction sets for *Escherichia coli* metabolism under different growth requirements and uptake environments. *Biotechnol. Prog.* **17,** 791–797 (2001).
9. Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol.* **2,** research0054 (2001).
10. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. & Ishikawa, H. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407,** 81–86 (2000).
11. van Ham, R. C. *et al.* Reductive genome evolution in *Buchnera aphidicola*. *Proc. Natl Acad. Sci. USA* **100,** 581–586 (2003).
12. Tamas, I. *et al.* 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296,** 2376–2379 (2002).
13. Akman, L. *et al.* Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genet.* **32,** 402–407 (2002).
14. Nakabachi, A. & Ishikawa, H. Provision of riboflavin to the host aphid, *Acyrthosiphon pisum*, by endosymbiotic bacteria, *Buchnera*. *J. Insect Physiol.* **45,** 1–6 (1999).
15. Baumann, P. *et al.* Genetics, physiology, and evolutionary relationships of the genus *Buchnera*—intracellular symbionts of aphids. *Ann. Rev. Microbiol.* **49,** 55–94 (1995).
16. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).
17. Kumari, S., Tishel, R., Eisenbach, M. & Wolfe, A. J. Cloning, characterization, and functional expression of *acs*, the gene which encodes acetyl coenzyme A synthetase in *Escherichia coli*. *J. Bacteriol.* **177,** 2878–2886 (1995).
18. Zientz, E., Dandekar, T. & Gross, R. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiol. Mol. Biol. Rev.* **68,** 745–770 (2004).
19. Nogge, G. Significance of symbionts for the maintenance of an optimal nutritional state for successful reproduction in haematophagous arthropods. *Parasitology* **82,** 101–104 (1981).
20. Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. *Introduction to Algorithms* (MIT Press, Cambridge, MA, 2001).
21. Burgard, A. P., Nikolaev, E. V., Schilling, C. H. & Maranas, C. D. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res.* **14,** 301–312 (2004).
22. Travisano, M., Mongold, J. A., Bennett, A. F. & Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **267,** 87–90 (1995).
23. Mushegian, A. R. & Koonin, E. V. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl Acad. Sci. USA* **93,** 10268–10273 (1996).
24. Gil, R., Silva, F. J., Pereto, J. & Moya, A. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68,** 518–537 (2004).
25. Westers, H. *et al.* Genome engineering reveals large dispensable regions in *Bacillus subtilis*. *Mol. Biol. Evol.* **20,** 2076–2090 (2003).
26. Kolisnychenko, V. *et al.* Engineering a reduced *Escherichia coli* genome. *Genome Res.* **12,** 640–647 (2002).
27. Hutchison, C. A. *et al.* Global transposon mutagenesis and a minimal *Mycoplasma genome*. *Science* **286,** 2165–2169 (1999).
28. Nilsson, A. I. *et al.* Bacterial genome size reduction by experimental evolution. *Proc. Natl Acad. Sci. USA* **102,** 12112–12116 (2005).
29. Oliver, S. G. From DNA sequence to biological function. *Nature* **379,** 597–600 (1996).
30. Mira, A. & Moran, N. A. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic bacteria. *Microb. Ecol.* **44,** 137–143 (2002).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** We thank C. von Mering for providing early access to the updated STRING database. C.P., B.P. and P.C. are supported by the Hungarian Scientific Research Fund (OTKA). C.P. is also supported by an EMBO Long-term Fellowship. B.P. is a Fellow of the Human Frontier Science Program. M.J.L. acknowledges financial support by the Deutsche Forschungsgemeinschaft. Work on systems biology in S.G.O.'s laboratory is supported by the Biotechnology and Biological Sciences Research Council.

**Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to L.D.H. (l.d.hurst@bath.ac.uk).



© 2006 Nature Publishing Group

# Network-level architecture and the evolutionary potential of underground metabolism

Richard A. Notebaart[a,1,2], Balázs Szappanos[b,1], Bálint Kintses[b,1], Ferenc Pál[b], Ádám Györkei[b], Balázs Bogos[b], Viktória Lázár[b], Réka Spohn[b], Bálint Csörgő[b], Allon Wagner[c], Eytan Ruppin[c,d], Csaba Pál[b,2], and Balázs Papp[b,2]

[a]Radboud Institute for Molecular Life Sciences, Centre for Bioinformatics and Systems Biology, Radboud University Medical Centre, 6525 GA, Nijmegen, The Netherlands; [b]Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, 6726, Szeged, Hungary; and [c]The Blavatnik School of Computer Science and [d]Sackler School of Medicine, Tel-Aviv University, Tel-Aviv 69978, Israel

Edited by Jeffrey H. Miller, University of California, Los Angeles, CA, and accepted by the Editorial Board July 3, 2014 (received for review April 3, 2014)

**A central unresolved issue in evolutionary biology is how metabolic innovations emerge. Low-level enzymatic side activities are frequent and can potentially be recruited for new biochemical functions. However, the role of such underground reactions in adaptation toward novel environments has remained largely unknown and out of reach of computational predictions, not least because these issues demand analyses at the level of the entire metabolic network. Here, we provide a comprehensive computational model of the underground metabolism in *Escherichia coli*. Most underground reactions are not isolated and 45% of them can be fully wired into the existing network and form novel pathways that produce key precursors for cell growth. This observation allowed us to conduct an integrated genome-wide in silico and experimental survey to characterize the evolutionary potential of *E. coli* to adapt to hundreds of nutrient conditions. We revealed that underground reactions allow growth in new environments when their activity is increased. We estimate that at least ∼20% of the underground reactions that can be connected to the existing network confer a fitness advantage under specific environments. Moreover, our results demonstrate that the genetic basis of evolutionary adaptations via underground metabolism is computationally predictable. The approach used here has potential for various application areas from bioengineering to medical genetics.**

enzyme promiscuity | evolutionary innovation | molecular evolution | network evolution | phenotype microarray

How do new molecular pathways evolve? In the best-studied molecular networks, small-molecule metabolism, the prevailing paradigm is that new pathways are patched together from preexisting enzymes borrowed from different parts of the network (1–3). Central to this "patchwork" model of pathway evolution is the notion that many enzymes have limited substrate specificities and can catalyze, albeit at low rates, reactions other than those for which they have evolved (also referred to as enzyme promiscuity) (4). These so-called underground (5) or side activities are prevalent (6–8) and were shown to serve as starting points for the evolution of novel functions both in directed evolution experiments (9) and in the diversification of gene families in the wild (7). However, how the underground catalytic repertoire encoded in the genome can generate novelties within the context of the existing metabolic network remains unknown. Do underground reactions remain isolated, or can they potentially be wired into the native network and allow the organism to survive in novel environments? Furthermore, would it be possible to computationally predict the genetic basis of phenotypic evolution based on a detailed knowledge of the organism's underground metabolism? Answering these questions requires both large-scale data on underground enzyme activities and systems-level approaches to analyze metabolic capabilities. Although systematic detection of underground activities by unbiased high-throughput approaches is not feasible at present, the accumulated knowledge of enzyme biochemistry in the well-studied prokaryote *Escherichia coli* provides a valuable resource of such nonnative enzyme activities (10). Thus, to explore the architecture of underground metabolism and its evolutionary potential, we compiled a comprehensive set of experimentally reported side activities of *E. coli* enzymes and integrated these reactions into a global metabolic network reconstruction of the same organism (11). Analysis of this extended network revealed that a substantial fraction of underground reactions can be fully integrated into the existing metabolism and participate in potential pathways that produce key precursors for cell growth. Using metabolic modeling, we then predicted specific environmental conditions under which such biologically relevant underground reactions confer a growth advantage, and hence deliver a phenotypic novelty. Our analyses revealed that the set of known underground reactions has a significant potential both to increase fitness in existing environments and to exploit new nutrient sources. A genome-wide gene overexpression screen across hundreds of carbon sources showed a good agreement with the model's predictions, which illustrates that the genetic basis of phenotypic novelties can be predicted based on the knowledge of underground metabolism.

## Results

**Reconstructing the Underground Metabolic Network of *E. coli*.** To reconstruct a global metabolic network of *E. coli* that also

### Significance

**Understanding how new metabolic pathways emerge is one of the key issues in evolutionary and systems biology. The prevailing paradigm is that evolution capitalizes on the weak side activities of preexisting enzymes (i.e. underground reactions). However, the extent to which underground reactions provide novelties in the context of the entire cellular system has remained unexplored. In this study, we present a comprehensive computational model of the underground metabolism of *Escherichia coli*. Together with a high-throughput experimental survey across hundreds of nutrient environments we predicted and confirmed new functional states of metabolism in which underground reactions allow growth when their activity is increased. Our approach has important implications for biotechnological and medical applications, such as understanding gain-of-function mutations in tumor development.**

Author contributions: R.A.N., B.S., B.K., C.P., and B.P. designed research; R.A.N., B.S., B.K., F.P., A.G., and B.B. performed research; B.B., V.L., R.S., B.C., A.W., and E.R. contributed new reagents/analytic tools; R.A.N., B.S., B.K., F.P., A.G., and B.P. analyzed data; and R.A.N., B.S., B.K., E.R., C.P., and B.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.H.M. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

[1]R.A.N., B.S., and B.K. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: Richard.Notebaart@radboudumc.nl, cpal@brc.hu, or pappb@brc.hu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1406102111/-/DCSupplemental.



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

incorporates underground reactions, we collected information for each *E. coli* enzyme on catalytic activities that have been detected in vitro and involve nonnative substrates (i.e., metabolites not considered as primary substrates of the enzyme) by database and literature mining (*Materials and Methods*). Altogether, we compiled 262 underground reactions and 277 novel compounds not present in the native network (Fig. 1 *A* and *B* and Dataset S1). Two analyses suggest that the assembled underground reactions occur at very low rates in *E. coli,* and hence they conceptually differ from enzyme multispecificity (12), where several reactions are catalyzed with similar efficiency (4). First, side reactions are catalyzed with significantly lower catalytic efficiency ($k_{cat}/K_m$) than native reactions of the same enzymes (~220-fold pairwise difference, $P < 0.001$, Fig. 1*C* and *SI Appendix*, Table S1). Second, by compiling data from the *E. coli* Metabolome Database (13) (*Materials and Methods*), we found that metabolites introduced into the network via underground activities only are strongly underrepresented among empirically observed metabolites (22% versus 81% for native metabolites, $P < 10^{-15}$, Fig. 1*D*; other metabolomics datasets yielded similar results, *SI Appendix*, Fig. S1). This suggests that these novel metabolites are either absent or present at very low abundances in the cell, hence remaining physiologically irrelevant. Because underground reactions occur at very low rates, they are unlikely to have essential roles in the wild-type background. Nevertheless, these side activities could potentially be enhanced by adaptive mutations (4) and thus provide raw material for network expansion.

**Underground Reactions Can Often Be Wired into the Native Network.** In principle, novel biochemical reactions can introduce cross-wirings, create dead-ends, or remain isolated from the rest of the metabolic network (Fig. 2*A*). Our reconstruction suggests that enzyme side activities most often create cross-wirings: Forty-five percent of underground reactions can be fully connected to the network (see Fig. 1*B* for examples), and only 22% of them are completely isolated from native metabolism. What factors influence the network positions of cross-wiring underground reactions? Side activities are most frequently caused by substrate ambiguity (14), that is, when an enzyme catalyzes the same transformation on multiple structurally related substrates. This chemical constraint indeed yields nonrandom positioning at the network level: Native and underground activities of the same enzyme tend to participate in the same metabolic subsystem and are separated by fewer reaction steps than expected by chance ($P < 10^{-4}$ and $P = 0.0066$, respectively, Fig. 2 *B* and *C*; also see *SI Appendix*, Fig. S2). Importantly, this result is based on a subset of the underground network that comes from a systematic substrate specificity screen (6) and hence is not distorted by potential investigation bias (*Materials and Methods*).

**Underground Reactions Potentially Contribute to Biologically Relevant Pathways Akin to Native Ones.** The observation that many underground reactions can be wired into the native network raises the question of their potential biological relevance. Do they have pathway-level properties akin to native ones, and can they potentially contribute to the formation of key precursors needed for growth (biomass components)? We used elementary flux mode (EFM) analysis (15) to investigate these questions. EFM is a mathematical approach to decompose the network into

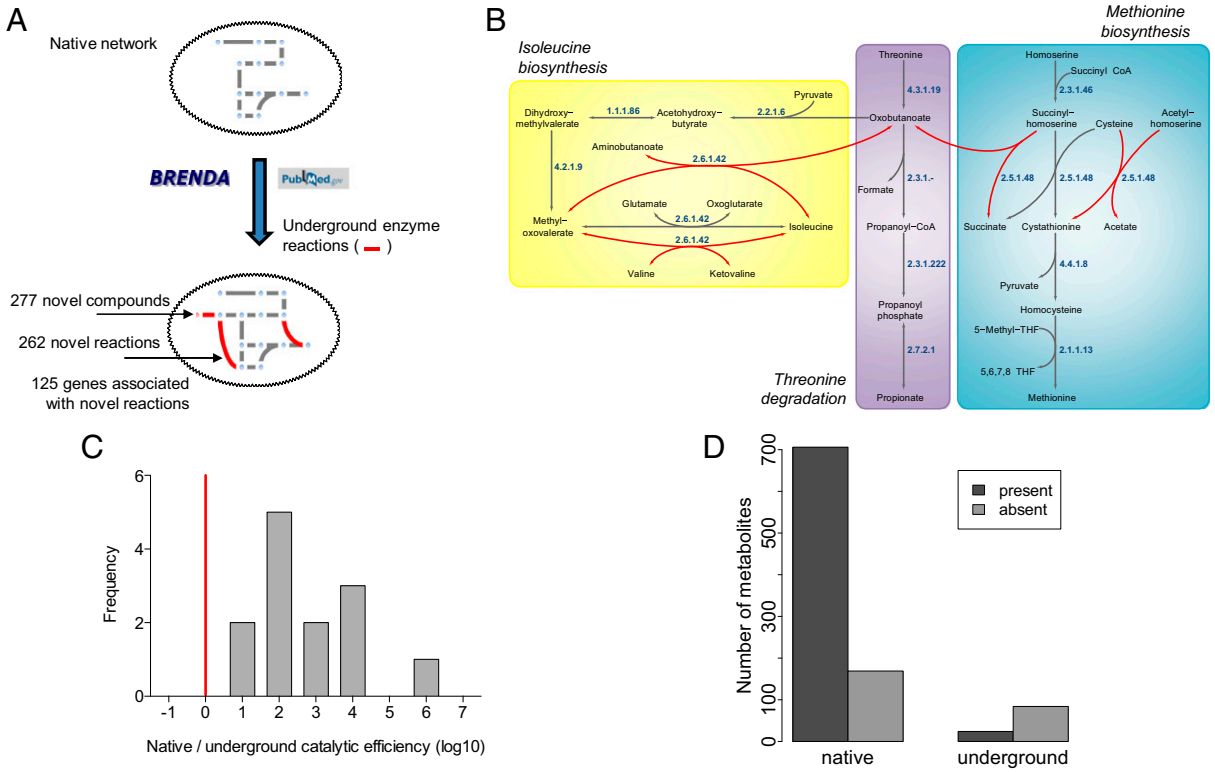



**Fig. 1.** Reconstruction and validation of the *E. coli* underground metabolic network. (*A*) Schematic overview of the reconstruction process. Underground enzyme reactions were added to the native iAF1260 metabolic reconstruction of *E. coli* using the BRENDA (10) database and literature information. (*B*) Example of integrating underground reactions (red arrows) into the native network (gray arrows) as demonstrated in three interconnected amino acid metabolism pathways. Cofactors and small molecules ($H_2O$, $CO_2$, etc.) are not shown. Numbers next to the arrows denote the Enzyme Commission numbers associated with the corresponding reactions. (*C*) Distribution of the relative catalytic efficiency between native and underground substrates of the same enzyme (logarithm of the ratio between the $k_{cat}/K_m$ values of the native and underground substrates). Red line represents equal catalytic efficiency. See also *SI Appendix*, Table S1. (*D*) Number of native and underground metabolites that are present/absent in the *E. coli* Metabolome Database (13). The difference is highly significant ($P < 10^{-15}$, Fisher's exact test). Underground metabolites are defined as those that are consumed and/or produced by underground reactions only.



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

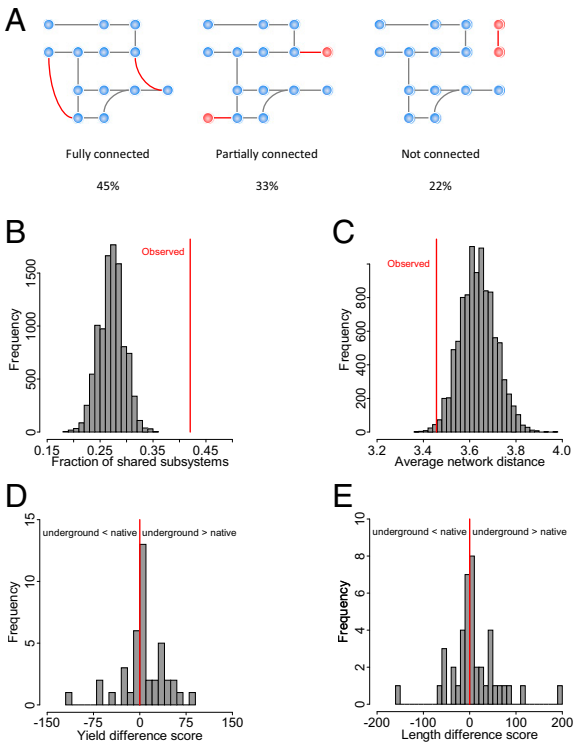



**Fig. 2.** Network properties of underground reactions. (*A*) Connectedness of underground reactions in the native network. Reactions are either fully connected (cross-wiring) (i.e., both substrates and products are present in the native network), partially connected (i.e., either the substrates or the products are present), or isolated from the rest of the network (not connected). Connectedness was assessed for each underground reaction one by one. Nodes denote metabolites (blue, native; red, those associated with underground reactions only), and edges indicate reactions (gray, native; red, underground). (*B* and *C*) Distribution of the fraction of shared metabolic subsystem (*B*) and average network distance (*C*) in 10,000 samples of randomly assigned underground–native reaction pairs. Red lines indicate observed values for underground–native reaction pairs annotated to the same enzymes in our reconstruction. (*D* and *E*) Distribution of pairwise differences in yield (*D*) and length (*E*) of elementary pathways formed by native and underground reactions of the same enzyme. For each enzyme we calculated a score measuring the yield (and length) difference between two samples of pathways (one containing the underground and another one the native reaction of the enzyme). See *Materials and Methods*. Red line represents the null expectation of no difference. $P = 0.21$ and $P = 0.45$, respectively.

biochemically relevant pathways that can operate in steady state from nutrient uptake to biomass component production. A sampling of such elementary paths showed that all underground reactions that can be fully wired into the native network and can carry flux in standard glucose medium also take part in at least one biomass-forming pathway. The chemical yields and lengths of pathways formed by native and underground reactions of the same enzyme are comparable to each other in standard glucose medium (Fig. 2 *D* and *E*). Taken together, a substantial fraction of the known underground catalytic repertoire of *E. coli* can be seamlessly integrated into the existing network and participate in biologically potentially relevant pathways.

However, the fact that underground reactions occur at very low rates suggests that they have not been exploited by evolution so far. Why should this be so? This issue is relevant because tradeoffs between novel and existing enzymatic functions are generally weak (4, 9) and can be readily resolved by gene duplication (16). We consider two alternative hypotheses to resolve this issue. First, underground reactions might interfere with existing processes (17) and are therefore disfavored by selection. Second, underground reactions might endow the cell with novel capabilities, but only under specific environmental conditions that the population has not regularly encountered during its evolutionary history.

**No Evidence for the Detrimental Nature of Underground Reactions.** To test whether underground reactions tend to introduce harmful metabolites, we focused on metabolite toxicity, which has been implicated in the interference between a novel pathway and native metabolism (17). Toxicity of metabolites, as measured by $IC_{50}$ values (half maximum inhibitory concentration), were predicted using a chemoinformatics tool trained on data measuring the susceptibility of *E. coli* against a diverse set of chemicals (18). Our analysis revealed no significant difference in toxicity between novel compounds introduced by underground reactions and native compounds associated with the same enzymes ($P = 0.81$, Fig. 3*A*). Thus, metabolite-induced damage is unlikely to pose a general limit on network expansion. A second possibility is that activation of underground reactions would incur a fitness cost by diverting metabolites from existing biomass-producing pathways. To address this scenario, we applied the EDGE algorithm (19), which identifies metabolic reactions that decrease growth when enforced to be active (i.e., higher flux). We found no support for this scenario: Underground reactions are not more likely to decrease growth when enforced to be active compared with native nonessential reactions of the same enzymes ($P = 0.22$; *Materials and Methods*).

**Predicting the Adaptive Potential of Underground Metabolism.** To test the impact of underground reactions on adaptation to specific environments we first systematically predicted growth phenotypes using flux balance analysis (FBA) (20) across a diverse set of environments. FBA is a modeling approach for analyzing metabolite flows from nutrient uptake toward production of metabolites in large-scale metabolic networks without the need for enzyme kinetic information. This modeling framework has been shown to be successful in predicting the growth capacity of wild-type *E. coli* across nutrient conditions and the viability of single-gene disruptions (11). The set of 2,754 environments defined here encompasses the complete range of carbon, nitrogen, sulfur, and phosphorus sources that can be imported into the network (Dataset S2) and thereby represents a comprehensive sample of the external nutrient space. As a baseline, we report that the native *E. coli* network shows in silico growth in 645 of these environments (Dataset S2). Next, we predicted the nutrient utilization profile for a metabolic network extended with all cross-wiring underground reactions. We note that because FBA seeks the most optimal flux distribution a larger network is expected to display slightly more optimal behavior under most conditions (i.e., pathways with somewhat higher biomass production can potentially be found in larger networks). To avoid such artifacts, we only considered fitness gains that are of at least 5% (*Materials and Methods*). The exhaustive list of investigated conditions enabled the identification of 19 otherwise non-utilizable nutrients on which the underground network allowed growth, hence introducing an innovation (21), and a further 31 environments where it provides a clear quantitative growth advantage (Dataset S2; see Fig. 3*B* for in silico fitness gains detected in aerobic carbon sources). Remarkably, incorporating the set of cross-wiring underground reactions into the native metabolic network increased its reaction content by 10.8% (i.e., from 1,257 to 1,393 intracellular reactions capable of carrying a flux) while concurrently expanding its scope of utilizable nutrients by 2.9% (from 645 to 664), underscoring the innovative potential of underground metabolism.

Next, we determined which underground reactions contribute to these novelties (*Materials and Methods*). We found that ∼15% of cross-wiring underground reactions confer an advantage when added individually to the network (Fig. 3*B*), and a further 5% do so in combination with other underground activities (Dataset S2). Furthermore, ∼11% contributed to metabolic innovations by increasing the scope of utilizable nutrients. In summary, these



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

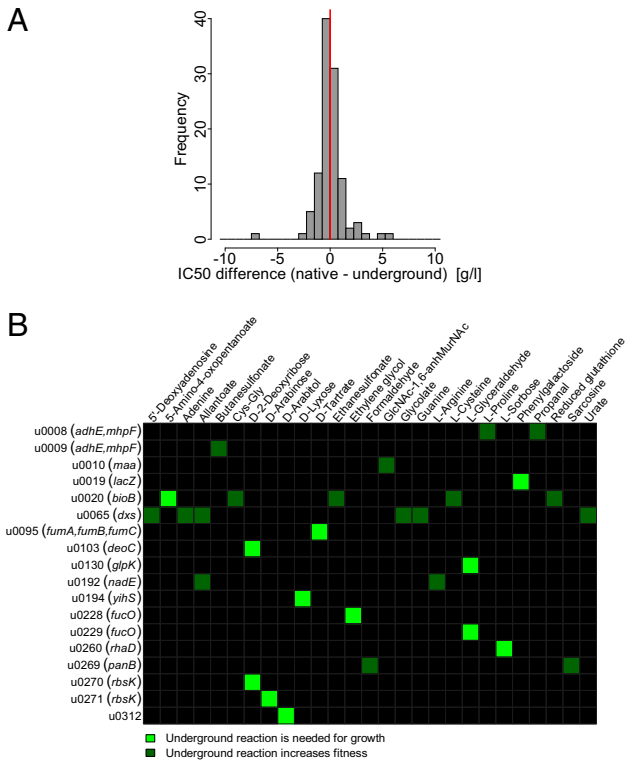



**Fig. 3.** Functional consequences of introducing underground reactions into the native network. (*A*) Metabolite toxicity. The plot shows the distribution of the pairwise differences in predicted $IC_{50}$ between existing native and novel underground metabolites of the same enzymes. The vertical red line indicates the null expectation of no difference. $P = 0.81$, Wilcoxon rank-sum test (see *Materials and Methods* for details). (*B*) Heat map of in silico fitness advantages gained by adding single underground activities to the native network across different nutrient conditions. Only aerobic carbon sources are depicted here. Bright green squares indicate innovation, that is, utilization of a nutrient on which the native network showed no growth at all (i.e., fitness was zero). We note that u0192, although not being innovative under aerobic carbon sources, allows growth on L-arginine as a sole carbon source under anaerobic condition.

simulations suggest that an important fraction of the biochemically feasible evolutionary raw material in *E. coli* could potentially contribute to adaptation to novel nutrient environments with respect to growth.

**Genome-Wide Experimental Screen Identifies Novel Phenotypes Conferred by Underground Activities.** To experimentally assess the role of underground reactions in adaptation to nutrient environments we performed an in vivo genome-wide screen to identify genes that enable growth on a new carbon source when strongly overexpressed. Our approach rests on the assumption that amplification of a single underground activity can confer a new phenotype under specific conditions. Indeed, strong gene overexpression through the viral promoter and high-copy plasmid that we apply here (typically three to four orders of magnitude increase in protein level; see *SI Appendix*) has been previously used to identify underground activities that provide metabolic suppression (22, 23). Following an established protocol (22, 24), a pooled collection of every *E. coli* ORF cloned into an expression vector (25) was tested for the provision of growth advantage in a large array of diverse carbon sources (~4,300 ORFs in 194 conditions; see *SI Appendix*, Fig. S3 for a workflow). Following verification, we identified 17 ORFs whose overexpression improved growth in at least one of 17 specific carbon sources (9% of the investigated conditions; Table 1, Dataset S3, and *SI Appendix*, Fig. S4). Importantly, more than half of these novel growth phenotypes were conferred by enzyme-encoding genes with underground activities that are either biochemically already described for the corresponding *E. coli* enzyme or hypothesized based on reaction chemistry or evidence from orthologous enzymes (Fig. 4*A*; for more details see Dataset S4). Furthermore, six of these enzymes not only improved growth but also produced a metabolic innovation in one of the five carbon sources where wild-type *E. coli* was unable to grow (Fig. 4*A*). We therefore estimate that amplifying single underground enzyme activities expands the scope of utilizable carbon sources by ~6% in this species (from 85 to 90 of the carbon sources experimentally tested here). This figure is expected to substantially underestimate the true evolutionary potential of underground metabolism in nutrient adaptations for at least four reasons. First, our screen captures only those innovations that can be accessed by increasing the activity of a single underground reaction. In principle, phenotypic novelties might also rely on multistep pathways and hence would go undetected in our screen. Indeed, our in silico screen identified three environments in which more than one underground reaction is jointly needed to confer a fitness benefit (Dataset S2). Second, underground activities with relatively modest beneficial effects might remain undetected in our assay. For example, there must be a lower threshold for fitness gains that is necessary for a clone within the pool of overexpression strains to overgrow the negative control. Third, overexpression is unlikely to cover the full dynamic range within which mutations can increase the catalytic efficiency of underground activities in nature. Finally, it is also possible that an underground activity with a potential fitness advantage remains silent because some of the required native reactions are unavailable for regulatory reasons.

**Experimentally Determined Metabolic Novelties Show Good Agreement with in Silico Predictions.** Computational predictions and results of the genome-wide survey showed a highly significant overlap ($P < 10^{-13}$, *SI Appendix*, Table S2). In particular, modeling successfully predicted 44% of the carbon sources that can be used by the amplification of enzyme side activities (Fig. 4). For example, D-lyxose occurs rarely in nature (26) and wild-type *E. coli* is unable to degrade it. Our model predicts that establishing a pathway to use this nutrient only requires a single metabolic reaction, the side activity of mannose isomerase (YihS). Our experimental screen confirms that only the overexpression of *yihS*, but not of other genes, enables growth on D-lyxose. Novel metabolic phenotypes missed by our model are related to catalytic functions that have not been described in *E. coli* (Dataset S4), revealing that several underground reactions are yet to be discovered. Finally, we note that repeating the computational predictions using a reconstruction based on a more recent version of the *E. coli* native network (27) also yielded highly significant prediction accuracy ($P < 10^{-7}$; see *SI Appendix*, Table S3 for details).

## Discussion

The specificity of enzymes is inherently limited and the catalytic side activities stemming from this imperfection are thought to provide the raw material for the evolution of novel enzymatic functions (4). However, a hitherto uncharacterized portion of the catalytic raw material might be isolated from the rest of the metabolic network, produce harmful metabolites (17), or only contribute to the formation of pathways that are redundant with existing network parts (23). Because such underground activities are unlikely to contribute directly to adaptive novelties at the phenotypic level, they may never be realized by evolution. Our study attempts to systematically assess the biological relevance and evolutionary potential of underground reactions within the context of the entire metabolic network. We report that known underground reactions of *E. coli* can often be integrated into the native metabolic network and contribute to pathways producing precursors for cellular growth, with efficiencies comparable to native ones. Furthermore, as opposed to a previous case study (17), we found no evidence that underground reactions have





Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

**Table 1. Experimental results of the genome-wide overexpression screen**

| Putative mode of action | No. of carbon sources |
|---|---|
| Native catalytic activity (*pepQ, rihB*) | 2 |
| Underground catalytic activity(*bglB, dmlA, fumA, fumB, lacZ, leuB, mhpF, rbsK, ybfF, yihS*) | 9 |
| Regulator of metabolic operon (*bglJ*) | 2 |
| Stress response (*sspA, ycgZ*) | 2 |
| Unknown mechanism (*frdD, ygeN*) | 2 |

Verified list of ORFs conferring a growth advantage in specific carbon sources when overexpressed. ORFs are grouped according to the putative mechanism of fitness gain. See Dataset S3 for more details.

a tendency to introduce harmful intermediates into the network or divert resources from growth-supporting pathways.

We found strong computational and experimental support for the notion that evolution can capitalize on underground reactions both to enhance growth in existing environments and to exploit completely novel nutrient sources. Furthermore, our in silico results suggest that the known underground repertoire of *E. coli* enzymes can substantially increase the range of utilizable carbon sources available to this organism (i.e., an ∼11% increase in network size expands the scope of available nutrients by ∼3%).

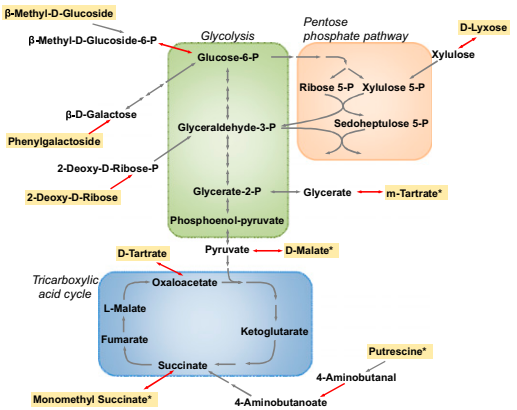
We speculate that the contribution of underground metabolism to adaptation to new environments might be even more pronounced in eukaryotes, where metabolic network expansion by means of horizontal gene transfer has a less prevalent role compared with bacteria (28).

Underground metabolism could be exploited for additional functions beyond using novel nutrients for cellular growth. First, because many underground reactions generate a cross-wiring between existing network parts, these reactions might allow the network to react rapidly to perturbations in metabolite or enzyme concentrations (29). Second, evolution of new pathways via underground reactions may be important for the production of novel secondary metabolites (30).

A central challenge of evolutionary systems biology is to predict the phenotypic effect of mutations and potential routes of evolution (31, 32). Although important progress has been made in predicting the fitness effect and epistatic interactions of deleterious mutations in large-scale metabolic networks (31), the genetic basis of adaptive novelties has remained more elusive. Our work demonstrates that, based on the knowledge of underground metabolism, it is possible to predict both the range of novel metabolic phenotypes available to an organism in one mutational step and their genetic bases. Although our present information on underground metabolism is far from being complete, the overlap between in silico and in vivo identified genotype–phenotype pairs (Fig. 4*A* and *SI Appendix*, Table S2) suggests that our underground network already covers a significant part of

A

| Carbon source | Enzyme | Known or proposed underground reaction | Phenotype | *In silico* |
|---|---|---|---|---|
| D-Lyxose | Mannose isomerase (*yihS*) | D-Lyxose ↔ D-Xylulose | ++ | ● |
| D-2-Deoxyribose | Ribokinase (*rbsK*) | D-2-Deoxyribose + ATP → D-2-Deoxyribose 5-P + ADP + $H^+$ | ++ | ● |
| D-Tartrate | Fumarase A (*fumA*), Fumarase B (*fumB*) | D-Tartrate ↔ Oxaloacetate + $H_2O$ | ++ | ● |
| Phenylgalactoside | β-galactosidase (*lacZ*) | Phenylgalactoside + $H_2O$ → D-Galactose + Phenol | + | ● |
| m-Tartrate | D-malate dehydrogenase (*dmlA*) * | m-Tartrate ↔ D-Glycerate + $CO_2$ | ++ | ○ |
| Putrescine | Acetaldehyde dehydrogenase (*mhpF*) * | 4-Aminobutanal + H-R + NAD ↔ 4-Aminobutanoyl-R + NADH + $H^+$ (R = CoA,. OH) | ++ | ○ |
| D-Malate | 3-isopropylmalate dehydrogenase (*leuB*) * | D-Malate + NAD ↔ Pyruvate + $CO_2$ + NADH + $H^+$ | + | ○ |
| β-Methyl-D-Glucoside | 6-phospho-beta-glucosidase (*bglB*) | β-Methyl-D-Glucoside 6-P + $H_2O$ ↔ β-D-Glucose 6-P + Methanol | + | ○ |
| Monomethyl Succinate | Carboxylic ester hydrolase (*ybfF*) * | Monomethyl Succinate + $H_2O$ ↔ Succinate + Methanol | + | ○ |




B

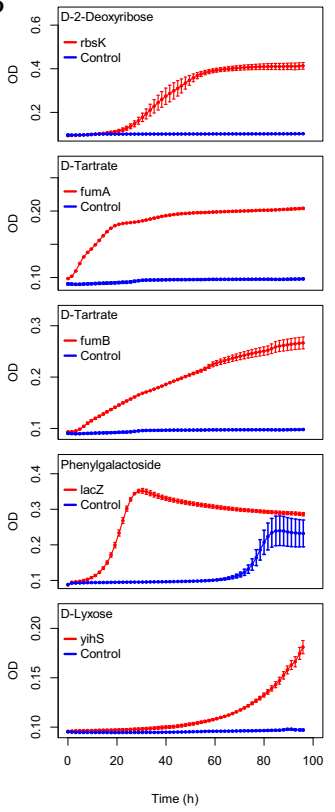



**Fig. 4.** Underground enzyme activities conferring growth advantage in new carbon sources. (*A*) List of enzymes with catalytic side activities that enable (++) or improve (+) the utilization of specific carbon sources when amplified. Forty-four percent of the experimentally confirmed phenotypes were also predicted in silico (●), and the rest were identified experimentally (○). (*B*) Growth experiments validating the five computationally predicted carbon source–enzyme pairs presented in *A*. Red curves show the growth of cells overexpressing the predicted ORFs, and the blue one is the negative control (cells harboring the empty plasmid). Each curve represents the average of three biologically independent replicates and their SE. We note that growth dynamics is variable between carbon sources, akin to what is observed on native carbon sources (35). (*C*) Schematic view of central carbon metabolism (gray arrows) with underground reactions (red arrows) that confer a growth advantage on specific carbon sources (yellow highlight) when amplified. Asterisks in *A* and *C* denote cases where the carbon source is channeled into the native network via an underground reaction with indirect evidence. For details, see Dataset S4.



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

the catalytic raw material available for short-term adaptation to novel environments.

Finally, the ability to computationally predict novel phenotypes based on knowledge of underground reactions also has important practical implications. For instance, systematic screens for enzyme side activities coupled with computational modeling could be used to reveal new pathways for industrially relevant compounds in new economically attractive growth environments. In addition, similar approaches might increase our understanding of the role of catalytic side activities and gain-of-function enzyme mutations in tumor evolution (33).

## Materials and Methods

We reconstructed the underground metabolism of *E. coli* K-12 MG1655 (hereby termed iRN1260u) by extending the genome-scale metabolic network iAF1260 (11) with weak underground reactions from published experimental studies based on the BRENDA database (10) and literature (Datasets S1 and S5). The reconstruction is available as a computational SBML model (Dataset S6, also downloadable from http://group.szbk.u-szeged.hu/sysbiol/papp-balazs-lab-resources.html). Reactions were considered as underground reactions if they were listed in the BRENDA "substrate," but not in the "natural substrates" section. Each reaction was examined as a whole for correct stoichiometry (i.e., mass and charge balance). To evaluate the correctness of the classification, we examined kinetic efficiency by $k_{cat}$ and $K_m$ values and metabolomics datasets for the occurrence of metabolites. Samples of elementary flux modes (i.e., minimal steady-state pathways) containing a reaction of interest were obtained using a modified algorithm of Kaleta et al. (34). We investigated the toxicity of both native and underground metabolites using a quantitative structure–activity relationship model developed to predict compound toxicity specifically in *E. coli* (EcoliTox web server) (18). The algorithm predicts $IC_{50}$ based on molecule structure with high accuracy ($R^2 = 0.71$). We applied FBA (20) to predict the contribution of underground metabolism to the adaptation to novel nutrient environments. Predictions were compared with results from a high-throughput gene-overexpression screen using the ASKA library (25) across ∼200 carbon sources following the protocol of Soo et al. (24) with modifications.

Detailed procedures of (*i*) the reconstruction and evaluation of the *E. coli* underground network, (*ii*) calculation of network distance and shared subsystems between native and underground reactions, (*iii*) identification of reactions capable of carrying a flux, (*iv*) elementary flux mode analysis, (*v*) metabolite toxicity analysis, (*vi*) EDGE analysis, and (*vii*) in silico and experimental surveys to identify novel phenotypes conferred by underground reactions are described in *SI Appendix*.

**ACKNOWLEDGMENTS.** We thank István Nagy for DNA sequencing, Christoph Kaleta for useful suggestions on elementary flux mode sampling, and Martijn Huynen and Steve Oliver for insightful comments. This work was supported by a Netherlands Organisation for Scientific Research Veni grant (to R.A.N.), the Lendület Programme of the Hungarian Academy of Sciences, the Wellcome Trust (B.P. and C.P.), European Research Council (C.P.), the FP7 Initial Training Network METAFLUX (Metabolic Flux Analysis and Cancer) (F.P. and B.P.), the Hungarian Scientific Research Fund PD (B.K. and B.C.), Hungarian Academy of Sciences Postdoctoral Fellowship Programme SZ-039/2013 (B.B.), and Társadalmi Megújulás Operatív Program Grant 4.2.4. A/2-11-1-2012-0001 (to B.S.).

1. Jensen RA (1976) Enzyme recruitment in evolution of new function. *Annu Rev Microbiol* 30:409–425.
2. Schmidt S, Sunyaev S, Bork P, Dandekar T (2003) Metabolites: A helping hand for pathway evolution? *Trends Biochem Sci* 28(6):336–341.
3. Rison SC, Teichmann SA, Thornton JM (2002) Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol* 318(3):911–932.
4. Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505.
5. D'Ari R, Casadesús J (1998) Underground metabolism. *BioEssays* 20(2):181–186.
6. Kuznetsova E, et al. (2006) Genome-wide analysis of substrate specificities of the Escherichia coli haloacid dehalogenase-like phosphatase family. *J Biol Chem* 281(47): 36149–36161.
7. Huang R, et al. (2012) Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc Natl Acad Sci USA* 109(8):2966–2971.
8. Macchiarulo A, Nobeli I, Thornton JM (2004) Ligand selectivity and competition between enzymes in silico. *Nat Biotechnol* 22(8):1039–1045.
9. Aharoni A, et al. (2005) The 'evolvability' of promiscuous protein functions. *Nat Genet* 37(1):73–76.
10. Scheer M, et al. (2011) BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 39(Database issue):D670–D676.
11. Feist AM, et al. (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121.
12. Nam H, et al. (2012) Network context and selection in the evolution to enzyme specificity. *Science* 337(6098):1101–1104.
13. Guo AC, et al. (2013) ECMDB: The E. coli Metabolome Database. *Nucleic Acids Res* 41(Database issue):D625–D630.
14. Khersonsky O, Malitsky S, Rogachev I, Tawfik DS (2011) Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. *Biochemistry* 50(13): 2683–2690.
15. Schuster S, Fell DA, Dandekar T (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol* 18(3):326–332.
16. Näsvall J, Sun L, Roth JR, Andersson DI (2012) Real-time evolution of new genes by innovation, amplification, and divergence. *Science* 338(6105):384–387.
17. Kim J, Copley SD (2012) Inhibitory cross-talk upon introduction of a new metabolic pathway into an existing metabolic network. *Proc Natl Acad Sci USA* 109(42): E2856–E2864.
18. Planson AG, Carbonell P, Paillard E, Pollet N, Faulon JL (2012) Compound toxicity screening and structure-activity relationship modeling in Escherichia coli. *Biotechnol Bioeng* 109(3):846–850.
19. Wagner A, et al. (2013) Computational evaluation of cellular metabolic costs successfully predicts genes whose expression is deleterious. *Proc Natl Acad Sci USA* 110(47):19166–19171.
20. Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat Rev Microbiol* 2(11):886–897.
21. Wagner A (2011) *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems* (Oxford Univ Press, Oxford).
22. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I (2007) Multicopy suppression underpins metabolic evolvability. *Mol Biol Evol* 24(12):2716–2722.
23. Kim J, Kershner JP, Novikov Y, Shoemaker RK, Copley SD (2010) Three serendipitous pathways in E. coli can bypass a block in pyridoxal-5′-phosphate synthesis. *Mol Syst Biol* 6:436.
24. Soo VW, Hanson-Manful P, Patrick WM (2011) Artificial gene amplification reveals an abundance of promiscuous resistance determinants in Escherichia coli. *Proc Natl Acad Sci USA* 108(4):1484–1489.
25. Kitagawa M, et al. (2005) Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): Unique resources for biological research. *DNA Res* 12(5):291–299.
26. Ahmed Z (2001) Production of natural and rare pentoses using microorganisms and their enzymes. *Electron J Biotechnol* 4(2):1–16.
27. Orth JD, et al. (2011) A comprehensive genome-scale reconstruction of Escherichia coli metabolism—2011. *Mol Syst Biol* 7:535.
28. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet* 37(12):1372–1375.
29. Wagner A, Fell DA (2001) The small world inside large metabolic networks. *Proc Biol Sci* 268(1478):1803–1810.
30. Weng JK, Philippe RN, Noel JP (2012) The rise of chemodiversity in plants. *Science* 336(6089):1667–1670.
31. Papp B, Notebaart RA, Pál C (2011) Systems-biology approaches for predicting genomic evolution. *Nat Rev Genet* 12(9):591–602.
32. Heckmann D, et al. (2013) Predicting C4 photosynthesis evolution: Modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell* 153(7):1579–1588.
33. Dang L, et al. (2009) Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. *Nature* 462(7274):739–744.
34. Kaleta C, de Figueiredo LF, Behre J, Schuster S, eds (2009) *EFMEvolver: Computing Elementary Flux Modes in Genome-Scale Metabolic Networks* (Gesellschaft für Informatik, Bonn), Vol 157, pp 179–189.
35. Vaas LA, Sikorski J, Michael V, Göker M, Klenk HP (2012) Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS ONE* 7(4):e34846.



Downloaded at MTA Szegedi Biologiai Kozpont on October 4, 2021

dc_1908_21
nature COMMUNICATIONS

ARTICLE

Received 2 Dec 2015 | Accepted 12 Apr 2016 | Published 20 May 2016

DOI: 10.1038/ncomms11607 OPEN

# Adaptive evolution of complex innovations through stepwise metabolic niche expansion

Balázs Szappanos[1,*], Jonathan Fritzemeier[2,*], Bálint Csörgő[1,*], Viktória Lázár[1], Xiaowen Lu[3], Gergely Fekete[1], Balázs Bálint[4], Róbert Herczeg[4], István Nagy[4,5], Richard A. Notebaart[3,6], Martin J. Lercher[2], Csaba Pál[1] & Balázs Papp[1]

A central challenge in evolutionary biology concerns the mechanisms by which complex metabolic innovations requiring multiple mutations arise. Here, we propose that metabolic innovations accessible through the addition of a single reaction serve as stepping stones towards the later establishment of complex metabolic features in another environment. We demonstrate the feasibility of this hypothesis through three complementary analyses. First, using genome-scale metabolic modelling, we show that complex metabolic innovations in *Escherichia coli* can arise via changing nutrient conditions. Second, using phylogenetic approaches, we demonstrate that the acquisition patterns of complex metabolic pathways during the evolutionary history of bacterial genomes support the hypothesis. Third, we show how adaptation of laboratory populations of *E. coli* to one carbon source facilitates the later adaptation to another carbon source. Our work demonstrates how complex innovations can evolve through series of adaptive steps without the need to invoke non-adaptive processes.

[1] Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Temesvári krt. 62, Szeged H-6726, Hungary. [2] Department for Computer Science, Heinrich Heine University, Universitätsstraße 1, Düsseldorf D-40221, Germany. [3] Department of Bioinformatics (CMBI), Radboud University Medical Centre, Geert Grooteplein Zuid 26-28, Nijmegen 6525 GA, The Netherlands. [4] SeqOmics Biotechnology Ltd, Vállalkozók útja 7, Mórahalom H-6782, Hungary. [5] Sequencing Platform, Institute of Biochemistry, Biological Research Centre of the Hungarian Academy of Sciences, Temesvári krt. 62, Szeged H-6726, Hungary. [6] Department of Internal Medicine, Radboud University Medical Center, Geert Grooteplein Zuid 8, Nijmegen 6525 GA, The Netherlands. * These authors contributed equally to this work. Correspondence and requests for materials should be addressed to C.P. (email: cpal@brc.hu) or to B.P. (email: pappb@brc.hu).

Evolutionary novelties frequently depend on the fixation of multiple, highly specific mutations, where intermediate stages of evolution seemingly provide little or no benefit[1]. Such complex adaptations are widespread in molecular networks and include the origin of multimeric protein machineries, establishment of interactions between transcription factors and their binding sites, receptor–ligand interactions and multi-step metabolic pathways[2–4]. According to the notion that evolutionary adaptation proceeds by the sequential fixation of single beneficial mutations[5], complex adaptations are expected to occur only sporadically. One theory suggests that many evolutionary innovations, that is, qualitatively new adaptive traits, have non-adaptive origins, where neutral mutations prepare the ground for later beneficial mutations that lead to innovations[6,7]. Evidence for this process comes from laboratory evolution of RNA enzymes[8], but its role in the establishment of complex molecular pathways remains unclear. In the case of metabolic networks, the theory proposes that 'many additions of individual reactions to a metabolic network will not change a metabolic phenotype until a second added reaction connects the first reaction to an already existing metabolic pathway'[7]. However, this non-adaptive process is expected to be extremely slow, and furthermore, there is no direct empirical support for this scenario in bacteria, which are especially prolific in producing metabolic innovations. Although free-living bacteria increase their genome size through horizontal gene transfer and gene duplication, their genomes remain compact, and non-functional sequences appear to be rare compared with most eukaryotes[9]. Genes under relaxed selection are rapidly inactivated and subsequently lost in free-living bacteria, not least because there is a pervasive mutational bias towards deletions of genomic segments[9]. Consequently, genes encoding functionally completely intact enzymes that provide no immediate fitness advantage are generally unlikely to be maintained for long periods. Even under a scenario where the neutral intermediate-step mutation is not required to reach high population frequencies (that is, 'stochastic tunnelling'[10]), evolution is expected to be slower than traversing purely adaptive trajectories through natural selection. Thus, understanding the evolution of complex innovations remains a formidable challenge.

Previous population genetic models[11] and computer simulations of genetic circuits and RNA molecules[12] offer a potential solution to the problem of complex adaptations. These works indicate that complex or temporally fluctuating conditions can facilitate adaptation, partly by allowing populations to escape fitness plateaus and reach new adaptive peaks. Similarly, a study on digital organisms revealed that populations often evolve complex features by building on simpler functions that had evolved earlier[13]. However, the extent to which these abstract considerations apply to specific cellular subsystems remained unknown, partly due to the shortage of systems-level analysis that would combine computational modelling and evolutionary experiments.

In this work, we focus on bacterial metabolic networks to examine how novel nutrient utilization phenotypes can be acquired via the addition of new reactions to an organism's enzyme repertoire. While not all complexity at the level of molecular systems are expected to provide a functional advantage[14,15], metabolic pathways utilizing novel nutrients arguably qualify as adaptive traits. The problem of the evolution of novel metabolic pathways has two complementary aspects, relating to their origin and subsequent evolutionary establishment across multiple species. Previous works were largely concerned by how novel biochemical reactions arise first during the course of evolution[16,17]. In this paper, we ask how existing enzymatic reactions can assemble to form a novel metabolic pathway in an organism that already harbours a complex metabolic network. We extend and generalize an early suggestion that varying nutrient environments play a prominent role in the establishment of biosynthetic pathways[16].

Specifically, we employ detailed simulations on a pan-genome scale to demonstrate that complex metabolic innovations can evolve via the successive acquisition of single biochemical reactions that each confers a benefit to utilize specific nutrients. Thus, temporal changes in nutrient availability or complex environments (where multiple nutrients are available) can facilitate adaptive evolution of metabolic pathways through the step-by-step expansion of metabolic niches. Gene acquisition patterns across bacterial genomes and *de novo* laboratory evolution of nutrient utilizations in *Escherichia coli (E. coli)* provide clear support for the hypothesis.

## Results

**Most metabolic innovations demand only a few novel reactions.** In this work, we systematically studied the expansion of metabolic networks. We specifically asked whether metabolic innovations can evolve in a purely Darwinian manner through series of adaptive steps. Our starting point was the previously reconstructed metabolic network of *E. coli* K-12, arguably the best studied and most reliable reconstruction of a genome-scale metabolic system, composed of 2,077 unique reactions, including transport processes[18]. Previous studies showed that bacterial networks expand largely by acquiring genes involved in the transport and catalysis of external nutrients, driven by adaptations to changing environments[19]. On the basis of these observations, here we studied the potential selective advantages conferred by the addition of new metabolic reactions to the *E. coli* network. We compiled a data set of 2,566 known enzymatic and 159 transport reactions across the three kingdoms of life ('universal reaction set') absent from the *E. coli* model[20] (see Methods). We next defined a comprehensive sample of the external nutrient space, consisting of 1,776 environments comprised of nutrient sources that can potentially be imported into the network (Supplementary Data 1). We focused on minimal media that differ from each other in a single carbon, nitrogen, sulphur or phosphorus source, thereby maximizing the variability between conditions while remaining computationally feasible (Methods). We determined the phenotypic impact of adding one or more reactions from the universal reaction set to the *E. coli* network in each of these environments using flux balance analysis (FBA)[21]. FBA identifies a steady-state flux distribution that maximizes the production of biomass (a weighted combination of major biosynthetic components) from a given set of available nutrients. This framework successfully predicts the growth capacity of wild-type *E. coli* across nutrient conditions[18], and it is biologically more realistic than graph-theoretical approaches[22].

Before the addition of novel reactions, the reconstructed *E. coli* metabolic network was unable to grow (that is, the rate of biomass production was zero) in 321 environments in which the network expanded by the complete universal reaction set was able to grow (Supplementary Data 1). Using a mixed integer linear programming (MILP) algorithm, we determined the minimal number of reactions from the universal reaction set that need to be added to the *E. coli* network to support growth in these novel environments. Strikingly, growth in additional environments required the addition of only one to three enzymatic and transport reactions in 74% of the cases (239 out of 321 environments; see Fig. 1). In 21.5% of the novel environments, acquisition of only one reaction was sufficient for growth (69 out of 321 environments, see Supplementary Data 2). These results suggest that in the genotype space around the *E. coli* metabolic

dc_1908_21

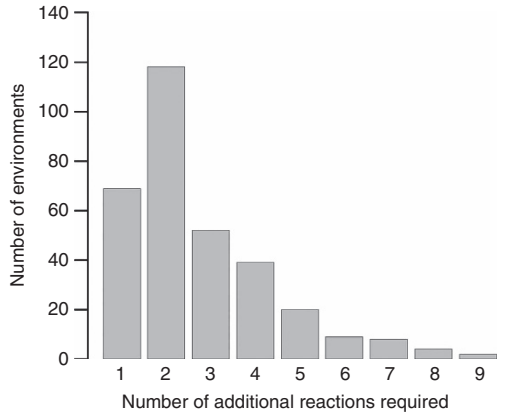



**Figure 1 | Metabolic innovations in the genotype space around the *E. coli* network.** Only few reactions need to be added to the *E. coli* metabolic network to enable growth in metabolic environments where the wild-type cannot grow. The histogram shows the distribution of additional minimal reaction set sizes needed for biomass production in 321 novel nutrient environments.

network, most metabolic innovations are only a few gene acquisition steps away.

**Complex innovations can arise via changing environments.** One can envisage a simple adaptationist hypothesis by which complex metabolic innovations can arise. A metabolic phenotype accessible through the addition of a single reaction may serve as an exaptation[23] from which metabolic phenotypes that demand the acquisition of multiple reactions can be developed. A major corollary of this hypothesis is that evolutionary adaptation to temporally varying environmental conditions facilitates the expansion of metabolic networks (see also ref. 16). In the parlance of fitness landscapes, varying environments result in dynamic landscapes with moving peaks which can be more easily tracked by hill-climbing evolution (see Fig. 2a,b).

To test the feasibility of the stepwise network expansion scenario, we focused on reaction pairs that are jointly required to provide a fitness benefit in at least one environment (for a list of the 538 such reaction pairs, see Supplementary Data 3). Next, we added each of the corresponding reactions individually into the network and asked whether their presence alone provides a selective advantage across the set of 321 novel environments. Consistent with the hypothesis, we found that in 40% of the 538 growth-promoting reaction pairs, one of the reactions enables growth on its own in at least one environmental setting, which therefore can serve as stepping stones along adaptive trajectories. For example, while the ability to metabolize chorismate demands the simultaneous acquisition of two reactions, one of them also confers L-phenylalanine utilization when added individually to the network (Fig. 2c). We note that many growth-promoting reaction pairs are phenotypically equivalent (that is, confer growth in the same environment) and share the same stepping-stone reaction (Supplementary Data 3). As a result, in total 8.5% of the 118 novel environments that require the simultaneous addition of two reactions becomes accessible through purely adaptive walks.

To more generally assess the potential for exaptation, we examined for each novel environment if its growth-promoting reactions are involved in adaptation to another (intermediate) environment. To this end, for each environment, we enumerated all possible minimal reaction sets that can support growth when added to the *E. coli* network from the universal reaction set. On average, 26% of the alternative minimal reaction sets required for growth in a given environment are also entirely present in at least one minimal growth-promoting reaction set of a second environment. This finding indicates that some of the growth-promoting reaction sets contribute to growth in multiple environments as parts of larger reaction sets. These figures are likely underestimates due to incomplete knowledge of available enzymatic reactions (including promiscuous side activities in the *E. coli* metabolic network[24]) and environmental conditions. We conclude that traversing complex evolutionary trajectories can be facilitated by exaptations when the environment varies.

**Metabolic gene acquisition patterns support the hypothesis.** The model predicts that acquisition of new metabolic genes during bacterial evolution should be contingent on the presence of other genes providing specific adaptations to intermediate environments. It has been established that a major source of metabolic network expansion is horizontal gene transfer in bacteria[19,25]. Genes recently acquired by *E. coli* through horizontal gene transfer confer condition-specific advantages and contribute to growth only in specific environments[19]. To test whether acquisition of an enzyme pair that is potentially accessible via adaptive steps occurs via a defined order, we used genomic data from 943 bacteria to reconstruct gene-gain events along the corresponding phylogeny using parsimony (Fig. 3a, Methods). As expected under the hypothesis, enzymes that are predicted to confer fitness benefits on their own and can hence serve as stepping stones towards two-step adaptations *in silico* tend to be gained on an earlier branch of the phylogenetic tree than their partner enzyme (in 65% of cases, $N=33$, as opposed to 50% expected by chance, $P=0.037$, one-tailed one-sample Wilcoxon signed-rank test, see Methods). We note that this pattern holds for different parameter values of the gene-gain reconstruction procedure (see Supplementary Table 1).

In contrast to such cases, growth-promoting enzyme pairs not accessible gradually are the most likely candidates for co-gain via horizontal gene transfer. In agreement with this expectation, such enzyme pairs show a much higher co-gain fraction, that is, number of co-gain relative to single gain events, compared with random gene pairs and growth-promoting gene pairs predicted to be accessible gradually through adaptive evolution via environmental changes ($P<0.001$, randomization analysis and $P=0.0038$, one-sided Wilcoxon rank test, respectively, $N=21$, Fig. 3b, see Methods). Also consistent with the hypothesis, growth-promoting enzyme pairs that are accessible gradually through adaptive evolution via environmental changes, have very low co-gain fractions that are indistinguishable from that of random gene pairs ($P=0.64$, randomization analysis, $N=40$, Fig. 3b, see Methods). These conclusions are robust to changes in parameter values of the gene-gain reconstruction procedure (see Supplementary Tables 2,3).

**Experimental evolution of a complex metabolic innovation.** New metabolic pathways can evolve not only through the acquisition of full-blown enzymes from other organisms but also through the enhancement of weak side activities of existing enzymes[3,24]. Thus, a further prediction of the hypothesis is that evolutionary adaptation to a specific nutrient via accumulating mutations in endogenous genes can influence the accessibility of adaptive paths towards the utilization of other nutrients. An early work[26] suggests that acquiring the ability to grow on ethylene glycol (EG, ethane-1, 2-diol) and propylene glycol (PG, (S)-propane-1, 2-diol), two related carbon sources unavailable for utilization by wild-type *E. coli* K12, might depend on one another in a contingent manner. Specifically, according to the anecdotal report, *E. coli* mutants able to grow

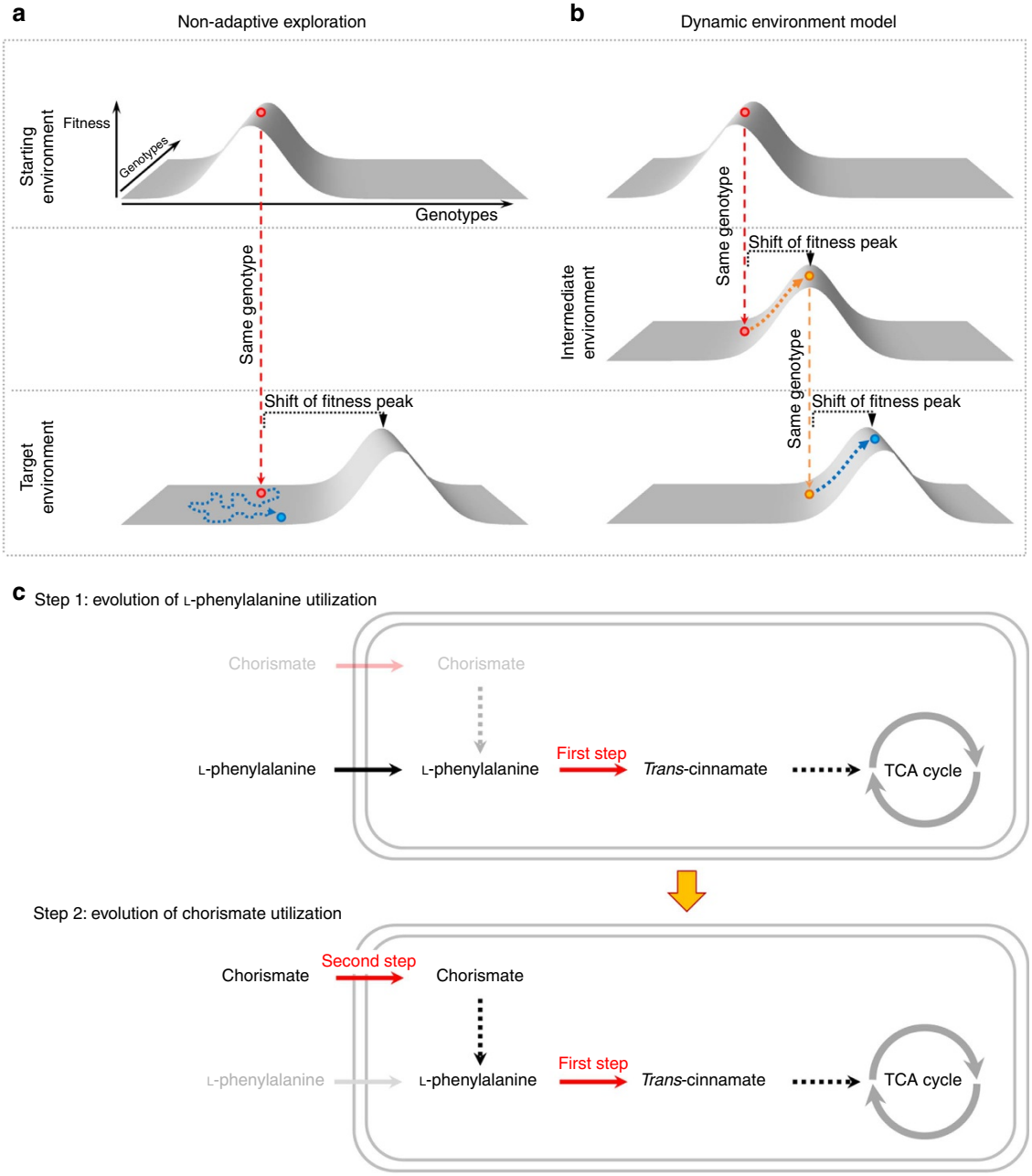



**Figure 2 | Evolution in varying environments is expected to facilitate the establishment of complex metabolic traits.** (**a**, top) Hypothetical fitness landscape over a two-dimensional genotype space. The red genotype is well-adapted, that is, it is located on the fitness peak of this starting fitness landscape. A change to the target environment shifts the fitness peak, so that the red genotype is no longer of high fitness (bottom). Adaptation to the shifted peak now cannot proceed purely through adaptive steps (that is, hill climbing); it requires the non-adaptive exploration of the neutral part of the landscape, illustrated by the yellow dotted line. (**b**) Depicting the same situation, but with an intermediate environment in which the fitness peak is only slightly shifted relative to the starting environment. The red genotype is located at the foot of the shifted fitness peak in this intermediate environment and can thus progress through purely adaptive steps, culminating in the yellow high-fitness genotype. When the environment now changes to the same target environment as in **a**, the blue genotype represents an exaptation, such that it can now progress towards the target fitness peak through purely adaptive steps. While **b** only shows one intermediate environment, the same reasoning applies to more complex scenarios including dynamic landscapes with moving peaks. (**c**) Example from simulated metabolic network expansions. *E. coli* K-12 is unable to utilize chorismate and L-phenylalanine as sole carbon sources. Simulations show that while chorismate utilization demands the simultaneous addition of two reactions to the network, one of these reactions (first step; catalysed by phenylalanine ammonia lyase) also confers L-phenylalanine utilization when added individually.

on EG could be obtained from mutants that could grow on propylene glycol[26]. Using these phenotypes as a test bed we aimed at directly testing the stepwise metabolic niche expansion scenario by examining (i) whether mutations that enable growth on propylene glycol *per se* increase adaptation rates to EG and (ii) whether the mutations conferring these two distinct growth phenotypes exhibit epistasis on EG, as predicted by the hypothesis.

First, we attempted to isolate mutants that can grow on EG (EG+) or propylene glycol (PG+) from large populations of bacteria (Supplementary Methods). No EG+ or PG+ cells were isolated from $\sim 10^{11}$ cells with wild-type mutation rate (Table 1),

dc_1908_21

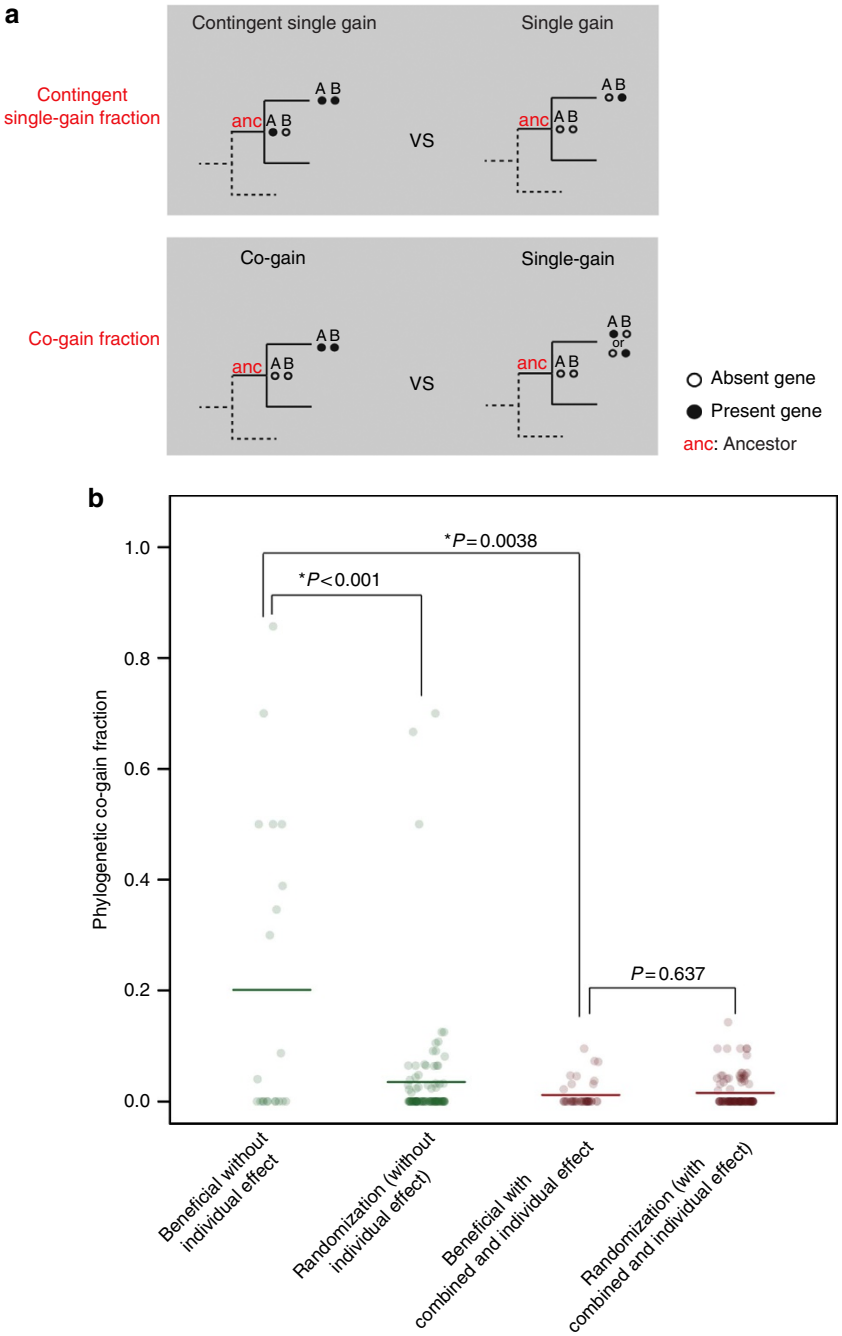



**Figure 3 | Evolutionary history of gene gains supports the dynamic environment model.** (**a**) Schematic representation of the phylogenetic comparisons to study the interdependence between gene-gain events. According to the dynamic environment model, if initial adaptation via a single gain of gene *A* serves as a stepping-stone for complex adaptation via a gain of gene *B*, then acquisition of *B* is expected to occur more frequently with gene *A* being present (contingent gain) compared with *A* being absent in the ancestral branch points of the bacterial tree (upper panel). Furthermore, enzyme pairs that confer a growth benefit only when present together are expected to be more frequently co-gained along branches of the bacterial tree in comparison to a gain of only one of the two (lower panel). Detailed description of the procedures is presented in Methods. (**b**) Phylogenetic co-gain measure (see Methods) of jointly beneficial enzymes based on analysis of hundreds of bacterial genomes. Orthologs of enzyme pairs that are beneficial jointly but not accessible gradually ('beneficial without individual effect', $N=21$) tend to be co-gained on the same branch of the phylogenetic tree. This trend is statistically significant when compared both with randomized pairs and to enzymes that are growth-promoting as a pair but are accessible gradually through adaptive evolution via varying environments ('beneficial with combined and individual effect', $N=40$), $P<0.001$ (randomization analysis) and $P=0.0038$ (one-sided Wilcoxon rank test), respectively. In addition, such 'accessible' pairs are not more likely to be co-gained than expected by chance ($P=0.637$).

demonstrating that these substrates demand the acquisition of one or more very rare specific mutations. Next, we employed an *E. coli* strain with an approximately 1,000-fold increased mutation rate[27]. In this case, PG+ cells occurred at a low, but detectable frequency of $1.5\times10^{-9}$, but still no EG+ mutants were found (Table 1). As discussed, the evolution of EG utilization might be facilitated by prior adaptation to PG[26]. This was indeed so: EG-utilizing cells were detected in PG+ populations at a frequency of $\sim3.8\times10^{-9}$ (Table 1), indicating an increase in adaptation rate of at least two orders of magnitude.

**Table 1 | Adaptation frequencies of different strains to PG and EG.**

| Strain | Frequency of cells growing on PG | Frequency of cells growing on EG |
|---|---|---|
| MG1655 | Up to $1.6 \times 10^{-11}$ | Up to $1.6 \times 10^{-11}$ |
| MG1655 mutD5 | $1.5 \times 10^{-9}$ | Up to $3.1 \times 10^{-11}$ |
| MG1655 mutD5 adapted to PG | Grows on PG | $3.8 \times 10^{-9}$ |
| MG1655 + *fucO* overexpressed | Grows on PG | $2.1 \times 10^{-8}$ |

EG, ethylene glycol; PG, propylene glycol.
MG1655 is the reference wild-type strain, while MG1655 mutD5 refers to a strain with an approximately 1,000-fold increased mutation rate. Values are averages of three parallel replicates when PG+ or EG+ cells were observed and upper estimates[35] when no growing cells were obtained (see Supplementary Methods).

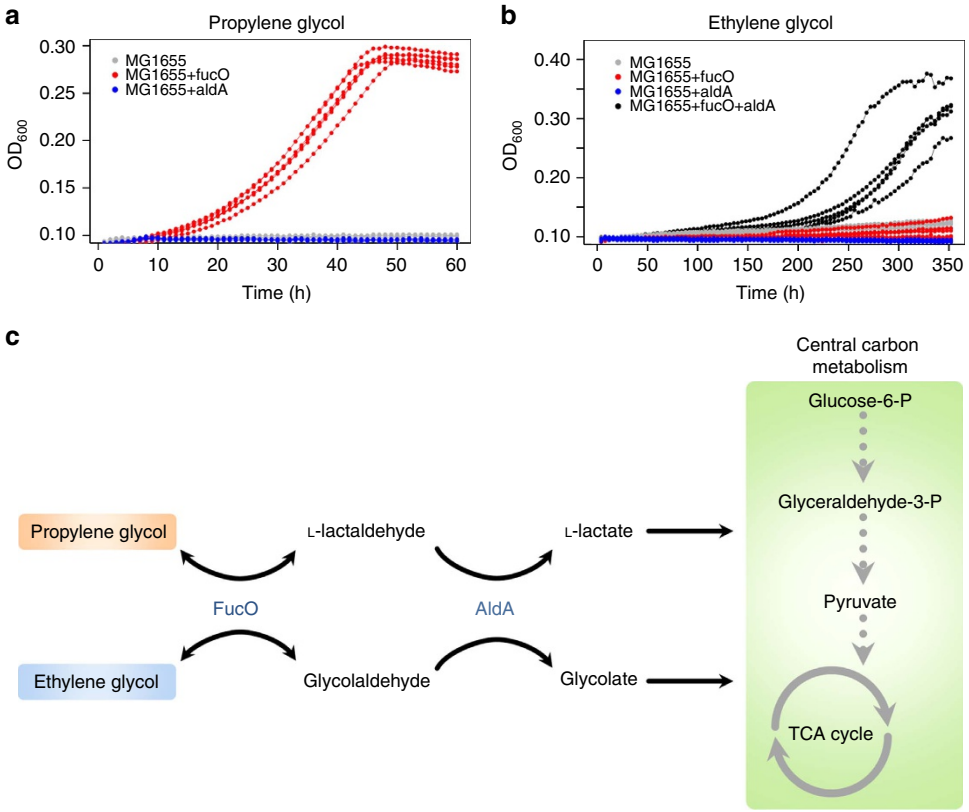



**Figure 4 | Utilization of propylene glycol increases adaptation rates towards growth on EG in the laboratory.** (**a**) Growth curve measurements demonstrating that overexpression of *fucO* (red) is sufficient for growth in propylene glycol. Wild-type MG1655 strain is depicted in grey. $OD_{600}$ measurements of six independent replicates were taken every 60 min. (**b**) Growth curve measurements demonstrating that joint overexpression of both *fucO* and *aldA* is required for growth on EG (black). Neither *fucO* (red) nor *aldA* (blue) can achieve this when overexpressed individually. Wild-type MG1655 strain is depicted in grey. $OD_{600}$ measurements of six independent replicates were taken every 240 min. One replicate population with joint overexpression of *fucO* and *aldA* failed to grow for unknown reason and is not shown. (**c**) Schematic pathway diagram representing the role of FucO and AldA enzymes in the utilization of PG and EG. In the first step, FucO catalyses the oxidation of PG and EG to glycolaldehyde and L-lactaldehyde, respectively. We note that the native activity of FucO operates in the reverse direction by reducing L-lactaldehyde to PG during the catabolism of L-fucose and L-rhamnose. In the next step, AldA oxidizes the products of FucO to hydroxycarboxylic acids which can be wired into central carbon metabolism following further enzymatic modifications. The affinity of AldA for L-lactaldehyde (PG utilization) is higher than for glycolaldehyde (EG utilization)[30], potentially explaining why growth on EG requires multiple copies of *aldA*.

It has been reported that constitutive activation of *fucO*, a gene encoding an enzyme involved in fucose and rhamnose catabolism, is a prerequisite for growth in PG[28]. We therefore hypothesized that *fucO* upregulation acts as a stepping-stone mutation towards EG utilization. To test this scenario, we overexpressed *fucO* from a strong constitute promoter in wild-type background[29]. As expected, *fucO* overexpression conferred the ability to utilize PG (Fig. 4a). Remarkably, employing a *fucO* overexpressed PG+ strain yielded EG-utilizing cells at a frequency of $\sim 2 \times 10^{-8}$ (Table 1). As this strain retained a wild-type mutation rate (Supplementary Fig. 1), this finding shows that the ability to metabolize PG *per se* promotes adaptation to EG. Whole-genome sequencing of an EG-utilizing strain suggested that ~10-fold amplification of a genomic segment encoding *aldA* might underlie EG utilization (Supplementary Table 4). Indeed, simultaneous overexpression of both *fucO* and *aldA* in wild-type background conferred the ability to grow on EG (Fig. 4b) with a growth kinetics akin to the strain adapted to EG (Supplementary Fig. 2). Furthermore, as neither *fucO* nor *aldA* alone conferred growth on EG, this finding provides evidence that the two overexpression mutations act epistatically, as predicted by the stepwise metabolic niche expansion hypothesis.

How do these two enzymes, FucO and AldA, contribute to EG utilization? FucO likely acts on EG in addition to its native

substrate to produce glycolaldehyde from EG[26]; AldA, an enzyme with broad substrate specificity, further converts glycolaldehyde to glycolate[30] (Fig. 4c). Interestingly, in addition to their role in EG metabolism, both enzymes are involved in PG utilization as well, indicating that regulatory rewiring of the same enzyme toolkit can produce multiple qualitatively different phenotypes.

## Discussion

Explaining the origin of evolutionary innovations that require the simultaneous acquisition of multiple mutations, none of which seemingly confer a benefit individually, remains a central challenge in evolutionary biology. On the basis of prior theoretical considerations[11,12,16], here we propose that metabolic innovations accessible through the addition of a single reaction serve as stepping stones towards the later establishment of complex metabolic features in another environment. We provided several lines of evidences in support of the hypothesis by focusing on the most well-studied molecular network, cellular metabolism, and by employing three complementary approaches. First, we simulated the adaptation of the *E. coli* metabolic network to novel environments. We revealed that new complex pathways can evolve via the successive acquisition of single biochemical reactions that allow the utilization of specific nutrients. Second, by reconstructing the evolutionary history of gene gains in bacteria, we demonstrated that complex metabolic pathways are indeed often established in a defined order as predicted. Finally, we conducted a laboratory evolution study of *E. coli* adaptation to two novel carbon sources; evolving the ability to utilize one nutrient remarkably facilitated later adaptation to the other. Thus, complex metabolic traits can emerge without the need to invoke neutral exploration of genotype space, a view that is in sharp contrast to non-adaptive scenarios of evolutionary innovation that rely on the accumulation of neutral intermediate mutations[6,7,31].

Taken together, our study demonstrates that complex metabolic innovations can evolve by adaptive means through the step-by-step expansion of nutrient utilization capacities. An important prediction is that metabolic innovations should be intertwined in nature: the ability to metabolize certain nutrients should act as a stepping stone towards the utilization of other nutrient sources[32]. A preliminary systems-level analysis based on nutrient utilization of 168 *E. coli* strains[33] suggests that it may indeed be so (Supplementary Fig. 3). Experimental case studies on the evolution of the catabolism of β-galactoside sugars[34] and citrate utilization[35] are also consistent with the scenario, but it remains to be seen how general these findings are. In addition, it is important to note that functionally linked enzymes frequently cluster in the bacterial genome or are encoded in the same operon and tend to be acquired together during evolution[19]. Future systematic works should study the extent to which simultaneous uptake of multiple physiologically linked reactions by horizontal gene transfer speeds up the evolution of metabolic networks.

We speculate that the major barrier to the dynamic environment model of complex adaptation may be the absence of relevant series of environmental conditions. This restriction could be lifted when multiple novel substrates are simultaneously present in a single environment and evolution proceeds by successively acquiring the capacity to utilize them. We emphasize that other conceptually different mechanisms might also contribute to the adaptive expansion of metabolic networks. For example, stepping-stone reactions might evolve as repair processes in an adaptive response to metabolite damage[36], to degrade toxic environmental chemicals[3], or to produce novel secondary metabolites[37].

Our work has important ramifications for understanding genetic interaction networks and the development of industrially useful microbes. First, epistatic interactions between metabolic genes of the same pathway should often be environment-specific: our results suggest that in many cases, one of the genes should provide fitness benefits independently of the other in at least one environment. Large-scale mapping of genetic interactions across a broad range of environmental conditions would provide a direct way to test this prediction[38]. Second, we anticipate that evolutionary engineering of microbes to obtain desired phenotypes could be facilitated by temporally varying the traits under selection[39].

Finally, our study could have important implications beyond the evolution of metabolism. Earlier studies claimed that varying environments accelerate evolutionary adaptation in genetic circuits and RNA molecules[12]. In computer science, standard genetic algorithms have a tendency to quickly converge to a local solution, and hence frequently fail to identify more promising regions of the search space[40]. Application of dynamically changing 'environments' offers a natural strategy to maintain the diversity required to explore the adaptive surface[41].

## Methods

**Reconstruction of the universal reaction set.** To study the potential adaptive value of adding new reactions to the *E. coli* metabolic network, we compiled a data set of metabolic reactions reported from species across the three kingdoms of life (universal reaction set) and absent from *E. coli*. First, we mapped the metabolites of the manually curated *E. coli* genome-scale metabolic model[18] to the Model SEED database[20] (and http://blog.theseed.org/model_seed/), a comprehensive resource for automatically generated genome-scale metabolic network reconstructions. Because Model SEED does not contain the most recent version (iJO1366 (ref. 42)) of the *E. coli* network reconstruction, we used an earlier version (iAF1260 (ref. 18)) that is widely utilized and has been extensively tested[43]. As a second step, we added all mass-balanced biochemical reactions from the Model SEED database to the *E. coli* model. From this draft network, we removed duplicate reactions. Next, we removed 'perpetuum mobile' cycles, that is, flux distributions capable of producing energy without consuming any nutrients (see Supplementary Methods and Supplementary Table 5). Finally, we removed unconditionally blocked reactions (that is, those unable to carry a flux under any condition). The resulting curated universal reaction network contains 4,949 metabolic reactions and 444 nutrient uptake reactions, of which 2,566 and 159 are not present in the *E. coli* network, respectively. The universal network is available as a computational Systems Biology Markup Language (SBML) model (Supplementary Data 4).

For more details on the reconstruction of the universal reaction set, see Supplementary Methods.

**Defining novel *in silico* nutrient environments.** We first defined a comprehensive set of nutrient environments by starting from a glucose minimal medium for *E. coli*. For each environment, we replaced the carbon (C), nitrogen (N), phosphate (P) or sulfur (S) source by an alternative one. To obtain a list of environments that is both representative of novel nutrient compounds and computationally tractable, we focused on only those growth media that differed from glucose minimal medium by one compound instead of enumerating all possible combinations of C, N, P and S-sources, as in previous works[24,31]. Although this approach does not take into account more complex conditions, it allowed us to focus on single C, N, P and S-sources and to maximize the variability between conditions. See Supplementary Data 1 for the list of resulting 1,776 conditions.

Next, we determined the viability of both the *E. coli* network and the universal network across these conditions using FBA[21]. A network was deemed inviable in a given environment if its maximum biomass production was zero. Before adding novel reactions, the reconstructed *E. coli* metabolic network was unable to grow in 321 environments in which the network expanded by the universal reaction set allowed growth (Supplementary Data 1). We considered these 321 conditions as the set of available novel environments to which *E. coli* can possibly adapt by adding reactions from other species.

**Finding growth-promoting reaction sets in new environments.** To calculate the minimum number of active, non-coli reactions in a particular environment we applied a MILP-based algorithm on the universal metabolic model similar to the problem of finding the shortest elementary flux mode[44]. The basis of the MILP problem was the steady-state assumption:

$$S\mathbf{v} = 0$$

Where $S$ is the stoichiometric matrix and $\mathbf{v}$ is the flux vector for all reactions. The reactions of the model were handled differently depending on whether they are part of the *E. coli* model or they can be added to the coli model

during evolution. The flux constraints on the *E. coli* reactions were the same as in FBA:

$$\mathbf{l}_i \leq \mathbf{v}_i \leq \mathbf{u}_i$$

Next, for each environment in which the universal network was viable but the wild-type *E. coli* network was not able to grow we set the nutrient uptake constraints to mimic the environment ($\mathbf{l}_i$ of the exchange reactions). The lower bound of the biomass production reaction was then constrained to $10^{-4}$ as the minimal growth requisite:

$$\mathbf{l}_{\text{biomass}} = 1e^{-4}$$

The reversible non-coli reactions of the universal network were decomposed into two opposing irreversible reactions. This way the fluxes of the non-coli reactions can only take positive values. Let $N'$ be the number of non-coli reactions. We assigned a binary variable to each non-coli reaction, $\mathbf{b}_i$, which tells whether the non-coli reaction $\mathrm{r}'_i$ *($i=1, \ldots, N'$)* is active ($\mathbf{b}_i=1$) or not ($\mathbf{b}_i=0$). The following equations ensure these rules:

$$\mathbf{v}'_i \geq \varepsilon\, \mathbf{b}_i$$
$$\mathbf{u}'_{i,\max}\, \mathbf{b}_i \geq \mathbf{v}'_i$$

Where $\mathbf{v}'_i$ is the flux and $\mathbf{u}'_i$ is the maximal possible flux of reaction $\mathrm{r}'_i$, while $\varepsilon$ is the minimal flux value (in our calculations $\varepsilon = 10^{-8}$). Also to avoid having two opposing reactions derived from the same reversible reaction being active simultaneously we introduced the following constraint:

$$\mathbf{b}_i + \mathbf{b}_j \leq 1;\ (i,j) \in \{\text{set of opposing reaction pairs}\}$$

Finally, the objective of the MILP problem was to minimize the active non-coli reactions:

$$\text{minimize} \sum \mathbf{b}_i; i \in \{1, .., N'\}$$

The result of this minimization is the minimum number of non-coli reactions need to be added to the coli model to allow growth in a particular environment.

**Enumerating all possible minimal reaction sets *in silico*.** The MILP optimization model described above not only provide the minimal number of reactions that support growth in new environments but also the list of the non-coli reactions involved in this solution: one of the minimal reaction sets. However, multiple equivalent minimal sets might exist for any given environment. To identify another minimal reaction set we extended the MILP problem with a new constraint which prevents the algorithm to find the same solution again:

$$\sum (B_i \mathbf{b}_i) \leq \sum B_i - 1; i \in \{1, .., N'\}$$

Where $B_i$ is the binary solution of the first minimal reaction set, and $B_i$ equals to 1 or 0 if reaction $\mathrm{r}'_i$ was active or inactive in the first solution, respectively. This constraint is fulfilled only if the two solutions differ in at least one active reaction. We can harvest more minimal reaction sets in an iterative way where after each solution we add a new constraint and we run the algorithm again. Our algorithm stopped when the new solution had more active reactions than the size of the minimal reaction sets, that is, when we collected all minimal reaction sets. This algorithm is based on the method of finding the k-shortest elementary flux modes[44].

**Defining growth-promoting reaction pairs using modelling.** To systematically test the dynamic environment model, we investigated all possible two-step adaptation scenarios. First, we inactivated all non-coli reactions in the universal reaction network. Next, we activated two non-coli reactions at a time and applied FBA to calculate the fitness of the model in each environment where the native *E. coli* model cannot grow. By repeating this procedure we probed all possible reaction pairs in the universal reaction set and identified those that provide growth in at least one environmental condition (3,290,895 reaction pairs in total, 538 are beneficial in at least one condition). As a next step, we determined if the identified two-reaction adaptations can be accessed by the consecutive addition of single beneficial reactions to the network, that is, whether at least one of the two reactions provide a fitness benefit on its own in any of the environments. For this purpose, we repeated the above procedure but instead of activating reaction pairs we activated single reactions and evaluated their fitness effect across environments using FBA. The list of 538 reaction pairs and corresponding environments can be found in Supplementary Data 3.

**Software and computation used in metabolic network analyses.** All simulations were implemented in GNU R (ref. 45) using the sybil package for constraint-based modelling[46]. As optimizer for linear programming and MILP we used ILOG CPLEX 12.5. The linear programming was done on a 64-bit Ubuntu Linux system with an Intel Core-i7 quadcore processor. MILP problems were solved on a Red Hat Enterprise Linux Server release 6.2 with 96 Intel Xeon central processing units.

**Phylogenetic analysis of gene-gain events.** To investigate contingent gain and co-gain in the evolutionary history of genes, we first generated the phylogenetic presence and absence profile across the present-day species for each reaction by mapping the profiles from gene to reaction level. Presence and absence profiles of orthologous genes across 943 bacterial species were obtained from EggNOG v3.0 (ref. 47). Reactions catalysed by enzyme complexes consisting of multiple gene products ('AND' relationships) are considered to be present in a species only when all genes of the complex are present in the genome. Reactions catalysed by isoenzymes ('OR' relationships) are considered to be present when at least one isoenzyme is encoded in the genome.

Next, we inferred the most parsimonious ancestral presence/absence states of each reaction by using a phylogenetic tree of the 943 eubacteria, retrieved from STRING v9.05 (http://string905.embl.de/newstring_download/species.tree.v9.05.txt) (ref. 48). Reaction presence and absence states across branch points along the phylogenetic tree, that is, the ancestral states, are calculated by using the tree and the present-day presence/absence state of the reaction. The ancestral state is inferred by minimizing the number of gene-gain and loss events across the tree that matches the present-day state. Such a maximum parsimony strategy is commonly used as it allows for the analysis of gene histories on a genome-wide scale in a computationally efficient manner, and has shown to be successful in explaining patterns in genome content and evolution[19,49,50]. Calculations were carried out using PAUP[51] with a gain/loss penalty ratio of 2/1 (ref. 52) and a delayed transition assumption (DELTRAN)[49]. We note that our results are robust against variations in PAUP parameter values (see Supplementary Tables 1–3).

**Contingent gain analysis.** For each stepping-stone reaction pair A–B, A is defined as the reaction that is beneficial in a given nutrient environment without B, while a gain of B is only beneficial in another environment when A is already present. For each A–B pair we calculated the phylogenetic contingent gain fraction ($f$), defined as $f=p1/(p1+p2)$, where $p1$ is calculated by dividing the number of evolutionary events where B is gained in the descendent (d) when A is already present in the ancestor (a) (a10_d11) by the total number of all possible gain and loss scenarios taking place in the descendant when A is present but B is absent in the ancestor (a10_dXX, where X = 0 or 1), and p2 is calculated by dividing the number of evolutionary events where B is gained in the descendant when A is absent in the ancestor (a00_d01) by the total number of all possible gain and loss scenarios taking place in the descendant when both A and B are absent in the ancestor (a00_dXX, where X = 0 or 1). The observed distribution of fractions was then compared with the null-hypothesis that a gain of B is independent of the presence of A, that is, $f=0.5$, using a one-tailed one-sample Wilcoxon signed-rank test.

**Co-gain analysis.** For the phylogenetic co-gain analysis we calculated for reaction pairs the co-gain fraction, defined as $f=n1/(n1+n2)$, where $n1$ is the number of evolutionary events where both reactions were absent in the ancestor (a) and both were gained in the descendent (d) (a00_d11), and $n2$ is the number of evolutionary events where both reactions were absent in the ancestor and only one was gained in the descendent (a00_d10 or a00_d01). We compared the fractions ($f$) from reaction pairs that are predicted to be beneficial for growth only when they are simultaneously gained, referred to as 'beneficial without individual effect', with the fractions from reaction pairs that are beneficial for growth in a specific environment when co-gained, but at least one of the reactions is also beneficial on its own in a different environment (beneficial with combined and individual effect) (see Fig. 3b in main text). A one-sided Wilcoxon rank test was used. In addition, we compared the fractions from 'beneficial without individual effect' reaction pairs with the expected co-gain fraction by chance (randomization (without individual effect)). To do that, we broke the pairing between reactions and shuffled the reactions into new pairs, thereby generating a new list of gene pairs. This was repeated 1,000 times. Then we determined for each of the 1,000 reaction pair list if the mean co-gain fraction is higher than that of the 'beneficial without individual effect' and summed these (n1). $P$-value was calculated as $P=(n1+1)/1{,}001$. The randomization analysis was also carried out for reaction pairs that are beneficial for growth in a specific environment when co-gained, but at least one of the reactions is also beneficial on its own in a different environment (beneficial with combined and individual effect versus randomization (combined and individual effect)).

**Strains and plasmids and primers for laboratory adaptation.** *E. coli* K-12 MG1655 was considered as the wild-type strain in our experiments. MG1655 mutD5 was constructed using a suicide plasmid-based genome engineering method incorporating a C->T mutation at position 236,110 on the genome (within the *dnaQ* gene) resulting in a T15I mutation of the encoded enzyme described previously[27]. Standard steps and plasmids (pST76-A, pSTKST) of this methodology have been described[53]. Briefly, an approximately 800-bp-long targeting DNA fragment carrying the desired point mutation in the middle was synthesized by PCR, then cloned into a thermosensitive suicide plasmid (pST76-A). This plasmid construct was then transformed into the cell, where it was able to integrate into the chromosome by way of a single crossover between the mutant allele and the corresponding chromosomal region. The desired co-integrates were selected by the antibiotic resistance carried on the plasmid at a non-permissive temperature for plasmid replication (42 °C). Next, the pSTKST helper plasmid was transformed,

then induced within the cells, resulting in the expression of the I-SceI meganuclease enzyme, which cleaves the chromosome at the 18 bp recognition site present on the integrated plasmid. The resulting chromosomal gap is repaired by way of RecA-mediated intramolecular recombination between the homologous segments in the vicinity of the broken ends. The recombinational repair results in either a reversion to the wild-type chromosome, or in a markerless allele replacement, which can be distinguished by sequencing the given chromosomal region. See Supplementary Table 6 for the primers used for the mutation construction.

For the overexpression of FucO, the pCA24N plasmid containing the *fucO* gene was selected from the ASKA library[29] and isolated from the host strain, then electroporated into the MG1655 strain. Overexpression of the gene was achieved by the addition of 50 μM IPTG.

For the simultaneous overexpression of *fucO* and *aldA*, the chloramphenicol resistance cassette ($Cm^R$) of the pCA24N-*aldA* plasmid from the ASKA library was exchanged to the kanamycin resistance marker ($Km^R$), resulting in pCA24N-*aldA*-Km. The pCA24N-*aldA* plasmid was first linearized by inverse PCR amplification using the pCA24N_frame_1 and pCA24N_frame_2 primer pair flanking the $Cm^R$ cassette. The PCR product was treated with DpnI for 60 min at 37 °C and purified using the DNA Clean & Concentrator-5 Kit (Zymo Research #D4004). The $Km^R$ marker was PCR amplified from a pSTKST template using the ASKA-Gibson_Kan_Fw and ASKA-Gibson_Kan_Rev primers. The PCR fragment was then isolated from 1% agarose gel using the GeneJET Gel Extraction Kit (Thermo Scientific #K0691). The resulting DNA fragments were assembled using Gibson assembly cloning (Gibson Assembly Master Mix, New England Biolabs #E2611), according to the manufacturer's protocol, then electroporated into electrocompetent *E. coli* DH10B cells. Correct assemblies were verified by colony PCR using the ASKA-S2 and aldA-1 primer pair. Sequences of primers used in this construction are listed in Supplementary Table 7.

**Media used in laboratory adaptation.** Minimal salts (MS) medium was used as described previously[34], supplemented either with 0.4% glycerol, 30 mM (S)-propane-1, 2-diol (propylene glycol, PG), or 30 mM ethane-1, 2-diol (EG). Antibiotics were employed in the following working concentrations: 50 μg ml$^{-1}$ ampicillin (Ap), 25 μg ml$^{-1}$ chloramphenicol (Cm) and 25 μg ml$^{-1}$ kanamycin (Km).

**Adaptation of strains for growth on PG and EG.** Three replicates of each individual strain were started from single colonies grown on MS + 0.4% glycerol agar plates (with Cm added where the *fucO* overexpression plasmid was present) at 30 °C. An MG1655 strain carrying the pCA24N-*fucO* plasmid was previously found to grow at 30 °C in 2 ml MS media supplemented with 30 mM PG (with 25 μg ml$^{-1}$ Cm and 50 μM IPTG added). This culture was subsequently plated onto MS + 0.4% glycerol (+ Cm) agar plates, from which the PG + colonies, starters for selection for EG-utilization, were isolated. We opted for glycerol as a base carbon source to avoid catabolite repression (that is, the inhibition of utilization of various other carbon sources) as in ref. 28. Starter cultures were then grown in 2 ml MS + 0.4% glycerol (+ Cm where needed), from which 250 μl was then transferred to 25 ml fresh liquid MS media + 0.4% glycerol (and Cm where needed). Cultures were grown to stationary phase at 30 °C, after which total cell count was determined by plating of appropriate dilutions onto MS + 0.4% glycerol agar plates. The remainder of the cultures were then harvested and resuspended in 400 μl MS media without carbon source and finally plated in two halves onto MS agar plates supplemented with either 30 mM PG or 30 mM EG (with Cm and 50 μM IPTG added where the *fucO* overexpression plasmid was present). Plates were then incubated at 30 °C for 40 days after which adapted colonies were counted and isolated. The plates were placed in plastic bags for the duration of the incubation to prevent significant drying of the agar media. Rates of adaptive mutations were calculated based on three replicate experiments as follows. When adapted colonies were observed, we simply calculated the average ratio of the number of adapted colonies per total cell number. In cases where no growing colonies were obtained, we calculated an upper limit to the adaptive mutation rate following the approach presented in ref. 35. Specifically, we made use of the fact that the Poisson distribution has a 5% probability of yielding zero events when the expected number of events is three. Thus, assuming no more than three adaptive mutations among all the cells tested in the three replicate experiments gives an upper bound on the adaptive mutation rate per cell per generation.

**Growth curve measurements.** Individual colonies of strains MG1655, MG1655 + pCA24N-*fucO*, MG1655 + pCA24N-*aldA*-Km and MG1655 + pCA24N-*fucO* + pCA24N-*aldA*-Km were grown and isolated from MS + 0.4% glycerol plates carrying the desired antibiotic for the given plasmids. Starter cultures from single colonies were grown in 5 ml liquid MS media supplemented with 0.4% glycerol, as well as 50 μM IPTG and 25 μg ml$^{-1}$ Cm and/or 25 μg ml$^{-1}$ Km in the case of plasmid-harbouring strains. Cultures were grown until saturation after which 10 ml MS media supplemented with 30 mM of either PG or EG as well as 50 μM IPTG and 25 μg ml$^{-1}$ Cm and/or 25 μg ml$^{-1}$ Km where needed, were inoculated with the overnight cultures at a 100 × dilution. A total of 100 μl of these samples were then placed in six separate wells on a 96-well tissue culture plate (Jet Biofil), and placed in a PowerWave XS2 (BioTek) microplate spectrophotometer and grown at 30 °C. The edges of the plate were sealed with Breathe-Easy gas permeable sealing membrane (Diversified Biotech) to prevent evaporation.

**Mutation rate measurements.** We estimated mutation frequencies of BW25113 (wild-type) and BW25113 overexpressing the FucO protein from the pCA24N_fucO plasmid. Briefly, cells resistant to rifampicin (carrying mutations in *rpoB* (ref. 54)) were selected and counted. After overnight growth at 37 °C, ten tubes of 1 ml LB (+ 25 μg ml$^{-1}$ chloramphenicol in the case of pCA24N_fucO carrying samples) were inoculated with approximately $10^4$ cells each. FucO overexpression was induced by adding 50 μM IPTG after 2 h of growth, and cultures were grown to early stationary phase, all at 37 °C. Appropriate dilutions were spread onto non-selective LB agar plates and LB agar plates containing rifampicin (100 μg ml$^{-1}$). The samples were incubated at 37 °C and colony counts were performed after 24 or 48 h, respectively. Mutation rates were calculated with the Ma–Sandri–Sarkar maximum-likelihood method[55] using the FALCOR web tool[56].

**Ion Torrent library construction for whole-genome sequencing.** Fragment libraries were constructed from purified genomic DNA using NEBNext Fast DNA Fragmentation & Library Prep Set for Ion Torrent (New England Biolabs) according to manufacturer's instructions. Briefly, genomic DNA was enzymatically digested and the fragments were end-repaired. Ion Xpress Barcode Adaptors (Life Technologies) were than ligated and the template fragments size-selected using AmPure beads (Agencourt). Adaptor ligated fragments were then PCR amplified, cleaned-up using AmPure beads, quality checked on D1000 ScreenTape and Reagents using TapeStation instrument (Agilent) and finally quantitated using Ion Library TaqMan Quantitation Kit (Life Technologies). The library templates were prepared for sequencing using the Life Technologies Ion OneTouch protocols and reagents. Briefly, library fragments were clonally amplified onto Ion Sphere Particles (ISPs) through emulsion PCR and then enriched for template-positive ISPs. More specifically, PGM emulsion PCR reactions utilized the Ion OneTouch 200 Template Kit (Life Technologies), and as specified in the accompanying protocol, emulsions and amplification were generated using the Ion OneTouch System (Life Technologies). Enrichment was completed by selectively binding the ISPs containing amplified library fragments to streptavidin-coated magnetic beads, removing empty ISPs through washing steps, and denaturing the library strands to allow for collection of the template-positive ISPs. For all reactions, these steps were accomplished using the Life Technologies ES module of the Ion OneTouch System. Template-positive ISPs were deposited onto the Ion 318 chips (Life Technologies); finally, sequencing was performed with the Ion PGM Sequencing Kit (Life Technologies).

**Ion PGM sequencing data processing and mutation calling.** The PGM sequencing data was processed using Ion Torrent Suite v4.2.1 in order to perform signal processing and base calling. Read mapper module of Torrent Suite (tmap) was used to align raw reads to the *E. coli* K12 MG1655 genome sequence (U00096.3). Torrent Variant caller (tvc) module of Torrent Suite was subsequently applied to detect single nucleotide mutations as well as small in/del variants. Variant caller was programmed to run in high stringency mode requesting at least 12 × read coverage and at least 66% mutation frequency. Only those variants were taken into account that were supported by sequencing on both strands. BAM alignment files were imported in CLC Genomics Workbench v7.5.1 (CLCBio) and variant regions were manually inspected in all strains. Large genomic rearrangements (deletions or amplifications with lengths above 10 kb) were manually identified using CLC Genomics Workbench Tool.

Sequencing data of the ancestral and evolved strains are deposited in the NCBI SRA database (accession numbers SRX1167076 and SRX1167031).

## References

1. Lynch, M. & Abegg, A. The rate of establishment of complex adaptations. *Mol. Biol. Evol.* **27,** 1404–1414 (2010).
2. Lynch, M. Scaling expectations for the time to establishment of complex adaptations. *Proc. Natl Acad. Sci. USA* **107,** 16577–16582 (2010).
3. Copley, S. D. Evolution of efficient pathways for degradation of anthropogenic chemicals. *Nat. Chem. Biol.* **5,** 559–566 (2009).
4. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14,** 559–571 (2013).
5. Orr, H. A. Adaptation and the cost of complexity. *Evolution* **54,** 13–20 (2000).
6. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* **9,** 965–974 (2008).
7. Wagner, A. *The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems* (Oxford University Press, 2011).
8. Hayden, E. J., Ferrada, E. & Wagner, A. Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* **474,** 92–95 (2011).
9. Mira, A., Ochman, H. & Moran, N. A. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* **17,** 589–596 (2001).
10. Iwasa, Y., Michor, F. & Nowak, M. A. Stochastic tunnels in evolutionary dynamics. *Genetics* **166,** 1571–1579 (2004).
11. Coyne, J. A., Barton, N. H. & Turelli, M. Perspective: A critique of Sewall Wright's shifting balance theory of evolution. *Evolution* **51,** 643–671 (1997).

12. Kashtan, N., Noor, E. & Alon, U. Varying environments can speed up evolution. *Proc. Natl Acad. Sci. USA* **104,** 13711–13716 (2007).
13. Lenski, R. E., Ofria, C., Pennock, R. T. & Adami, C. The evolutionary origin of complex features. *Nature* **423,** 139–144 (2003).
14. Gray, M. W., Lukes, J., Archibald, J. M., Keeling, P. J. & Doolittle, W. Irremediable complexity? *Science* **330,** 920–921 (2010).
15. Finnigan, G. C., Hanson-Smith, V., Stevens, T. H. & Thornton, J. W. Evolution of increased complexity in a molecular machine. *Nature* **481,** 360–364 (2012).
16. Horowitz, N. H. On the Evolution of Biochemical Syntheses. *Proc. Natl Acad. Sci. USA* **31,** 153–157 (1945).
17. Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30,** 409–425 (1976).
18. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).
19. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* **37,** 1372–1375 (2005).
20. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat. Biotechnol.* **28,** 977–982 (2010).
21. Price, N. D., Reed, J. L. & Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2,** 886–897 (2004).
22. Papp, B., Notebaart, R. A. & Pal, C. Systems-biology approaches for predicting genomic evolution. *Nat. Rev. Genet.* **12,** 591–602 (2011).
23. Gould, S. J. & Vrba, E. S. Exaptation-a missing term in the science of form. *Paleobiology* **8,** 4–15 (1982).
24. Notebaart, R. A. *et al.* Network-level architecture and the evolutionary potential of underground metabolism. *Proc. Natl Acad. Sci. USA* **111,** 11762–11767 (2014).
25. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405,** 299–304 (2000).
26. Boronat, A., Caballero, E. & Aguilar, J. Experimental evolution of a metabolic pathway for ethylene glycol utilization by Escherichia coli. *J. Bacteriol.* **153,** 134–139 (1983).
27. Fijalkowska, I. J. & Schaaper, R. M. Mutants in the Exo I motif of Escherichia coli dnaQ: defective proofreading and inviability due to error catastrophe. *Proc. Natl Acad. Sci. USA* **93,** 2856–2861 (1996).
28. Lee, D. H. & Palsson, B. O. Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a nonnative carbon source, L-1, 2-propanediol. *Appl. Environ. Microbiol.* **76,** 4158–4168 (2010).
29. Kitagawa, M. *et al.* Complete set of ORF clones of Escherichia coli ASKA library (a complete set of E. coli K-12 ORF archive): unique resources for biological research. *DNA Res.* **12,** 291–299 (2006).
30. Baldoma, L. & Aguilar, J. Involvement of lactaldehyde dehydrogenase in several metabolic pathways of Escherichia coli K12. *J. Biol. Chem.* **262,** 13991–13996 (1987).
31. Barve, A. & Wagner, A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500,** 203–206 (2013).
32. Maslov, S., Krishna, S., Pang, T. Y. & Sneppen, K. Toolbox model of evolution of prokaryotic metabolic networks and their regulation. *Proc. Natl Acad. Sci. USA* **106,** 9743–9748 (2009).
33. Sabarly, V. *et al.* The decoupling between genetic structure and metabolic phenotypes in Escherichia coli leads to continuous phenotypic diversity. *J. Evol. Biol.* **24,** 1559–1571 (2011).
34. Hall, B. G. The EBG system of E. coli: origin and evolution of a novel beta-galactosidase for the metabolism of lactose. *Genetica* **118,** 143–156 (2003).
35. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc. Natl Acad. Sci. USA* **105,** 7899–7906 (2008).
36. Linster, C. L., Van Schaftingen, E. & Hanson, A. D. Metabolite damage and its repair or pre-emption. *Nat. Chem. Biol.* **9,** 72–80 (2013).
37. Weng, J. K., Philippe, R. N. & Noel, J. P. The rise of chemodiversity in plants. *Science* **336,** 1667–1670 (2012).
38. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330,** 1385–1389 (2010).
39. Sauer, U. Evolutionary engineering of industrially important microbial phenotypes. *Adv. Biochem. Eng. Biotechnol.* **73,** 129–169 (2001).
40. O'Neill, M., Vanneschi, L., Gustafson, S. & Banzhaf, W. Open issues in genetic programming. *Genet. Program. Evolvable Mach.* **11,** 339–363 (2010).
41. Das, S., Mandal, A. & Mukherjee, R. An adaptive differential evolution algorithm for global optimization in dynamic environments. *IEEE Trans. Cybern.* **44,** 966–978 (2014).
42. Orth, J. D. *et al.* A comprehensive genome-scale reconstruction of Escherichia coli metabolism--2011. *Mol. Syst. Biol.* **7,** 535 (2011).
43. McCloskey, D., Palsson, B. O. & Feist, A. M. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.* **9,** 661 (2013).
44. de Figueiredo, L. F. *et al.* Computing the shortest elementary flux modes in genome-scale metabolic networks. *Bioinformatics* **25,** 3158–3165 (2009).
45. R Core Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2015).
46. Gelius-Dietrich, G., Amer Desouki, A., Fritzemeier, C. J. & Lercher, M. J. sybil - Efficient constraint-based modelling in R. *BMC Syst. Biol.* **7,** 125 (2013).
47. Powell, S. *et al.* eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.* **40,** D284–D289 (2012).
48. Szklarczyk, D. *et al.* The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.* **39,** D561–D568 (2011).
49. Notebaart, R. A., Kensche, P. R., Huynen, M. A. & Dutilh, B. E. Asymmetric relationships between proteins shape genome evolution. *Genome Biol.* **10,** R19 (2009).
50. Lu, X., Kensche, P. R., Huynen, M. A. & Notebaart, R. A. Genome evolution predicts genetic interactions in protein complexes and reveals cancer drug targets. *Nat. Commun.* **4,** 2124 (2013).
51. Swofford, D. L. *{PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4.}* (Sinauer Associates, 2003).
52. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12,** 17–25 (2002).
53. Fehér, T. *et al.* Scarless engineering of the Escherichia coli genome. in *Microbial Gene Essentiality: Protocols and Bioinformatics* (Springer, 2008).
54. Jin, D. J. & Gross, C. A. Mapping and sequencing of mutations in the Escherichia coli rpoB gene that lead to rifampicin resistance. *J. Mol. Biol.* **202,** 45–58 (1988).
55. Sarkar, S., Ma, W. T. & Sandri, G. H. On fluctuation analysis: a new, simple and efficient method for computing the expected number of mutants. *Genetica* **85,** 173–179 (1992).
56. Hall, B. M., Ma, C. X., Liang, P. & Singh, K. K. Fluctuation analysis CalculatOR: a web tool for the determination of mutation rate using Luria-Delbruck fluctuation analysis. *Bioinformatics* **25,** 1564–1565 (2009).

## Acknowledgements

We acknowledge the insightful comments of the anonymous reviewers on a previous version of the paper. This work was supported by the 'Lendület' Programme of the Hungarian Academy of Sciences and The Wellcome Trust (B.P. and C.P.), European Research Council (C.P.), the Hungarian Scientific Research Fund PD 109572 (B.C.), the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/2-11-/1-2012-0001 'National Excellence Program' (B.S.), the János Bolyai Research Scholarship of the Hungarian Academy of Sciences (I.N.), the Hungarian Academy of Sciences Postdoctoral Fellowship Programme (V.L.), the Seventh Framework Programme (FP7) of the European Union through the GENCODYS Consortium, FP7-HEALTH-241995 (XL), German Research Foundation (CRC 680) (MJL) and a fellowship from the graduate school E-Norm of the Heinrich-Heine University (J.F.). Computational support of the Zentrum für Informations- und Medientechnologie (ZIM) at the Heinrich-Heine University is gratefully acknowledged.

## Author contributions

B.P., C.P. and M.J.L. conceived and supervised the project; B.C., V.L., B.S. and I.N. designed the experiments; B.C., V.L., I.N., B.B. and R.H. performed the experiments; B.S., J.F., X.L., R.A.N. and G.F. performed computational data analysis; and B.S., J.F., B.C., M.J.L., R.A.N., C.P. and B.P. wrote the paper.

## Additional information

**Accession codes:** Sequencing data of the ancestral and evolved strains are deposited in the NCBI SRA database with accession codes SRX1167076 and SRX1167031.

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**How to cite this article:** Szappanos, B. *et al.* Adaptive evolution of complex innovations through stepwise metabolic niche expansion. *Nat. Commun.* 7:11607 doi: 10.1038/ncomms11607 (2016).

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/