MODELLING

Systems-biology approaches for predicting genomic evolution

Balázs Papp[*‡], *Richard A. Notebaart*[§] *and Csaba Pál*[*]

Abstract | Is evolution predictable at the molecular level? The ambitious goal to answer this question requires an understanding of the mutational effects that govern the complex relationship between genotype and phenotype. In practice, it involves integrating systems-biology modelling, microbial laboratory evolution experiments and large-scale mutational analyses — a feat that is made possible by the recent availability of the necessary computational tools and experimental techniques. This Review investigates recent progresses in mapping evolutionary trajectories and discusses the degree to which these predictions are realistic.

Epistatic interaction
Non-independent effect of mutations on a phenotype. Epistasis is negative when a genotype with two mutations has a lower phenotype value or positive when it has a higher value than would be expected from the product of the single mutant values.

*Synthetic and Systems Biology Unit, Institute of Biochemistry, Biological Research Center, Temesvári krt. 62, H-6726 Szeged, Hungary.
‡Cambridge Systems Biology Centre and Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK.
§Departments of Bioinformatics (CMBI) and Systems Biology (CSBB), Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, P.O. BOX 9101, 6500 HB Nijmegen, The Netherlands.
Correspondence to C.P.
e-mail: cpal@brc.hu
doi:10.1038/nrg3033
Published online 2 August 2011

The integration of Mendelian genetics into evolutionary biology in the early twentieth century allowed a better understanding of a broad range of biological phenomena and unified several previously isolated fields. Despite the enormous success of the modern synthesis, certain key issues have remained unanswered. Most notably, although evolutionary biology successfully interprets molecular and cellular phenotypes as a result of diverse evolutionary forces that acted in the past, it rarely builds an explicit theoretical framework to predict potential routes of evolution[1,2]. Why is this issue important? First, it could help to establish the degree to which evolution is repeatable. Although long-term microbial evolutionary experiments have provided numerous examples of parallel phenotypic and genetic evolution[3], it is unclear how predictable evolution is at the level of genomes and molecular networks. Second, such a framework has the potential to permit informed decisions in medicine[4], biotechnology[5] and environmental issues[6]. For example, although *in vitro* methods have been developed to forecast the evolution of antibiotic resistance to newly developed drugs at the protein level[7], no such general tool exists for larger subsystems or whole organisms.

In this Review, we demonstrate that it is possible to predict, rather than simply interpret, past evolution by synthesizing evolutionary theory, systems biology and molecular data. Even under constant selection, predicting evolutionary change is challenging for two main reasons. First, evolution is a complex mixture of deterministic and chance events: the occurrence, order and fixation of mutations in populations are all partially stochastic. Second, predicting evolution requires a detailed knowledge of the range of available mutations and their fitness effects, an issue that could be best addressed by combining organism-specific mechanistic models and large-scale mutational analyses. There are three layers of prediction that we consider in this Review (TABLE 1): predicting the distribution of mutational effects and epistasis (that is, parameters that influence many central issues in evolutionary genomics); explaining the driving forces of sequence and expression evolution on a genomic scale; and understanding why particular evolutionary trajectories are realized, whereas others are not.

The recent availability of systematic gene-deletion studies[8,9], genome-scale epistatic interaction maps[10] and detailed mutation analyses of individual proteins[11] provides valuable insights into these problems. However, most established experimental approaches are limited, either because they focus on individual genes instead of large gene networks, or because they study a restricted set of environments and mutation types (TABLE 1). Systems biology can help to resolve these issues by allowing the analysis of large cellular subsystems and providing molecular explanations with clear links to changes in environmental conditions. The Review focuses on genome-scale models of microbial metabolic networks[12] owing to their large-scale, predictive power coupled with mechanistic insights and wide usage.

The Review starts with a brief summary of metabolic network modelling; we emphasize the data used for model reconstructions, the model-building steps and the reliability and limitations of these models. We then discuss how these models can be used to study the three layers of prediction described above (TABLE 1). Last, we demonstrate how computational and experimental approaches can be more tightly integrated by

Table 1 | **Three major issues in evolutionary systems biology**

Layer of prediction	Importance	Current state of knowledge	Difficulty	New tools and knowledge needed for integration
Distribution of mutational effects and epistatic interactions	General architecture of adaptation. Robustness against mutations	Wealth of systematic gene deletion studies and epistasis maps	Existing fitness landscape models are not biologically detailed. High-throughput experiments are restricted to a few environmental conditions, or they only consider null mutations	Realistic systems-biology models offer new predictions on mutational effects and mechanistic insights. New types of experimental data (for example, fitness profiling of point mutations or gene overexpression studies)
General patterns of genome evolution	Evolutionary forces driving protein and expression divergence, gene loss, horizontal gene transfer and gene duplicability	Impact of post-genomic features (for example, gene expression or network position)	No clear relationship between fitness and the post-genomic gene features studied	Predictions and measuring most relevant physiological data (for example, range of neutrality, optimal gene activity or physiological coupling between genes)
Specific evolutionary trajectories	Relative importance of chance and necessity in evolution. Predictive tools for applications	Map of adaptive landscape for single proteins. Insight from experimental evolutionary studies	Difficult to map adaptive landscapes for large cellular subsystems empirically. Interpretations dominate over predictions	Modelling the outcome of adaptive evolution at the molecular level. New experimental technologies to map adaptive landscapes

considering new technological advances in the fields of experimental evolution, genome engineering and automated model reconstruction.

Genome-scale metabolic models

Properties and advantages. The three layers of predictions listed above require biologically detailed computational models that estimate the impact of mutations and environmental changes on fitness. Models that are most suited for evolutionary studies are based on sound biochemical principles: they should capture the functional states of the cell and compute phenotypes (for example, growth rate) that serve as fitness correlates. These models include detailed kinetic models of specific metabolic pathways[13] or regulatory circuits (for example, the cell cycle[14]), logical models of signalling networks[15] and constraint-based models of genome-scale metabolic networks[16] (TABLE 2). Organism-specific kinetic models are generally highly accurate and realistic but require detailed experimental data, which are rarely available for large systems. However, constraint-based models allow integration of high-throughput post-genomic data but generally offer no information about metabolite concentrations or about the temporal dynamics of the system (but see REFS 17–19).

Genome-scale metabolic models have been useful, as they rely on high-quality metabolic network reconstructions[12]. These reconstructions are primarily based on a sequenced genome and are generally built manually using information from metabolic databases — such as KEGG[20] and BRENDA[21] — and the primary literature. Next, the network reconstruction is converted into a mathematical model that can be analysed using constraint-based approaches (BOX 1). By mimicking nutrient conditions used in prior experimental studies, the model is validated against high-throughput data, and existing discrepancies are resolved by new sets of experiments.

Genome-scale metabolic models therefore have at least two conceptual advantages over other approaches (TABLE 2). First, whereas most other modelling approaches focus on small-scale biochemical systems (that is, individual pathways), constraint-based models aim to calculate the metabolic behaviour of moderately large systems (that is, 600–1,300 genes). Thus, these models allow comparisons to be made with the results of high-throughput genomic data. Second, in contrast to most statistical or graph-theoretical approaches[22], these models are far more detailed and realistic, as they infer the functional states of the network as a function of nutrient availability in the environment.

Genome-scale metabolic models have already proven to be successful in several applications: distinguishing between essential and non-essential genes across environmental conditions[23]; identifying epistatic interactions[24]; predicting growth properties[25]; guiding metabolic engineering[16]; and charting the functional dependence (coupling) between genes[26]. However, several important problems remain to be addressed (BOX 1), not least because these models generally do not incorporate enzyme kinetic information and cannot capture the nonlinear relationship between enzyme level and metabolic flux (but see REFS 18,19,27,28). These are the major reasons why predicting minor mutational effects on enzyme activity or predicting metabolite concentrations remains challenging.

Model applications. Despite these current limitations, evolutionary biologists have recognized the potential of these models and have used them to reach three goals. First, they have been used to estimate the overall patterns of epistasis[29–31] and distribution of mutational effects[32,33], and second, the models can infer interspecies differences in metabolic gene content and hence can explain general trends of genome evolution[34,35]. But perhaps the most inspiring aspect of this framework is its capacity to make specific and reliable predictions on the outcome of metabolic evolution, both in short-term laboratory evolution and on macroevolutionary time scales.

Distribution of mutational effects and epistasis

Recent large-scale gene deletion analyses demonstrated that mutations with weak phenotypic effects are

Graph-theoretical approaches
The study of graphs. A graph provides an abstract representation of a biological or physical system in which components are represented by nodes that are connected to each other by edges (links).

Table 2 | **Modelling approaches to study genotype–phenotype relationships**

Modelling approach	Examples	Scale	Data requirement	Phenotypes predicted	Advantages	Disadvantages	Refs
Digital organisms (that is, self-replicating computer programs)	Avida platform	Potentially large	No biological data required	Replication rate; performance of mathematical operations	Infers general principles of evolution	No direct connection to specific biological systems	82,83
Graph-theoretical models	Wide range of cellular interaction networks	Large	Large-scale molecular interaction data	Network properties, such as diameter and centrality	Low data requirement; insights into similarities of network architecture across species	Unclear how network architecture relates to cellular physiology and fitness	22
Kinetic biochemical models	Metabolic pathways, gene regulation and cell cycle	Small	Detailed: molecular interactions and kinetic information	Reaction fluxes; component concentrations	Conceptual understanding; realistic; quantitative; captures dynamics	Only available for small-scale systems	13,14, 84
Logical models	Cell cycle, signalling and metabolism	From medium to large	Qualitative knowledge of molecular interactions	Activity states; viability; dynamic behaviour	Low data requirement; captures dynamics to some extent	Difficult to capture continuous molecular response in a discrete model	15,85, 86
Constraint-based models	Flux balance analysis of genome-scale metabolic networks	From medium to large	Network reconstruction based on omics data; biochemical and physiological studies	Growth properties; reaction fluxes across conditions	No enzyme kinetic information required; testable predictions on a genomic scale	Basic models lack dynamics in time; metabolite concentrations are not captured	12

common[8,36], and epistasis between mutations is widespread[10]. It has been intensely debated why mutations with overt phenotypes are so rare, and it remains largely unexplored to what extent the distribution of epistatic interactions changes across loci and environmental conditions (TABLE 1). These issues are expected to influence several problems in evolutionary genomics, starting from the architecture of adaptation[37], the accumulation of deleterious mutations[38], the evolution of sexual reproduction[39] and the extent of robustness against harmful mutations[40]. We focus on a specific aspect of the last problem, referred to as the 'gene knockout paradox', and then briefly discuss recent progress in understanding epistasis networks (FIG. 1). Attempts to resolve this paradox demonstrate the power and challenges of evolutionary systems biology.

Causes of gene dispensability. One of the most surprising discoveries of the post-genomic era has been the extent to which organisms can tolerate inactivation of their genes (FIG. 1A). Large-scale single-gene deletion screens suggest that nearly 80% of protein-coding genes in *Saccharomyces cerevisiae* are not essential for viability under standard laboratory conditions[8] — an observation that tallies with results from similar analyses performed in other organisms[36]. These findings raise questions about the mechanistic basis of gene dispensability and about whether this tolerance to inactivation is the result of an evolved capacity of genetic networks to compensate for mutations. There are at least three mutually non-exclusive explanations for the knockout paradox. First, gene dispensability may be more apparent than real: these genes are important under other natural environmental settings that are not yet investigated in the laboratory. Second, gene deletions may be compensated for by a gene duplicate with a redundant function[41]. Third, reorganization of metabolic fluxes across alternative pathways[42] may buffer gene loss. Although clear examples exist for all three scenarios, it is difficult to establish which of these mechanisms explains the majority of dispensable genes. As genome-scale metabolic models correctly predict knockout viability in 80–90% of the genes studied[33] and make inferences about reaction activities, they also hold the promise to test different scenarios to resolve this problem.

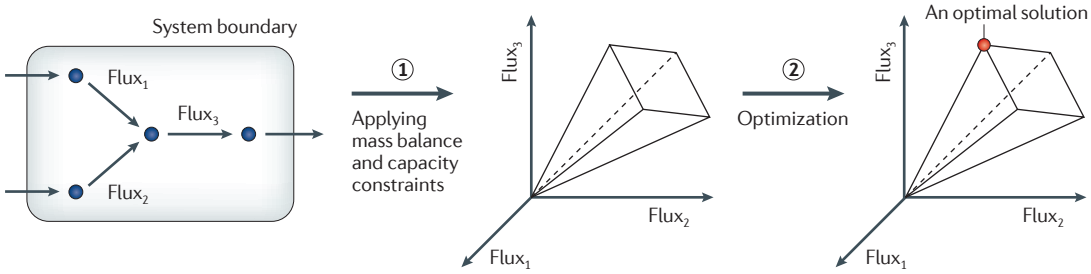
A computational analysis of *S. cerevisiae* metabolism showed that a large fraction of non-essential enzyme-encoding genes catalyse reactions that are inactive under the tested condition (that is, they carry zero flux). Furthermore, by simulating gene deletions under several different nutrient conditions, the model claimed that many functionally inactive genes would become essential under some other conditions[32]. The model indicates that most of the apparently dispensable genes (37–68%) belong to this category, whereas redundant gene duplicates (15–28%) and alternative pathways (4–17%) can only explain a few cases. These computational predictions also gained strong support from comparative and experimental studies. First, as might be expected, these condition-specific genes have limited phylogenetic distribution[32]. Second, experimental measurements of reaction fluxes in the same species showed remarkable agreement with general predictions of the model[43]. Approximately 50% of reactions are estimated to be inactive under laboratory conditions, whereas redundancy through duplicate genes was the major (37.5%) molecular mechanism behind gene dispensability, and alternative pathways constituted the minor mechanisms

Robustness
Mutational robustness describes the resilience of phenotypes to genetic perturbations.

Gene dispensability
A measure that is inversely related to the overall importance of a gene. It is usually approximated by the fitness of the corresponding gene-knockout strain under laboratory conditions.

Metabolic fluxes
Turnover rate of substrates through metabolic reactions or pathways.

Box 1 | **Genome-scale metabolic network models and their limitations**



Genome-scale metabolic networks are studied within the framework of constraint-based modelling: that is, by identifying and applying constraints to define ranges of allowable phenotypes without the need for enzyme kinetic information[12]. The approach relies on a mathematical representation of the network and its interaction with the environment and uses physicochemical constraints to describe the potential behaviour of the network. At least two constraints are applied: a mass balance constraint and a capacity constraint. Assuming that the system will reach a steady state, the mass balance constraint specifies that each internal metabolite is consumed and produced at the same rate, whereas the capacity constraints set maximum and minimum bounds on fluxes (step **1** in the figure). The nutrient environment is specified by allowing certain metabolites to enter the network via transport processes. Imposing constraints defines the space of allowable metabolic flux states (solution space), making it possible to query where the physiological solution lies.

Flux balance analysis (FBA) is commonly used to predict a physiologically meaningful steady state and is based on the idea that organisms adapt towards maximal growth efficiency. Importantly, in FBA, growth is explicitly represented as a reaction in which all biomass compounds that are necessary for growth are drained from the system. FBA then uses optimization techniques to identify a flux distribution that maximizes growth (step **2** in the figure).

Despite its predictive power, the basic FBA suffers from several limitations. For example, it cannot predict metabolite concentrations, and it is restricted to studying steady states. Thus, certain types of evolutionary predictions — for instance, those requiring metabolic dynamics — cannot be made. Furthermore, in its simplest form, it does not account for regulatory mechanisms. Given the complexity of the metabolic network and the small amount of information used to make predictions, it is not surprising that FBA performs better at predicting growth efficiency (a proxy for fitness) than the underlying intracellular flux distribution and that it can best capture major genetic changes (that is, addition or removal of genes). There are at least three major complementary areas of development aimed at overcoming these limitations. First, basic models are being extended to incorporate other cellular subsystems, such as gene regulatory[60,61] and signalling networks[17]. Second, dynamic FBA models[17,18,87] have been developed with the intention of investigating temporal changes in metabolic behaviour. Third, *in vivo*, high-throughput data (metabolomics, fluxomics, proteomics and transcriptomics) can be integrated to increase the predictive value of the model[19,27,88,89].

(12.5%). More generally, a recent large-scale chemical genomic assay in yeast demonstrated that 97% of the single-gene deletions exhibited a measurable growth phenotype in at least one of the hundreds of tested conditions[9].

Thus, these studies indicate that environmental specificity is the dominant explanation for apparent dispensability. However, it is important to emphasize that theories on gene dispensability are not mutually exclusive, as the capacity to compensate for null mutations may vary substantially between different nutritional environments. One theory suggests that availability of nutrients across conditions determines the number of parallel metabolic pathways that can produce a specific key cellular metabolite[24]. A unique prediction of the theory is that the phenotypic effect of single- and double-gene deletions should vary across conditions. In agreement with expectation, computational models coupled with experimental studies demonstrated that most synthetic lethal interactions between loci are restricted to certain environments owing to lack of compensation under some nutrient conditions[24]. These results indicate that robustness against null mutations is unlikely to be a directly selected trait but that it is a side effect of adaptation to survive in changing conditions[44].

Epistatic interactions. Despite its influence on a number of evolutionary processes, the consequences of epistasis have been mainly examined using simplified and unrealistic assumptions about the distribution of interactions between mutations[45]. For example, most population genetics studies assumed that all genes interact in a uniform way. Genome-scale metabolic models offer realistic insights into the architecture of epistatic interactions on a global scale, and recent studies in yeast and bacteria have started to explore several of the properties of epistatic interaction networks. Although the methods used are far from perfect, these studies (along with complementary high-throughput experimental studies[10,31]) clearly indicate that more realistic theoretical models should be used to explore the evolutionary consequences of epistasis. These works have reached four important conclusions:

- First, the distribution of epistatic interactions across pairs of loci is clearly non-uniform and probably trimodal[29]. Trimodality means that, after controlling for average mutational effects, pairs of loci show either very strong negative or positive interactions or no interactions at all. This pattern demands a re-evaluation of the influence of epistasis on evolutionary processes, such as the rate of accumulation of deleterious mutations and the evolution of recombination[45].

Flux balance analysis
(FBA). A mathematical approach for analysing the behaviour of large-scale metabolic networks. It does not require knowledge of metabolite concentration or enzyme kinetic details.

Synthetic lethal interaction
A form of epistasis between two genes in which the double mutant shows a no-growth phenotype that is not exhibited by either single mutant.

Trimodal
Trimodality is a statistical term for a distribution that has three modes.

- Second, epistatic interactions are generally plastic across environmental conditions[24,46]; that is, they are present in some, but not all, environments. Thus, evolutionary theories that rely on a constant adaptive landscape (for example, theories on the origins of robustness against mutations or the impact of compensatory mutations on evolution) are far too simplistic.
- Third, a few genes ('hubs') exhibit an especially large number of epistatic interactions, whereas the majority of genes display few interactions[31]. Systems modelling illuminated that there is a strong link between the degree of epistatic interaction and the extent of pleiotropy (FIG. 1B): hub genes contribute to multiple biological processes, and, as such, the phenotypic effect that occurs on their deletion can potentially be modulated by a large number of other genes, resulting in numerous epistatic links[31]. Because hub genes are likely to influence the phenotypic effects of mutations in many other genes, one might speculate that their loss may markedly alter which mutational paths are available for adaptation.
- Fourth, epistatic interactions are frequent between biochemical pathways or modules[31]. In the context of evolutionary biology, a module should be tightly integrated by strong pleiotropic effects and should be largely independent from other such modules (that is, pleiotropic effects of genes within a module are mostly restricted to phenotypic traits that are associated with that module)[47]. Thus, it is unclear how far biochemical modules fulfil these criteria, and the claim that modularity facilitates evolution by minimizing pleiotropic constraints[48] needs to be re-evaluated.

Most of the above conclusions are based on studying null mutations only. Future work should reveal how far they hold when the range of mutations is extended to include point mutations and minor regulatory changes, as well as interactions between beneficial mutations[49].

Understanding general patterns of evolution

Until recently, studies of evolution have concentrated on estimating relevant parameters and understanding the cellular mechanisms behind them. With recent advances in systems-biology modelling coupled to comparative genomics, it has become possible to move a step forward and study genome evolution. More specifically, these models can explain broad patterns of gene conservation, gene duplicability and sequence and expression evolution on a genomic scale. For example, systematic surveys showed that the extent of protein sequence conservation is governed by global biochemical and cellular features, including expression pattern, centrality in biological networks and genomic position[50]. Nevertheless, it has remained difficult to draw mechanistic conclusions from these patterns because these features have no clear links to cellular physiology and hence to fitness. This problem can be overcome by investigating functionally more relevant features that can be computed by systems modelling. In the case of metabolic genes, these features include: the impact of gene deletion on the performance of the network[51] (which is a proxy for essentiality); the extent and direction of functional dependencies between enzymes[52] (that is, identifying functional modules); optimal flux distribution[53]; and the range of optimal flux levels[54].

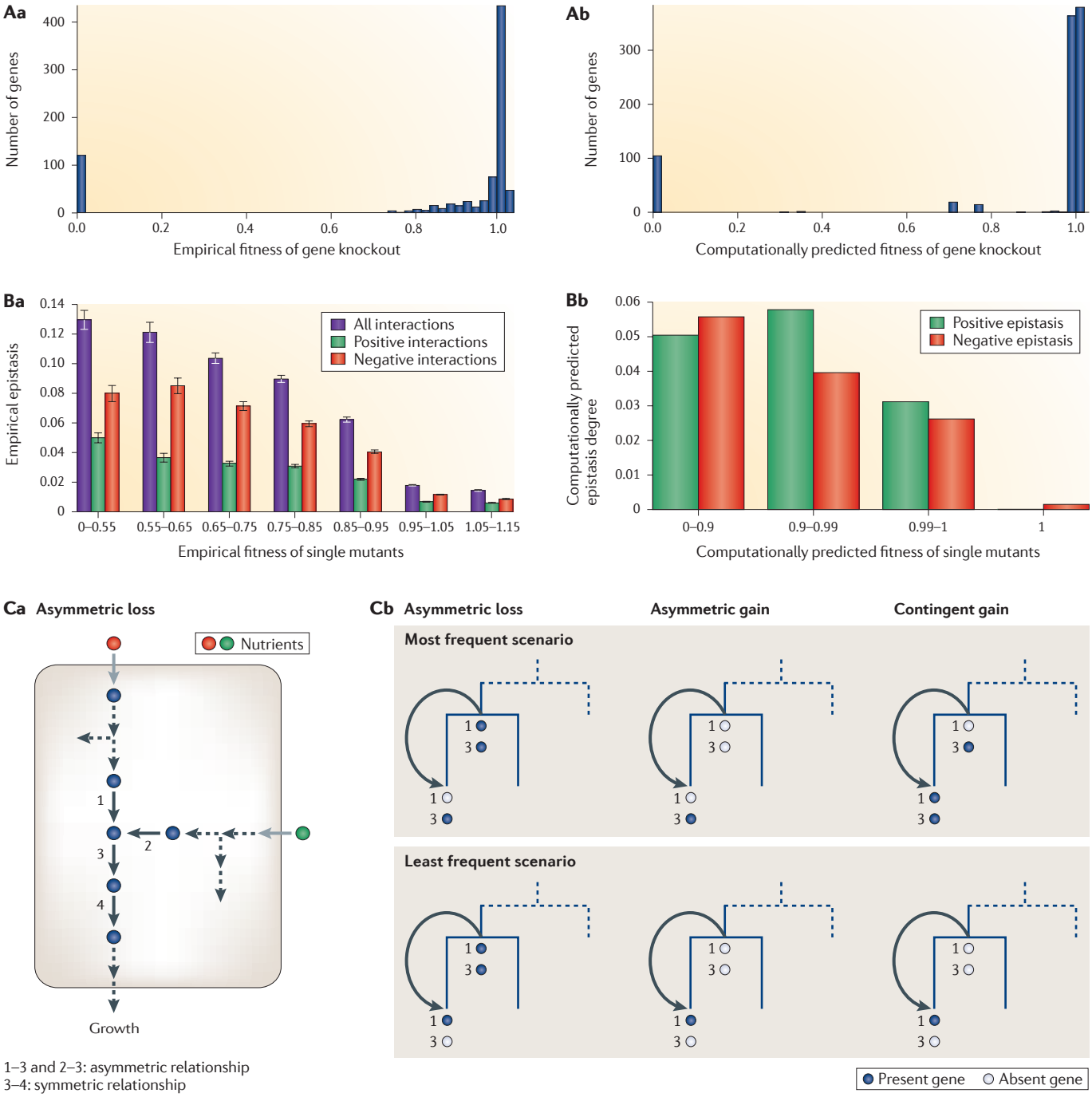
Although these features are expected to have a large impact on the evolution of protein-coding genes in metabolic networks, measuring them experimentally on a reasonably large scale is an enormous challenge. Moreover, as these gene features depend on the functional state of the cell, they cannot be inferred by studying single proteins or short metabolic pathways in isolation. Instead, genome-scale metabolic models can be used to derive computational estimates of these features for each gene in the network (TABLE 3) across a wide range of nutrient conditions. These estimates can then be correlated to different aspects of genomic evolution using standard tools of comparative genomics.

Evolution of gene content. Such an integrated approach has recently been applied to develop an understanding of the evolutionary forces driving the expansion and shrinkage of metabolic networks[34]. It has been demonstrated that most recent changes to bacterial metabolic networks are due to horizontal gene transfer rather than gene duplication events (see also REF. 55). The acquired genes are generally integrated at the periphery of the network, leaving the central part of the network intact for millions of years of evolution. Remarkably, computational modelling showed that these changes are driven by adaptation to new environmental conditions rather than optimization of performance under routine growth conditions[34].

Are genes added or lost from metabolic networks one at a time, or does evolution proceed through simultaneous gain and loss of whole sets of genes? Metabolic models offer unbiased and systematic calculations of various physiological dependencies between reactions and enzymes[26], ultimately leading to a reconstruction of functional modules. The simplest situation is when two proteins mutually depend on each other for their function. In the case of metabolism, this happens when fluxes through two or more reactions are fully correlated, such as seen in the case of unbranched linear metabolic pathways. As expected, the encoding genes are often transferred together through horizontal transfer, frequently in operons[34].

A more complex situation arises when functional relationships between proteins are asymmetric. For instance, when multiple metabolic pathways (for example, 1 and 2) converge into one central pathway (for example, 3), fluxes through pathways 1 and 2 depend on functioning of the central pathway 3 but not vice versa (FIG. 1Ca). As most functional relationships between metabolic genes are asymmetric[35], they can potentially have important implications for genomic evolution. Indeed, evolutionary analysis of functionally asymmetric gene pairs revealed that genes in pathways 1 or 2 can be easily lost in the presence of pathway 3,

Pleiotropy
The phenomenon of one mutation affecting multiple traits.



1–3 and 2–3: asymmetric relationship
3–4: symmetric relationship

Figure 1 | **Model and data: mutational effects, epistasis and gene content evolution. A** | Distribution of experimentally measured (**a**) and computationally predicted (**b**) single-gene-knockout fitness values in *Saccharomyces cerevisiae* metabolism. Empirical fitness data were taken from REF. 97 and refer to competitive fitness in a rich medium (YPD). Although the histogram is restricted to metabolic genes, the same trend is observed when all *S. cerevisiae* genes are considered. *In silico* fitness values, which closely match the empirical distribution, were computed using a genome-scale metabolic model of yeast[98]. **B** | Degree of epistasis of a gene (that is, its number of epistatic interactions with other genes, whether positive or negative) correlates inversely with its single mutant fitness both experimentally (**a**) and computationally (**b**). **C** | Panel **a** shows a schematic representation of a simplified metabolic network in which the 1–3 reaction pair (and the 2–3 reaction pair) is asymmetrically dependent, whereas the 3–4 reaction pair is symmetrically dependent in functional states of the network. Because reactions 3 and 4 are in the same linear pathway, their activities are completely correlated, hence their symmetric relationship. By contrast, the activities of 1 and 2 depend on 3, but 3 does not exclusively depend on either 1 or 2, therefore their relationships are asymmetric. Panel **b** shows that the predicted asymmetric functional relationships between enzyme-encoding genes (1 and 3) are reflected in their evolutionary trajectories, as inferred from comparative genomic analyses[35]. In asymmetric loss, 1 is more frequently lost in situations where only one gene is lost. In asymmetric gain, 3 is more frequently gained in situations where only one gene is gained. In contingent gain, 1 is more frequently gained if 3 is already present in the ancestor. Panel **Ba** is modified, with permission, from REF. 10 © (2010) American Association for the Advancement of Science. Panel **Bb** is modified, with permission, from REF. 31 © (2011) Macmillan Publishers Ltd. All rights reserved.

Table 3 | **Patterns of genome evolution explained by metabolic modelling**

Patterns of genome evolution to be explained	Evolutionary scenario	Prediction	Variable predicted by the systems-biology model
Phylogenetic distribution of genes	Gene loss is governed by changing environmental conditions	Condition-specific genes should be more frequently lost	Conditional essentiality — that is, the number of conditions in which the gene is required for growth
Correlated loss and gain of genes	Coevolution at the genomic level reflects the extent and direction of functional coupling between genes	Loss or gain of a gene in a specific metabolic pathway or modules should alter the evolutionary trajectories of other functionally related genes	Flux coupling — that is, physiological dependence between enzymes
Gene duplicability	Evolutionary maintenance of gene duplicates with redundant functions is governed by selection favouring enhanced dosage of the same protein	Optimal enzymatic flux should be especially high for reactions catalysed by gene duplicates	Optimal enzymatic flux
Rate of protein evolution	Enzymes carrying high fluxes tend to have more central roles. Therefore, mutational reduction of enzymatic activity should be especially detrimental	Enzymes with high enzymatic fluxes should evolve slowly	Optimal enzymatic flux
Gene expression divergence	Rate of expression divergence is largely governed by neutral evolution. Genes for which expression can differ widely without major phenotypic effects should diverge more rapidly across species	Range of neutral expression variation should correlate with expression divergence	Range of neutral variation (estimated by optimal flux range)

but the reverse is not true[35] (FIG. 1Cb). Moreover, genes involved in pathways 1 or 2 only tend to be gained across evolutionary lineages when pathway 3 is present in the genome (contingent evolution). Thus, this framework can explain the order of gene-gain and gene-loss events in evolution.

Gene duplicability. In sharp contrast to bacteria, in which most major changes are due to horizontal gene transfer, in eukaryotic metabolic networks, evolutionary novelties are mainly due to gene duplicates[34]. The foremost difficulty for the evolution of metabolic diversity by duplicates is preservation of both copies prior to functional divergence. Surprisingly, eukaryotic metabolic networks contain several reactions that are catalysed by gene duplicates with highly overlapping functions. These duplicates may be retained to provide a shield against harmful mutations[56]. Alternatively, selection may favour enhanced dosage of the same protein to provide high enzymatic flux. A genome-scale model of yeast metabolism suggests that the second explanation is closer to the truth: gene duplicates that catalyse the same enzymatic reactions are not especially common for essential reactions but rather for reactions that require high fluxes[32].

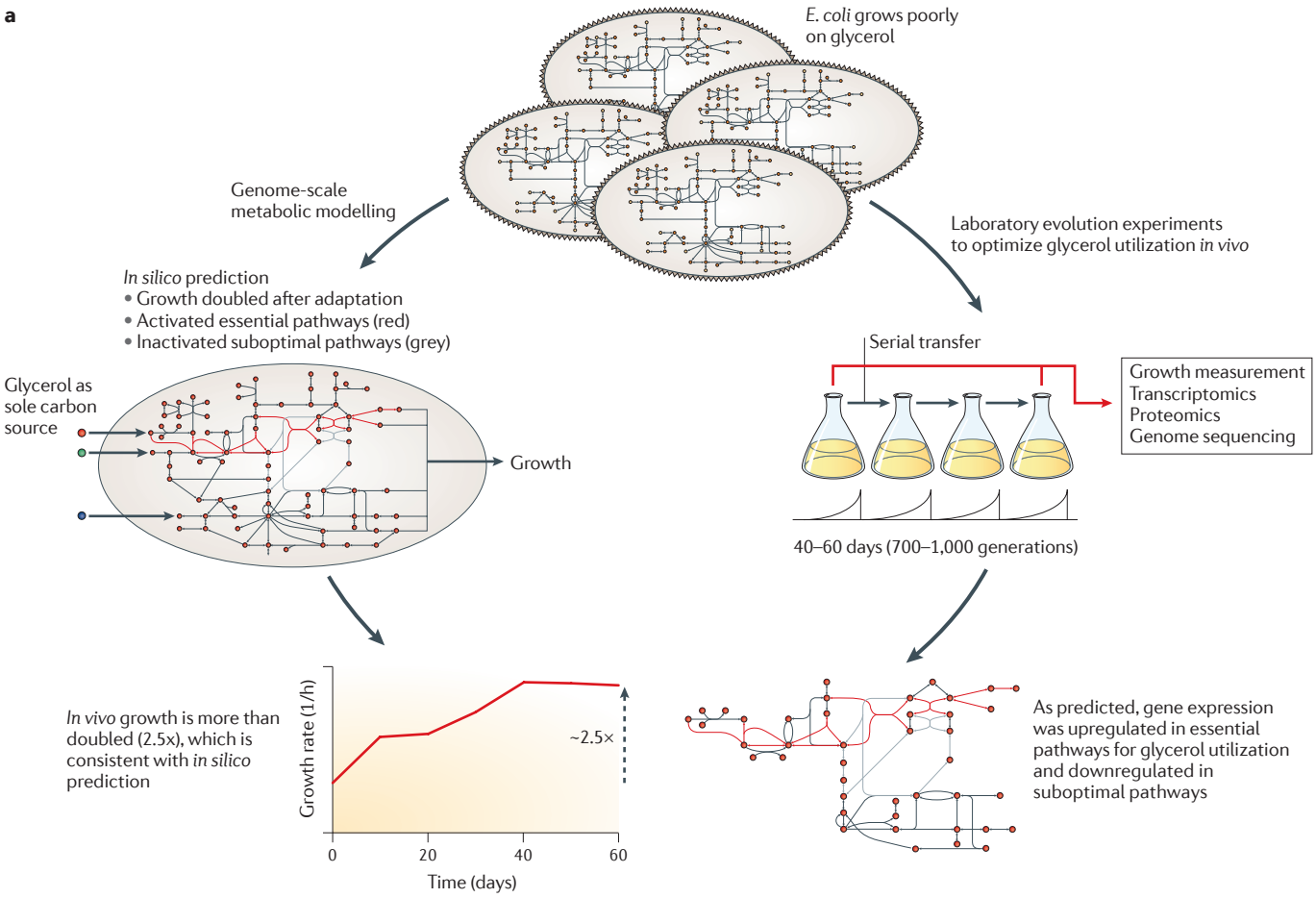
Sequence and expression evolution. Metabolic networks evolve not only by gene-gain and gene-loss events but also by minor changes in enzyme kinetics and gene dosage through point mutations. According to standard theories of molecular evolution, the rate of protein and gene expression evolution should first and foremost depend on the availability of (nearly) neutral mutations. Systems modelling offers insight into the evolutionary driving forces by calculating both the level of optimal flux[53] and the range of its neutral variation[54] across genes. A study conducted in yeast showed that enzymes with high predicted metabolic fluxes not only tend to undergo gene duplication events but also evolve more slowly[57]. One possible interpretation of this finding is that most mutations are detrimental[58], so they tend to reduce enzymatic activity and hence flux. Under the assumption that enzymes carrying high fluxes are more important, mutations in such enzymes should be especially detrimental, leading them to undergo few amino acid changes during evolution. In a similar vein, genes with narrow neutral flux ranges are also more conserved in sequence and display low variation in expression among different yeast strains[59]. Taken together, these results support the notion that the extent of sequence and expression conservation is not only influenced by the molecular properties of the gene but also by its activity and position in the metabolic network.

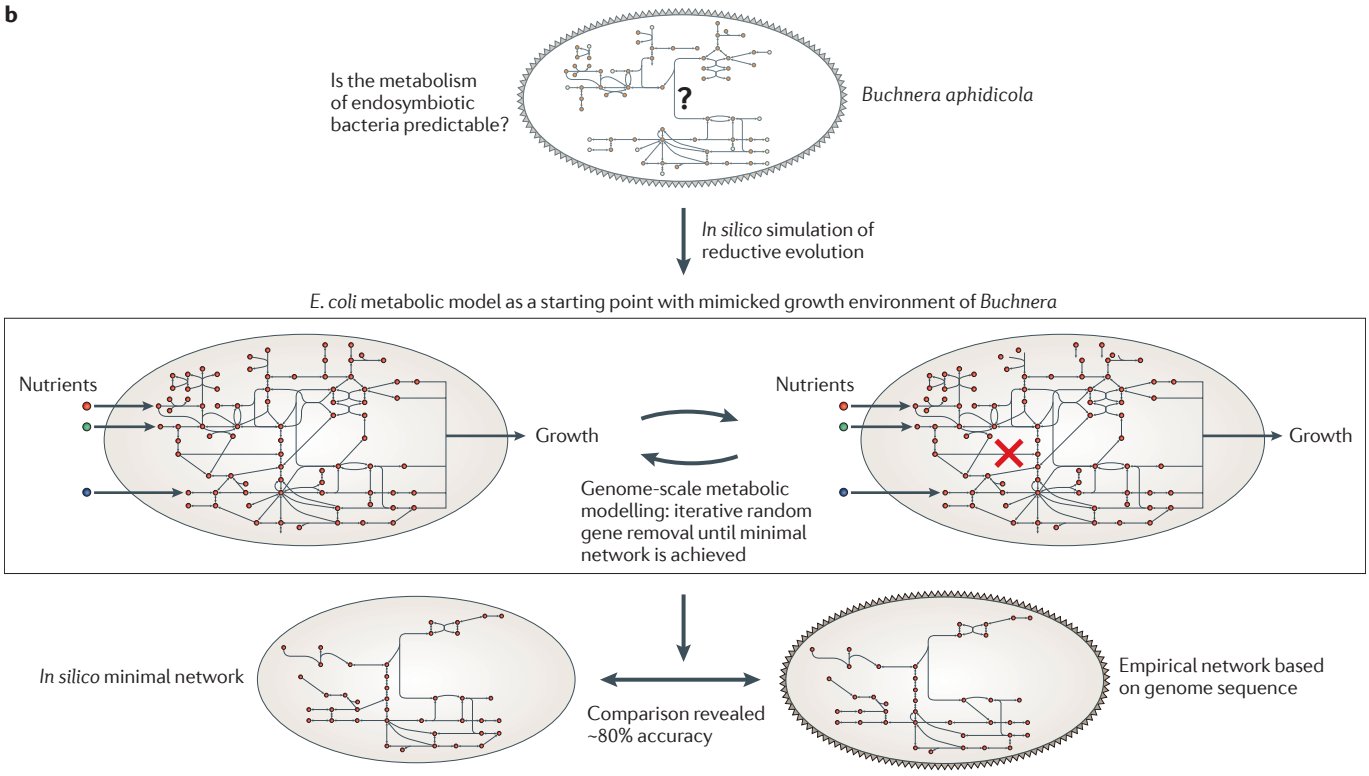
Based on the above considerations, one can claim that, by calculating gene features with more direct links to cellular physiology and hence fitness, systems modelling allows the testing of important issues in evolutionary genomics. Nevertheless, the scope of problems studied so far is still limited, not least because existing approaches consider functional variation in enzyme-encoding genes only. We anticipate that metabolic models incorporating regulatory details[60,61] (including non-coding RNAs) will allow a better understanding of the evolutionary forces shaping regulatory and signalling networks.

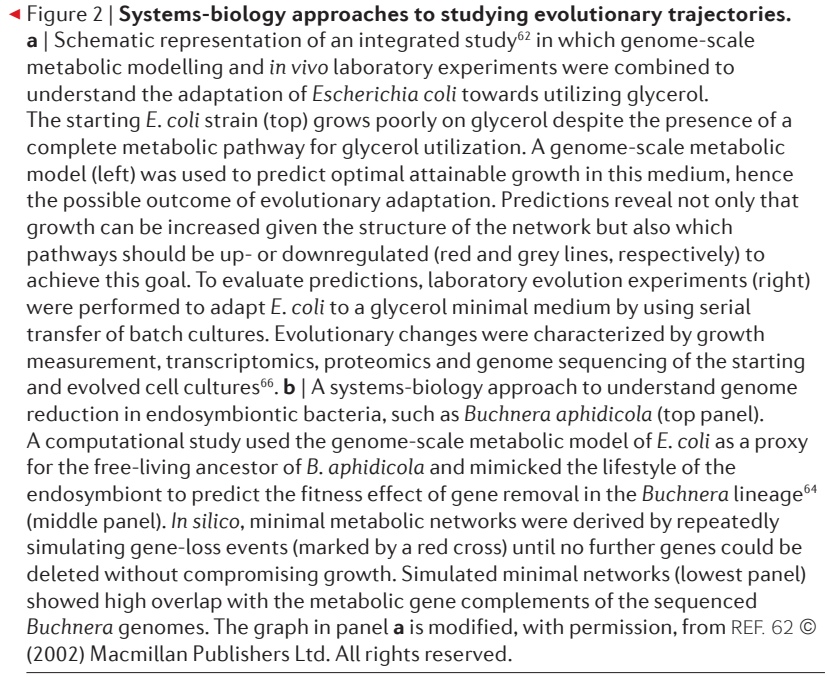
Inferring evolutionary trajectories

The above sections demonstrate that systems approaches successfully capture mutational effects and epistatic interactions and also contribute to our understanding of genomic evolution. Based on these initial successes, one might wonder whether these models could be even more specific and predict which genes are likely to be

(Nearly) neutral mutations
A neutral mutation is one that has no fitness effect. A mutation is 'nearly' neutral when its fitness effect is too small to be governed by selection, and hence its fate is determined largely by genetic drift.







◀ Figure 2 | **Systems-biology approaches to studying evolutionary trajectories.** **a** | Schematic representation of an integrated study[62] in which genome-scale metabolic modelling and *in vivo* laboratory experiments were combined to understand the adaptation of *Escherichia coli* towards utilizing glycerol. The starting *E. coli* strain (top) grows poorly on glycerol despite the presence of a complete metabolic pathway for glycerol utilization. A genome-scale metabolic model (left) was used to predict optimal attainable growth in this medium, hence the possible outcome of evolutionary adaptation. Predictions reveal not only that growth can be increased given the structure of the network but also which pathways should be up- or downregulated (red and grey lines, respectively) to achieve this goal. To evaluate predictions, laboratory evolution experiments (right) were performed to adapt *E. coli* to a glycerol minimal medium by using serial transfer of batch cultures. Evolutionary changes were characterized by growth measurement, transcriptomics, proteomics and genome sequencing of the starting and evolved cell cultures[66]. **b** | A systems-biology approach to understand genome reduction in endosymbiontic bacteria, such as *Buchnera aphidicola* (top panel). A computational study used the genome-scale metabolic model of *E. coli* as a proxy for the free-living ancestor of *B. aphidicola* and mimicked the lifestyle of the endosymbiont to predict the fitness effect of gene removal in the *Buchnera* lineage[64] (middle panel). *In silico*, minimal metabolic networks were derived by repeatedly simulating gene-loss events (marked by a red cross) until no further genes could be deleted without compromising growth. Simulated minimal networks (lowest panel) showed high overlap with the metabolic gene complements of the sequenced *Buchnera* genomes. The graph in panel **a** is modified, with permission, from REF. 62

lost, to mutate or to change regulatory interactions over the course of evolution. Indeed, the ultimate goal of evolutionary systems biology is to develop a global understanding of why particular evolutionary trajectories are realized, whereas others are not (TABLE 1). Although only a handful of systems-biology studies have investigated this issue in detail, they have been successful: models can predict the short-term outcome of adaptive[62] and compensatory[63] evolution in the laboratory and replay 200 million years of genomic evolution[64] (TABLE 3).

Evolution in the laboratory. Predicting evolution at the molecular level requires a detailed map of adaptive landscapes — that is, a clear understanding of how multiple mutations affect fitness. Recent experimental mutational studies have provided insight into the properties of adaptive landscapes at the level of individual proteins[11]. However, exhaustive experimental exploration of mutations for larger cellular subsystems is unfeasible. Investigating different evolutionary routes requires a combination of three tools: laboratory evolution experiments, metabolic modelling and post-genomic analyses (FIG. 2a). In a series of pioneering studies, Palsson and colleagues integrated these techniques to study the adaptive evolution of *Escherichia coli* K12 with glycerol as its sole carbon source[62,65,66]. Despite the presence of a complete metabolic pathway for glycerol utilization, starting bacterial populations only grew suboptimally on this medium. After 700 generations of evolution in the laboratory, the evolved lines reproducibly showed a massive increment in fitness. Remarkably, the corresponding changes in growth rates and nutrient uptake during laboratory evolution showed good agreement with predictions of the genome-scale metabolic model. Thus, strains have evolved towards the computationally predicted optimal metabolic state[62].

What are the underlying molecular mechanisms of glycerol adaptation? Initial slow growth in this medium may be due to a suboptimal genomic transcriptional program or allosteric inhibition of key enzymatic steps in the glycerol utilization pathway. Genomic comparison of ancestral lines with five evolved lines has found several recurring patterns, and they support both scenarios[65]. Identified mutations fell into two major categories: those affecting specific rate-limiting enzymatic steps (for example, glycerol kinase) and those affecting global transcription patterns. Quantitative proteomic and transcriptomic data obtained from ancestral and laboratory-evolved strains revealed that hundreds of genes and proteins became differentially expressed[66]. Importantly, the metabolic model not only successfully explains the observed trends of gene expression evolution but also provides a mechanistic explanation for why down- or upregulation of certain enzymes and pathways is advantageous[66]. The optimal enzyme usage predicted by the model shows excellent agreement with genome-scale expression data obtained from the evolved lines. Furthermore, both models and experiments indicate that regulatory adaptation has two complementary aspects. First, bottlenecks from dosage limitations in genes that are essential for growth (or needed for optimal glycerol utilization) were overcome by upregulation of these genes. Second, non-functional pathways were shut down for optimal growth in order to reduce costs of producing unnecessary proteins and metabolites.

In summary, evolved strains upregulate genes within pathways needed for optimal growth and downregulate genes outside the optimal growth solutions. Some of these laboratory-evolved lines also showed partial fitness loss in rich media, hinting at an evolutionary trade-off[65]. It remains to be seen how far metabolic network models can capture such trade-offs and their underlying molecular mechanisms.

Evolution of genome minimization. The above studies show that it is becoming increasingly possible to predict the outcome of laboratory evolution. Can models also replay evolution at the molecular level on a macroevolutionary timescale? This would be an important first step towards predicting genomic evolution rather than simply interpreting it in retrospect. This is an ambitious goal, as it requires knowledge of relevant selection pressures, population size, growth conditions and the range of possible mutations. In sharp contrast to laboratory evolution, where these parameters can be monitored (or even controlled) throughout the experiment, these parameters cannot generally be inferred from the current lifestyle of the organism and genomic data, let alone from fossil records.

Despite the above obvious difficulties, it has been shown that certain aspects of genome evolution (and the corresponding intermediate steps) can be predicted with knowledge of distant ancestors and estimates of relevant selection pressures[64,67] (FIG. 2b). These studies concentrated on understanding the process of genome reduction in endosymbiotic bacteria such as *Buchnera* and *Wigglesworthia* species. These species are relatives of

Adaptive landscapes
Visualizations of the relationship between genotype and fitness. The plane of the landscape contains all possible genotypes in such a way that similar genotypes are located close to each other on the plane and the height of the landscape reflects the fitness of the corresponding genotype.

Trade-off
Two traits are in a trade-off relationship when an increase in fitness owing to a change in one trait is opposed by a decrease in fitness owing to a concomitant change in the second trait.

Historical contingency
This term describes the situation in which future evolutionary alternatives of a population depend on its prior history.

E. coli and have evolved from free-living ancestors since the two lineages split approximately 200 million years ago and lost 75% of their protein-coding genes[68]. Despite reaching a nearly minimal set of genes needed to sustain life, even closely related *Buchnera aphidicola* strains contain substantial variation in their gene contents. Three processes may underlie this apparent genomic diversity: differences in initial genetic makeup, variation in lifestyle and chance deletion of alternative genetic pathways. To examine how lifestyle variation and chance deletion shape the evolution of metabolic gene contents, one study[64] combined evolutionary simulations and metabolic modelling (FIG. 2b). The study used a genome-scale model of *E. coli* metabolism and mimicked the lifestyle of the endosymbiont. The authors repeatedly simulated successive gene-loss events computationally, until no further genes could be deleted without compromising *in silico* growth. Comparison of the gene content of simulated minimal networks with the *B. aphidicola* genome revealed that the model has reached an overall 80% prediction success. Furthermore, by integrating metabolic modelling and phylogenetic reconstruction of intermediate evolutionary steps, a recent study showed that the order and timing of loss events are also predictable[67].

However, these studies also recognized a theoretical limit on predictability. Although independently simulated minimal networks preserved a core metabolism (also over-represented in strict intracellular bacteria), they were variable in both gene content and number. Thus, evolutionary paths are contingent on the order of gene deletion events owing to the presence of parallel metabolic pathways in the ancestral genomes. These results suggest the hypothesis that some of the differences in gene content between intracellular bacteria may reflect alternative solutions to reach similar goals rather than organism-specific adaptations. We anticipate that the role of historical contingency will be especially important when networks evolve in more complex environments or when horizontal gene transfer has a large role in adaptation.

Box 2 | **Key issues in network evolution**

Integrating targeted genome engineering with laboratory evolution and computational modelling could considerably increase our understanding of the following open issues in network evolution.

Impact of network rewiring on metabolic functioning
What is the adaptive value of introducing new enzymatic reactions or rewiring regulatory links in particular environments? Systematic network modifications by means of genome engineering[90] will allow researchers to map the fitness landscape of metabolic networks and also explore the space of plausible alternative molecular circuits.

Neutral evolution and emergence of key innovations
How does the neutral evolution of metabolic networks influence the emergence of evolutionary innovations[91]? A computational study showed that the presence of alternative metabolic circuits with the same phenotype is a key facilitator of evolutionary novelty (that is, the ability to utilize new nutrients)[92]. In principle, this prediction can be tested experimentally by measuring the fitness of alternative network circuits under various environmental conditions.

Role of promiscuous enzyme activities in network evolution
Promiscuous functions — weak activities for which the enzyme is not directly selected — have been suggested to have important roles as raw materials for future adaptive evolution[93,94]. Generating large pools of mutations in numerous targeted promiscuous enzymes and exposing the mutant strains to repeated rounds of selection will shed light on how novel promiscuous pathways evolve.

Importance of regulatory versus structural mutations in adaptive evolution
Phenotypic changes could arise through mutations in *cis*-regulatory sequences or coding regions, but their relative importance remains intensely debated[95]. This issue could be addressed by directed evolution *in vitro*[81] by modifying the targets of available genetic variation.

Convergent evolution of network structure and function
How frequent is convergent evolution at the network level[96]? Replaying adaptive network evolution in the laboratory would allow the prevalence of convergence to be estimated and computational predictions on the availability of alternative evolutionary trajectories to be tested.

Outlook and future directions

The emerging field of evolutionary systems biology reinvestigates central issues in evolutionary biology by using realistic and organism-specific models of cellular subsystems. The goal of the corresponding computational analyses is at least threefold. First, they calculate evolutionarily relevant variables that are difficult to estimate experimentally on a large-scale or across environmental conditions. Second, they provide mechanistic insights into complex evolutionary phenomena from the causes of gene dispensability and the adaptation of global transcriptional programs to the evolution of minimized genomes. Third, these models also hold the promise to transform evolutionary biology into a more predictive discipline. We wish to emphasize that conclusions derived from these analyses should be fully compatible with established population genetics mechanisms and molecular knowledge[69]. This goal can be best achieved by applying hybrid models, in which dynamic, system-level analysis is integrated into an explicit evolutionary framework (see REF. 44 for an example of this hybrid approach). More generally, systems biology and population genetics should calculate the mutational effects and fixation rate of mutations, respectively.

We see at least two major directions for progress, which are discussed in turn. First, new modelling frameworks are required, and existing ones should be improved (BOX 1). The functional analysis of genome-scale metabolic networks currently has no counterpart in any other large cellular subsystem, but this limitation is expected to change in the near future. Existing genome-scale metabolic models also frequently fail to capture minor mutational effects and interactions between them. We anticipate that solving this difficult problem requires integration of SNP data on enzyme property variations[70] and more complex models that incorporate some details on enzyme kinetics as well[18].

Moreover, reliable genome-scale metabolic models are currently only available for a handful of microbial species[16], and even existing ones have several limitations. However, there has recently been enormous progress towards the automated, high-throughput generation, optimization and analysis of genome-scale metabolic models[71]. Thus, there is a real possibility that the development of new models will soon keep pace with genome sequencing of new microbial species, a much awaited step towards 'comparative systems biology' (REFS 72,73).

Cross-feeding
This describes the situation in which one species or strain degrades a primary resource and secretes a chemical compound that is used as a substrate by another species or strain.

Multiplex automated genome engineering
An automated and efficient experimental technique to simultaneously modify many targeted genomic locations.

Such advances would be important for the following reasons. For the first time, it would be possible to investigate the evolution of system-level properties of metabolic networks — such as growth properties or epistatic interactions — across related microbial species and how these properties depend on changes in genome architecture and ecological conditions[73,74]. Along with comparative genomics approaches, it will pave the way for network archaeology: that is, the reconstruction and analysis of the functional properties of ancestral cellular networks[75]. It will allow systematic predictions of ecological interactions (for example, mutualism) between species and will shed new light on topics such as the evolution of cross-feeding and cooperative behaviour[76,77]. In a similar vein, it could lay a rigorous foundation for 'reverse ecology'. Reverse ecology aims to gain insight into the habitats in which organisms have evolved based on comparison of networks across a wide range of species[78,79]. More practically, comparative systems-biology models could provide enormous help in the identification of new drug targets shared by numerous related pathogenic bacteria[80].

Second, there is an urgent need for new experimental technologies to investigate mutational effects and evolution in a high-throughput manner. Given the limited timescale of microbial laboratory evolution experiments, only a few mutations are fixed in most laboratory-evolved populations. Moreover, studying the evolution of a particular cellular subsystem (for example, metabolism) is hindered by the fact that beneficial mutations can occur outside the subsystem under investigation. Novel genome-engineering techniques may provide the key to solving these problems: it has recently become possible to generate huge diversity at specific loci in the genome (through multiplex automated genome engineering[81]). We anticipate that these novel experimental techniques, along with computational models of specific cellular subsystems, will allow researchers to reinvestigate key issues in network evolution (BOX 2).

1. Stern, D. L. & Orgogozo, V. The loci of evolution: how predictable is genetic evolution? *Evolution* **62**, 2155–2177 (2008).
2. Stern, D. L. & Orgogozo, V. Is genetic evolution predictable? *Science* **323**, 746–751 (2009).
3. Barrick, J. E. *et al.* Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* **461**, 1243–1247 (2009).
4. Nesse, R. M. & Stearns, S. C. The great opportunity: evolutionary applications to medicine and public health. *Evol. Appl.* **1**, 28–48 (2008).
5. Johannes, T. W. & Zhao, H. Directed evolution of enzymes and biosynthetic pathways. *Curr. Opin. Microbiol.* **9**, 261–267 (2006).
6. Collins, S. & Bell, G. Phenotypic consequences of 1,000 generations of selection at elevated CO_2 in a green alga. *Nature* **431**, 566–569 (2004).
7. Hall, B. G. Predicting the evolution of antibiotic resistance genes. *Nature Rev. Microbiol.* **2**, 430–435 (2004).
8. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418**, 387–391 (2002).
9. Hillenmeyer, M. E. *et al.* The chemical genomic portrait of yeast: uncovering a phenotype for all genes. *Science* **320**, 362–365 (2008).
10. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
11. Dean, A. M. & Thornton, J. W. Mechanistic approaches to the study of evolution: the functional synthesis. *Nature Rev. Genet.* **8**, 675–688 (2007).
12. 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).
13. Teusink, B., Walsh, M. C., van Dam, K. & Westerhoff, H. V. The danger of metabolic pathways with turbo design. *Trends Biochem. Sci.* **23**, 162–169 (1998).
14. Chen, K. C. *et al.* Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell* **15**, 3841–3862 (2004).
15. Christensen, T. S., Oliveira, A. P. & Nielsen, J. Reconstruction and logical modeling of glucose repression signaling pathways in *Saccharomyces cerevisiae*. *BMC Syst. Biol.* **3**, 7 (2009).
16. Oberhardt, M. A., Palsson, B. O. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5**, 320 (2009). **This excellent review summarizes the wide range of current and potential applications of flux balance analysis and related methods.**
17. Lee, J. M., Gianchandani, E. P., Eddy, J. A. & Papin, J. A. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput. Biol.* **4**, e1000086 (2008).
18. Covert, M. W., Xiao, N., Chen, T. J. & Karr, J. R. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* **24**, 2044–2050 (2008).
19. Jamshidi, N. & Palsson, B. O. Mass action stoichiometric simulation models: incorporating kinetics and regulation into stoichiometric models. *Biophys. J.* **98**, 175–185 (2010).
20. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
21. Schomburg, I., Chang, A. & Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **30**, 47–49 (2002).
22. Barabasi, A. L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nature Rev. Genet.* **5**, 101–113 (2004).
23. 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).
24. 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).
25. Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotech.* **19**, 125–130 (2001).
26. 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).
27. Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. & Shlomi, T. Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics* **26**, i255–i260 (2010).
28. Smallbone, K., Simeonidis, E., Broomhead, D. S. & Kell, D. B. Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J.* **274**, 5576–5585 (2007).
29. Segrè, D., Deluna, A., Church, G. M. & Kishony, R. Modular epistasis in yeast metabolism. *Nature Genet.* **37**, 77–83 (2005).
30. He, X., Qian, W., Wang, Z., Li, Y. & Zhang, J. Prevalent positive epistasis in *Escherichia coli* and *Saccharomyces cerevisiae* metabolic networks. *Nature Genet.* **42**, 272–276 (2010).
31. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genet.* **43**, 656–662 (2011). **This study was the first global systems-biology analysis of epistatic interactions in metabolic networks.**
32. Papp, B., Pál, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429**, 661–664 (2004).
33. Kuepfer, L., Sauer, U. & Blank, L. M. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430 (2005).
34. Pál, C., Papp, B. & Lercher, M. J. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nature Genet.* **37**, 1372–1375 (2005).
35. Notebaart, R. A., Kensche, P. R., Huynen, M. A. & Dutilh, B. E. Asymmetric relationships between proteins shape genome evolution. *Genome Biol.* **10**, R19 (2009).
36. Hurst, L. D. & Pál, C. in *Evolutionary Genomics and Proteomics* (eds Pagel, M. & Pomiankowski, A.) 141–165 (Sinauer Associates Inc., Sunderland, Massachusetts, 2007).
37. Orr, H. A. The genetic theory of adaptation: a brief history. *Nature Rev. Genet.* **6**, 119–127 (2005).
38. Kondrashov, A. S. Mullers ratchet under epistatic selection. *Genetics* **136**, 1469–1473 (1994).
39. de Visser, J. A. & Elena, S. F. The evolution of sex: empirical insights into the roles of epistasis and drift. *Nature Rev. Genet.* **8**, 139–149 (2007).
40. Wagner, A. *Robustness and Evolvability of Living Systems* (Princeton Univ. Press, Princeton, 2005).
41. Gu, Z. *et al.* Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63–66 (2003).
42. Wagner, A. Robustness against mutations in genetic networks of yeast. *Nature Genet.* **24**, 355–361 (2000).
43. Blank, L. M., Kuepfer, L. & Sauer, U. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* **6**, R49 (2005).
44. Soyer, O. S. & Pfeiffer, T. Evolution under fluctuating environments explains observed robustness in metabolic networks. *PLoS Comput. Biol.* **6**, e1000907 (2010).
45. Phillips, P. C., Otto, S. P. & Whitlock, M. C. in *Epistasis and the Evolutionary Process* (eds Wolf, J. B., Brodie, E. D. & Wade, M. J.) 20–38 (Oxford Univ. Press, New York, 2000).
46. Bandyopadhyay, S. *et al.* Rewiring of genetic networks in response to DNA damage. *Science* **330**, 1385–1389 (2011).
47. Wagner, G. P. Homologues, natural kinds and the evolution of modularity. *American Zoologist* **36**, 36–43 (1996).
48. Wagner, G. P. & Altenberg, L. Complex adaptations and the evolution of evolvability. *Evolution* **50**, 967–976 (1996).
49. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332**, 1193–1196 (2011).
50. Pál, C., Papp, B. & Lercher, M. J. An integrated view of protein evolution. *Nature Rev. Genet.* **7**, 337–348 (2006).
51. Forster, J., Famili, I., Palsson, B. O. & Nielsen, J. Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *Omics* **7**, 193–202 (2003).
52. Burgard, A. P. & Maranas, C. D. Probing the performance limits of the *Escherichia coli* metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* **74**, 364–375 (2001).

53. Schuetz, R., Kuepfer, L. & Sauer, U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol. Syst. Biol.* **3**, 119 (2007).
54. Mahadevan, R. & Schilling, C. H. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab. Eng.* **5**, 264–276 (2003).
55. Lerat, E., Daubin, V., Ochman, H. & Moran, N. A. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3**, e130 (2005).
56. Nowak, M. A., Boerlijst, M. C., Cooke, J. & Maynard Smith, J. Evolution of genetic redundancy. *Nature* **388**, 167–171 (1997).
57. Vitkup, D., Kharchenko, P. & Wagner, A. Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7**, R39 (2006).
58. Eyre-Walker, A. & Keightley, P. D. The distribution of fitness effects of new mutations. *Nature Rev. Genet.* **8**, 610–618 (2007).
59. Bilu, Y., Shlomi, T., Barkai, N. & Ruppin, E. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS Comp. Biol.* **2**, e106 (2006).
This paper proposes that the range of neutral metabolic flux variation has a large impact on sequence and expression evolution.
60. 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).
61. Covert, M. W. & Palsson, B. O. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* **277**, 28058–28064 (2002).
62. Ibarra, R. U., Edwards, J. S. & Palsson, B. O. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* **420**, 186–189 (2002).
This paper demonstrates how systems modelling can be used to predict the outcome of laboratory experimental evolution.
63. Fong, S. S. & Palsson, B. O. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nature Genet.* **36**, 1056–1058 (2004).
64. Pál, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670 (2006).
This paper shows that 200 million years of genomic evolution is predictable by combining metabolic network analysis and evolutionary dynamics.
65. Herring, C. D. *et al.* Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nature Genet.* **38**, 1406–1412 (2006).
66. Lewis, N. E. *et al.* Omic data from evolved *E. coli* are consistent with computed optimal growth from genome-scale models. *Mol. Syst. Biol.* **6**, 390 (2010).
67. Yizhak, K., Tuller, T., Papp, B. & Ruppin, E. Metabolic modeling of endosymbiont genome reduction on a temporal scale. *Mol. Syst. Biol.* **7**, 479 (2011).
68. Moran, N. A. & Mira, A. The process of genome shrinkage in the obligate symbiont *Buchnera aphidicola*. *Genome Biol.* **2**, RESEARCH0054 (2001).
69. Loewe, L. A framework for evolutionary systems biology. *BMC Syst. Biol.* **3**, 27 (2009).
70. Jamshidi, N., Wiback, S. J. & Palsson, B. B. *In silico* model-driven assessment of the effects of single nucleotide polymorphisms (SNPs) on human red blood cell metabolism. *Genome Res.* **12**, 1687–1692 (2002).
71. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotech.* **28**, 977–982.
72. Teusink, B., Westerhoff, H. V. & Bruggeman, F. J. Comparative systems biology: from bacteria to man. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2**, 518–532 (2010).
73. Oberhardt, M. A., Puchalka, J., Martins dos Santos, V. A. & Papin, J. A. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput. Biol.* **7**, e1001116 (2011).
74. Harvey, P. H. & Purvis, A. Comparative methods for explaining adaptations. *Nature* **351**, 619–624 (1991).
75. Navlakha, S. & Kingsford, C. Network archaeology: uncovering ancient networks from present-day interactions. *PLoS Comput. Biol.* **7**, e1001119 (2011).
In this paper, the authors propose several novel algorithms to reconstruct the evolutionary history of microbial cellular networks.
76. Klitgord, N. & Segre, D. Environments that induce synthetic microbial ecosystems. *PLoS Comput. Biol.* **6**, e1001002 (2010).
77. Wintermute, E. H. & Silver, P. A. Emergent cooperation in microbial metabolism. *Mol. Syst. Biol.* **6**, 407 (2010).
By combining metabolic network modelling and laboratory experiments, the authors of this paper show mutual compensation of metabolic mutants by cross-feeding of essential metabolites.
78. Borenstein, E., Kupiec, M., Feldman, M. W. & Ruppin, E. Large-scale reconstruction and phylogenetic analysis of metabolic environments. *Proc. Natl Acad. Sci. USA* **105**, 14482–14487 (2008).
79. Borenstein, E. & Feldman, M. W. Topological signatures of species interactions in metabolic networks. *J. Comput. Biol.* **16**, 191–200 (2009).
80. Lee, D. S. *et al.* Comparative genome-scale metabolic reconstruction and flux balance analysis of multiple *Staphylococcus aureus* genomes identify novel antimicrobial drug targets. *J. Bacteriol.* **191**, 4015–4024 (2009).
81. Wang, H. H. *et al.* Programming cells by multiplex genome engineering and accelerated evolution. *Nature* **460**, 894–898 (2009).
This paper is major step towards the automated, large-scale generation of combinatorial genomic diversity for directed evolution of cells.
82. Ofria, C. & Wilke, C. O. Avida: a software platform for research in computational evolutionary biology. *Artif. Life* **10**, 191–229 (2004).
83. Adami, C. Digital genetics: unravelling the genetic basis of evolution. *Nature Rev. Genet.* **7**, 109–118 (2006).
84. Dekel, E. & Alon, U. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**, 588–592 (2005).
85. Li, F., Long, T., Lu, Y., Ouyang, Q. & Tang, C. The yeast cell-cycle network is robustly designed. *Proc. Natl Acad. Sci. USA* **101**, 4781–4786 (2004).
86. Whelan, K. E. & King, R. D. Using a logical model to predict the growth of yeast. *BMC Bioinformatics* **9**, 97 (2008).
87. Mahadevan, R., Edwards, J. S. & Doyle, F. J. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys. J.* **83**, 1331–1340 (2002).
88. Shlomi, T., Cabili, M. N., Herrgard, M. J., Palsson, B. O. & Ruppin, E. Network-based prediction of human tissue-specific metabolism. *Nature Biotech.* **26**, 1003–1010 (2008).
89. Kummel, A., Panke, S. & Heinemann, M. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol. Syst. Biol.* **2**, 2006.0034 (2006).
90. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452**, 840–845 (2008).
91. Wagner, A. Neutralism and selectionism: a network-based reconciliation. *Nature Rev. Genet.* **9**, 965–974 (2008).
In this paper, the author proposes a reconciliation in which neutral mutations prepare the ground for later evolutionary adaptations.
92. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5**, e1000613 (2009).
93. D'Ari, R. & Casadesus, J. Underground metabolism. *Bioessays* **20**, 181–186 (1998).
94. Khersonsky, O. & Tawfik, D. S. Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505 (2010).
95. Hoekstra, H. E. & Coyne, J. A. The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016 (2007).
96. Conant, G. C. & Wagner, A. Convergent evolution of gene circuits. *Nature Genet.* **34**, 264–266 (2003).
97. Deutschbauer, A. M. *et al.* Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. *Genetics* **169**, 1915–1925 (2005).
98. 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).

Acknowledgements
We wish to thank the anonymous reviewers and S. G. Oliver for their valuable comments on the manuscript. This work was supported by grants from the European Research Council (C.P.), the Wellcome Trust (C.P.), the 'Lendület Program' of the Hungarian Academy of Sciences (B.P.), the International Human Frontier Science Program Organization and the Hungarian Scientific Research Fund (B.P and C.P). R.N. is supported by The Netherlands Genomics Initiative (NGI – Horizon grant) and The Netherlands Organisation for Scientific Research (NWO - VENI Grant).

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Balázs Papp and Csaba Pál's homepage: www.brc.hu/sysbiol
BiGG, a database of large-scale metabolic reconstructions: http://bigg.ucsd.edu/
BRENDA, the enzyme information system: http://www.brenda-enzymes.info/
Kyoto Encyclopedia of Genes and Genomes (KEGG): http://www.genome.jp/kegg/
***Nature Reviews Genetics* series on Modelling:** http://www.nature.com/nrg/series/modelling/index.html
The openCOBRA Project, a Matlab toolbox for constraint-based modelling: http://opencobra.sourceforge.net/

ALL LINKS ARE ACTIVE IN THE ONLINE PDF