nature genetics

Adaptive evolution of bacterial metabolic networks by horizontal gene transfer

Csaba Pál[1,2], Balázs Papp[2,3] & Martin J Lercher[1,4]

Numerous studies have considered the emergence of metabolic pathways[1], but the modes of recent evolution of metabolic networks are poorly understood. Here, we integrate comparative genomics with flux balance analysis to examine (i) the contribution of different genetic mechanisms to network growth in bacteria, (ii) the selective forces driving network evolution and (iii) the integration of new nodes into the network. Most changes to the metabolic network of *Escherichia coli* in the past 100 million years are due to horizontal gene transfer, with little contribution from gene duplicates. Networks grow by acquiring genes involved in the transport and catalysis of external nutrients, driven by adaptations to changing environments. Accordingly, horizontally transferred genes are integrated at the periphery of the network, whereas central parts remain evolutionarily stable. Genes encoding physiologically coupled reactions are often transferred together, frequently in operons. Thus, bacterial metabolic networks evolve by direct uptake of peripheral reactions in response to changed environments.

Although horizontal gene transfer shapes bacterial genomes[2,3], most large-scale analyses have ignored its influence on the evolution of biological networks. Theoretical models[1] and systematic analyses[4–6] of the evolution of metabolic networks concentrate on the effects of gene duplicates. Similarly, the selective forces that influence the growth of biochemical networks are largely unknown. Here, we analyze these issues using the previously reconstructed metabolic network[7] of *Escherichia coli* K-12, composed of 904 proteins and 931 unique biochemical reactions, including coenzymes and transport processes of specified external nutrients.

In eukaryotes, gene duplicates are the main source of evolutionary novelties. Is gene duplication also the dominant genetic mechanism contributing to growth of bacterial biochemical networks? In sharp contrast to the eukaryotic yeast *Saccharomyces cerevisiae*, *E. coli* K-12 contains few duplicated enzymes in its metabolic network, almost all of which seem to be ancient (**Fig. 1**). Detailed phylogenetic analysis (**Supplementary Methods** online) indicated that only 1 of 451 investigated duplicated enzymes in *E. coli* arose since the divergence from *Salmonella* ~100 million years ago[8], despite vast differences in lifestyle and genome content between those two species[9]. Moreover, this one duplicate pair (ornithine carbamoyltransferase 1 and 2) functions in the same enzymatic reaction. Therefore, gene duplication had little effect on the topology of the *E. coli* metabolic network over the last 100 million years.

An alternative source of network growth is horizontal gene transfer. To identify transfer events, we first established the phylogeny of 51 proteobacteria species including *E. coli* K-12 and several of its close relatives, using 5 additional species to root the phylogenetic tree. The

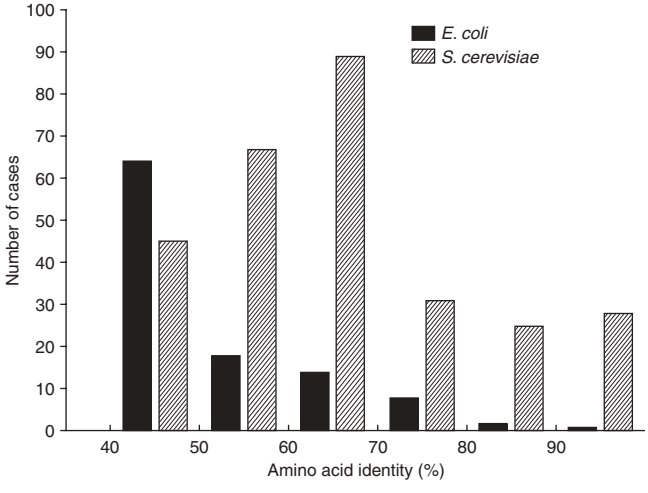

Figure 1 Comparison of duplicate genes in the metabolic networks of yeast (*S. cerevisiae*)[29] and *E. coli*. The distributions of amino acid sequence similarities differ strongly between the two species ($N = 107$ genes (*E. coli*), $N = 285$ genes (*S. cerevisiae*), $P < 0.001$ from Kolmogorov-Smirnov two-sample test). Amino acid sequence similarities between all gene pairs in each network were computed by BLAST, retaining all pairs with more than 40% amino acid similarity. The result remains after excluding remnants of genome duplicates in yeast[30] ($N = 107$ (*E. coli*), $N = 243$ (*S. cerevisiae*), $P < 0.001$ from Kolmogorov-Smirnov two-sample test).

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, D-69012 Heidelberg, Germany. [2]MTA, Theoretical Biology and Ecology Research Group, Eötvös Loránd University, Budapest H-1117, Hungary. [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. Correspondence should be addressed to M.J.L. (lercher@embl.de).

Received 6 May; accepted 8 September; published online 20 November 2005; doi:10.1038/ng1686

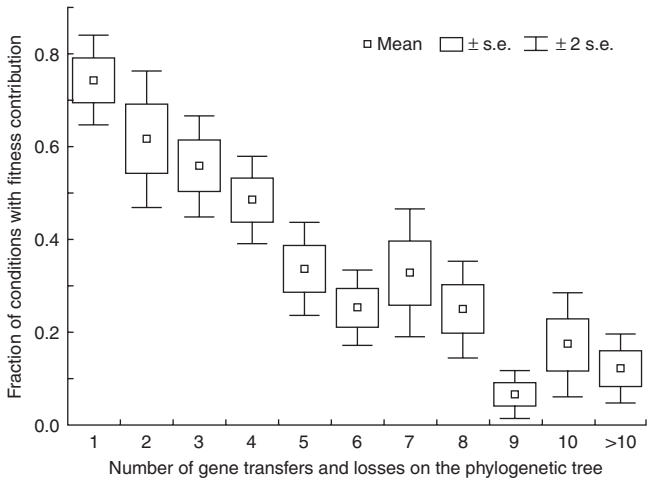

Figure 2 Environment-specificity of proteins increases with the frequency of horizontal transfer and loss events of the encoding genes ($N = 689$; ANOVA: $P < 10^{-7}$, $F = 16.32$, d.f. = 10). s.e., standard error.

maximum-likelihood tree reconstructed from 47 concatenated protein sequences was well-supported by bootstrap analyses (**Supplementary Fig. 1** online), and comparison with four independent phylogenetic studies confirmed the branching order of all previously investigated species sets (**Supplementary Methods**). Following earlier studies[10–12], we then used the presence or absence of proteins among the 51 species to identify the most parsimonious scenarios for horizontal gene transfers and gene losses across the reconstructed tree. We did this for each of 2,325 orthologous families with members in *E. coli* K-12 (numbers of inferred transfers at each node are listed in **Supplementary Table 1** online). Our results rely on the biologically reasonable assumption that gene losses are approximately twice as likely to occur as are transfer events[10,11] (gain/loss penalty ratio = 2); we obtained very similar results with other parameter settings (**Supplementary Methods**).

Consistent with expectations and earlier observations[13], a large fraction (30%) of the most recently transferred genes are annotated[14] with virus- or transposon-related functions (**Supplementary Fig. 2** online). For recently acquired genes, our results are in good agreement with those from complementary approaches[13] based on irregular GC content and use of suboptimal codons (**Supplementary Table 2** online). We found a gradual decay of both GC and codon usage irregularities with the age of the transfer event (**Supplementary Fig. 3** online), providing support for the previously hypothesized 'amelioration' of compositional biases over evolutionary time[13]. Under realistic parameter settings, we estimated that 15–32 genes were transferred horizontally into the *E. coli* metabolic network since its divergence from the *Salmonella* lineage, vastly outnumbering the one identified gene duplication over the same period.

Although gene duplication may have been an important source for network changes during the early evolution of pathways[1], the above analyses suggest that horizontal gene transfer was the dominant genetic mechanisms in the recent expansion of metabolic networks in bacteria. Which forces may be responsible for the low contribution of gene duplication to bacterial network growth? The foremost difficulty for the expansion of gene families is preserving both copies until they develop functionally distinct roles[2]. Moreover, the initial preservation of duplicated genes probably depends on the effect of enhanced gene dosage, which will be deleterious except under specific selection pressures[15]. Most gene duplicates are quickly removed from bacterial populations[16].

What are the selective pressures driving the acquisition of foreign genes? In comparisons with a systematic experimental gene knockout study[17], we found that only 7% of the genes horizontally transferred into the metabolic network of *E. coli* are essential under nutrient-rich laboratory conditions, compared with 23% of other genes ($N = 761$ genes, $\chi^2 = 26.53$, degrees of freedom (d.f.) = 1, $P < 5 \times 10^{-7}$). This observation is consistent with at least two hypotheses. First, transferred genes may provide small but evolutionarily important contributions to fitness, even under the examined routine growth conditions[18]. Alternatively, horizontal gene transfers might confer condition-specific advantages, facilitating adaptation to new environments. To assess the fitness contribution of all metabolic *E. coli* K-12 genes under different environments *in silico*, we carried out flux balance analyses of the metabolic network[19] (very similar results were obtained with minimization of metabolic adjustment analyses[20]; **Supplementary Methods** and **Supplementary Table 2**). Assuming a steady state of metabolite concentrations, we determined the flux distribution that maximized the production of a physiological combination of major biosynthetic components, the biomass, for a given set of available nutrients (**Supplementary Methods**).

Using a previously described protocol[19], we investigated systematically the effect of gene deletions on fitness in different environments, approximating fitness by the rate of biomass production. We examined 136 simulated environments, characterized by their main carbon source and the availability of oxygen, which had been shown *in silico* to support aerobic and/or anaerobic growth[21] (**Supplementary Table 3**). Those genes that contributed most to the evolution of metabolic networks (*i.e.*, that were frequently gained or lost during the

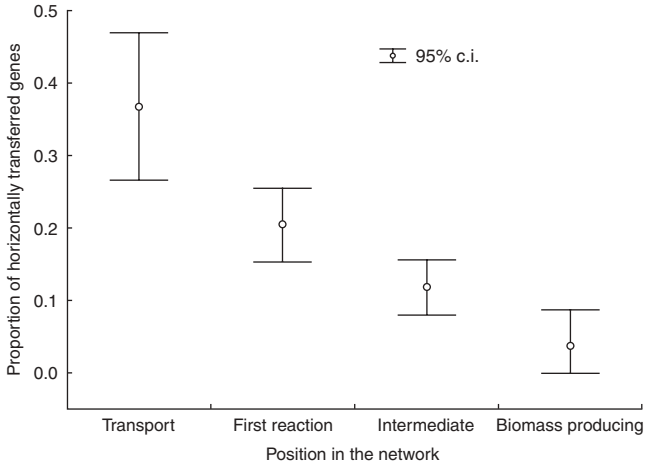

Figure 3 Proteins at the periphery of the metabolic network are much more likely to have undergone horizontal gene transfer into the *E. coli* lineage since its split from the *Vibrio* lineage. Genes are divided into the following groups: (i) transport proteins involved in nutrient uptake (87 genes); (ii) enzymes catalyzing the first reaction after uptake (240 genes); (iii) enzymes catalyzing internal reactions (271 genes); and (iv) enzymes producing major biosynthetic components (55 genes). $P < 10^{-7}$, $\chi^2 = 37.03$, d.f. = 3. Cofactors and metabolites involved in large numbers of reactions were excluded from the metabolic map, including NAD^+, NADH, NADPH, $NADP^+$, H^+, ATP, ADP, orthophosphate, CO_2, pyrophosphate, FAD, $FADH_2$ and H_2O. Genes with ambiguous network positions were excluded from the analysis. Results remain for gain/loss penalty = 1 ($\chi^2 = 65.39$, d.f. = 3, $P < 10^{-13}$), as well as when transfer events across the whole phylogenetic tree are considered (data not shown). c.i., confidence interval.

Table 1 Physiologically coupled enzyme pairs are frequently transferred or lost together

Interaction	Event	Individual events	Fraction of co-events	OR (95% c.i.)
Fully coupled	Transfer	59	37%	64.6 (24.2–168.8)
Fully coupled	Loss	1,624	53%	50.0 (41.8–59.6)
Directionally coupled	Transfer	78	30%	60.3 (24.3–147.2)
Directionally coupled	Loss	2,833	21%	9.6 (8.3–11.1)

'Individual events' is the total number of individual gene gains (or losses) investigated. 'Fraction of co-events' is the fraction of the total gains (or losses) of genes involved in physiologically coupled pairs that occur together with their coupled partner. Only branches originating from an ancestral node in which both genes are absent (or present) were considered in the analysis of 'co-gains' (or 'co-losses'). Odds ratios (ORs) quantify how much more likely gain (or loss) of a gene is when its coupled partner gene is gained (or lost) along the same phylogenetic branch; all odds ratios are highly significant (Fisher's exact test, $P < 10^{-12}$). Similar results were obtained with different model settings (**Supplementary Table 2**). c.i., confidence interval.

evolution of proteobacteria) were generally environment-specific, whereas those genes that were invariant among proteobacteria contributed to fitness in most environments (**Fig. 2**). Previous analyses showed quantitative agreement between the predictions of the flux balance model and experimentally measured nutrient uptake, enzymatic fluxes and the effects of gene deletions under several conditions in *E. coli*[19], but detailed predictions under several other conditions remain to be validated. The above result remained valid when the analysis was restricted to conditions where growth of *E. coli* K-12 was experimentally shown; it was also robust to changes in the method to identify gene transfers and losses and in the optimization protocols to calculate the impact of gene deletions[19,20] (**Supplementary Table 2**). The environment-specificity of horizontally transferred genes might explain why most of them are not translated into proteins under laboratory conditions[22]. We conclude that the evolution of the network is largely driven by adaptation to new environments and not by optimization in fixed environments.

Having established the genetic mechanisms and the selective forces that govern network evolution, we next turned to the topological effect of horizontal gene transfer on the network. The above results suggest that addition and deletion of reactions might be concentrated on those network parts that interact with the environment. The number of independent horizontal transfer events was highly variable across different enzymatic pathways (**Supplementary Table 2**), and genes in the central pathways of the network (*e.g.*, glycolysis) had undergone few transfer events across the tree. To analyze further the relationship between network position and gene transfers, we classified proteins according to their involvement in nutrient uptake, first reactions after uptake, intermediate steps of metabolism and production of major biosynthetic components. As predicted, proteins contributing to peripheral reactions (nutrient uptake and first metabolic step) were more likely to be transferred, whereas enzymes catalyzing central reactions (intermediate steps and biomass production) were largely invariant across species (**Fig. 3**).

Are genes added or lost from metabolic networks one at a time, or does network evolution proceed by steps involving whole sets of genes simultaneously? Modules of physiologically coupled genes might be the best candidates for simultaneous acquisition or loss during evolution. We identified physiologically coupled enzyme pairs by flux-coupling analysis[23]. Two special cases were considered: fully and directionally coupled enzyme pairs. In fully coupled enzyme pairs, the flux catalyzed by one protein is always the same as that catalyzed by the other except for a constant factor, as in linear pathways. Fully coupled enzymatic pairs provide a very rigorous and stringent definition of biochemical modules, as only together can such pairs fulfill their metabolic function. Directional coupling indicates that removal of one enzyme shuts down flux through the other but not vice versa. As predicted, both fully and directionally coupled enzymes were much more often gained or lost together on the same branch of the proteobacterial phylogenetic tree than would be expected by chance (**Table 1**). This suggests that physiological modules tend to be conserved during evolution, contrary to previous results based on more loosely defined modules[24].

Moreover, 30% of the fully coupled pairs are encoded in the same operon in *E. coli*[25], a fraction much higher than would be expected for randomly chosen pairs (0.5%). The fraction of pairs sharing the same operon rises to at least 75% when considering only fully coupled pairs that were gained together during evolution leading to *E. coli*. These latter results confirm that the gains of physiologically fully coupled pairs together most likely occurred in one step, the uptake of at least part of an operon.

Future studies will aim to characterize the molecular details of the evolutionary network dynamics, for example, by analyzing how the enzymatic composition of the network affects its ability to adapt to

Table 2 Some operons containing horizontally transferred genes with physiologically coupled, environment-specific functions

Operon name	Predicted nutrient for which operon is required	Literature information	Genes in operon and physiological coupling
atoDAE	Acetoacetate, butyrate	Short fatty acids	**atoE** ↔ **atoD/atoA***
codBA	NA	Cytosine	**codB** → codA
cynTSX	NA	Cyanate	**cynX** ↔ **cynS** → cynT
fucPIKUR	Fucose	Fucose	**fucP** ↔ **fucI** ↔ **fucK***
melAB	Melibiose	Melibiose	**melB** ↔ **melA**
mtlADR	Mannitol	Mannitol	**mtlA** ↔ **mtlD***
AraBAD	Arabinose	Arabinose	**araA** ↔ **araB** ↔ araD

The operons listed are required for the uptake or catalysis (mostly the first or second step after uptake) of specific nutrients. Members of these operons are physiologically coupled. Nutrient conditions where the operons are required are derived from the model and from literature compiled from RegulonDB and Ecocyc[14]. Genes in the operons are listed as ordered on the metabolic map. Genes involved in transport processes are underlined. Genes horizontally transferred to *E. coli* are depicted in bold. Physiological coupling between genes is denoted by arrows (↔, fully coupled; →, directionally coupled). Genes that are not part of the metabolic network are excluded from the analysis. Unless otherwise indicated (*), evidence for horizontal transfer is consistent under all investigated parameter settings. NA, not analyzed.

new environments. As a first step, we examined whether the gradual evolution of metabolic pathways can be understood by analyzing the details of physiological coupling between enzymes. For example, one might expect that an enzyme whose function depends on the presence of another enzyme would have been acquired by the *E. coli* genome more recently than its partner. This is indeed observed in 70% of those directionally coupled gene pairs in which both genes were acquired on different branches leading to *E. coli* ($N = 386$, sign test $P < 10^{-16}$). Future studies will also have to examine how the number of physiological interactions influences the probability of successful gene transfer[26]. Furthermore, given that the physiological adaptation to new environments is accompanied by major flux reorganizations along the high-flux backbone of the metabolic network[27], the role of horizontally transferred genes in these reorganizations needs to be examined.

In summary, metabolic networks in bacteria evolve in response to changing environments, not only by changes in enzyme kinetics through point mutations, but also by the uptake of peripheral genes and operons through horizontal gene transfers (a list of examples is given in **Table 2**). Our results indicate that systems biology cannot stop at the boundaries of the metabolic network: to understand network evolution, we need to extend our analysis to the environment, both inanimate (providing nutrients) and animate (providing genetic material).

METHODS

Gene gains and losses. Based on gene presence and absence obtained from STRING[28], we reconstructed the most parsimonious scenarios[10–12] for gene loss and horizontal transfer events (gene gains) on the rooted phylogeny using generalized parsimony as implemented in PAUP* (**Supplementary Methods**). All results were obtained using relative penalties for horizontal gene transfer and deletions of 2:1 (gain/loss penalty = 2)[10,11]; different settings gave similar results (**Supplementary Table 2**).

To analyze CDgains, we started with nodes of the phylogenetic tree in which the two investigated enzymes (*e.g.*, A and B) were absent. We then constructed a contingency table by counting the occurrence of the four possible evolutionary scenarios (gain of A, gain of B, gain of A and B, and no gain) along all branches starting from these nodes. The odds ratio quantifies how much more likely the gain of a gene A is if its physiologically coupled partner gene B is gained along the same phylogenetic branch. We used an analogous procedure for loss events, analyzing all branches starting from nodes in which both A and B were present. Gene families with more than one member in *E. coli* K-12 were excluded from the analysis.

Metabolic network analysis. We examined the reconstructed metabolic network (*i*JR904 GSM/GPR) of *E. coli* K-12. We followed previously established protocols[19,20] to investigate the effect of gene deletions under 136 environmental conditions. Flux balance analysis involves two fundamental steps: (i) specification of mass balance constraints around intracellular metabolites and (ii) maximization of the production of biomass components (the list of environments and biomass components is given in **Supplementary Table 3** online). Physiologically coupled reactions and blocked reactions were identified as described previously[23]. We found 772 fully coupled reaction pairs and 1,542 directionally coupled reaction pairs.

More methodological details (including ortholog identification, inference of phylogenetic genome tree and age estimation of gene duplicates) are given in **Supplementary Methods**.

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

ACKNOWLEDGMENTS

We thank P. Bork, L. Hurst and S. McWeeney for suggestions on previous versions of the manuscript; C. von Mering for discussions and providing early access to the updated STRING database; and E. Nikolaev for discussions on flux coupling analysis. C.P. and B.P. are supported by the Hungarian Scientific Research Fund. B.P. is a Fellow of the Human Frontier Science Program. C.P. acknowledges support by an EMBO Long-Term Fellowship. M.J.L. acknowledges support by the Deutsche Forschungsgemeinschaft.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Schmidt, S., Sunyaev, S., Bork, P. & Dandekar, T. Metabolites: a helping hand for pathway evolution? *Trends Biochem. Sci.* **28**, 336–341 (2003).
2. Lawrence, J.G. & Hendrickson, H. Lateral gene transfer: when will adolescence end? *Mol. Microbiol.* **50**, 739–749 (2003).
3. Lerat, E., Daubin, V., Ochman, H. & Moran, N.A. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol.* **3**, e130 (2005).
4. Teichmann, S.A. *et al.* The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *J. Mol. Biol.* **311**, 693–708 (2001).
5. Rison, S.C., Teichmann, S.A. & Thornton, J.M. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J. Mol. Biol.* **318**, 911–932 (2002).
6. Alves, R., Chaleil, R.A. & Sternberg, M.J. Evolution of enzymes in metabolism: a network perspective. *J. Mol. Biol.* **320**, 751–770 (2002).
7. 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).
8. Lawrence, J.G., Hartl, D.L. & Ochman, H. Molecular considerations in the evolution of bacterial genes. *J. Mol. Evol.* **33**, 241–250 (1991).
9. Ochman, H. & Groisman, E.A. The origin and evolution of species differences in *Escherichia coli* and *Salmonella typhimurium*. *EXS* **69**, *479–493 (1994).*
10. Snel, B., Bork, P. & Huynen, M.A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* **12**, 17–25 (2002).
11. Mirkin, B.G., Fenner, T.I., Galperin, M.Y. & Koonin, E.V. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* **3**, 2 (2003).
12. Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.A. & Andersson, S.G. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc. Natl. Acad. Sci. USA* **101**, 9722–9727 (2004).
13. Lawrence, J.G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl. Acad. Sci. USA* **95**, 9413–9417 (1998).
14. Keseler, I.M. *et al.* EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.* **33**, D334–D337 (2005).
15. 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).
16. Hooper, S.D. & Berg, O.G. On the nature of gene innovation: duplication patterns in microbial genomes. *Mol. Biol. Evol.* **20**, 945–954 (2003).
17. Gerdes, S.Y. *et al.* Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684 (2003).
18. Thatcher, J.W., Shaw, J.M. & Dickinson, W.J. Marginal fitness contributions of nonessential genes in yeast. *Proc. Natl. Acad. Sci. USA* **95**, 253–257 (1998).
19. 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).
20. Segre, D., Vitkup, D. & Church, G.M. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. USA* **99**, 15112–15117 (2002).
21. Reed, J.L. & Palsson, B.O. Genome-scale *in silico* models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797–1805 (2004).
22. Taoka, M. *et al.* Only a small subset of the horizontally transferred chromosomal genes in *Escherichia coli* are translated into proteins. *Mol. Cell. Proteomics* **3**, 780–787 (2004).
23. 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).
24. Snel, B. & Huynen, M.A. Quantifying modularity in the evolution of biomolecular systems. *Genome Res.* **14**, 391–397 (2004).
25. Salgado, H. *et al.* RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32**, D303–D306 (2004).
26. Jain, R., Rivera, M.C. & Lake, J.A. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc. Natl. Acad. Sci. USA* **96**, 3801–3806 (1999).
27. Almaas, E., Kovacs, B., Vicsek, T., Oltvai, Z.N. & Barabasi, A.L. Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature* **427**, 839–843 (2004).
28. von Mering, C. *et al.* STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* **33**, D433–D437 (2005).
29. Forster, J., Famili, I., Fu, P., Palsson, B.O. & Nielsen, J. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253 (2003).
30. Kellis, M., Birren, B.W. & Lander, E.S. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature* **428**, 617–624 (2004).