

Chance and necessity in the evolution of minimal metabolic networks

Csaba Pál^{1,2*}, Balázs Papp^{3*}, Martin J. Lercher^{1,4}, Péter Csermely⁵, Stephen G. Oliver³ & Laurence D. Hurst⁴

It is possible to infer aspects of an organism's lifestyle from its gene content¹. Can the reverse also be done? Here we consider this issue by modelling evolution of the reduced genomes of endosymbiotic bacteria. The diversity of gene content in these bacteria may reflect both variation in selective forces and contingency-dependent loss of alternative pathways. Using an *in silico* representation of the metabolic network of *Escherichia coli*, we examine the role of contingency by repeatedly simulating the successive loss of genes while controlling for the environment. The minimal networks that result are variable in both gene content and number. Partially different metabolisms can thus evolve owing to contingency alone. The simulation outcomes do preserve a core metabolism, however, which is over-represented in strict intracellular bacteria. Moreover, differences between minimal networks based on lifestyle are predictable: by simulating their respective environmental conditions, we can model evolution of the gene content in *Buchnera aphidicola* and *Wigglesworthia glossinidia* with over 80% accuracy. We conclude that, at least for the particular cases considered here, gene content of an organism can be predicted with knowledge of its distant ancestors and its current lifestyle.

Naturally evolved, nearly minimal gene sets in closely related intracellular symbionts contain substantial differences². The diversity of these evolved minimal gene sets may be the product of three fundamental processes: differences in initial genetic makeup; variation in selective forces within host cells; and differences in the order of gene deletions, resulting in a choice between alternative cellular pathways². By modelling the reductive evolution of a detailed metabolic network, we first explore the evolutionary significance of the last of these alternatives.

Using the metabolic network of *Escherichia coli* K12 (ref. 3) as our model system has several advantages. First, the best evidence for the presence of alternative pathways within and across species comes from studies of metabolic networks⁴. Second, flux balance analysis provides a rigorous modelling framework for studying the impact of gene deletions^{4,5}; the method relies on optimizing the steady-state use of the metabolic network to produce biomass components. Third, not only is the metabolic network of *E. coli* K12 one of the best studied cellular subsystems, but this organism is also a close relative of several endosymbiotic organisms⁶, including *Buchnera aphidicola* and *Wigglesworthia glossinidia*. Cellular domestication has resulted in the elimination of 70–75% of the ancestral genome in these latter organisms⁷.

The previously reconstructed metabolic network of *E. coli*³ consists of 904 genes and 931 unique biochemical reactions, and incorporates external nutrients and the corresponding transport processes. The composition of a 'minimal reaction set' has been previously shown to

depend strongly on the given environmental conditions⁸. Gradual evolution towards minimal genomes and the role of chance in this process, however, have remained unexplored. The smallest sets of genes that are compatible with cellular life will relate to the most favourable conditions, in which most nutrients are available from the environment. This situation is approximated by organisms with a strict intracellular lifestyle, where the host provides most of their nutrients². Accordingly, we first characterized the simulated evolution of the network under nutrient-rich conditions (Supplementary Tables 1–3).

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

The resulting networks share on average 77% of their reactions, whereas only 25% would be shared by randomly deleting the same number of genes (Fig. 1a). This suggests that both selective constraints and historical contingencies influence the reductive evolution of metabolic networks. Owing to alternative metabolic pathways in the original *E. coli* network, numerous functionally equivalent minimal networks are possible, even under identical selective conditions. For the same reason, only 55% of the reactions are recoverable by single-gene deletion studies (Fig. 1b). The number of genes in the minimal networks is also variable (Fig. 1b), suggesting that there are differences in the number of enzymatic steps between alternative pathways. Deletions at the early stages of genome reduction may affect large genomic regions rather than single genes⁹. However, additional simulations showed that, although allowing such block deletions reduces the number of independent gene-loss events, it has no effect on the size and average similarity of the networks evolved (Supplementary Methods and Supplementary Table 4).

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

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

*These authors contributed equally to this work.

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

To investigate further how accurately the model describes reductive evolution in nature, we focused our simulations on three fully sequenced genomes of *B. aphidicola* strains^{10–12} and *W. glossinidia*¹³. These are close relatives of *E. coli* with an evolved intracellular

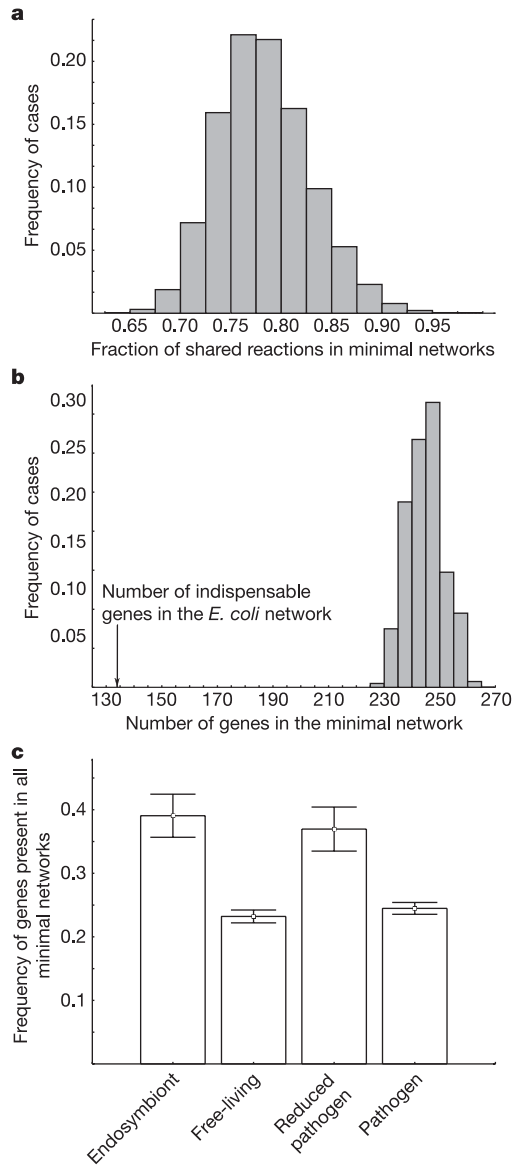


Figure 1 | General properties of evolved minimal networks. **a**, Distribution of the fraction of shared metabolic reactions between all possible pairs among 500 simulated minimal networks. Only reactions with annotated enzyme-encoding genes are shown. The resulting networks share $77 \pm 4.4\%$ (mean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

endosymbiotic lifestyle. Gene acquisition must have been a negligible factor in the evolution of these lineages (Supplementary Methods), providing a unique opportunity to study reductive evolution. Setting boundary conditions that mimic the relevant nutrient conditions and selective forces (Supplementary Tables 2 and 3), we performed simulations as described above.

Detailed physiological studies have shown that *Buchnera* supply their aphid hosts with riboflavin¹⁴ and essential amino acids¹⁵ that are lacking in their hosts' diets. To quantify the agreement between our predictions and the observed reductive evolution in *Buchnera*, while considering gene-content variation in simulated minimal genomes, we used a combined measure of sensitivity and specificity¹⁶. For each possible cutoff (that is, the minimal fraction of simulated genomes in which a gene must be present to predict its presence in *Buchnera*), Fig. 2a shows the fraction of true-positive predictions (sensitivity) plotted against the fraction of false-positive predictions ($1 - \text{specificity}$). The area under the resulting curve gives a cutoff-independent measure of predictive accuracy¹⁶. For each of the *Buchnera* strains, the accuracy of the model is $\sim 80\%$ as compared with the 50% expected by chance (Fig. 2a). The above results remain valid when genes putatively transferred horizontally into *E. coli* since its split

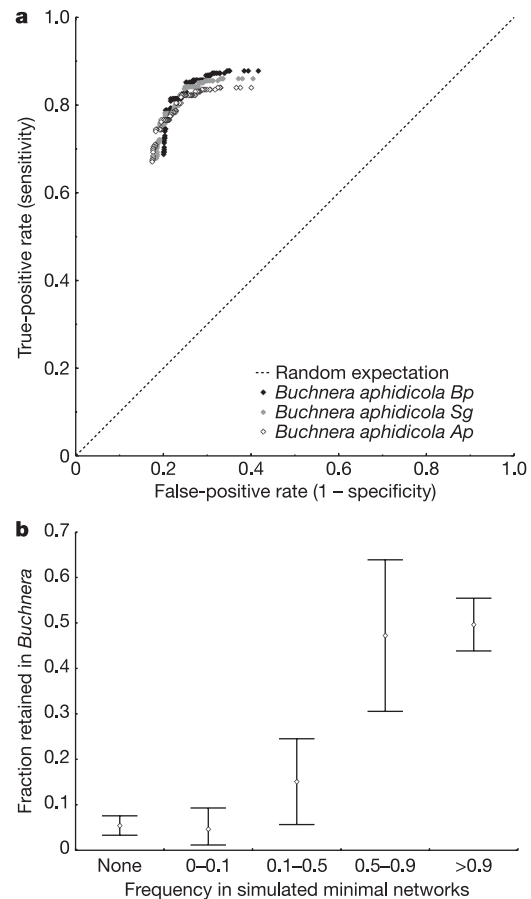


Figure 2 | Comparison of reaction content of simulated and *Buchnera* metabolic networks. **a**, Predictive accuracy for all possible cutoffs (receiver operating characteristic curve)¹⁶. *Bp*: *B. aphidicola*, endosymbiont of *Baizongia pistaciae*; *Sg*: *B. aphidicola*, endosymbiont of *Schizaphis graminum*;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

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

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

The above analysis relied on detailed knowledge of the lifestyle of *Buchnera*. Is it possible to predict gene content of an organism with much less information on lifestyle? *Wigglesworthia*, another endosymbiont and close relative of *E. coli*, is an obvious choice. *Wigglesworthia* provides some cofactors and vitamins for its host, the tsetse fly¹⁸. On the basis of the available physiological information¹⁹, it is possible to model the evolution of the metabolic network of this organism with nearly 76% accuracy for the reaction content (Supplementary Fig. 2 and Table 3). It is likely that the available experiments underestimate the number of cofactors produced by the endosymbiont. We thus elaborated a systematic protocol to find the most likely set of cofactors synthesized by *Wigglesworthia* (Supplementary Methods). Based on the idea of greedy algorithms²⁰, the protocol iteratively adds biosynthetic components that must be produced for the host and calculates the impact on the accuracy of predicting the real reaction content of *Wigglesworthia*. In each round, the cofactor resulting in the best prediction is kept and a new round of simulations is started, adding again each of the remaining compounds one at a time (Supplementary Methods). The method substantially increases model accuracy up to 84% (Supplementary Table 5). It also results in a series of non-trivial predictions on the metabolic capability of *Wigglesworthia*. For example, it suggests that this organism retained the ability to synthesize not only protohaem, but also another related cofactor, haem O (Supplementary Methods).

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

Each loss of a reaction reduces the space available for further reductive evolution. This is most obvious for physiologically fully coupled reactions (such as those in linear pathways), which can only fulfil their metabolic function together²¹. As predicted, members of pairs are either lost or retained together in the investigated endosymbionts in 74–84% of cases, whereas only ~50–55% would be expected by chance (Supplementary Table 10).

Deviations between the model predictions and gene content of endosymbionts might be due to incomplete biochemical knowledge or inaccuracies in modelling the types and relative amounts of nutrient conditions and biosynthetic components required by the endosymbiont or the host cell. Finally, hosts and endosymbionts

interact in ways that are not completely understood, and biomass production may be only a rough proxy for endosymbiont fitness. These caveats aside, our approach might be considered a step towards a predictive theory of gene-content evolution. Complementary to traditional approaches, in which lifestyle is inferred from genomic data, it seems possible to take an organism's ecology and to predict which genes it should have by *in silico* network analysis. Moreover, we find that evolutionary paths are contingent on prior gene deletion events, resulting in networks that generally do not represent the most economical solution in terms of the number of genes retained. Thus, history and chance seem to have significant roles not only in adaptive²² but also in reductive evolution of genomes.

These results also have implications for the search for a minimal genome. By using comparative genomics^{23,24} and systematic gene knock-out studies^{25–27}, traditional analyses of minimal gene sets aim to define a repertoire of genes that is necessary and sufficient to support cellular life². The theoretical foundations of the minimal genome concept have remained, however, largely unexplored. We have established that the catalogue of essential genes in free-living species identified by single-gene deletion studies will underestimate the minimal gene set for metabolic system by about 45% (Fig. 1b). Such considerations, and the simulation techniques used to reach these conclusions, should inform attempts by experimentalists to construct minimal genomes by gradual evolution in the laboratory^{28,29}.

METHODS

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

Flux balance analysis of the *E. coli* network. A reconstructed metabolic network (*iJR904* GSM/GPR)³ of *E. coli* K12 was used in this study. The model consists of 931 unique biochemical reactions (including transport processes) and 904 genes. The metabolic reconstruction gives accurate information on the stoichiometry and direction of enzymatic reactions, on the presence of isoenzymes, and on enzymatic complexes. Details of flux balance analysis of the *E. coli* metabolic network have been described elsewhere^{4,5}. In brief, it involves two fundamental steps: first, specification of mass balance constraints around intracellular metabolites; and second, maximization of the production of biomass components. The assumption of a steady state of metabolite concentrations specifies a series of linear equations of individual reaction fluxes, which is written in the form $Sv = 0$, where S is the mn stoichiometric matrix (m being the number of metabolites and n being the number of reactions) and v is the vector of individual fluxes through the network. An individual element S_{ij} gives the contribution of the j -th reaction to metabolite i . A biomass reaction describes the relative contribution of metabolites to the cellular biomass. Availability of nutrients and directions of individual reactions were included as boundary conditions (Supplementary Tables 1–3). Using the linear programming package CPLEX 9.0.0, we identified the flux distribution that maximizes the rate of biomass production.

Simulations on reductive evolution. Following previously elaborated protocols⁵, we start by investigating the behaviour of the *E. coli* metabolic network model under a given environmental condition (Supplementary Tables 1–3). Next, we remove a randomly chosen enzyme from the network and calculate the impact of this deletion on the production of biomass components (for a list, see Supplementary Tables 1–3). Enzyme deletions were simulated by constraining the flux of the corresponding reactions to zero and calculating the corresponding knockout flux configuration by established protocols^{4,5}. A gene was classified as having no fitness effect if the biomass production rate of the knockout strain was reduced by less than a given cutoff; different cutoffs led to very similar results (Supplementary Table 5). Deletions of isoenzymes were considered to have no impact on fitness as long as at least one member remained. By contrast, deletion of any of the subunits of a protein complex was considered to result in zero flux through the corresponding reactions. Reactions with no annotated encoding genes were retained throughout the simulations. If the fitness effect of a simulated gene deletion was below the cutoff, the deletion was assumed to be viable and the enzyme was considered to be permanently lost. Otherwise, the gene was restored to the network. The procedure was repeated until no further enzymes could be deleted. This simulation was repeated 500 times; each run provided an independent evolutionary outcome.

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

must be produced for the host (Supplementary Table 2). Besides amino acids, mononucleotides and fatty acids, among others, the biomass components that must be synthesized also include riboflavin. A previous study³⁰ estimated the population size of *Buchnera* as $N_e \approx 10^2$ – 10^3 . Gene deletions are effectively neutral and can thus spread through a population if $|N_e s| < 1$, where s is the selective effect of the gene deletion. Accordingly, the cutoff for the fitness effect of simulated gene deletions was set to 10^{-2} . A less stringent cutoff (0.1) gave very similar results (Supplementary Table 6). For details of *Wigglesworthia* uptake and selective conditions, see Supplementary Table 3.

Received 7 November; accepted 27 December 2005.

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

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

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

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