LETTER

doi:10.1038/nature13827

Long-term phenotypic evolution of bacteria

Germán Plata[1,2], Christopher S. Henry[3] & Dennis Vitkup[1,4]

For many decades comparative analyses of protein sequences and structures have been used to investigate fundamental principles of molecular evolution[1,2]. In contrast, relatively little is known about the long-term evolution of species' phenotypic and genetic properties. This represents an important gap in our understanding of evolution, as exactly these proprieties play key roles in natural selection and adaptation to diverse environments. Here we perform a comparative analysis of bacterial growth and gene deletion phenotypes using hundreds of genome-scale metabolic models. Overall, bacterial phenotypic evolution can be described by a two-stage process with a rapid initial phenotypic diversification followed by a slow long-term exponential divergence. The observed average divergence trend, with approximately similar fractions of phenotypic properties changing per unit time, continues for billions of years. We experimentally confirm the predicted divergence trend using the phenotypic profiles of 40 diverse bacterial species across more than 60 growth conditions. Our analysis suggests that, at long evolutionary distances, gene essentiality is significantly more conserved than the ability to utilize different nutrients, while synthetic lethality is significantly less conserved. We also find that although a rapid phenotypic evolution is sometimes observed within the same species, a transition from high to low phenotypic similarity occurs primarily at the genus level.

Analyses of phenotypic evolution, such as the morphological variation of beaks in Darwin's finches[3], provided the original impetus and context for understanding natural selection. Because the evolutionary importance and physiological role of specific phenotypic traits change over time, it is often difficult to connect genotype to phenotype to fitness across long evolutionary distances, especially for metazoan organisms. For microbial species, on the other hand, the ability to metabolize different nutrient sources, although clearly not the only important phenotype, always remains an essential determinant of their fitness and lifestyle. Even though a large-scale comparative analysis of microbial phenotypes—such as growth on different nutrients or the impact of genetic perturbations—is currently challenging owing to a relative paucity of experimental data, we rationalized that thoroughly validated computational methods can be used to investigate the phenotypic evolution of diverse bacterial species. Flux balance analysis (FBA)[4], in particular, has been previously used to accurately predict gene and nutrient essentiality, growth yields, and evolutionary adaptations to environmental and genetic perturbations[5]. Notably, the accuracy of FBA methods has

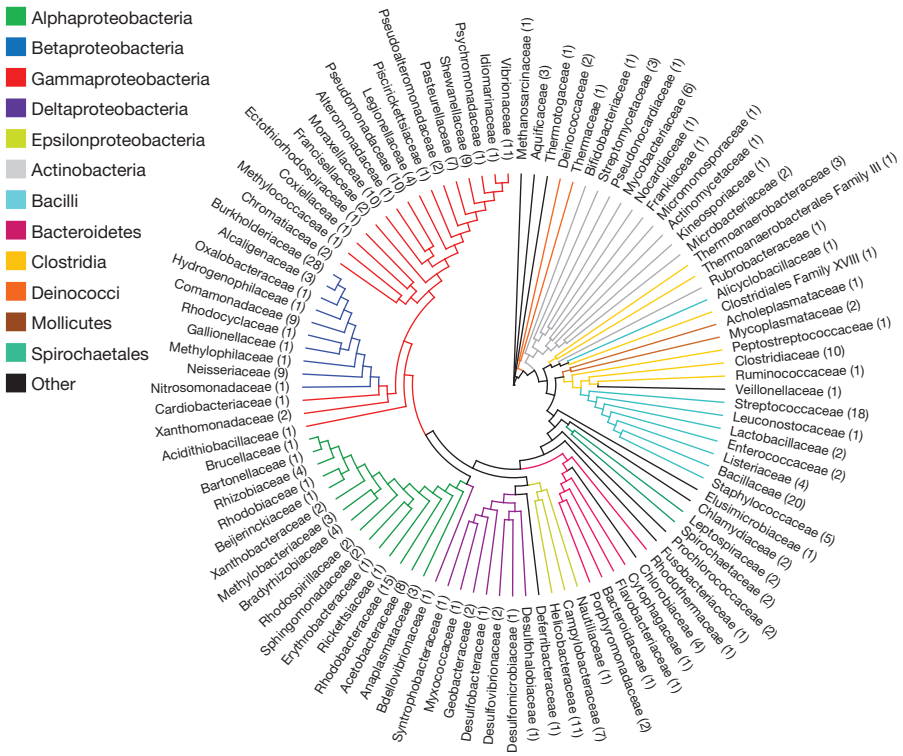


Figure 1 | Diversity of considered bacterial families. The cladogram shows the evolutionary relationship between the 100 bacterial families that include the 322 species considered in our study. The tree is based on the average 16S ribosomal RNA (rRNA) genetic distances between species in each family (see Methods). The numbers of considered species in each family are shown in parentheses. Different colours represent different bacterial classes. The tree was rooted using the *Methanosarcina barkeri* rRNA sequence.

[1]Department of Systems Biology, Center for Computational Biology and Bioinformatics, Columbia University, New York, New York 10032, USA. [2]Integrated Program in Cellular, Molecular, Structural and Genetic Studies, Columbia University, New York, New York 10032, USA. [3]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois 60439, USA. [4]Department of Biomedical Informatics, Columbia University, New York, New York 10032, USA.

©2014 Macmillan Publishers Limited. All rights reserved

been independently demonstrated for many dozens of species encompassing diverse phylogenetic distributions and growth environments[6]. We selected for our analysis more than 300 phylogenetically diverse bacteria (Fig. 1) for which genome-scale metabolic models were reconstructed using a recently developed protocol[7] (see Methods).

To investigate the long-term evolution of growth phenotypes, we considered 62 carbon sources that are commonly used by microbial species for growth and energy production[8]. For each considered species we used FBA to determine a subset of the compounds that could be used for biomass synthesis or ATP generation—two key metabolic objectives of bacterial growth[9]. This analysis resulted in binary phenotypic profiles that describe the ability of each microbial species to use each of the considered compounds (see Methods). The evolution of these phenotypic profiles—that is, the change in phenotypic similarity as a function of species divergence (genetic distance)—is shown in Fig. 2a, (see also Extended Data Fig. 1). Notably, this analysis demonstrated that the average long-term evolution of growth phenotypes can be approximated well by an exponential decay (see Methods and Extended Data Table 1). A three-parameter exponential model fits the data in Fig. 2a significantly better than simpler alternative models (Extended Data Table 2). Similar divergence trends were observed for larger sets of carbon source compounds (Extended Data Fig. 2), and for compounds that could be used as a source of nitrogen (Fig. 2b). The observed trend was also robust towards subsampling or removal of specific species and families used in the analysis (Extended Data Fig. 3).

The observed exponential trends suggest that as microbial species diverge over planetary timescales and adapt to different environmental niches, approximately similar fractions of phenotypic properties change per unit time. For species separated by more than 1 billion years of evolution (~0.2 genetic distance in Fig. 2), the divergence of growth phenotypes approaches saturation around a similarity of 21% (Fig. 2a), which is higher than the value expected by chance (~12%) given the average number of carbon compounds used by the models. This difference is likely due to a widespread utilization of common nutrient sources across bacterial species (see Extended Data Table 3)[10].

Notably, before the evolution of growth phenotypes settles into the aforementioned average trend, a much higher rate of phenotypic evolution is observed for pairs of bacteria at very close genetic distances (<0.01, or ~50 million years[11]). Our computational analysis predicts ~71% phenotypic similarity for closely related bacteria (Fig. 2, Extended Data Table 1), which agrees well with available experimental data on intra-species phenotypic similarity: for example, 75% for the utilization of carbon sources in *Escherichia coli*[12] and 69% for *Campylobacter jejuni*[13]. The diversity of bacteria observed at close distances reflects a well-documented genetic and phenotypic variability within bacterial pan-genomes[14]. The observed patterns also suggest that phenotypic evolution proceeds in two different stages, namely through fast phenotypic diversification of closely related strains followed by a slower exponential divergence lasting billions of years. Notably, patterns of multi-stage and hierarchical evolution have been observed in other systems, for example in bacterial and eukaryotic developmental networks[15,16].

To validate experimentally the predicted patterns of long-term phenotypic evolution of bacteria, we obtained GENIII Biolog Phenotype Microarrays[8] data for 40 diverse microbial species (Extended Data Fig. 4 and Supplementary Data 1). Phenotype Microarrays data were used to determine the ability of each considered bacteria to utilize the 62 different carbon sources used in the simulations (Fig. 2c; see Methods). In agreement with previous results[7], FBA predicted microbial growth phenotypes

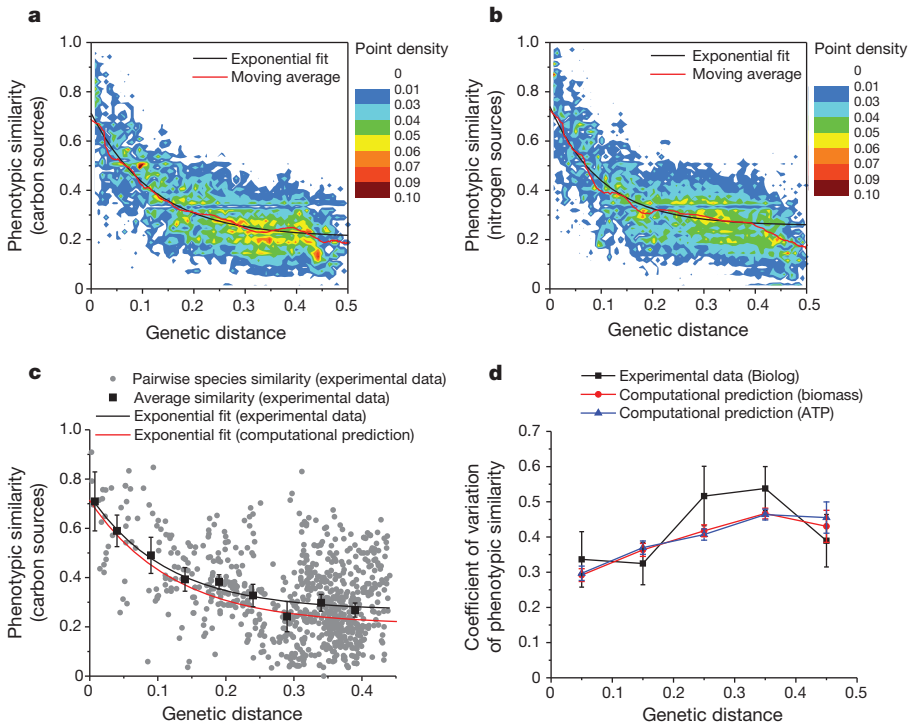


Figure 2 | Evolution of bacterial metabolic growth phenotypes. Genetic distances in the figure are based on bacterial 16S rRNA sequences. **a**, The evolution of phenotypic similarity in the usage of carbon sources for biomass synthesis. The colours represent the point density at a given genetic distance for all pairwise comparisons between metabolic models ($n = 26{,}106$). The black line shows a three-parameter exponential fit to the computational predictions; the red line shows a moving average of the predictions. **b**, Like **a**, but phenotypic similarity in the usage of nitrogen sources for biomass production across metabolic models ($n = 36{,}856$). **c**, Experimental analysis of the long-term phenotypic divergence trend. Grey points represent pairwise comparisons of carbon usage phenotypes (Biolog data) between 40 bacterial species ($n = 780$). The black squares represent the average values of experimental phenotypic similarity at different divergence distances. The black line represents an exponential fit to the experimental phenotypic similarity data; the red line represents an exponential fit to the computationally predicted phenotypic similarity data for biomass synthesis. **d**, The variability of experimental and computationally predicted phenotypic similarity at different divergence distances. The variability was quantified by the coefficient of variation, defined as the ratio of the standard deviation to the mean. Error bars in **c** and **d** represent s.e.m. obtained on the basis of 10,000 bootstrap re-samplings of the considered species.

©2014 Macmillan Publishers Limited. All rights reserved

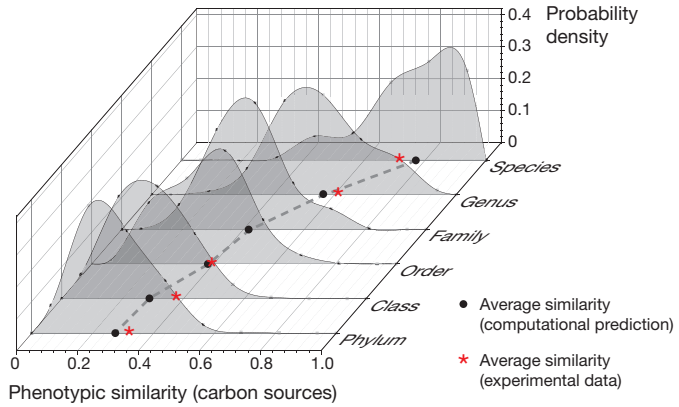


Figure 3 | Distribution of phenotypic similarity at different levels of bacterial taxonomic classification. The distributions of phenotypic similarity in the usage of carbon sources for biomass synthesis were obtained based on computational simulations of metabolic models ($n = 26{,}106$). The dashed line connects the average values (black dots) of computational predictions at each taxonomic level. The red asterisks in the figure indicate the average values of experimental data obtained using Biolog arrays (see Methods).

with an average accuracy of ~70%. Importantly, the experimental results agree well with the computationally predicted average trend describing phenotypic bacterial divergence (Fig. 2c). Also, for the experimental data, as well as for the computational predictions, the three-parameter exponential model fitted the data significantly better than simpler alternative models (Extended Data Table 2). The comparison of computational and experimental values for the coefficient of variation of phenotypic similarity shows that computational predictions capture well not only the average trend but also the variability of phenotypic similarity for bacteria at different genetic distances (Fig. 2d). Overall, the analyses of experimental data suggest that although individual models need to be further validated and improved, high-throughput metabolic reconstructions can be used for comparative functional studies across a large number of diverged species.

We next investigated the diversity of metabolic growth phenotypes at different levels of conventional taxonomic classification (Fig. 3). Although bacteria from the same species show mostly similar phenotypic properties, the long left tail of the top distribution in Fig. 3 suggests that some organisms have substantial phenotypic differences even at this basic taxonomic level. At the genus level the distribution is very broad, with an average similarity of ~60%; this suggests that transitions from high to low phenotypic similarity usually occur at the level of genera. On the contrary, much lower conservation levels are observed for taxonomic ranks beyond the level of families, where the differences between the ranks are relatively small. This analysis suggests that computational approaches similar to the one presented here could be useful in refining bacterial taxonomy.

To complement the analysis of metabolic growth phenotypes, we used FBA to investigate the long-term evolution of gene deletion phenotypes (Fig. 4). Specifically, we considered the evolution of metabolic gene essentiality and synthetic lethality (see Methods). First, we confirmed a high (~76%) accuracy of FBA gene essentiality predictions for considered species with available experimental data[7] (Extended Data Table 4). Second, our analysis demonstrated that the average long-term evolution of gene essentiality can also be approximated by an exponential divergence (Fig. 4a). Notably, the average rate of evolution of metabolic gene essentiality is substantially faster and saturates at closer genetic distances than the evolution of growth phenotypes (Fig. 2, see Extended Data Table 1). Even at long evolutionary distances, for an average pair of microbial species more than half of the conserved essential genes in one species usually remain essential in the other. Reassuringly, the predicted average trend (Fig. 4a, red line) is consistent with available experimental data (Extended Data Table 4) for microbial species with genome-wide gene deletion screens (Fig. 4a, black dots/black line).

In contrast to gene essentiality, our analysis revealed a very low conservation of synthetic lethality between metabolic genes (Fig. 4b). Following a common definition, we considered a pair of non-essential genes to be synthetic lethal if simultaneous *in silico* deletion of the corresponding reactions from FBA models made biomass synthesis infeasible. Even at close genetic distances (<0.01 in Fig. 4b) synthetic lethality is conserved, on average, for only ~30% of orthologous metabolic gene pairs. At close distances there is also a substantial variability in the conservation of synthetic lethality across species. As bacterial species diverge further, the average conservation of synthetic lethality drops to ~5%. This suggests that synthetic lethality is much more sensitive to changes in microbial genotypes than gene essentiality and metabolic growth phenotypes. Only several comprehensive studies, none of them in bacteria, have been performed to assess experimentally the conservation of genetic interactions and synthetic lethality[17,18]. Comparison of fitness data from budding and fission yeast revealed a conservation of epistatic gene pairs of ~29% (ref. 17) (corresponding to ~17% similarity). On the other hand, ~5% of the orthologues of synthetic lethal gene pairs in yeast were also found to be synthetic lethal in *Caenorhabditis elegans*[18] (~2.5% similarity). Although these data were obtained in eukaryotic species

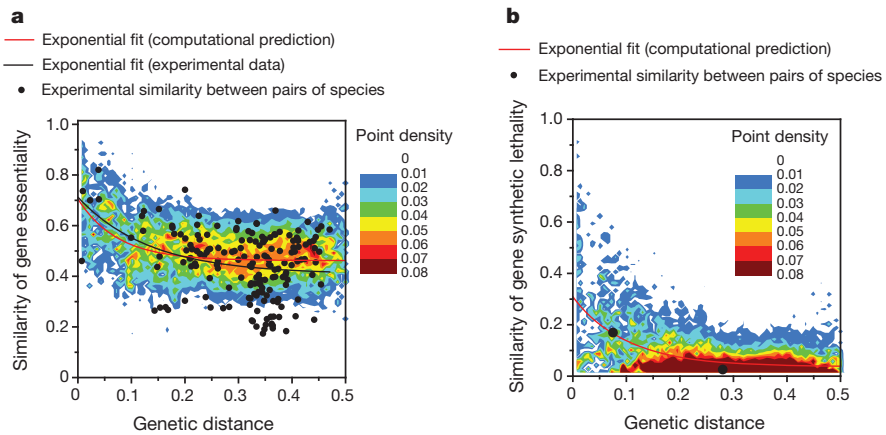


Figure 4 | Evolution of bacterial genetic phenotypes. Genetic distances are based on bacterial 16S rRNA sequences. **a**, The evolution of similarity in gene essentiality across the considered bacterial species. The colours represent the point density at a given genetic distance for pairwise comparisons among considered models ($n = 48{,}920$). The red line shows a three-parameter exponential fit to the computational predictions. Black points represent the available gene essentiality experimental data for 21 bacterial species (see Extended Data Table 4) ($n = 173$); the black line shows an exponential fit to the experimental data. **b**, Like **a**, but for the evolution of similarity in synthetic lethality ($n = 39{,}616$). Black points represent the experimentally determined similarity in synthetic sick/synthetic lethality for two pairs of eukaryotic species (see Methods).

©2014 Macmillan Publishers Limited. All rights reserved

and the FBA accuracy for predictions of genetic interactions is lower than for essentiality or growth phenotypes[19], the available experimental results (Fig. 4b, black dots) are generally consistent with the average divergence trend predicted in our bacterial simulations.

We note that the observed behaviour of long-term phenotypic divergence is somewhat reminiscent of the molecular clock in protein evolution[1]. Similar to protein evolution, it is likely that the phenotypic divergence trends are due to both bacterial adaptation to diverse environmental niches and neutral changes[20]. The relative contribution of adaptive and neutral changes is likely to be different in each particular linage and evolutionary context. Our analysis shows that growth phenotypes, gene essentiality, and synthetic lethality diverge with different rates and have different sensitivities to bacterial genotypes. It is likely that many other phenotypic properties, such as the ability to synthesize different compounds, interact with other species, or withstand specific environmental perturbations, will also show distinct evolutionary patterns. We believe that the accelerating pace of genomic and metagenomic sequencing, and continuous improvement in computational annotation methods[21], will soon allow mapping of the evolution of various phenotypic properties across the entire bacterial tree of life.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 6 May; accepted 2 September 2014.
Published online 26 October 2014.

1. Zuckerkandl, E. & Pauling, L. in *Evolving Genes and Proteins* (eds Bryson, V. & Vogel, H.) 97–166 (Academic, 1965).
2. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217,** 624–626 (1968).
3. Darwin, C. *The Origin of Species* (Barnes & Noble Classics, 2008).
4. Orth, J. D., Thiele, I. & Palsson, B. O. What is flux balance analysis? *Nature Biotechnol.* **28,** 245–248 (2010).
5. Oberhardt, M. A., Palsson, B. O. & Papin, J. A. Applications of genome-scale metabolic reconstructions. *Mol. Syst. Biol.* **5,** 320 (2009).
6. Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J. & Lee, S. Y. Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr. Opin. Biotechnol.* **24,** 617–623 (2011).
7. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnol.* **28,** 977–982 (2010).
8. Bochner, B. R. Global phenotypic characterization of bacteria. *FEMS Microbiol. Rev.* **33,** 191–205 (2009).
9. Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M. & Sauer, U. Multidimensional optimality of microbial metabolism. *Science* **336,** 601–604 (2012).
10. Peregrin-Alvarez, J. M., Sanford, C. & Parkinson, J. The conservation and evolutionary modularity of metabolism. *Genome Biol.* **10,** R63 (2009).
11. Moran, N. A., Munson, M. A., Baumann, P. & Ishikawa, H. A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc. R. Soc. Lond. B* **253,** 167–171 (1993).
12. Sabarly, V. *et al.* The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *J. Evol. Biol.* **24,** 1559–1571 (2011).
13. Gripp, E. *et al.* Closely related *Campylobacter jejuni* strains from different sources reveal a generalist rather than a specialist lifestyle. *BMC Genomics* **12,** 584 (2011).
14. Monk, J. M. *et al.* Genome-scale metabolic reconstructions of multiple *Escherichia coli* strains highlight strain-specific adaptations to nutritional environments. *Proc. Natl Acad. Sci. USA* **110,** 20338–20343 (2013).
15. de Hoon, M. J., Eichenberger, P. & Vitkup, D. Hierarchical evolution of the bacterial sporulation network. *Curr. Biol.* **20,** R735–R745 (2010).
16. Kirschner, M. W. & Gerhart, J. C. *The Plausibility of Life: Resolving Darwin's Dilemma* (Yale Univ. Press, 2005).
17. Dixon, S. J. *et al.* Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc. Natl Acad. Sci. USA* **105,** 16653–16658 (2008).
18. Tischler, J., Lehner, B. & Fraser, A. G. Evolutionary plasticity of genetic interaction networks. *Nature Genet.* **40,** 390–391 (2008).
19. Szappanos, B. *et al.* An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nature Genet.* **43,** 656–662 (2011).
20. Barve, A. & Wagner, A. A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500,** 203–206 (2013).
21. Plata, G., Fuhrer, T., Hsiao, T. L., Sauer, U. & Vitkup, D. Global probabilistic annotation of metabolic networks enables enzyme discovery. *Nature Chem. Biol.* **8,** 848–854 (2012).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank B. Bochner and Biolog for providing the experimental phenotypic growth data. We also thank members of the Vitkup laboratory for discussions. This work was supported in part by the National Institute of General Medical Sciences GM079759 grant to D.V. and the U54CA121852 grant to Columbia University. The work by C.S.H. was supported by the Department of Energy contract DE-AC02-06CH11357, as part of the SB Knowledgebase.

Author Contributions G.P. and D.V. conceived the study and performed the research and data analysis. C.S.H. built the metabolic models. D.V. directed the research. G.P. and D.V. wrote the manuscript. All authors read and edited the manuscript.

Author Information The 322 models used in this study are available at http://vitkuplab.c2b2.columbia.edu/phenotypes. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.V. (dv2121@columbia.edu).

©2014 Macmillan Publishers Limited. All rights reserved

METHODS

Metabolic models. We obtained 322 genome-scale metabolic models using a recently published protocol for automatic network reconstruction[7]. To minimize possible biases due to computational gap-filling and network auto-completion, we only considered models in which more than 80% of the reactions were directly based on available gene annotations. To prevent biases related to uneven sampling of bacterial phylogenic space, we did not use models from the order Enterobacteriales, which contains a significantly higher number of sequenced genomes compared with other bacterial linages. The exact identity of the considered species did not have a significant impact on the predicted average trends (Extended Data Fig. 3).

Computation of genetic distances. For the considered bacteria, 16S rRNA gene sequences were obtained from GenBank[22]. Sequences were aligned using Clustal Omega[23] to a reference alignment of small subunit rRNA sequences from the SILVA database[24]. Genetic distances were then calculated on the basis of the multiple sequence alignment using the Dnadist program in the Phylip software package[25]; the F84 model of nucleotide substitution, with default parameter values, was used. The cladograms in Fig. 1 and Extended Data Fig. 4 were computed from the distance matrix using the Fitch program in Phylip with default parameters; the *M. barkeri* rRNA was used as the outgroup sequence. The taxonomic classification used in Fig. 3 was obtained from the National Center for Biotechnology Information taxonomy database[26].

Prediction of growth phenotypes. FBA allows one to determine feasible values of metabolic reaction fluxes subject to reaction stoichiometry constraints and the assumption of metabolic steady state[4]. Additional constraints, such as upper and lower bounds for metabolic fluxes or flux combinations, can be applied to the models. All FBA calculations in our manuscript used the COBRA toolbox[27]. We used the following procedure to determine the ability of each species to utilize the considered nutrient sources. First, we identified a set of compounds required by the models to simulate growth on different carbon or nitrogen sources (Supplementary Data 2); these compounds include various vitamins, nucleotides and amino acids[8] as well as several model-specific requirements (see below). To test the ability of each model to use different carbon sources, an *in silico* growth medium was defined where the aforementioned substances were constrained to a maximum combined uptake of 10 mmol of carbon per gram dry weight, and all other carbon compounds were removed. All carbon-free compounds were made available in the simulated medium (with the maximum uptake rate of 1 mol g^{-1} dry weight). An analogous procedure was used to define a growth medium for testing nitrogen sources. Second, for each considered carbon or nitrogen source (compound), we used FBA to calculate the maximum biomass or ATP synthesis rate when the compound was made available in the corresponding *in silico* medium (maximum uptake rate of 1 mol g^{-1} dry weight). Similar to the treatment of experimental data (see experimental procedures below), biomass or ATP flux values were normalized on a scale of 0–100 corresponding to the minimum (no carbon or nitrogen source tested) and maximum flux values across considered compounds, respectively. Similar to experimental measurements, carbon and nitrogen sources scoring 10 or above were considered as being positive for growth. The metabolic compounds tested in our analysis correspond to carbon and nitrogen sources that are commonly used by multiple bacteria (Extended Data Table 3) and are assayed in Biolog MicroPlates[8]. To prevent low phenotypic similarities arising because of models with a very low overall number of positive growth phenotypes, we only considered models that could synthesize biomass on more than five of the tested compounds; the exact value of this cutoff had little effect on the observed average trends (Extended Data Fig. 5a). In total, 229 and 272 models were used for the analysis of carbon and nitrogen sources for biomass synthesis, respectively.

Definition of an *in silico* growth medium. Similar to bacterial growth *in vivo*, many of the metabolic models used in our analysis are auxotrophic for specific compounds, beyond the main carbon and nitrogen sources tested for their ability to support microbial growth (62 carbon and 68 nitrogen compounds); that is, the models can simulate biomass or ATP synthesis only if small amounts of additional nutrients are available in the simulated growth media. To define a single minimal medium, used across all models to test growth on the main carbon sources, we used the following procedure. First, in addition to the main carbon sources (available with a maximum uptake of 1 mol g^{-1} dry weight), all metabolic compounds that could be imported by the models were made available in the simulated media with a maximum combined carbon uptake of 10 mmol g^{-1} dry weight; we note that this maximum uptake rate is only 1% of the maximum uptake of the main carbon sources. Second, we determined which of the main carbon sources could support growth under these conditions. Third, for each model the additional carbon compounds were sequentially removed (in a random order) from the simulated media until no compound could be further removed while allowing growth on the main carbon sources determined in the second step. Fourth, the additional carbon sources required for growth in more than 75% of the tests with a positive growth phenotype were combined across all models; this resulted in the carbon-containing component of the minimal media. Fifth, the same procedure was used to determine the nitrogen-containing component of the minimal media. Sixth, the carbon- and nitrogen-containing components were combined to produce the minimal media used in the study (Supplementary Data 2). Notably, very similar results were observed using different minimal media obtained from independent runs of the aforementioned procedure. Very similar results (Extended Data Fig. 6) were also obtained when the *in silico* growth medium, used for all tests, contained all possible nutrients (without any removals) with maximum uptakes rates of 1 mol g^{-1} dry weight for the main carbon or nitrogen sources, and with combined maximum uptake rates of 10 mmol g^{-1} dry weight (1% of the uptake for the main nutrients) for additional carbon or nitrogen compounds, respectively.

Prediction of essentiality and synthetic lethality. To determine essential genes, we first established the association between every gene and corresponding metabolic reactions. We then simulated gene deletions by setting the maximal fluxes through corresponding reactions that cannot be catalysed by the products of other genes to zero. If such an *in silico* deletion of a gene made it impossible, on the basis of FBA calculations, to produce a non-zero biomass, the gene was considered to be essential. A pair of non-essential genes was considered to be synthetic lethal if simultaneous deletion of the two genes made it impossible to produce a non-zero biomass. All FBA simulations for testing gene essentiality and synthetic lethality were performed, similar to common experimental procedures, using an *in silico* rich medium: that is, non-zero fluxes were allowed through every transport reaction in the models. Genes associated with lumped reactions, such as 'protein synthesis', were not considered in the calculations. To prevent low phenotypic similarities arising because of models with a very small number of essential genes, only models with more than ten predicted essential genes or synthetic lethal gene pairs with mapped orthologues were considered in the analysis; 314 and 290 models were used for the analysis of essentiality and synthetic lethality, respectively.

Quantifying phenotypic similarity. For a given set of features, namely carbon sources, essential genes, or synthetic lethal gene pairs, similarities between species were quantified by Jaccard's similarity index. Jaccard's index is defined as the size of the intersection between two given sets divided by the size of the union of the two sets; for example, if A is the set of all carbon sources that can be used by species *a*, and B the set that can be used by species *b*, then Jaccard's carbon source similarity between *a* and *b* is defined as $J(A,B) = |A \cap B|/|A \cup B|$. Importantly, to calculate the similarity of gene essentiality and synthetic lethality between species, we only considered orthologous genes and gene pairs that are shared between corresponding metabolic networks. Orthologous genes were identified using bi-directional BLASTP[28] hits (with expect (*E*) values <0.05) between the species' genomes.

Analysis of experimental Biolog data. A collection of 40 microbial species spanning a wide range of phylogenetic distances (Extended Data Fig. 4) was used to confirm the computationally predicted trends. The ability of these species to metabolize the 62 carbon sources used in the computational analysis was determined using Biolog GENIII Phenotype Microarrays[8]; all data were obtained directly from the Biolog GEN III database (Biolog). The Phenotype Microarrays technology is based on the reduction of a tetrazolium dye, which allows determination of the usage of different nutrient sources across multiple growth conditions[8]. Biolog assays were performed essentially as described in the GEN III MicroPlate instruction manual. Colorimetric measurements after a 24 h incubation period for each species and each carbon source were expressed on a scale of 0–100, representing the average colour density (across at least five biological replicates) in each well of the Biolog plate relative to the negative and positive controls; only scores of 10 or above were considered as evidence that a tested compound was used by a species. This cutoff value was obtained on the basis of the bimodal-like distribution of the data[12] (Extended Data Fig. 5b); similar results were obtained using other cutoff values (Extended Data Fig. 5c). The accuracy of FBA in predicting microbial growth phenotypes was evaluated using the nine species that were present both in the computational and experimental analyses. The experimental values in Fig. 3 (red stars) were based on data from the aforementioned 40 species, and the intra-species similarity studies in refs 12 and 13.

Experimental gene essentiality data. Gene essentiality data were compiled for 19 species with genome-wide gene deletion screens (Extended Data Table 4). Species from the genus *Mycoplasma* were excluded from this analysis because of their very small genomes and a very high (~80% (ref. 29)) fraction of essential genes. For every pair of species, orthologous genes were identified using bi-directional BLASTP hits (with *E* values <0.05). The similarity of gene essentiality was determined for genes with identified orthologues annotated as enzymes in the KEGG database[30]. To estimate the similarity of gene essentiality at close genetic distances, we also considered partial essentiality data for *Streptococcus pneumoniae* R6 and *Staphylococcus aureus* N315, which were compared with essential genes in *Streptococcus sanguinis* SK36 and *S. aureus* NCTC 8325, respectively. To use these incomplete data, and given the similarity of the species' genome sizes, we assumed symmetry of essential gene conservation: that is, that the number of essential genes in one species that are

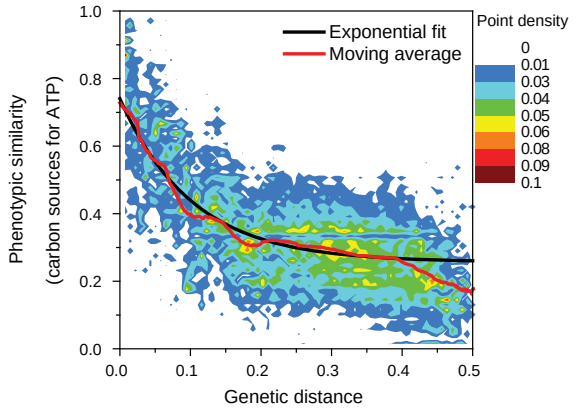
©2014 Macmillan Publishers Limited. All rights reserved

not essential in the other is the same for both bacteria. A similar approach was used to estimate Jaccard's similarity of genetic interactions between eukaryotic species on the basis of published data[17,18] (Fig. 4b).

Exponential model fits. Pairwise divergence of bacterial phenotypic similarity (y) as a function of genetic distance (t) was fitted using the following equation: $y = a + be^{-ct}$, where the parameter a represents the saturation level for phenotypic divergence at long genetic distances, $(a + b)$ represents the level phenotypic similarity at close genetic distances, and the parameter c quantifies the divergence rate, namely the phenotypic similarity decrease per unit of genetic distance (time). Larger values of c correspond to faster divergence of the phenotypic similarity. The parameter a was not considered in the nested two-parameter exponential model used for model comparison. To quantify the genetic distance between bacterial species, we used the divergence between their 16S rRNA sequences; 1% 16S rRNA distance approximately corresponds to 50 million years since divergence from a common ancestor[11].

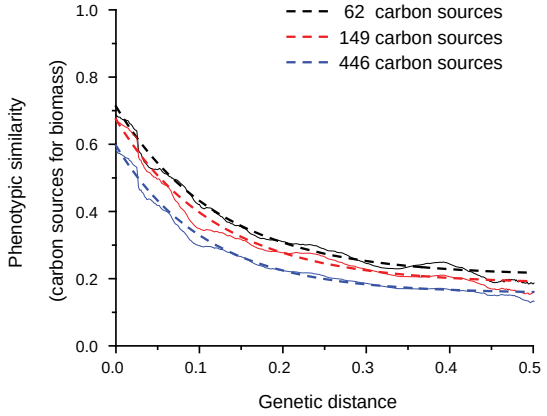
22. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. *Nucleic Acids Res.* **37,** D26–D31 (2009).
23. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
24. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41,** D590–D596 (2013).
25. Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.6. *Cladistics* **5,** 164–166 (1989).
26. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **37,** D5–D15 (2009).
27. Becker, S. A. *et al.* Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Protocols* **2,** 727–738 (2007).
28. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
29. Glass, J. I. *et al.* Essential genes of a minimal bacterium. *Proc. Natl Acad. Sci. USA* **103,** 425–430 (2006).
30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).

©2014 **Macmillan Publishers Limited. All rights reserved**



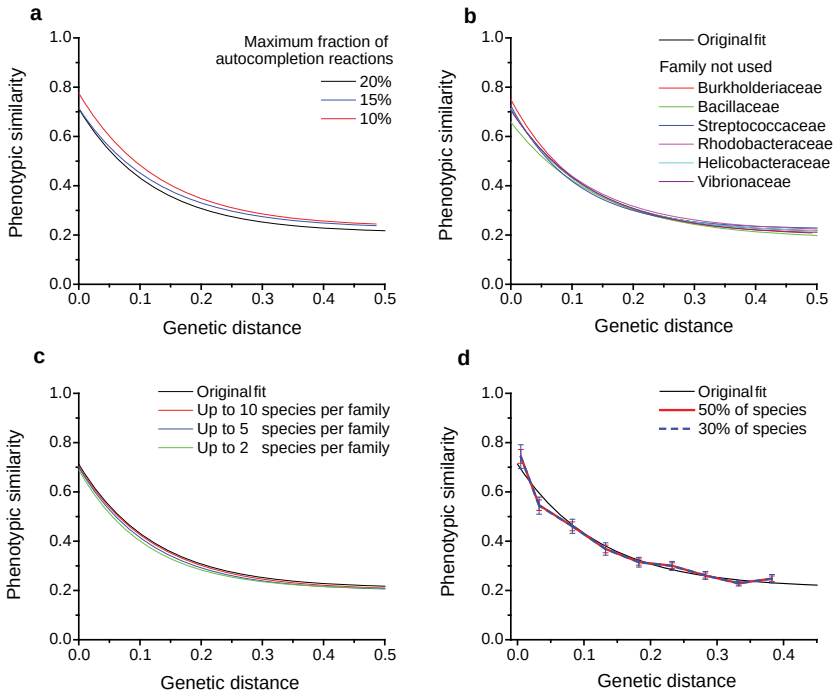
Extended Data Figure 1 | The evolution of phenotypic similarity in the usage of carbon sources for ATP production. Genetic distances are based on 16S bacterial rRNA sequences. The colours represent the point density at a given genetic distance for all pairwise comparisons between metabolic models ($n = 20{,}910$). The black line shows a three-parameter exponential fit to the computational predictions; the red line shows a moving average of the predictions.

©2014 Macmillan Publishers Limited. All rights reserved



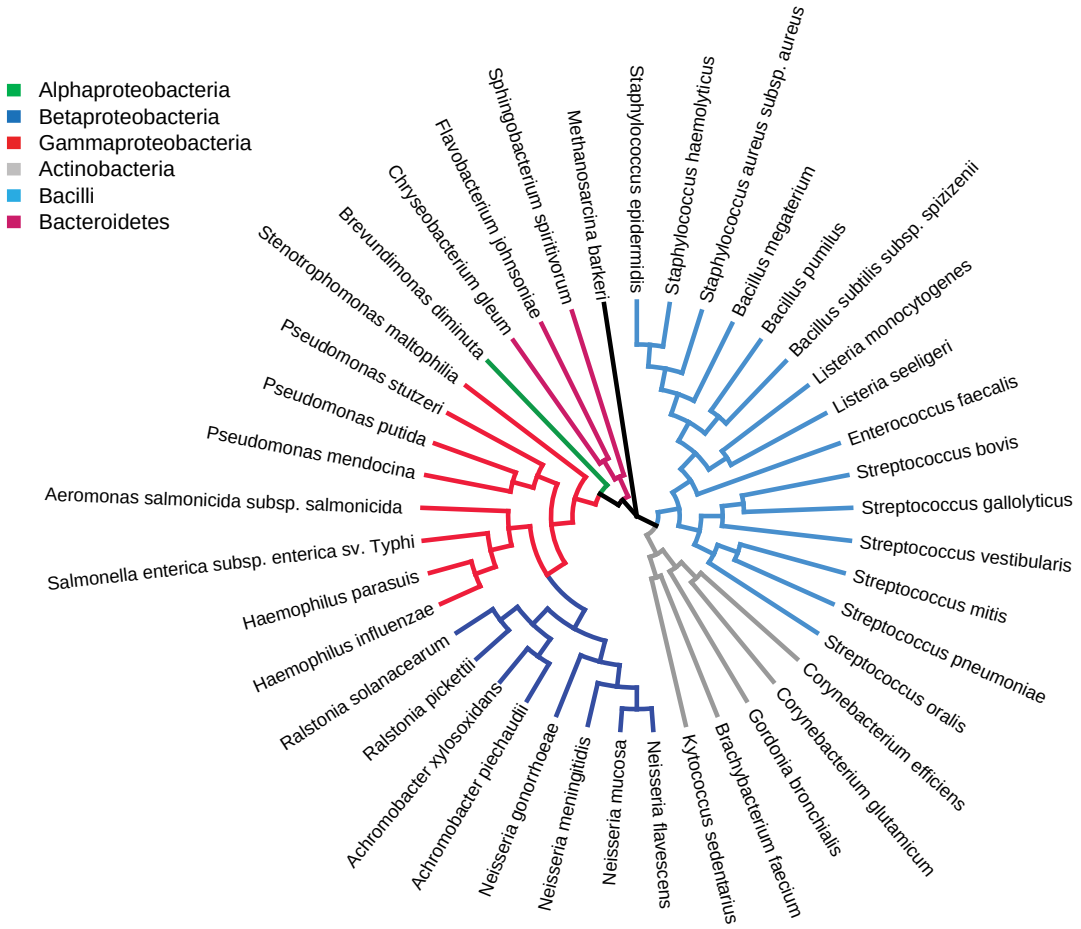
Extended Data Figure 2 | The evolution of phenotypic similarity for different sets of carbon sources. Genetic distances are based on 16S bacterial rRNA sequences. Phenotypic similarities of biomass synthesis are shown for 62, 149, and 446 carbon sources. Solid lines represent moving averages (using a 0.05 genetic distance window) of computational predictions; dashed lines represent exponential fits to the data.

©2014 Macmillan Publishers Limited. All rights reserved



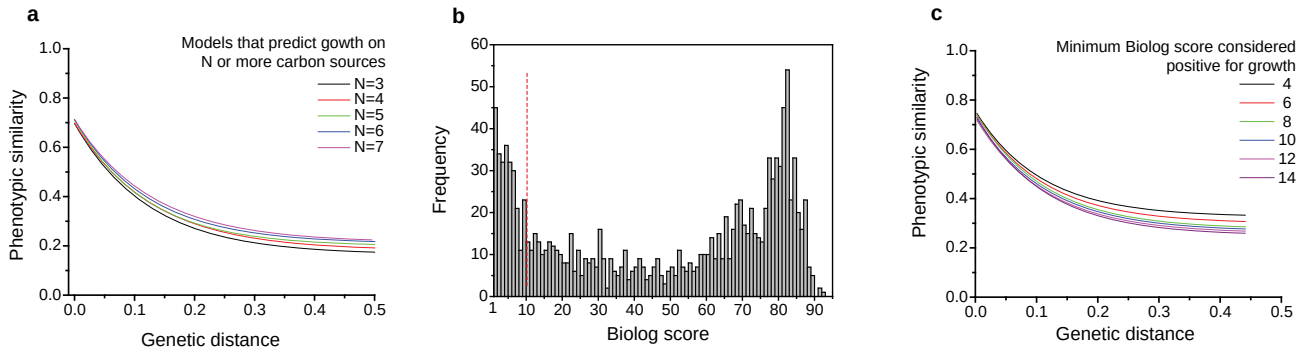
Extended Data Figure 3 | The effect of species selection on observed patterns of phenotypic divergence. The black lines in all panels (marked 'Original fit') represent the exponential fit of the phenotypic similarity (carbon source utilization) as a function of genetic distance for all pairs of considered models, that is, models with fewer than 20% auto-completion reactions and more than five predicted carbon sources for growth. The observed trends of phenotypic evolution remain similar when (**a**) only models with a smaller fraction of auto-completion reactions are considered, (**b**) models from individual families that include more than ten modelled species are excluded from the analysis, (**c**) only a maximum number of species per family is considered, and (**d**) a subset of species is chosen at random from the pool of all considered models. In **d**, the average values at different genetic distance bins are shown for 1,000 random samples of a given number of species; error bars represent the s.e.m. obtained on the basis of the 1,000 replicates.

©2014 Macmillan Publishers Limited. All rights reserved



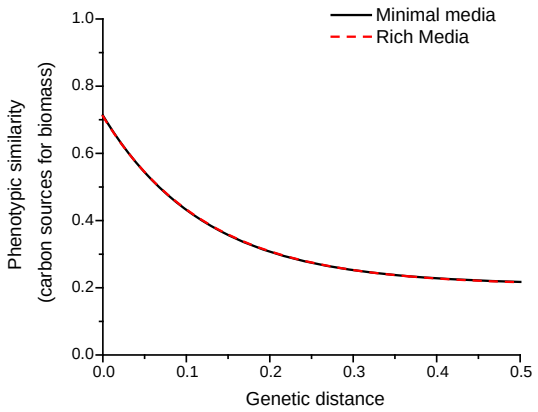
Extended Data Figure 4 | Evolutionary relationship between 40 species for which experimental phenotype microarray data were considered. The cladogram is based on 16S bacterial rRNA sequences. Different colours indicate different bacterial classes. The tree was rooted in the *M. barkeri* rRNA sequence.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 5 | The effect of cutoff selection on the computational and experimental phenotypic similarity trends. **a**, The exponential fits of phenotypic similarity as a function of genetic distance for metabolic models that predict growth on more than a given number of carbon sources. **b**, The frequency distribution of normalized Biolog scores for 40 species across 62 experimental growth conditions. The figure shows a bimodal pattern (scores of 0 are not plotted). The dashed red line shows the cutoff score of 10 used in the main analysis. **c**, The effect of different cutoffs used to define positive growth in the Biolog data. Different lines represent exponential fits to the experimental values of phenotypic similarity based on different values of the Biolog cutoff score.

©2014 Macmillan Publishers Limited. All rights reserved



Extended Data Figure 6 | Effect of growth media on the predicted phenotypic similarity trends. The black line shows an exponential fit to the predicted phenotypic similarity in the usage of carbon sources calculated on the basis of the *in silico* minimal media used in the study of growth phenotypes (see Methods and Supplementary Data 2) ($n = 26{,}106$). The dashed red line shows an exponential fit to the predicted phenotypic similarity using the *in silico* medium in which all carbon sources that could be imported by the models were made available, with the total combined uptake of carbon constrained to a maximum value of 10 mM g^{-1} dry weight ($n = 27{,}261$).

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 1 | Parameters of the exponential divergence models, describing the evolution of growth and genetic phenotypes

Phenotype	*a*	*b*	*c*
Carbon source (biomass)	0.21	0.50	8.16
Carbon source (ATP)	0.21	0.50	7.86
Nitrogen source (biomass)	0.25	0.48	9.75
Essentiality	0.46	0.24	13.6
Synthetic lethality	0.037	0.27	9.98

Values in the table show the parameters (a, b, c) of the divergence model $y = a + be^{-ct}$, where y represents the phenotypic similarity and t represents the genetic distance between species. Parameter a represents the saturation level of phenotypic divergence at long genetic distances, ($a + b$) represents the level phenotypic similarity at close genetic distances, and the parameter c quantifies the divergence rate.

©2014 **Macmillan Publishers Limited. All rights reserved**

Extended Data Table 2 | Model comparisons for predicted and experimentally determined phenotypic similarity as a function of genetic distance

Phenotype	3-parameter exponential vs. linear model (relative likelihood based on Akaike weights)	3-parameter exponential vs. 2-parameter exponential model (F-test P-value)
FBA carbon source similarity (biomass)	$>10^{20}$	$< 10^{-20}$
Experimental carbon source similarity (Biolog)	$9*10^{9}$	$4.6*10^{-8}$
FBA essentiality similarity	$>10^{20}$	$< 10^{-20}$
Experimental essentiality similarity	27.2	$8.5*10^{-3}$
FBA Synthetic lethality similarity	$>10^{20}$	$< 10^{-20}$

Comparisons between the three parameter exponential model and the linear model were performed on the basis of Akaike's Information Criterion (AIC). Values in the table represent the Akaike-based relative likelihoods of the three-parameter exponential model compared with the linear model. Comparisons between the three-parameter exponential model and the nested two-parameter exponential model were performed using the *F*-test; the corresponding *P* values reflect the probability that the nested two-parameter model fits the data as well as the more complex three-parameter model.

©2014 **Macmillan Publishers Limited. All rights reserved**

Extended Data Table 3 | The predicted frequency of carbon and nitrogen source usage across metabolic models

Carbon sources for biomass production		Nitrogen sources for biomass production	
Metabolite name	Number of models	Metabolite name	Number of models
L-Glutamic Acid	208	Ammonia	241
a-D-Glucose	199	Urea	241
D-Fructose	185	L-Proline	226
L-Malic Acid	165	L-Glutamic Acid	203
L-Lactic Acid	164	L-Valine	203
Maltose	149	L-Isoleucine	183
Glycerol	134	L-Leucine	178
L-Serine	128	Nitrate	155
L-Aspartic Acid	126	Nitrite	145
D-Mannose	115	L-Glutamine	136
L-Arginine	112	L-Serine	127
N-Acetyl-DGlucosamine	104	L-Aspartic Acid	124
D-Trehalose	97	Cytosine	123
Sucrose	89	L-Arginine	119
Inosine	87	L-Ornithine	115
a-Keto-GlutaricAcid	84	Uracil	111
L-Histidine	82	N-Acetyl-D-Glucosamine	103
L-Alanine	77	Cytidine	101
D-GlucuronicAcid	74	Adenosine	94
D-Serine	72	Glycine	93
a-D-Lactose	71	L-Methionine	92
Formic Acid	69	Xanthine	87
Acetoacetic Acid	68	Histamine	86
D-Cellobiose	67	L-Histidine	83
D-Mannitol	64	L-Tryptophan	80
D-Malic Acid	64	L-Alanine	79
D-Sorbitol	61	Ethanolamine	74
D-GalacturonicAcid	56	D-Alanine	73
Mucic Acid	55	D-Serine	72
D-Saccharic Acid	55	L-Lysine	61
D-Gluconic Acid	53	D-Glucosamine	61
D-Galactose	47	Putrescine	54
Citric Acid	33	Acetamide	54
D-Raffinose	29	L-Phenylalanine	49
Dextrin	27	Allantoin	47
g-Amino-ButyricAcid	27	L-Tyrosine	45
L-Rhamnose	26	Formamide	44
Salicin	23	Inosine	41
N-Acetyl-b-D-Mannosamine	20	L-Cysteine	25
Glucose-6-Phosphate	20	L-Asparagine	25
D-Melibiose	15	Thymidine	23
D-Aspartic Acid	15	Uridine	23
L-Fucose	13	Guanine	19
M-Inositol	11	Guanosine	18
Quinic Acid	11	N-Acetyl-D-Mannosamine	18
Propionic Acid	11	Xanthosine	16
Stachyose	9	D-Aspartic Acid	15
N-Acetyl-Neuraminic Acid	7	Methylamine	10
D-Arabitol	7	Thymine	9
Acetic Acid	7	L-Threonine	6
Fructose-6-Phosphate	5	D-Lysine	5
N-Acetyl-DGalactosamine	4	Adenine	4
a-Keto-ButyricAcid	4	N-Acetyl-D-Galactosamine	4
Gentiobiose	0	L-Citrulline	3
b-Methyl-D-Glucoside	0	D-Galactosamine	3
D-Fucose	0	D-Glutamic Acid	2
Gelatin	0	Tyramine	2
L-PyroglutamicAcid	0	b-Phenylethylamine	2
Pectin	0	L-Homoserine	1
L-GalactonicAcid-g-Lactone	0	Uric Acid	1
P-HydroxyPhenyl AceticAcid	0	Agmatine	0
a-HydroxyButyric Acid	0	Hydroxylamine	0
		Ethylamine	0
		L-Pyroglutamic Acid	0
		D-Asparagine	0
		D-Mannosamine	0
		D-Valine	0
		Biuret	0

Numbers in the table represent the total number of models, out of 322, predicted to use the corresponding carbon or nitrogen source. Metabolites are ranked from most to least frequent across models.

©2014 Macmillan Publishers Limited. All rights reserved

Extended Data Table 4 | Bacteria with experimental genome-wide data used to analyse the conservation of gene essentiality

Species name	Number of essential genes	Reference PubMed ID
Acinetobacter baylyi ADP1	499	18319726
Bacillus subtilis subsp. subtilis str. 168	276	14602916, 12682299
Bacteroidetes thetaiotaomicron VPI-5482	325	19748469
Burkholderia thailandensis E264	406	23382856
Caulobacter crescentus NA1000	480	21878915
Escherichia coli K-12	302	16738554
Francisella novicida U112	396	17215359
Haemophilus influenzae Rd KW20	667	11805338
Helicobacter pylori 26695	336	15547264
Mycobacterium tuberculosis H37Rv	689	23028335
Porphyromonas gingivalis ATCC 33277	463	23114059
Pseudomonas aeruginosa PAO1	774	14617778
Salmonella enterica subsp. enterica serovar Typhi str. Ty2	331	23470992
Salmonella enterica serovar Typhimurium str. SL1344	355	23470992
Shewanella oneidensis MR-1	403	22125499
Sphingomonas wittichii RW1	572	23601288
Staphylococcus aureus subsp. aureus N315	168*	11952893
Staphylococcus aureus subsp. aureus NCTC 8325	351	19570206
Streptococcus pneumoniae	134†	15995353
Streptococcus sanguinis	218	22355642
Vibrio cholerae O1 biovar El Tor str. N16961	344	23901011

* Only incomplete data available, used to estimate conservation relative to *S. aureus* NCTC 8325.
† Only incomplete data available, used to estimate conservation relative to *S. sanguinis*.

©2014 **Macmillan Publishers Limited. All rights reserved**