doi:10.1038/nature12301

A latent capacity for evolutionary innovation through exaptation in metabolic systems

Aditya Barve[1,2] & Andreas Wagner[1,2,3]

Some evolutionary innovations may originate non-adaptively as exaptations, or pre-adaptations, which are by-products of other adaptive traits[1–5]. Examples include feathers, which originated before they were used in flight[2], and lens crystallins, which are light-refracting proteins that originated as enzymes[6]. The question of how often adaptive traits have non-adaptive origins has profound implications for evolutionary biology, but is difficult to address systematically. Here we consider this issue in metabolism, one of the most ancient biological systems that is central to all life. We analyse a metabolic trait of great adaptive importance: the ability of a metabolic reaction network to synthesize all biomass from a single source of carbon and energy. We use novel computational methods to sample randomly many metabolic networks that can sustain life on any given carbon source but contain an otherwise random set of known biochemical reactions. We show that when we require such networks to be viable on one particular carbon source, they are typically also viable on multiple other carbon sources that were not targets of selection. For example, viability on glucose may entail viability on up to 44 other sole carbon sources. Any one adaptation in these metabolic systems typically entails multiple potential exaptations. Metabolic systems thus contain a latent potential for evolutionary innovations with non-adaptive origins. Our observations suggest that many more metabolic traits may have non-adaptive origins than is appreciated at present. They also challenge our ability to distinguish adaptive from non-adaptive traits.

How evolutionary adaptations and innovations originate is one of the most profound questions in evolutionary biology. Previous work[1,2] emphasizes the importance of exaptations, also sometimes called pre-adaptations, for this origination. These are traits whose benefits to an organism are unrelated to the reasons for their origination; they are features that originally serve one (or no) function, and become later co-opted for a different purpose[1–5]. Although examples of exaptations occur from the macroscopic scale to the molecular[1–6] and abound also in human evolution[7], no number of examples could reveal how important exaptations are in the origination of adaptations in general. This limitation of case studies can be overcome in those biological systems where it is possible to study systematically many genotypes and the phenotypes they form[8–12].

One of these systems is metabolism. The metabolic genotype of an organism encodes a metabolic reaction network with hundreds of enzyme-catalysed chemical reactions. One of metabolism's fundamental tasks is to synthesize small biomass precursor molecules from environmental molecules, such as different organic carbon sources. An organism or metabolic network is said to be viable on a carbon source if it is able to synthesize all biomass molecules from this source. Viability on a new carbon source can be an important adaptation, and anecdotal evidence shows that this ability can originate as a pre-adaptation[13,14]. For example, laboratory evolution of *Pseudomonas putida* for increased biomass yield on xylose as a carbon source produces strains that utilize arabinose as efficiently as they do xylose, even though the ancestral strains did not utilize arabinose[14]. Thus, viability on arabinose can be a by-product of increased viability on xylose. We here analyse systematically whether such exaptations are typical or unusual in metabolic systems.

Our analysis relies on the ability to predict a metabolic phenotype from a metabolic genotype with the constraint-based method of flux balance analysis (Methods), to study not just one metabolic network but to explore systematically a vast space of possible metabolic networks. The members of this space can be described as follows. The currently known 'universe' of biochemical reactions comprises more than 5,000 chemical reactions with well-defined substrates and products. In the metabolic network of any one organism, however, only a fraction of these reactions take place, enabling us to describe this network through a binary presence/absence pattern of enzyme-catalysed reactions in the known reaction universe. Recent methods based on Markov chain Monte Carlo (MCMC) sampling (Methods) allow a systematic exploration of this space; that is, they permit the creation of arbitrarily large and uniform samples of networks with a given phenotype[12]. This sampling is based on long random walks through metabolic network space, where each step in a walk adds or eliminates a metabolic reaction from a metabolic network, with the only constraint being that the network remains viable on a focal carbon source. The starting point of the MCMC random walk is the *Escherichia coli* metabolic network, which we know a priori to be viable on different carbon sources[15]. Here we use this approach to create random samples of metabolic networks that are viable on a given set of carbon sources. We refer to such networks as random viable networks.

Our analysis focuses on 50 biologically relevant and common carbon sources[15] (Supplementary Table 1). For each carbon source *C*, we create a sample of 500 random viable networks that are viable on *C* if it is provided as the sole carbon source. We then use flux balance analysis to determine the viability of these networks on each of the 49 other carbon sources. This approach allows us to ask whether viability on *C* usually entails viability on other carbon sources. The answers to this and related questions show that potential exaptations are ubiquitous in metabolism.

We began our analysis with a sample of 500 random networks that were viable on glucose as the sole carbon source (Methods). Each network can synthesize the 63 essential biomass precursors of *E. coli*—many of which are important for most organisms[15,16]—in an aerobic minimal environment (Methods) containing glucose as the only carbon source. Importantly, we did not require that these 500 networks be viable on any carbon source except glucose.

We first examined whether these networks were viable on each of the 49 other carbon sources. The information resulting from this analysis can be represented, for each network, as a binary 'innovation vector' whose *i*th entry equals 1 if the network is viable on carbon source C_i, and equals 0 otherwise (Fig. 1a). We define the innovation index, I_{Glucose}, of a network to be the number of additional carbon sources on which each network is viable. The distribution of this index is shown in Fig. 1b. Ninety-six per cent of networks are viable on other carbon sources in addition to glucose ($I_{\mathrm{Glucose}} > 0$). The mean innovation index

[1]Institute of Evolutionary Biology and Environmental Sciences, Building Y27, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland. [2]The Swiss Institute of Bioinformatics, Bioinformatics, Quartier Sorge, Bâtiment Genopode, 1015 Lausanne, Switzerland. [3]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

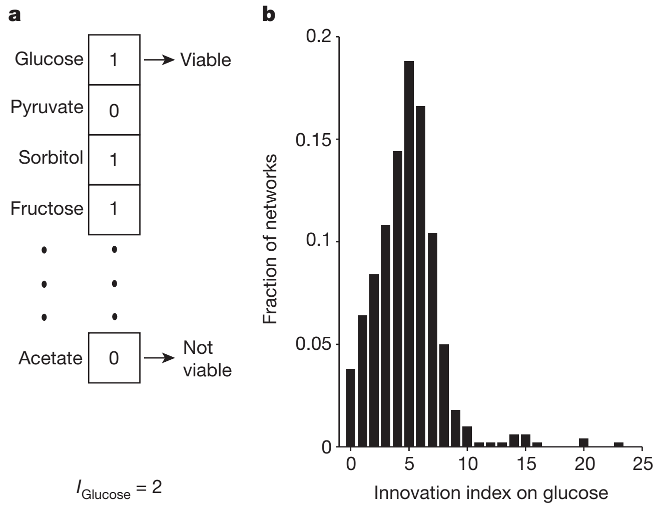

Figure 1 | Viability on glucose entails viability on multiple other carbon sources. **a**, The binary innovation vector of a hypothetical metabolic network that is viable on glucose. The vector shows that the random network is viable (labelled by 1) on glucose, sorbitol and fructose, but not viable (labelled by 0) on pyruvate and acetate. The innovation index of this network ($I_{\text{Glucose}} = 2$) is the number of additional carbon sources on which the network is viable. **b**, The distribution of innovation indices for 500 random networks viable on glucose. Only 4% of networks have $I_{\text{Glucose}} = 0$, meaning that they are viable only on glucose.

is $I_{\text{Glucose}} = 4.86$ (standard deviation, 2.83 carbon sources). This means that networks viable on glucose typically are also viable on almost 5 additional carbon sources. Ninety-four networks (18.8%) are viable on exactly 5 new carbon sources, and 187 networks (37.4%) are viable on 6 or more carbon sources. Viability on each such carbon source is a potential exaptation. This viability is merely a by-product of viability on glucose and could become an adaptation whenever this carbon source is the sole carbon source. We also found that different random viable networks differ in the additional carbon sources to which they are pre-adapted (Supplementary Figs 1 and 2). Most of the 50 carbon sources we study confer viability on at least one network in our sample (Supplementary Results). Moreover, a variation in our sampling procedures that allows only reactions already connected to a metabolism to be altered further increases the incidence of exaptation (Methods and Supplementary Fig. 3). Finally, complex metabolic networks that have more reactions have greater potential for exaptation (Supplementary Fig. 4).

We next asked whether the ability to grow on multiple additional carbon sources is a peculiarity of networks viable on glucose. To this end, we sampled, for each of our remaining 49 carbon sources, 500 random metabolic networks viable on this carbon source (for a total of 49 × 500 = 24,500 sampled networks). We then computed the distribution of the innovation index, I_C, for each carbon source C. Figure 2a shows the mean of this distribution (bars) and its coefficient of variation (vertical lines), that is, the ratio of the standard deviation to the mean. The figure shows that glucose (highlighted in red) is not unusual. Eighteen carbon sources (36%) have a greater average innovation index than glucose. For example, acetate allows viability on the greatest number (9.75) of additional carbon sources. Conversely, some carbon sources, such as adenosine ($I_{\text{Adenosine}} = 0.27$) and deoxyadenosine ($I_{\text{Deoxyadenosine}} = 0.1$), allow growth on fewer additional carbon sources than glucose. Carbon sources with a small average innovation index—entailing viability on few additional carbon sources—are also more variable in innovation index (Spearman's $\rho = -0.82$, $P < 10^{-101}$; see also Supplementary Fig. 5). Even though any one carbon source may confer growth on only few additional carbon sources in any one network (Fig. 2a), when considering all networks in a sample, it may still allow pre-adaptation to most other carbon sources (Supplementary Fig. 6).

In summary, viability on any one carbon source, C, usually entails viability on multiple other carbon sources, whose number and identity can vary with C. Viability on carbon sources never before encountered is thus a typical metabolic property. Environmental generalists capable of surviving on multiple carbon sources may be viable on many more carbon sources than occur in their environment (Supplementary Tables 2 and 3 and Supplementary Fig. 10).

We next asked whether metabolically close carbon sources show the highest potential for pre-adaptation. The centre path of Fig. 2b shows a hypothetical metabolic pathway that leads from one carbon source, C, to another, C_{new} (boxed area), and from there through (possibly multiple) further metabolic reactions to the synthesis of biomass. Figure 2c shows the same scenario, except that C and C_{new} are separated by several further reactions. It is possible that random networks viable on C are more likely to be viable also on C_{new} if C_{new} is closer to C, that is, if they are separated by fewer metabolic reactions, as in Fig. 2b. In this case, metabolite C_{new} may be less easy to bypass through an alternative pathway that originates somewhere between C and C_{new} (right-hand sequence of arrows in Fig. 2c).

To test this hypothesis (Supplementary Results), we analysed our 50 samples of 500 random metabolic networks, where networks in each sample were required to be viable on a different one of our 50 carbon sources. For each sample (carbon source C) and for each of the other 49 possible carbon sources, C_{new}, we asked whether the metabolic distance between C and C_{new} is correlated with the fraction of networks that are also viable on C_{new}. To answer this, we used metabolic networks that were selected for growth on C and were additionally viable on C_{new} (Methods). We then computed the mean metabolic distance and binned the distances. The results, pooled for all networks, are shown on the vertical axis of Fig. 2d, whose horizontal axis shows the mean metabolic distance (binned into nine bins). The closer C_{new} is to C, the more networks viable on C are also viable on C_{new} (Spearman's $\rho = -0.42$, $P = 10^{-87}$, $n = 1{,}990$). However, the figure also shows that the association is highly noisy, especially at low metabolic distances. Taking reaction irreversibility into account yields the same result (Spearman's $\rho = -0.39$, $P = 10^{-57}$, $n = 1{,}601$), as does a different way of computing distances between pairs of carbon sources (Methods and Supplementary Results). The association is noisy, because metabolism is highly reticulate (Supplementary Results).

Although metabolic 'nearness' cannot explain exaptations involving two carbon sources, biochemical similarities help explain why a network viable on C might be viable on one additional carbon source, C_{n1}, but not on another source, C_{n2}. Indeed, exaptations often involve carbon sources with broadly defined biochemical similarities (Supplementary Figs 7 and 8). For example, glycolytic carbon sources are more likely to entail exaptations for growth on other glycolytic carbon sources, and likewise for gluconeogenic carbon sources, as well as for carbon sources involved in nucleotide metabolism. Furthermore, we also find that pre-adaptation is synergistic; that is, the innovation index for a pair of carbon sources is greater than the sum of the innovation indices, I_{C1} and I_{C2} (Supplementary Fig. 9).

Our analysis has several limitations. First, it is based on present knowledge about the reaction universe. Future work may increase the number of known reactions, but this would not diminish, and could only enhance, the spectrum of possible exaptations. The reason is that additional reactions would allow the use of additional carbon sources by some metabolic networks. Second, most of our analysis focused on random networks that are viable on a specific carbon source, but selection in the wild can affect more than viability, which may affect the incidence of exaptations. Of special importance is selection that favours networks with a high rate of biomass synthesis. This particular selective constraint would not affect our conclusions, because we found that networks with high biomass synthesis rates have even greater potential for metabolic innovation than merely viable networks (Supplementary Table 4 and Supplementary Fig. 11). Third, we considered all necessary nutrient transporters to be present (Methods). If

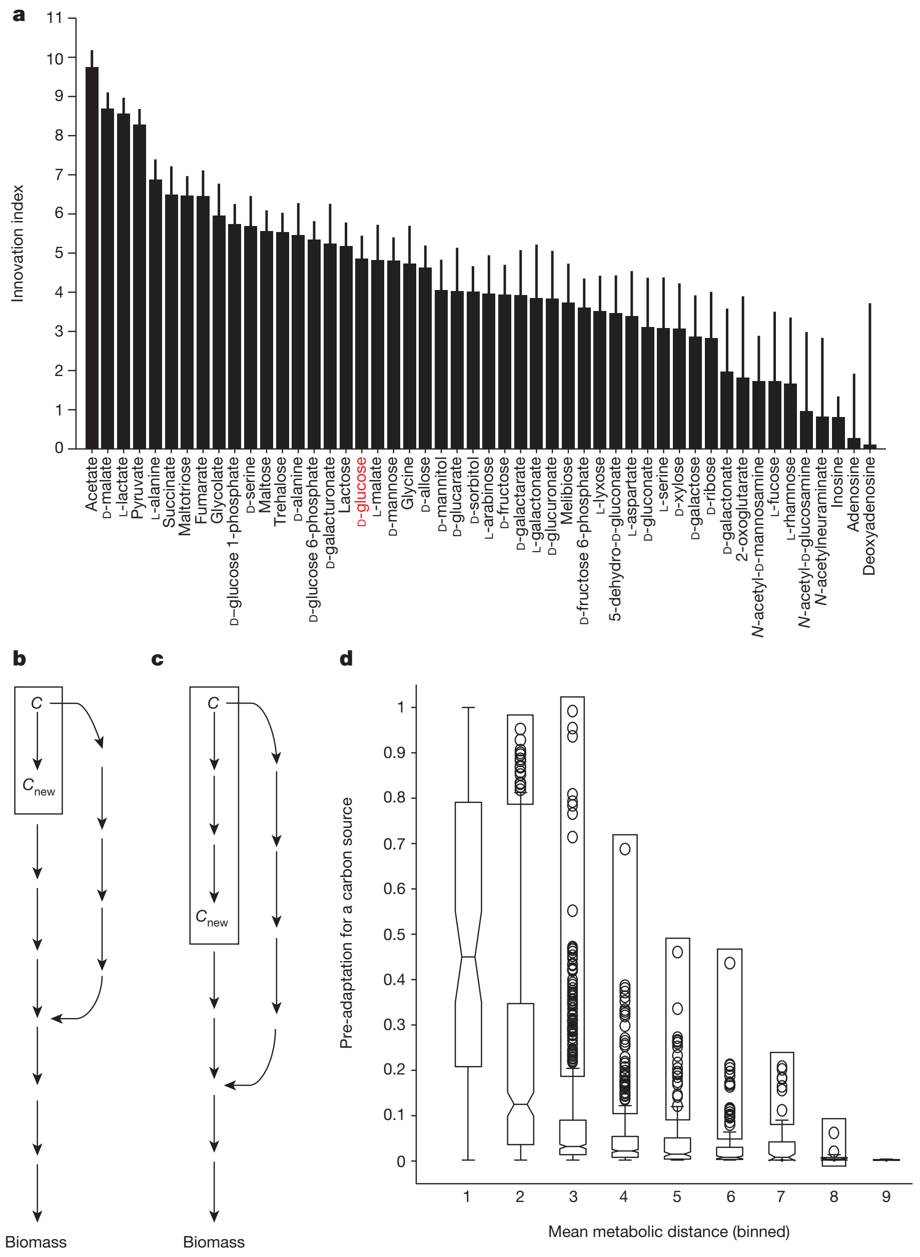

Figure 2 | Innovation varies with respect to the carbon source, *C*, and the mean metabolic distance between *C* and C_{new}. **a**, For each of 50 carbon sources (horizontal axis), the figure indicates the mean innovation index (bar) and its coefficient of variation (vertical line) for 500 random networks required to be viable on that carbon source. Note the broad distribution of the index. Some carbon sources, such as acetate, allow viability on more than nine additional carbon sources, on average, whereas others, such as deoxyadenosine, support viability on fewer than one additional carbon source. The innovation index of glucose (red) is typical compared with other carbon sources. **b**, A hypothetical carbon source, C_{new}, which can be synthesized from another carbon source, *C*, in one reaction (arrow), and which leads, through multiple further reactions, to the synthesis of biomass. Some metabolic networks may have an alternative metabolic pathway that bypasses C_{new} altogether (right-hand sequence of arrows). **c**, Like **b**, but with C_{new} and *C* separated by multiple reactions. The fewer reactions separate *C* and C_{new}, the more likely it is that C_{new} is not bypassed by some alternative metabolic pathway, and that viability on *C* therefore implies viability on C_{new}. **d**, Testing the hypothesis in **c**. The horizontal axis shows the mean number of reactions that separate *C* and C_{new} in networks that are viable on both *C* and C_{new}, binned into integer intervals according to the floor of this number (that is, the greatest smaller integer). The vertical axis shows the fraction of random metabolic networks required to be viable on carbon source *C* that are additionally viable on C_{new}. We note that the potential for innovation decreases with increasing distance. Box edges, 25th and 75th percentiles; central horizontal line in each box, median; whiskers, ±2.7 s.d.; open circles, outliers. Data are based on samples of 500 random viable networks for each of 50 carbon sources *C* ($n = 25{,}000$).

this is not the case, the incidence of exaptation may be reduced. In this regard, we note that 84% of *E. coli* transporters can transport multiple molecules[17] and that their substrate specificity can change rapidly[18], thus ameliorating this constraint. Fourth, real metabolic networks may contain more reactions connected to the rest of metabolism than do our randomly sampled networks. However, when restricting our analysis to networks in which all reactions are connected, we found an even greater incidence of exaptation than in random networks (Methods, Supplementary Results and Supplementary Fig. 3). Thus, our results provide a lower bound on the incidence of exaptations. Finally, most of our analysis is based on sampling a limited number of 500 networks viable on each carbon source, but sampling 5,000 random networks for select carbon sources yielded identical results (Supplementary Fig. 12).

Our observations show that latent metabolic abilities are pervasive features of carbon metabolism. They expose non-adaptive origins of potentially useful carbon-source utilization traits as a universal and inevitable feature of metabolism. The abundance of non-adaptive trait origins results from the complexity of metabolic systems, which have many enzyme parts that can jointly form multiple metabolic phenotypes, but this ability is not restricted to metabolic networks. Many enzymes are capable of using various substrates[17,19], which can further increase network complexity and the potential for exaptation. The ability to form multiple phenotypes also occurs in regulatory circuits[20], which can form different patterns of molecular activity, as well as in RNA molecules[21], which can form multiple conformations with different biological functions. Systematic analyses of genotype–phenotype relationships are becoming increasingly possible in such systems[22,23], and already hint at exaptive origins of molecular traits. If confirmed in systematic analyses like ours, the pervasiveness of non-adaptive traits may require a rethinking of the early origins of beneficial traits.

METHODS SUMMARY

We used MCMC random walks that utilize reaction swapping to sample random viable metabolic networks[12], and used flux balance analysis[24] to compute the viability of metabolic networks during the MCMC procedure. We performed all analyses for minimal aerobic growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulphate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc[15].

Full Methods and any associated references are available in the online version of the paper.

Received 6 August 2012; accepted 14 May 2013.

Published online 14 July 2013.

1. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* 6th edn (Murray, 1872).
2. Gould, S. J. & Vrba, E. S. Exaptation: a missing term in the science of form. *Paleobiology* **8,** 4–15 (1982).
3. True, J. R. & Carroll, S. B. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **18,** 53–80 (2002).
4. Zákány, J. & Duboule, D. Hox genes in digit development and evolution. *Cell Tissue Res.* **296,** 19–25 (1999).
5. Keys, D. N. *et al.* Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* **283,** 532–534 (1999).
6. Tomarev, S. I. & Piatigorsky, J. Lens crystallins of invertebrates. Diversity and recruitment from detoxification enzymes and novel proteins. *Eur. J. Biochem.* **235,** 449–465 (1996).
7. Pievani, T. & Serrelli, E. Exaptation in human evolution: how to test adaptive vs exaptive evolutionary hypotheses. *J. Anthropol. Sci.* **89,** 9–23 (2011).
8. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **255,** 279–284 (1994).
9. Lipman, D. J. & Wilbur, W. J. Modelling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B* **245,** 7–11 (1991).
10. Cowperthwaite, M. C., Economo, E. P., Harcombe, W. R., Miller, E. L. & Meyers, L. A. The ascent of the abundant: how mutational networks constrain evolution. *PLoS Comput. Biol.* **4,** e1000110 (2008).
11. Ferrada, E. & Wagner, A. A comparison of genotype-phenotype maps for RNA and proteins. *Biophys. J.* **102,** 1916–1925 (2012).
12. Samal, A., Matias Rodrigues, J. F., Jost, J., Martin, O. C. & Wagner, A. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* **4,** 30 (2010).
13. Poulsen, T. S., Chang, Y.-Y. & Hove-Jensen, B. D-allose catabolism of *Escherichia coli*: involvement of alsI and regulation of als regulon expression by allose and ribose. *J. Bacteriol.* **181,** 7126–7130 (1999).
14. Meijnen, J.-P., De Winde, J. H. & Ruijssenaars, H. J. Engineering Pseudomonas putida S12 for efficient utilization of D-xylose and L-arabinose. *Appl. Environ. Microbiol.* **74,** 5031–5037 (2008).
15. Feist, A. M. *et al.* A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3,** 121 (2007).
16. Neidhardt, F. & Ingraham, J. *Escherichia Coli and Salmonella Typhimurium: Cellular and Molecular Biology* Vol. 1, 13–16 (American Society for Microbiology, 1987).
17. Nam, H. *et al.* Network context and selection in the evolution to enzyme specificity. *Science* **337,** 1101–1104 (2012).
18. Aguilar, C. *et al.* Genetic changes during a laboratory adaptive evolution process that allowed fast growth in glucose to an *Escherichia coli* strain lacking the major glucose transport system. *BMC Genomics* **13,** 385 (2012).
19. Kim, J., Kershner, J. P., Novikov, Y., Shoemaker, R. K. & Copley, S. D. Three serendipitous pathways in *E. coli* can bypass a block in pyridoxal-5′-phosphate synthesis. *Mol. Syst. Biol.* **6,** 436 (2010).
20. Martin, O. C. & Wagner, A. Multifunctionality and robustness trade-offs in model genetic circuits. *Biophys. J.* **94,** 2927–2937 (2008).
21. Ancel, L. W. & Fontana, W. Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool.* **288,** 242–283 (2000).
22. Amitai, G., Gupta, R. D. & Tawfik, D. S. Latent evolutionary potentials under the neutral mutational drift of an enzyme. *HFSP J.* **1,** 67–78 (2007).
23. Isalan, M. *et al.* Evolvability and hierarchy in rewired bacterial gene networks. *Nature* **452,** 840–845 (2008).
24. Price, N. D., Reed, J. L., Palsson, B. & Ø. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Rev. Microbiol.* **2,** 886–897 (2004).

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

Acknowledgements We thank K. Raman, N. Sabath, J. L. Payne and J. Rodrigues for discussions. A.W. acknowledges financial support through Swiss National Science Foundation grant 315230-129708.

Author Contributions A.B. and A.W. designed the research; A.B. performed the research; A.B. contributed new reagents and analytic tools; A.B. and A.W. analysed the data; and A.B. and A.W. wrote the paper.

Author Information 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 A.W. (andreas.wagner@ieu.uzh.ch).

METHODS

Flux balance analysis. Flux balance analysis (FBA) is a constraint-based computational method[24,25] used to predict synthetic abilities and other properties of large metabolic networks, which are complex systems of enzyme-catalysed chemical reactions. FBA requires information about the stoichiometry of each molecular species participating in the chemical reactions of a metabolic network. This stoichiometric information is represented as a stoichiometric matrix, S, of dimensions $m \times n$, where m denotes the number of metabolites and n denotes the number of reactions in a network[24,25]. FBA also assumes that the network is in a metabolic steady state, such as would be attained by an exponentially growing microbial population in an unchanging environment. This assumption makes it possible to impose the constraint of mass conservation on the metabolites in the network. This constraint can be expressed as $Sv = 0$, where v denotes a vector of metabolic fluxes whose entries, v_i, describe the rate at which reaction i proceeds. The solutions, or 'allowable' fluxes, of this equation form a large solution space, but not all of these solutions may be of biological interest. To restrict this space to fluxes of interest, FBA uses linear programming to maximize a biologically relevant quantity in the form of a linear objective function Z (ref. 25). Specifically, the linear programming formulation of an FBA problem can be expressed as

$$\max\{Z\} = \max\{c^{\mathrm{T}}v \mid Sv = 0, a \leq v \leq b\}$$

The vector c contains a set of scalar coefficients that represent the maximization criterion, and the individual entries of vectors a and b respectively contain the minimal and maximal possible fluxes for each reaction in v; that is, each entry v_i is bounded from below by a_i and bounded from above by b_i.

We are here interested in predicting whether a metabolic network can sustain life in a given spectrum of environments, that is, whether it can synthesize all necessary small biomass molecules (biomass precursors) required for survival and growth. In a free-living bacterium such as *E. coli*, there are more than 60 such molecules, which include 20 proteinaceous amino acids, DNA and RNA nucleotide precursors, lipids and cofactors. We use the *E. coli* biomass composition[15] to define the objective function and the vector c, because most molecules in *E. coli*'s biomass would be typically found in free-living organisms. We used the package CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve the linear programming problems mentioned above.

Chemical environments. Along with the biomass composition and stoichiometric information about a metabolic network, it is necessary to define one or more chemical environments that contain the nutrients needed to synthesize biomass precursors. Here we consider only minimal aerobic growth environments composed of a sole carbon source, along with oxygen, ammonium, inorganic phosphate, sulphate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese and zinc[15]. When studying the viability of a metabolic network in different environments, we vary the carbon source while keeping all other nutrients constant. When we say, for example, that a particular network is viable on 20 carbon sources, we mean that the network can synthesize all biomass precursors when each of these carbon sources is provided as the sole carbon in a minimal medium. For reasons of computational feasibility, we restrict ourselves to 50 carbon sources (Supplementary Table 1). They are all carbon sources on which *E. coli* is known to be viable from experiments[15]. We chose these carbon sources because many of them are prominent, and because they are of known biological relevance, but we emphasize that our observations do not otherwise make a statement about the metabolism of *E. coli* or its close relatives. They apply to metabolic networks that vary much more broadly in reaction composition than any relative of *E. coli*, because of our network sampling approach described below, which effectively randomizes the reaction composition of a microbial metabolism.

The known reaction universe. The known reaction universe is a list of metabolic reactions known to occur in some organisms. For the construction of this universe, we used data from the LIGAND database[26,27] of the Kyoto Encyclopedia of Genes and Genomes[28,29]. The LIGAND database is divided into two subsets—the REACTION database and the COMPOUND database. These two databases together provide information about metabolic reactions, participating chemical compounds and associated stoichiometric information in an interlinked manner.

As we described earlier[12,30,31], we specifically used the REACTION and COMPOUND databases to construct our universe of reactions while excluding all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry; reactions involving glycans, owing to their complex structure; reactions with unbalanced stoichiometry; and reactions involving complex metabolites without chemical information[29]. The published *E. coli* metabolic model (*i*AF1260) consists of 1,397 non-transport reactions[15]. We merged all reactions in the *E. coli* model with the reactions in the LIGAND database, and retained only the non-duplicate reactions. After these procedures of pruning and merging, our universe of reactions consisted of 5,906 non-transport reactions and 5,030 metabolites.

Sampling of random viable metabolic networks. In an organism, a metabolic network can change through mutations. They can lead to addition of new reactions, by way of horizontal gene transfer, or through the evolution of enzymes with novel activities. They can also lead to loss of reactions through loss-of-function mutations in enzyme-coding genes. Natural selection can preserve those changed metabolic networks that are viable in a particular environment. Together, mutational processes and selection may change a metabolic network drastically on a long evolutionary timescale. Recent work has shown that even metabolic networks that differ greatly in their sets of reactions can have the same metabolic phenotype, that is, the same biosynthetic ability[32]. We here use a recently developed MCMC random-sampling[12,30,31,33,34] procedure to generate metabolic networks that are viable in specific environments, but that contain an otherwise random complement of metabolic reactions. Briefly, this procedure involves random walks in the space of all possible networks. During any one such random walk, a metabolic network can change through the addition and deletion of reactions. Although this process resembles the biological evolution of metabolic networks through horizontal gene transfer and (recombination-driven) gene deletions, we here use it for the sole purpose of creating random samples of metabolic networks from the space of all such networks[12,34].

In any one MCMC random walk, we keep the total number of reactions at the same number as in the starting *E. coli* network (1,397; ref. 15), to avoid artefacts due to varying reaction network size[12]. Specifically, each mutation step in a random walk involves the addition of a randomly chosen reaction from the reaction universe, followed by the deletion of a randomly chosen metabolic reaction from the metabolic network. We call such a sequence of reaction addition and deletion a reaction swap. Reaction addition does not abolish the viability of a network in any environment. However, reaction deletion might. Thus, after a reaction deletion, we use FBA to ask whether the network is still viable, that is, whether it can synthesize all biomass precursors, in the specified environment. If so, we accept the deletion; otherwise, we reject it and choose another reaction for deletion at random, until we have found a deletion that retains viability. After that, we accept the reaction swap, thus completing a single step in the random walk. We do not subject transport reactions to reaction swaps. These reactions are therefore present in all networks generated by our random walk.

Any MCMC random walk begins from a single starting network, in our case that of *E. coli*. The theory behind MCMC sampling[12,34] shows that it is important to carry out as many reaction swaps as possible for MCMC to 'erase' the random walker's similarity ('memory') to the initial network. The reason is that successive genotypes in a random walk are strongly correlated in their properties, because they differ by only one reaction pair. These correlations decrease as the number of reaction swaps increases. Because we are interested in analysing growth phenotypes of networks, correlations to the initial network would result in identification of growth on carbon sources similar to those of the starting network. In past work[12,30], we found that for the network sizes that we use (1,397 reactions), 3×10^3 reactions swaps are sufficient to erase the similarity of the final network to the starting network. To err on the side of caution, we thus carried out 5×10^3 reaction swaps before beginning to sample, and sampled a network every 5×10^3 reaction swaps thereafter. In this way, we generated samples of 500 random viable metabolic networks through an MCMC random walk of 2.5×10^6 reaction swaps. We carried out different random walks to sample networks viable on different carbon sources.

For some of our analyses, we also sampled random metabolic networks of sizes different from that of the *E. coli* metabolic network. To do this, we followed a previously established procedure[12,30,31] to create a starting network for an MCMC random walk that has the desired size. This procedure first converts the known universe of reactions into a 'global' metabolic network by including the *E. coli* transport reactions in it. Not surprisingly, this global network can produce all biomass components and is therefore viable on all carbon sources studied here. We used this global network to delete successively a sequence of randomly chosen reactions in the following way. After each reaction deletion, FBA was used to determine whether the network was still viable on a given carbon source. If so, the deletion was accepted; otherwise, another reaction was chosen at random for deletion. We deleted in this way as many reactions as needed to generate a network of the desired size. We then used this network as the starting network for an MCMC random walk, as described above, to generate samples of 500 random viable networks.

Identification of disconnected non-functional reactions. We performed some of our analysis with a version of the reaction universe that does not contain disconnected reactions. Reactions that are not connected to the rest of a metabolic network would be non-functional, because they cannot carry a non-zero steady-state metabolic flux, and thus could not contribute to the synthesis of biomass. The genes encoding them would eventually be lost from a genome. (We note that this loss could still take tens of thousands of years, given known deleterious mutation rates and generation times[35,36], which is enough for some for other genetic or environmental changes to render these reactions functional.) We define a disconnected reaction as a reaction that does not share any one substrate or any one product with any other reaction in the known reaction universe. We focus here on reactions in the universe rather than in one metabolic network, because an individual network can gain additional reactions that may connect previously disconnected reactions. We note that even this 'universal' definition of disconnectedness depends on our current knowledge of biochemistry, as well as on the environment, because the right environment could supply metabolites that connect previously disconnected reactions or pathways to the rest of a metabolic network. To identify the connected universe, we removed disconnected reactions. Because this removal may render other reactions disconnected, we repeated this process iteratively until no further reactions in the universe became disconnected. In this way, we found that 3,646 of the 5,906 reactions in the universe of reactions were connected. We used this connected universe in some analyses to generate network samples using the MCMC approach.

Estimation of the metabolic distance between carbon sources. To compute the metabolic distance between a pair of carbon sources, C and C_{new}, we used the 500 networks selected for growth on a specific carbon source, C. We first represented a network as a substrate graph[37]. In this graph, vertices correspond to metabolites. Two metabolites (vertices) are linked by an edge if the metabolites participate in the same metabolic reaction, be it as an educt or as a product. We excluded 'currency' metabolites from this substrate graph, which are metabolites that transfer small chemical groups and are involved in many reactions[38]. Specifically, we excluded protons, H_2O, ATP (adenosine triphosphate), ADP (adenosine diphosphate), AMP (adenosine monophosphate), NADP(H) (nicotinamide adenosine dinucleotide diphosphate), NAD(H) (nicotinamide adenosine dinucleotide), and P_i (inorganic phosphate), CoA (coenzyme A), hydrogen peroxide, ammonia, ammonium, bicarbonate, GTP (guanosine triphosphate), GDP (guanosine diphosphate), and PP_i (inorganic diphosphate) that occurred in both the cytoplasmic and periplasmic compartments[15]. In addition, we excluded oxidized and reduced forms of cofactors such as quinone, ubiquinone, glutathione, thioredoxin, flavodoxin and flavin mononucleotide. That is, we eliminated all vertices corresponding to these metabolites when constructing the substrate graph. For each metabolic network, we constructed two substrate graphs: one in which the reaction irreversibility was ignored and all reactions were considered reversible, and one in which irreversibility was taken into account. For a network selected for growth on carbon source C, we calculated the shortest distance from C to each exapted carbon source, C_{new}, in the substrate graph of that network, as computed by a breadth-first search[39]. We preformed this analysis for each network in our ensemble of 500 networks viable on C. The distance between C and C_{new} was then computed as the mean of the metabolic distances based on networks viable on both carbon sources.

We also computed the metabolic distance for any two carbon sources by representing the universe of reactions as a graph in the above manner. We again constructed two substrate graphs, as above. Taking irreversibility into account increases the maximal distance to infinity because some carbon sources are connected by irreversible reactions.

Clustering of carbon sources based on the innovation matrix. Entry I_{ij} of the innovation matrix, I, represents the fraction of random metabolic networks that we required to be viable on carbon source C_i and that was additionally viable on carbon source C_j. To cluster the entries of this matrix, we first computed for all pairs of rows in this matrix the quantity $d = 1 - \rho$, where ρ is the Spearman rank correlation coefficient between the row entries. This yielded a new distance matrix that describes the distances between all pairs of rows. We clustered the rows of I by applying UPGMA (unweighted pair group method with arithmetic means[40]), a hierarchical clustering method, to the distance matrix.

Hierarchical clustering with UPGMA classifies data such that the average distance between elements belonging to the same cluster is lower than the average distance between elements belonging to different clusters[12]. UPGMA identified two clusters of glycolytic and gluconeogenic carbon sources, and we wanted to know whether the distances between them were significantly different. To this end, we first calculated the distribution of distances $d = 1 - \rho$ for all pairs of row vectors of I within each of the two clusters. We called the resulting distance distribution the 'within-cluster' distance distribution. Similarly, we computed the distances between any pair of row vectors belonging to two different clusters. These formed a 'between-cluster' distance distribution. We then used the non-parametric Mann–Whitney U-test to check whether these two distributions were significantly different.

Estimation of carbon waste production. FBA determines the maximal biomass yield achievable by a network for a given carbon source[25]. However, even when a network produces the maximally achievable yield, not all of the carbon input into the network may be converted into biomass. The non-converted carbon input constitutes carbon waste. Such unused carbon can be secreted in the form of one or more metabolites. For example, in a glucose minimal environment *E. coli* secretes carbon dioxide and acetate into the extracellular compartment as carbon waste. FBA estimates the amount of each metabolite secreted per unit time[15,25]. To estimate the amount of carbon waste that a random network viable on glucose produces, we first identified the different metabolites that it secretes as waste and then computed the amount of carbon waste per metabolite as the product of carbon atoms in that metabolite and the amount of the metabolite secreted (millimoles per gram dry weight per hour). The total carbon waste produced by a network is computed as the sum of the above quantity for each of the secreted carbon-containing molecules. We repeated the above procedure for each random network in a sample of 500 random networks viable on glucose. We found a total of 62 metabolites that are secreted as waste metabolites in at least one network of our sample of networks viable on glucose.

We carried out all numerical analyses using MATLAB (Mathworks Inc.).

25. Kauffman, K. J., Prakash, P. & Edwards, J. S. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14,** 491–496 (2003).
26. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* **28,** 380–382 (2000).
27. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. & Kanehisa, M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* **30,** 402–404 (2002).
28. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
29. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M. & Hirakawa, M. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* **38,** D355–D360 (2010).
30. Barve, A., Rodrigues, J. F. M. & Wagner, A. Superessential reactions in metabolic networks. *Proc. Natl Acad. Sci. USA* **109,** E1121–E1130 (2012).
31. Matias Rodrigues, J. F. & Wagner, A. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5,** e1000613 (2009).
32. Henry, C. S. *et al.* High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nature Biotechnol.* **28,** 977–982 (2010).
33. Matias Rodrigues, J. F. & Wagner, A. Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Syst. Biol.* **5,** 39 (2011).
34. Binder, K. & Heerman, D. W. *Monte Carlo Simulation in Statistical Physics* (Springer, 2010).
35. Koskiniemi, S., Sun, S., Berg, O. G. & Andersson, D. I. Selection-driven gene loss in bacteria. *PLoS Genet.* **8,** e1002787 (2012).
36. Ochman, H., Elwyn, S. & Moran, N. A. Calibrating bacterial evolution. *Proc. Natl Acad. Sci. USA* **96,** 12638–12643 (1999).
37. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. R. Soc. Lond. B* **268,** 1803–1810 (2001).
38. Ma, H.-W. & Zeng, A.-P. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics* **19,** 1423–1430 (2003).
39. Moore, E. in *Proc. Internat. Symp. Theory Switching, Ann. Comput. Lab. Harvard Univ.* 285–292 (Harvard Univ. Press, 1959).
40. Sokal, R. R. & Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **28,** 1409–1438 (1958).