Biophysical Journal
Article

Biophysical Society

Constraint and Contingency Pervade the Emergence of Novel Phenotypes in Complex Metabolic Systems

Sayed-Rzgar Hosseini[1,2] and Andreas Wagner[1,2,3,*]
[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland; [2]The Swiss Institute of Bioinformatics, Bioinformatics, Lausanne, Switzerland; and [3]The Santa Fe Institute, Santa Fe, New Mexico

ABSTRACT An evolutionary constraint is a bias or limitation in phenotypic variation that a biological system produces. We know examples of such constraints, but we have no systematic understanding about their extent and causes for any one biological system. We here study metabolisms, genomically encoded complex networks of enzyme-catalyzed biochemical reactions, and the constraints they experience in bringing forth novel phenotypes that allow survival on novel carbon sources. Our computational approach does not limit us to analyzing constrained variation in any one organism, but allows us to quantify constraints experienced by any metabolism. Specifically, we study metabolisms that are viable on one of 50 different carbon sources, and quantify how readily alterations of their chemical reactions create the ability to survive on a novel carbon source. We find that some metabolic phenotypes are much less likely to originate than others. For example, metabolisms viable on D-glucose are 1835 times more likely to give rise to metabolisms viable on D-fructose than on acetate. Likewise, we observe that some novel metabolic phenotypes are more contingent on parental phenotypes than others. Biochemical similarities among carbon sources can help explain the causes of these constraints. In addition, we study metabolisms that can be produced by recombination among 55 metabolisms of different bacterial strains or species, and show that their novel phenotypes are also contingent on and constrained by parental genotypes. To our knowledge, our analysis is the first to systematically quantify the incidence of constrained evolution in a broad class of biological system that is central to life and its evolution.

INTRODUCTION

Individual organisms or populations cannot produce every conceivable kind of phenotypic variation. In other words, phenotypic evolution is to some extent constrained. More precisely, an evolutionary constraint is a bias or limitation in the emergence of phenotypic variation in a given biological system (1). Examples of constraints on the organismal level include the absence of photosynthesis in higher animals, the absence of birds that can give birth to live young instead of to eggs, the general lack of teeth in the lower jaw of frogs, and the absence of palm trees in cold climates (1,2). Other examples include constrained variation in segment number, orientation and identity in the fruit fly *Drosophila melanogaster* (3), and correlations among different characters, such as in allometric scaling (4). Molecular examples of phenotypic constraints include the absence of L-isomers in the 20 amino acids found in natural proteins (5), and a limited number of possible protein folds caused by the packing requirements of hydrophobic amino acids (6). It is useful to distinguish between absolute constraints, which occur when some phenotype cannot be produced, and relative constraints, when some phenotypes are more likely to arise than others.

A closely related concept is that of contingency. We speak of contingency when the origin of a novel phenotype depends on the history of a population, and specifically on preexisting genotypes or phenotypes (7,8). For example, experimental evolution of *Escherichia coli* has shown that the emergence of citrate-utilization as a novel metabolic phenotype is contingent on the genetic history of a population (9). Analogously to constraints, one can distinguish between phenotypes that are absolutely or relatively contingent on evolutionary history. Although many anecdotal examples of constraints and contingent evolution exist, such examples do not allow one to quantify the potential for either phenomenon in any one class of biological system.

Submitted December 5, 2016, and accepted for publication June 19, 2017.
*Correspondence: andreas.wagner@ieu.uzh.ch
Editor: Reka Albert.
http://dx.doi.org/10.1016/j.bpj.2017.06.034

We here undertake such a quantification using a computational approach applied to metabolic systems, which are ideal for this purpose for several reasons.

First, metabolic systems, and especially those of microbes, are an abundant source of new adaptations and innovations (i.e., qualitatively new adaptations). Especially important innovations are those that allow an organism to extract energy and chemical elements from new molecules, which can help it survive in new habitats. For instance, microorganisms have acquired the ability to utilize many nonnatural substances, such as polychlorinated biphenyls, chlorobenzenes, organic solvents, synthetic pesticides, and even antibiotics as food (10–14).

Second, experimentally validated computational methods such as flux balance analysis (FBA) provide efficient means to systematically predict metabolic phenotypes—the ability of an organism to survive on specific nutrients—from information about metabolic genotypes (15,16). A metabolic genotype is the part of a genome encoding metabolic enzymes. However, computational analyses of metabolic systems often use a more abstract and compact representation of such a genotype, referring to it as the collection of chemical reactions that a metabolic reaction network is able to catalyze (17–26).

Third, in metabolic systems, we are not restricted to studying the metabolism of any one organism, together with the constraints and contingencies it may be subject to. Instead, we can study the potential for contingency and constraint in entire classes of metabolic systems. To do so, we take advantage of Markov chain Monte Carlo (MCMC) algorithms (21,23) (see Materials and Methods) that allow us to create large numbers of metabolisms. Each such metabolism is a complex network of chemical reactions with a given phenotype, but its complement of metabolic reactions is otherwise sampled at random from a universe of metabolic reactions that are known to exist among prokaryotes (see Materials and Methods). We refer to such metabolisms as "random viable metabolic networks". The phenotypes we study are viability phenotypes, and specifically a metabolism's ability to synthesize all essential biomass precursors in a minimal medium that harbors only a single carbon source. We consider 50 such carbon sources, i.e., 50 different metabolic phenotypes.

When analyzing phenotypic variability, it is important to consider the kinds of genotype changes that cause this variability. We focus on recombination-like processes as a means for genotypic change, and do so for two reasons. First, recombination is a ubiquitous force of genetic change, not only in eukaryotes but also in prokaryotes whose genomes are being continually reorganized through horizontal gene transfer. Second, in contrast to smaller-scale genetic change, such as point mutations, recombination causes larger-scale genetic change with greater potential to create novel phenotypes (27–32). Thus, if we found that phenotypic evolution was constrained when recombination causes genotypic change, it would be even more constrained if point mutations caused such change.

In our simulations, we generated 1000 parental pairs of random viable metabolic networks for each of the 50 carbon utilization phenotypes. For each one of these 50,000 parental pairs, using a recombination-like process that mimics horizontal gene transfer in bacteria (see Materials and Methods), we generated 1000 offspring to obtain 50,000,000 recombinant metabolic networks. We focused on those recombinants that did not only retain viability on their parental carbon source, but also gained viability on at least one novel carbon source. For brevity, we will also refer to them as "innovative offspring". We analyze their phenotypes and how they depend on parental phenotypes. In addition, we also study recombination among metabolic networks of 55 bacterial species or strains.

We find little evidence for absolute constraints and contingencies. That is, the metabolic phenotypes we consider can be brought forth through recombination among some parental metabolic networks. However, relatively constraints and contingencies are pervasive. Differences in the biochemical relatedness of carbon sources, and the ensuing correlations among different carbon usage phenotypes, can help explain some of these constraints and contingencies.

MATERIALS AND METHODS

Genotype-phenotype representation in metabolic networks

The set of enzyme-catalyzed biochemical reactions that take place in an organism constitutes the organism's metabolic reaction network, i.e., its metabolism. Each such metabolism contains a subset of the reaction universe of all biochemical reactions that are known to occur in some organism within the biosphere. We have manually curated a representation of the prokaryotic reaction universe, which comprises 5906 reactions known to occur in prokaryotes (see Supporting Materials and Methods for details). In this framework, we represent an organism's metabolic genotype as a binary vector of length 5906, each entry of which corresponds to a given reaction in the universe, and is equal to 1 if the corresponding reaction is present in the network, and 0 otherwise. Hence, each genotype can be envisioned as a single member of a vast space of all possible metabolic networks, which contains 2^{5906} distinct genotypes. We determine the phenotype of a given metabolic genotype based on its ability to sustain life in one or more of 50 distinct minimal environments that differ only in the sole carbon source they contain (Supporting Materials and Methods). We consider a genotype viable on a given carbon source, if FBA (see Supporting Materials and Methods) predicts that it can produce all essential biomass precursors using this carbon source as its only carbon source (15). We used the biomass composition of the *E. coli* metabolic model iAF1260, because the sampling approach described in the next section starts from the *E.coli* metabolism (Supporting Materials and Methods). Our C++ implementation of FBA and the code necessary for the analyses in this article are available through this public github repository: https://github.com/rzgar/EMETNET.

Random sampling of parental metabolic network pairs from metabolic genotype space

We here employ a previously described in silico process that relies on MCMC random walks to generate randomly sampled viable metabolic networks, i.e., networks that are viable on a given carbon source, but that otherwise contain a random subset of reactions in the reaction universe

(Supporting Materials and Methods) (21,23). This procedure ensures uniform sampling from the set of all metabolic networks viable on a given carbon source. Our analyses required us to recombine pairs of parental metabolic networks (i.e., donor-recipient pairs) with particular features, such as a given genotypic distance (D), defined as the number of reactions differing between the parents. We used simultaneous genotype-converging MCMC random walks to generate pairs of metabolic networks with a given D (see Supporting Materials and Methods). We required parental metabolisms to be exclusively viable on a particular carbon source, i.e., to be inviable on all 49 other carbon sources we considered. In most of our analyses, we kept the number of reactions present in the metabolic networks constant and equal to that of *E. coli* with 2079 reactions.

Modeling a recombination-like process in metabolic networks

As in a previous contribution (32), we use a coarse-grained model of prokaryotic recombination that mimics the effects of horizontal gene transfer events between bacteria on metabolism (33–36). This model is motivated by the importance of horizontal gene transfer as a means of genetic change. Through its high incidence, horizontal gene transfer can change the gene content of genomes on short evolutionary timescales (33,37,38). It can also occur between very distantly related organisms (39,40). For several reasons, our recombination model also takes DNA deletions into consideration. The first is that during horizontal gene transfer, incorporating genes from a donor into a recipient genome relies on DNA rearrangements that can also delete resident genes (41). Second, the majority of newly acquired genes obtained via horizontal gene transfer reside in the genome only for short amounts of time (42). Third, the evolution of prokaryotic genomes is biased toward DNA deletions (43). Motivated by these observations, we here model prokaryotic recombination as a process where the transfer of biochemical reactions from a donor to a recipient is accompanied by concurrent deletion of reactions from the recipient metabolic network.

Specifically, to model recombination for each parental metabolic network pair, we generated 1000 recombinant offspring by 1) adding to the recipient metabolic network a given number $n/2$ of randomly chosen reactions that were present in the donor and absent from the recipient, followed by 2) deleting $n/2$ reactions randomly chosen from the recipient. Thus, the total number of reactions changed by a recombination event in the recipient is equal to n. In this contribution, we repeated most of our analyses by using three different values of n; namely $n = 10$, 20, and 30. Empirical observations also suggest that altering up to $n = 60$ reactions in a recombination event is biologically realistic, because horizontal gene transfer can affect long DNA regions (44). Importantly, the transferred material that is integrated into the host genome by recombination can constitute stretches of noncoding DNA, fragments of genes (45,46), entire genes (47), multiple adjacent genes (48,49), operons, transposable chromosomal elements, and plasmids, as well as other naturally occurring extrachromosomal elements (50). The length of contiguously transferred stretches may range from a few nucleotides (51) to >3 Mbp (44), i.e., some two-thirds of the length of the *E. coli* genome, which encodes >1300 reactions. In addition, some megabase-scale horizontally transferred DNA segments can become incorporated into a chromosome in the form of hundreds of smaller fragments (52). As we have discussed in a previous contribution ((32); Supporting Materials and Methods), the probability that a recombination event preserves viability exceeds 10^{-3} for values up to $n = 60$.

Modeling recombination in curated bacterial metabolic networks from the BiGG database

We used the R-package Sybil (53) to collect 55 well-annotated bacterial genome-scale metabolic networks available in the BiGG database (54). Each of these species or strains has its own biomass growth function, its own complement of reactions, and well-defined gene-reaction association rules that allowed us to model recombination on the level of genes instead of reactions. We used the genomic location of metabolic genes in these bacterial species or strains (55) to take gene linkage into account when modeling recombination.

To generate a recombinant metabolic network from a donor and a recipient organism, first a given stretch of DNA from the donor genome that contains a given number of metabolic genes is translated into reactions based on the gene-reaction association rules of the donor organism, and then the resulting reactions are added to the recipient metabolic network. Second, a given stretch of DNA from the recipient genome that contains a given number of metabolic genes is translated into reactions based on the gene-reaction association rules of the recipient organism, and then the resulting reactions are deleted from the recipient metabolic network.

In a recombination event between a pair of organisms, we set the number of genes in a given donor DNA stretch such that on average a given number of $n = 5$ reactions are added to the recipient metabolic network, and on average an equal number $n = 5$ of reactions are deleted from it. Because gene-reaction associations are not generally one-to-one and can be very complicated, and because most of the reactions that are encoded in a given stretch of DNA may already be present in the recipient metabolic network, the number of metabolic genes that needs to be added from donor to recipient genome, such that exactly n reactions are added to the recipient, will often be higher than n. In contrast, we found that the number of metabolic genes in a DNA stretch to be deleted from the recipient genome to eliminate n reactions from the recipient metabolic network is lower than n, because deletion of a single metabolic gene often causes elimination of multiple reactions.

More specifically, we modeled recombination among all distinct pairs of donor-recipient bacterial species or strains in our analysis (55×54 pairs). From each given pair we generated a recombinant offspring by adding a given (p) number of consecutive metabolic genes from the donor genome, followed by deleting a given (q) number of consecutive metabolic genes from the recipient genome. Importantly, we examined all possible combinations of (p) consecutive genes from the donor and (q) consecutive genes from the recipient. Thus, for a donor genome with n metabolic genes, and a recipient genome with m metabolic genes, we generated all $(n - p + 1) \times (m - q + 1)$ recombinant offspring, a number that exceeded 1,000,000 offspring for most pairs. Note that (p) and (q) are selected based on the gene-reaction association rules of the donor and recipient species or strains to ensure that any one recombination event adds on average five new reactions and deletes five reactions from the recipient metabolic network.

To study the effect of linkage on the emergence of novel phenotypes, we followed a second recombination procedure that neglects linkage between metabolic genes. That is, we added or deleted reactions randomly, just as we had done for randomly sampled metabolic networks, irrespective of the genomic position of the metabolic genes encoding these reactions. To do so, we examined all distinct donor-recipient pairs, and from each pair we generated the same number ($(n - p + 1) \times (m - q + 1)$) of recombinant offspring as in the linkage-based approach, ensuring that on average five randomly chosen reactions are added from the donor and deleted from the recipient metabolic network.

To identify innovative offspring among all the generated recombinants, we used 30 carbon-containing metabolites on which none of the 55 bacterial species or strains are predicted to be viable (listed in Supporting Materials and Methods). To predict viability of a recombinant metabolic network using FBA, we used the objective function of the recipient, because recombinants are much more similar to the recipient than to the donor.

RESULTS

All metabolic phenotypes can emerge through recombination

Our first analysis focused on the perhaps most fundamental question regarding absolute constraints: Do some parental

phenotypes not give rise to any offspring with novel phenotypes? To find out, we quantified for each carbon source C_i and for each of the 1000 parental pairs viable on C_i, the number $N_{C_i\rightarrow}$ of offspring gaining viability on some new carbon source (C_j, $j \neq$ i, among their 10^6 recombinant offspring, with $n = 10$ altered reactions relative to the parents). Fig. 1 *a* shows the distribution of this number, demonstrating that offspring with metabolic innovations can emerge from each of the 50 carbon usage phenotypes we analyzed. However, we also note that the number of offspring with a metabolic innovation varies greatly among different carbon usage phenotypes, ranging from 1433 for parents viable on adenosine to 61,835 for parents viable on D-galactose (per 1,000,000 offspring). We repeated this analysis by varying the number of reactions (n) altered during recombination, which shows that the relative abundance and the ranking of carbon sources in terms of the frequency of innovation stays almost the same for various n ((Figs. S1 and S2); $n = 10$ and 20, Spearman's $R = 0.9982$; $p < 10^{-60}$; $n = 10$ and 30, Spearman's $R = 0.9750$; $p < 10^{-33}$.) In sum, all parental phenotypes we consider can give rise to metabolic innovations.

Next, we asked whether different carbon usage phenotypes differ in their propensity to be found as novel offspring phenotypes, regardless of the parental phenotype. Fig. 1 *b* shows that this is indeed the case. But whereas all 50 carbon-usage phenotypes appear in the innovative offspring we analyzed, their prevalence ($N_{\rightarrow C_i}$) varies by a factor 16

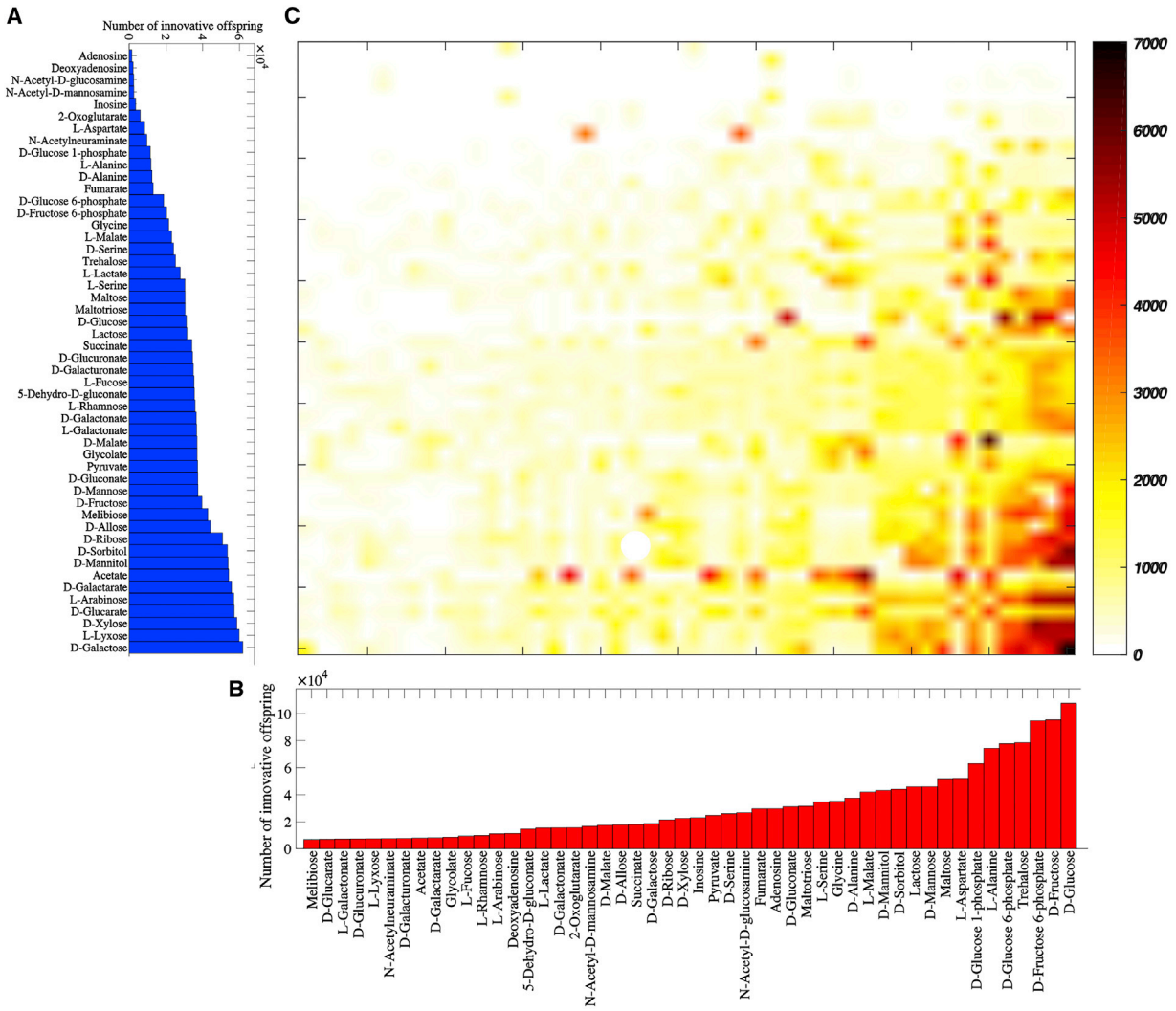


FIGURE 1 Recombination can create all 50 carbon-use phenotypes considered here. (*A*) The horizontal axis shows the number of innovative recombinant offspring (out of 1,000,000 offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. (*B*) Shown here is the number of innovative recombinants (per 1,000,000 offspring) gaining viability on the novel carbon source specified on the horizontal axis. (*C*) Shown here is the number of innovative recombinant (per 1,000,000 offspring, coded according to the *color legend*) resulting from recombination between parents viable exclusively on the carbon source specified in (*A*), which have gained viability on the novel carbon source specified in (*B*). In these analyses, parental metabolic networks contain $||G|| = 2079$ reactions, the same number as the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks during recombination.

among carbon sources, ranging from 6783 innovative offspring gaining viability on melibiose to 107,784 gaining viability on D-glucose (among 50×10^6 recombinant offspring, and a total of 1,556,237 innovative offspring). This variability is similarly great with a number (n) of recombined reactions different from $n = 10$ (Figs. S1 and S2). We noted a negative correlation between $N_{C_i \rightarrow}$ and $N_{\rightarrow C_i}$ (Fig. S3), i.e., carbon-usage phenotypes that give rise to more innovative offspring are found less frequently as products of recombinational innovation.

Finally, Fig. 1 *c* shows the variability among different pairs of carbon sources in terms of their propensity for generating innovative offspring. In 2038 pairs (81.52% among the possible 2500 pairs of carbon sources (C_i, C_j), fewer than 1000 innovative recombinant (among 1,000,000 offspring) gain viability on C_j from recombination between parents viable on C_i, and only in 17 pairs (0.68%) do more than 5000 innovative offspring emerge. The largest number of innovative offspring (7071) emerges when parents viable exclusively on D-galactose give rise to offspring that gain viability on D-glucose.

To find out whether parental genotypic distance and the number of reactions in a metabolic network might affect our observations, we repeated our analyses with more divergent parents ($D = 1000$) and smaller metabolic networks (1800 and 1600 reactions, as opposed to the 2079 reactions identical to the number in *E. coli*, which we had used so far). Although recombination gives rise to fewer innovative offspring at higher D and for smaller networks (Fig. S4), the general patterns (Figs. S5, S6, and S7) remain similar to that of Fig. 1.

Also, we had so far recombined parents that were viable on the same carbon source. To find out whether this could affect our observations, we generated recombinational offspring where one parent is viable on glucose, and the other is viable on a different carbon source. We found that recombination again results in fewer innovative offspring (Fig. S8), but leaves the patterns observed in Fig. 1 intact (Figs. S9 and S10).

In sum, each of the 50 carbon usage phenotypes we consider can give rise to metabolic innovations. Conversely, recombinants can acquire viability on each of 50 carbon sources. Thus, at least from this analysis, there is no evidence for absolute constraints on carbon usage phenotypes. However, different carbon usage phenotypes differ greatly in their propensity to arise as metabolic innovations, providing a first line of evidence for relative constraints on metabolic innovation by parental phenotypes.

Novel metabolic phenotypes are relatively constrained by parental phenotypes

Our next analysis goes to the heart of the question we pose. For each of the 50 focal carbon sources C_i, we examined all innovative offspring originating from parents viable on C_i to find out whether gaining viability on each of the other 49 carbon sources (C_j, $j \neq i$) is possible. For 43 of the 50 carbon sources C_i, this is the case (Fig. 2 *a*). That is, for such a parental carbon source C_i, at least one innovative offspring exists that gains viability on some new carbon source (C_j, $j \neq i$). Even for the remaining seven carbon sources C_i, this holds for the majority of the carbon sources C_j. That is, starting from viability on five of the seven carbon sources C_i, recombination can produce viability on more than 40 of the 49 carbon sources C_j. The remaining carbon sources C_i are deoxyadenosine and adenosine, where recombination can produce metabolisms viable on 30 and 26 other carbon sources, respectively. Similar observations emerge when we repeat this analysis by increasing the number of reactions exchanged during recombination (Figs. S11 *a* and S12 *a*). In sum, for a majority (43 of 50) of parental phenotypes, there are no absolute constraints on metabolic innovation, i.e., all novel metabolic phenotypes considered here can arise through recombination.

Our next analysis (Fig. 2 *b*) provides evidence for abundant relative constraints, that is, some carbon-usage phenotypes C_j are more likely to emerge as metabolic innovations than others from parents viable on a given carbon source C_i. For example, 65.13% of the innovative offspring emerging from parents viable on glucose, gain viability on only four other carbon sources: 16.37% on D-fructose 6-phosphate, 17.72% on D-glucose 6-phosphate, 15.15% on D-fructose, and 15.89% on D-gluconate. The other 34.87% of metabolic innovations are distributed among 45 other carbon sources (on average each receiving 0.77% of the innovative offspring). As another example, for parents viable on D-serine, 46% of the innovative offspring gain viability on glycine (9.71%), L-aspartate (11.4%), L-alanine (16.73%) or D-alanine (8.16%) and the rest of 54% innovations is distributed among the other 45 carbon sources (each on average 1.2%).

We then clustered the 50 carbon sources based on their relative innovation distance in Fig. 2 *b*, where two carbon sources (C_i, C_j) are more distant if parents viable on C_i give rise to fewer offspring viable on C_j. Fig. 2 *c* shows that all glycolytic carbon sources (see Text S3) form one major branch of the resulting tree (*colored red*), and 17 of the 20 gluconeogenic carbon sources (exceptions: D-galacturonate, L-galactonate, and D-glucoronate) form another major branch (*colored cyan*). Hence, the propensity for innovation between carbon sources belonging to the same class is higher than those belonging to different classes. This observation hints at a cause of the relative constraints we observe, which we discuss in more detail in On the Underlying Causes of Constraints and Contingencies.

We observe qualitatively identical patterns when we repeat this analysis with altered numbers of reactions exchanged during recombination (Figs. S11 and S12), with altered genotypic distances among metabolic networks (Fig. S13), with smaller metabolic networks (Figs. S14 and

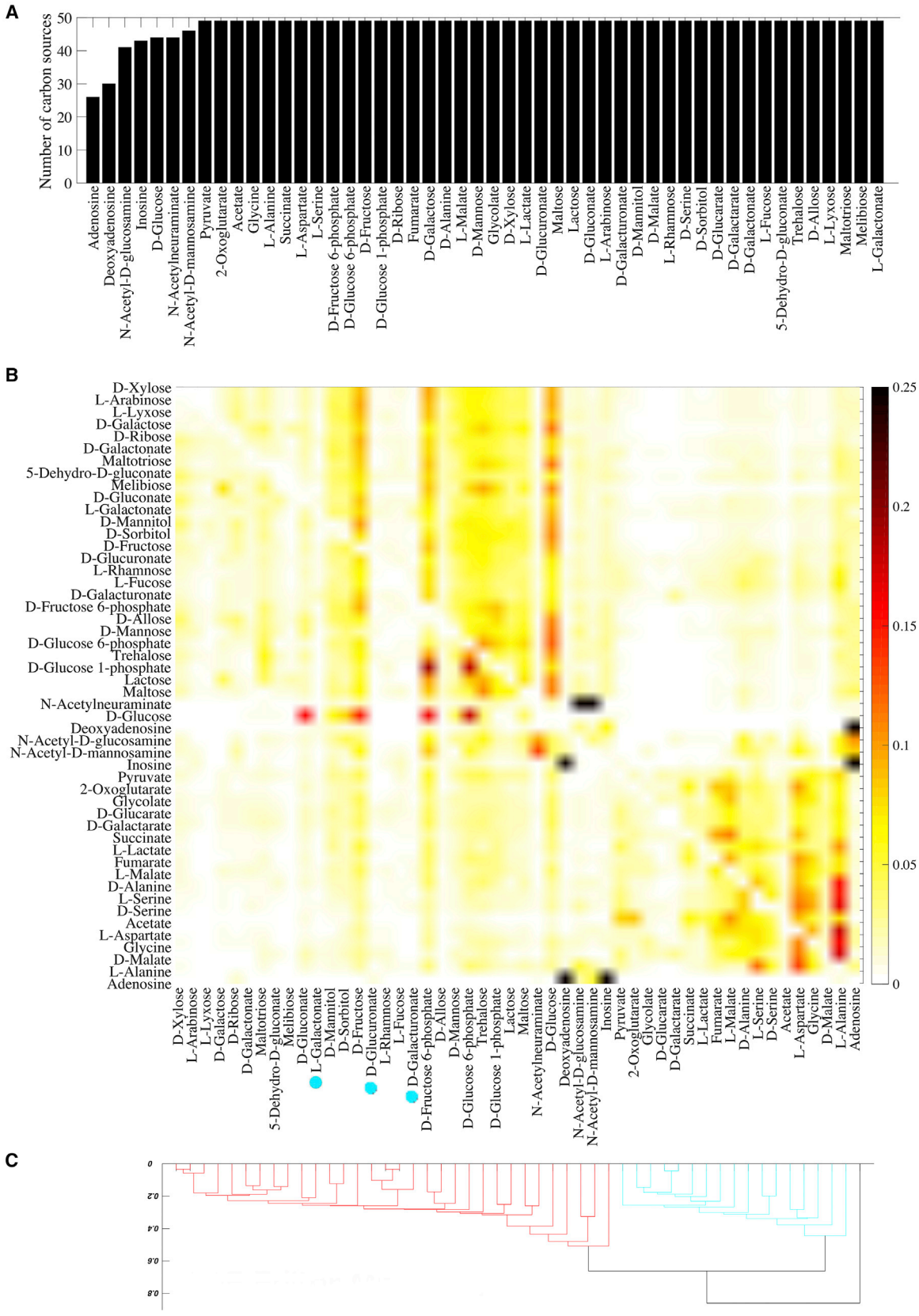


FIGURE 2 Emergence of innovative offspring can be constrained by parental phenotypes. (*A*) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least

(legend continued on next page)

S15), and with heterogeneous parental phenotypes (Figs. S16 and S17). However, in smaller metabolic networks, perhaps due to a substantially lower incidence of phenotypic innovation (Fig. S4), emergence of novel phenotypes is more constrained by parental phenotypes (Figs. S14 and S15). Moreover, for heterogeneous parental phenotypes where all the recipients are viable only on glucose and donors are viable on other carbon sources, carbon sources do not cluster according to innovation distance. The likely reason is that the recipient parental phenotype is constant in this analysis (Fig. S17).

In sum, different novel phenotypes are constrained in their evolution, because they originate with different probabilities from a given parental carbon-usage phenotype.

Emergence of innovative offspring is not absolutely, but only relatively, contingent on parental phenotypes

To complement our above analyses, we also studied whether some novel metabolic phenotypes are absolutely contingent on a specific parental phenotype. That is, can they only emerge from parents with this phenotype? To find out, we studied the parental phenotypes of all innovative offspring that have gained viability on a given carbon source C_j, and did so for all carbon sources C_j. Fig. S18 *a* shows that for all novel carbon-usage phenotypes C_j, innovative offspring can emerge from parents with at least 40 different phenotypes. Similar observations emerge when recombination alters a different number of reactions (Figs. S19 *a* and S20 *a*).

Although absolute contingency does therefore not exist in our study system, we observe relative contingency: different parental phenotypes C_i have a greater or lesser propensity to give rise to a given carbon-usage phenotype C_j (Fig. S18 *b*).

For example, 42.15% innovative offspring gaining viability on D-galactarate originate from parents viable only on four different carbon sources, namely D-malate (12.86%), D-galacturonate (11.99%), pyruvate (8.70%), and glycolate (8.61%). The other 57.85% originate from parents viable on the other 45 carbon sources (where each accounts for 1.28% of the innovative offspring on average). Another example regards viability on succinate, 20.8% of which originates from parents viable on acetate and the rest is distributed among other parental phenotypes (each contributing 1.65% on average).

And once again, classification of carbon sources based on their distance (Fig. S18 *b*) results in separation of glycolytic and gluconeogenic carbon sources (Fig. S18 *c*). We observe similar patterns when we repeat this analysis with a different number of reactions altered during recombination (Figs. S19 and S20), with higher genotypic distances among parental metabolisms (Fig. S21), with smaller metabolic networks (Figs. S22 and S23), and with heterogeneous parental phenotypes (Figs. S24 and S25). In smaller metabolic networks, perhaps due to the lower incidence of innovation (Fig. S8), relative contingency is most pronounced (Figs. S22 and S23).

In sum, although we do not observe absolute contingency, some parental phenotypes are much more likely than others to give rise to specific new metabolic phenotypes, which show relative contingency.

On the underlying causes of constraints and contingencies

As we observed in Figs. 2 *c* and S18 *c*, one specific measure of biochemical similarity among carbon sources can help explain the patterns of constraints and contingencies that we observed. That is, carbon sources can be broadly partitioned into glycolytic and gluconeogenic classes, where parents viable on a carbon source in one class are most likely to produce innovative offspring viable on a new carbon source in the same class. To provide complementary evidence that constraints increase with biochemical distance among carbon sources, we used two other biochemical similarity measures, and determined whether they are associated with the innovation distance between carbon sources.

The first defines the metabolic distance between a given pair of carbon sources (C_i, C_j) as the average shortest path between C_i and C_j in the substrate graph of 1000 metabolic networks viable on C_i (Supporting Materials and Methods). This network-based biochemical distance is significantly associated with the number of recombinants that are generated from parents viable on C_i, and that gain viability on carbon source C_j (Fig. S26, Pearson $r = -0.2722$, and $p < 10^{-41}$). A second quantifier of distance relies on the superessentiality index, the proportion of random viable networks in which a given reaction is essential for viability on a given carbon source (Supporting Materials and Methods). Here also, innovation declines with increasing biochemical distance among carbon sources (Pearson $r = -0.3935$, and $p < 10^{-83}$, Fig. S27 *a*; Supporting Materials and Methods).

Another complementary analysis involving biochemical distance focuses on the individual reactions that can be

one innovative offspring resulting from recombination between parental metabolic networks is viable. (*B*) Given here is the fraction of innovative recombinants (coded according to the *color legend*) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. (*C*) Dendrogram of carbon sources clustered based on their innovation distance defined by the data in (*B*). We used the unweighted pair group method with arithmetic means for clustering carbon sources. Branches colored in red and cyan correspond to glycolytic and gluconeogenic carbon sources, respectively (except D-galacturonate, L-galactonate, and D-glucoronate (*cyan circles*), which are the gluconeogenic carbon sources). In these analyses, parental metabolic networks contain $||G|| = 2079$ reactions, the same number as the *E. coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks during recombination.

transferred from donor to recipient, and that can lead to metabolic innovation. For this analysis, it is relevant that the majority of metabolic innovations is caused by the transfer of a single key reaction (32). We analyzed transferable reactions in greater depth, focusing on all 1000 parental donor metabolic networks viable on a given carbon source C_i, and on the ($D/2 = 50$) reactions that are present in the donor metabolic network, but are absent in the recipient, and so can potentially be transferred from the donor to the recipient. Specifically, we quantified the fraction of the 1000 parental donor metabolic networks viable on C_i in which at least one reaction among the ($D/2 = 50$) transferrable reactions can have C_j as a product or substrate, reasoning that such reactions may be especially prevalent among reactions causing viability on C_j. The number of innovative offspring that gain viability on C_j by recombining parents viable on C_i, increases significantly with the fraction of transferable reactions that involves C_j (Pearson $r = 0.163$, and $p < 10^{-15}$; Fig. S27 *b*). It is not difficult to see that this association can also be a consequence of the relatedness of two carbon sources. The reason is that metabolic networks viable on a given carbon source C_i are likely to already have some reactions involving metabolically related carbon sources C_j. In this case, it is more likely that addition of a single novel reaction leads to the completion of a pathway in the recipient that is needed to metabolize C_j. We note that these correlation coefficients, albeit statistically significant, are low in magnitude, implying that these properties cannot fully explain the mechanism underlying phenotypic constraint. A more detailed analysis of each pathway connecting different carbon sources may be required to fully understand the causes of constraints and contingencies. We leave such an analysis for future work.

Emergence of innovative offspring is constrained by, and contingent on, parental genotypes of both donors and recipients

Our analyses thus far were focused on parental metabolic networks with given phenotypes, which allowed us to analyze constraints and contingencies emerging from such phenotypes. However, the emergence of novel phenotypes may also depend on parental genotypes, and we next analyzed such constraints. Random viable metabolisms are less than ideal for such an analysis for two reasons. First, they do not derive from any one organism with its specific gene-reaction association, and they do therefore not allow us to define genotypes on the level of genes. Second, our simple model of recombination for such metabolisms neglects the linkage of metabolic genes on chromosomes.

To overcome these limitations, we focused our next analysis on curated metabolic networks of 55 distinct bacterial strains or species. Their metabolic genes, reactions, gene-reaction association rules, metabolic gene locations, and biomass reactions are well studied and available from the BiGG database (54). We used 30 carbon sources on which none of the 55 metabolisms are viable to study the emergence of novel phenotypes (Supporting Materials and Methods). We examined all 2970 ($= 55 \times 54$) distinct pairs of donor-recipient species or strains, and subjected them to recombination events that take into account metabolic gene linkage (see Modeling Recombination in Curated Bacterial Metabolic Networks from the BiGG Database). From each donor-recipient pair, we generated millions of recombinant offspring to identify innovative offspring, that is, offspring gaining viability on at least one of the 30 novel carbon sources.

We observed that the emergence of novel phenotypes is strongly contingent on the recombining parental genotypes. Among the 2970 pairs of recombining parental genotypes, only 347 pairs (11.68%) brought forth at least one innovative offspring (Fig. 3 *a*). In addition, these 347 pairs vary greatly in the number of innovative offspring that they can generate. The highest number of innovative offspring (56,461, or 1.17% of recombination events) emerges when the donor is *Staphylococcus aureus N315* and the recipient genotype is *E. coli DH1*, and the lowest number ((904), or 0.02% of recombination events) emerges when the donor is *E. coli BL21* and the recipient genotype is *Bacillus subtilis*.

The emergence of innovative offspring was also strongly constrained by the donor genotype (Fig. 3 *b*). A quantity of 97.84% of all innovative offspring identified in this analysis was generated from only six donor genotypes. The other 49 donors together were responsible only for 2.16% of all innovative offspring. Recombination involving *Staphylococcus aureus N315* donors caused an exceptionally large fraction of 45.97% of innovative offspring. Despite this strong relative constraint on donor genotypes, we did not observe any absolute constraints, because all 55 prokaryotic metabolisms generated at least one innovative offspring as donor genotypes—although the contributions of 49 metabolisms were so small that they are not visible in Fig. 3 *b*.

In contrast, the emergence of innovative offspring was not strongly constrained by the parental recipient genotype. That is, the majority of recipient metabolisms (48 out of 55) can generate approximately the same number of innovative offspring (Fig. 3 *c*). Only four of them generated considerably fewer innovative offspring, and three of them did not generate any innovative offspring as recipients (Fig. 3 *c*). Importantly, the potential of metabolic genotypes in generating innovative offspring when used as donors or recipients was highly asymmetric. For example, although *Staphylococcus aureus* and *Mycobacterium tuberculosis* accounted for most innovative offspring as donor genotypes, they did not generate any innovative offspring as recipient genotype. Similarly asymmetric biases emerged when we repeated the analysis with a recombination approach that does not take into account metabolic gene linkage, suggesting that such asymmetry is not caused by gene linkage but by the

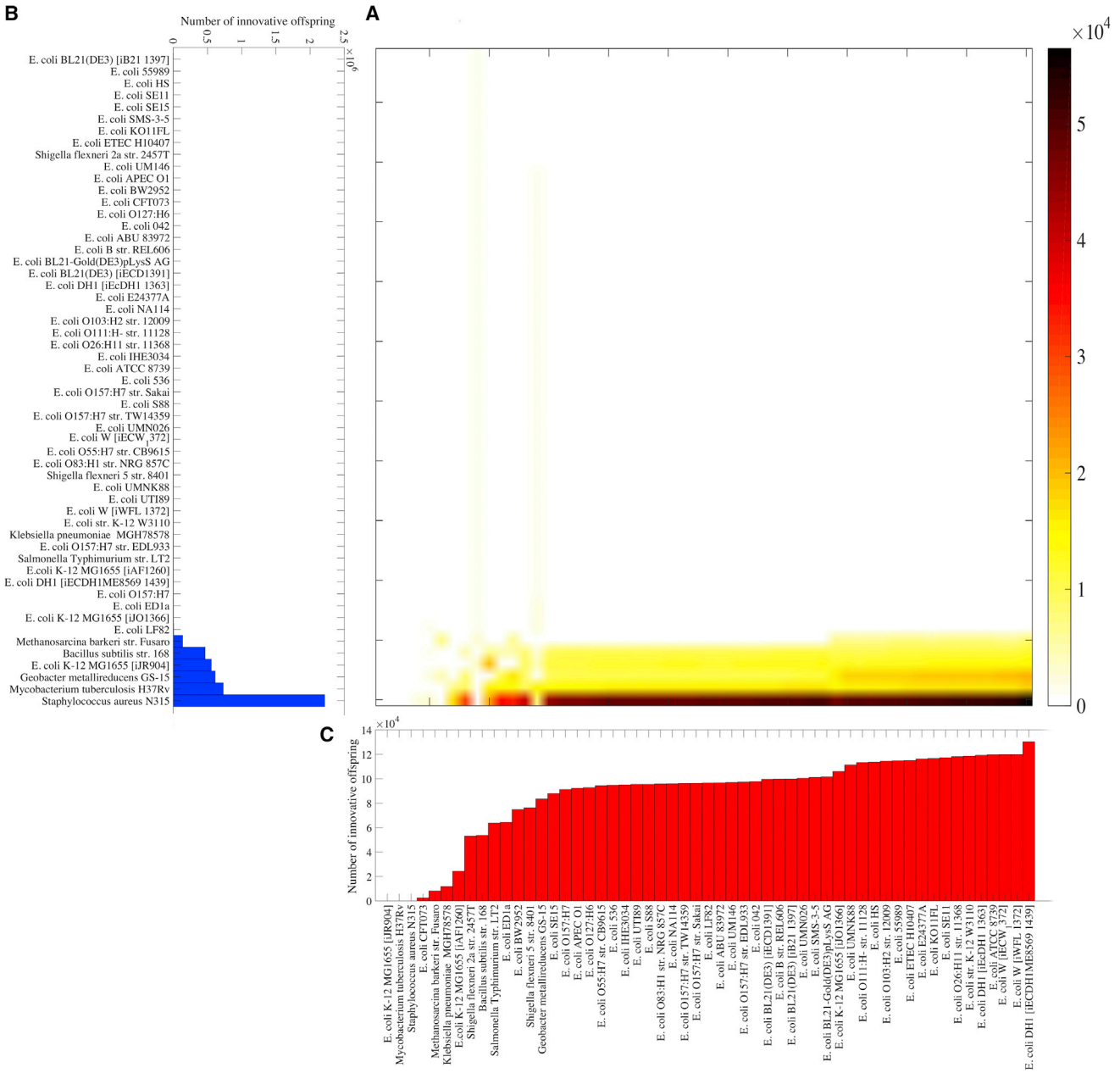


FIGURE 3 Emergence of innovative offspring is contingent on and constrained by parental genotypes. (*A*) Shown here is the number of innovative recombinant offspring resulting from linkage-based recombination between bacterial DNA donors specified on the vertical axis of (*B*), and the corresponding recipients specified on the horizontal axis of (*C*) (number of recombinants encoded according to the color legend). (*B*) Shown here is the total number of innovative recombinant offspring involving the donor genotype specified on the vertical axis. (*C*) Shown here is the total number of innovative recombinant offspring involving the recipient genotype specified on the horizontal axis.

metabolic gene content of genomes (Fig. S28). In sum, the emergence of innovative offspring is strongly contingent on the genotypes of parental donor-recipient pairs, and especially on donor genotypes.

DISCUSSION

In this study, we systematically analyzed the prevalence of constraint and contingency for emerging novel phenotypes in complex metabolic systems. We did so by computationally emulating recombination among thousands of parental metabolic network pairs with specific phenotypes, and created millions of recombinant metabolic networks.

Overall, we observed little evidence for absolute constraints in the origin of novel phenotypes, i.e., metabolic networks with most carbon-usage phenotypes can give rise to all 50 novel carbon-usage phenotypes we consider here. However, there is ample evidence for relative constraints,

that is, some carbon-usage phenotypes are much more likely to arise relative to others from any one parental carbon-usage phenotype.

Similarly, we observed no absolute contingency in the origin of novel phenotypes, i.e., recombinant metabolic networks with a given novel carbon-usage phenotype can originate from all 50 parental phenotypes. In contrast, relative contingency is pervasive. That is, a given novel carbon-usage phenotype is much more likely to originate from some parental phenotypes than from others. Importantly, our observations remain qualitatively unchanged when we alter various properties of parental genotypes, such as their genotypic distance, which suggests that the different extents of constraints we observe may be an inherent property of metabolic systems.

We also analyzed the causes of constraints and contingencies, where several complementary analyses point to the importance of biochemical similarities among carbon source pairs (C_i, C_j), where parents are viable on C_i, and recombinant offspring gain viability on C_j. First, if parents are viable on a carbon source that belongs to one of two major biochemical classes (glycolytic or gluconeogenic), then recombinant offspring tend to gain viability on a carbon source within the same class (Fig. 2 *c*; Fig S18 *c*). Second, the smaller the number of reactions that separate C_i and C_j in a metabolic network, the greater the likelihood that offspring gain viability on C_j. Third, offspring gain viability on C_j most often if a reaction transferred between donor and recipient involves C_j. This, in turn, is most likely if the recipient already harbors some reactions necessary to metabolize C_j, and thus if catabolizing C_i and C_j involves similar reactions. Our analysis used carbon sources that are not very heterogenous. Many of them, for example, are sugars that play important roles in central carbon metabolism. This biochemical similarity among carbon sources reduces constraints, and it may be responsible for the paucity of evidence for absolute constraints.

One strength of our approach is that it can address contingency and constraint in an entire class of system, and not just a single organism. However, the approach also has several limitations.

First, any study relying on sampling is sensitive to sample size. For example, if we had analyzed only 100 parental metabolic networks and 100 recombinants per pair, we would not have observed any innovative offspring for most parental carbon usage phenotypes. Thus, we would have misleadingly concluded that absolute constraints are frequent in our study system. And even though we had generated a (computationally expensive) sample of 1,000,000 offspring for each parental phenotype, we did see a small number of carbon sources showing evidence for absolute constraints. Such apparent absolute constraints may disappear at even higher sample sizes (Fig. S29 *a*). In contrast, our assertion that relative constraints exist is less sensitive to sample sizes (Fig. S29 *b*). Our current analysis generated fewer than 1000 innovative metabolisms for most (C_i, C_j) pairs, and larger sample sizes may help us find out why some pairs (C_i, C_j) are more or less involved in metabolic innovation.

Second, our work is based on FBA (15,16), which neglects the influence of gene and enzyme regulation. However, because regulatory changes toward optimal expression of enzymes readily occur, even on the short timescales of laboratory evolution, this limitation may not affect our main observations (Supporting Materials and Methods).

Third, a recent study showed that the genome-scale metabolic networks are likely to include thermodynamically impossible energy-generating cycles (EGCs), which are capable of charging energy metabolites without nutrient consumption (56). These EGCs can artificially inflate biomass flux and so may mislead evolutionary simulations. Most of our randomly sampled viable metabolisms indeed harbor EGCs (97.3% and 97.8% of sampled metabolisms viable on glucose and acetate, respectively; Supporting Materials and Methods). However, these EGCs do not strongly affect the emergence of novel phenotypes, nor do they substantially distort the patterns of relative constraint we observed (Figs. S30, S31, and S32; Supporting Materials and Methods).

Finally, in our simulations using random metabolic networks, following common practice in the field (17–26), we define metabolic genotypes on the level of biochemical reactions rather than on that of genes or DNA. This representation neglects potentially important information, and especially the linkage of related metabolic genes on chromosomes, which may affect the outcome of recombination. To address this limitation, we also modeled recombination among metabolisms of 55 prokaryotic species or strains in a way that includes gene linkage information. This analysis also demonstrates strong constraints and contingencies in the emergence of novel metabolic phenotypes.

A previous experimental evolution study suggested a strong relative constraint in the emergence of a novel citrate utilization phenotype, which required thousands of generations of laboratory evolution subject to mutation and selection to emerge (9). Although our simulations are not strictly commensurate with any experimental study, for example, because we do not consider DNA changes explicitly, we speculate that such relative constraints would be less pronounced in any system where recombination is abundant, because recombination can cause larger-scale changes than mere point mutations that would alter individual reactions or transport processes (9). This was one motivation to choose recombination as an agent of genetic change in the first place, reasoning that any constraints visible in the presence of recombination might be even stronger in the presence of less dramatic genetic changes.

Metabolic systems are one of the three classes of biological systems in which phenotypic variation is crucial for evolutionary adaptation and innovation (57). The other

two are macromolecules (protein and RNA) and regulatory systems. Predicting phenotypes in these systems is less straightforward than for metabolic systems (58–60). In proteins, for example, phenotypes form through a complex and incompletely understood 3D folding process (58), and in regulatory systems, gene expression phenotypes emerge from complex interactions among regulatory molecules (59,60). Our understanding of inherent biases in phenotypic variability will not be complete until we understand contingencies and constraints in these classes of systems as well, which remains an important task for future work.

SUPPORTING MATERIAL

Supporting Materials and Methods and thirty-two figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(17)30684-7.

AUTHOR CONTRIBUTIONS

S.-R.H. and A.W. designed research. S.-R.H. performed research. S.-R.H. and A.W. analyzed data and wrote the paper.

ACKNOWLEDGMENTS

We acknowledge support through Swiss National Science Foundation grant 31003A_146137, through an EpiphysX RTD grant from SystemsX.ch, as well as through the University Priority Research Program in Evolutionary Biology at the University of Zürich.

SUPPORTING CITATIONS

References (61–74) appear in the Supporting Material.

REFERENCES

1. Maynard-Smith, J., R. Burian, …, L. Wolpert. 1985. Developmental constraints and evolution. *Q. Rev. Biol.* 60:265–287.
2. Wagner, A. 2011. Genotype networks shed light on evolutionary constraints. *Trends Ecol. Evol. (Amst.).* 26:577–584.
3. Nüsslein-Volhard, C., and E. Wieschaus. 1980. Mutations affecting segment number and polarity in Drosophila. *Nature.* 287:795–801.
4. West, G. B., J. H. Brown, and B. J. Enquist. 1997. A general model for the origin of allometric scaling laws in biology. *Science.* 276:122–126.
5. Nelson, D. L., and M. M. Cox. 2004. Lehninger Principles of Biochemistry,, 3rd Ed. W. H. Freeman, New York, NY.
6. Levitt, M. 2009. Nature of the protein universe. *Proc. Natl. Acad. Sci. USA.* 106:11079–11084.
7. Gould, S. J. 1990. Wonderful Life: The Burgess Shale and the Nature of History. W. W. Norton, New York, NY.
8. Lobkovsky, A. E., and E. V. Koonin. 2012. Replaying the tape of life: quantification of the predictability of evolution. *Front. Genet.* 3:246.
9. Blount, Z. D., C. Z. Borland, and R. E. Lenski. 2008. Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA.* 105:7899–7906.
10. Copley, S. D. 2000. Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem. Sci.* 25:261–265.
11. Rehmann, L., and A. J. Daugulis. 2008. Enhancement of PCB degradation by *Burkholderia xenovorans* LB400 in biphasic systems by manipulating culture conditions. *Biotechnol. Bioeng.* 99:521–528.
12. van der Meer, J. R., Jr., C. Werlen, …, J. C. Spain. 1998. Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. *Appl. Environ. Microbiol.* 64:4185–4193.
13. Cline, R. E., R. H. Hill, Jr., …, L. L. Needham. 1989. Pentachlorophenol measurements in body fluids of people in log homes and workplaces. *Arch. Environ. Contam. Toxicol.* 18:475–481.
14. Dantas, G., M. O. A. Sommer, …, G. M. Church. 2008. Bacteria subsisting on antibiotics. *Science.* 320:100–103.
15. Orth, J. D., I. Thiele, and B. Ø. Palsson. 2010. What is flux balance analysis? *Nat. Biotechnol.* 28:245–248.
16. Edwards, J. S., R. U. Ibarra, and B. O. Palsson. 2001. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* 19:125–130.
17. Edwards, J. S., and B. O. Palsson. 2000. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. USA.* 97:5528–5533.
18. Feist, A. M., and B. Ø. Palsson. 2008. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat. Biotechnol.* 26:659–667.
19. McCloskey, D., B. Ø. Palsson, and A. M. Feist. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of *Escherichia coli*. *Mol. Syst. Biol.* 9:661.
20. Lewis, N. E., H. Nagarajan, and B. O. Palsson. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nat. Rev. Microbiol.* 10:291–305.
21. Matias Rodrigues, J. F., and A. Wagner. 2009. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLOS Comput. Biol.* 5:e1000613.
22. Edwards, J. S., and B. O. Palsson. 1999. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* 274:17410–17416.
23. Samal, A., J. F. Matias Rodrigues, …, A. Wagner. 2010. Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* 4:30.
24. Barve, A., S.-R. Hosseini, …, A. Wagner. 2014. Historical contingency and the gradual evolution of metabolic properties in central carbon and genome-scale metabolisms. *BMC Syst. Biol.* 8:48.
25. Hosseini, S.-R., A. Barve, and A. Wagner. 2015. Exhaustive analysis of a genotype space comprising 10(15)central carbon metabolisms reveals an organization conducive to metabolic innovation. *PLOS Comput. Biol.* 11:e1004329.
26. Hosseini, S.-R., and A. Wagner. 2016. The potential for non-adaptive origins of evolutionary innovations in central carbon metabolism. *BMC Syst. Biol.* 10:97.
27. Stemmer, W. P. 1994. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA.* 91:10747–10751.
28. Zhang, Y.-X., K. Perry, …, S. B. del Cardayré. 2002. Genome shuffling leads to rapid phenotypic improvement in bacteria. *Nature.* 415:644–646.
29. Crameri, A., G. Dawes, …, W. P. Stemmer. 1997. Molecular evolution of an arsenate detoxification pathway by DNA shuffling. *Nat. Biotechnol.* 15:436–438.
30. Chang, C. C., T. T. Chen, …, P. A. Patten. 1999. Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* 17:793–797.
31. Ness, J. E., M. Welch, …, J. Minshull. 1999. DNA shuffling of subgenomic sequences of subtilisin. *Nat. Biotechnol.* 17:893–896.
32. Hosseini, S.-R., O. C. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. *Proc. Biol. Sci.* 283:20161536.
33. Thomas, C. M., and K. M. Nielsen. 2005. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* 3:711–721.

34. Guttman, D. S., and D. E. Dykhuizen. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science.* 266:1380–1383.

35. Feil, E. J., E. C. Holmes, …, B. G. Spratt. 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. *Proc. Natl. Acad. Sci. USA.* 98:182–187.

36. Whitaker, R. J., D. W. Grogan, and J. W. Taylor. 2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol. Biol. Evol.* 22:2354–2361.

37. Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature.* 405:299–304.

38. Pál, C., B. Papp, and M. J. Lercher. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37:1372–1375.

39. Fraser, C., W. P. Hanage, and B. G. Spratt. 2007. Recombination and the nature of bacterial speciation. *Science.* 315:476–480.

40. Majewski, J., P. Zawadzki, …, C. G. Dowson. 2000. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. *J. Bacteriol.* 182:1016–1023.

41. Kowalczykowski, S. C., D. A. Dixon, …, W. M. Rehrauer. 1994. Biochemistry of homologous recombination in *Escherichia coli*. *Microbiol. Rev.* 58:401–465.

42. Kuo, C.-H., and H. Ochman. 2009. The fate of new bacterial genes. *FEMS Microbiol. Rev.* 33:38–43.

43. Mira, A., H. Ochman, and N. A. Moran. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17:589–596.

44. Lin, C. H., G. Bourque, and P. Tan. 2008. A comparative synteny map of *Burkholderia* species links large-scale genome rearrangements to fine-scale nucleotide variation in prokaryotes. *Mol. Biol. Evol.* 25:549–558.

45. Bork, P., and R. F. Doolittle. 1992. Proposed acquisition of an animal protein domain by bacteria. *Proc. Natl. Acad. Sci. USA.* 89:8990–8994.

46. Inagaki, Y., E. Susko, and A. J. Roger. 2006. Recombination between elongation factor 1α genes from distantly related archaeal lineages. *Proc. Natl. Acad. Sci. USA.* 103:4528–4533.

47. Hartl, D. L., E. R. Lozovskaya, and J. G. Lawrence. 1992. Nonautonomous transposable elements in prokaryotes and eukaryotes. *Genetica.* 86:47–53.

48. Igarashi, N., J. Harada, …, K. V. Nagashima. 2001. Horizontal transfer of the photosynthesis gene cluster and operon rearrangement in purple bacteria. *J. Mol. Evol.* 52:333–341.

49. Omelchenko, M. V., K. S. Makarova, …, E. V. Koonin. 2003. Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol.* 4:R55.

50. Chan, C. X., R. G. Beiko, …, M. A. Ragan. 2009. Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol. Evol.* 1:429–438.

51. Denamur, E., G. Lecointre, …, I. Matic. 2000. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. *Cell.* 103:711–721.

52. Didelot, X., M. Achtman, …, D. Falush. 2007. A bimodal pattern of relatedness between the *Salmonella paratyphi* A and Typhi genomes: convergence or divergence by homologous recombination? *Genome Res.* 17:61–68.

53. Gelius-Dietrich, G., A. A. Desouki, …, M. J. Lercher. 2013. Sybil—efficient constraint-based modelling in R. *BMC Syst. Biol.* 7:125.

54. King, Z. A., J. Lu, …, N. E. Lewis. 2015. BiGG Models: a platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Res.* 44:D515–D522.

55. Tatusova, T., S. Ciufo, …, I. Tolstoy. 2014. RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.* 42:D553–D559.

56. Fritzemeier, C. J., D. Hartleb, …, M. J. Lercher. 2017. Erroneous energy-generating cycles in published genome scale metabolic networks: identification and removal. *PLOS Comput. Biol.* 13:e1005494.

57. Wagner, A. 2011. The Origins of Evolutionary Innovations: A Theory of Transformative Change in Living Systems. Oxford University Press, Oxford, UK.

58. Dill, K. A., S. B. Ozkan, …, T. R. Weikl. 2008. The protein folding problem. *Annu. Rev. Biophys.* 37:289–316.

59. Karlebach, G., and R. Shamir. 2008. Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 9:770–780.

60. De Smet, R., and K. Marchal. 2010. Advantages and limitations of current network inference methods. *Nat. Rev. Microbiol.* 8:717–729.

61. Goto, S., T. Nishioka, and M. Kanehisa. 2000. LIGAND: chemical database of enzyme reactions. *Nucleic Acids Res.* 28:380–382.

62. Goto, S., Y. Okuno, …, M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.* 30:402–404.

63. Kanehisa, M., S. Goto, …, M. Hirakawa. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.* 38:D355–D360.

64. Feist, A. M., C. S. Henry, …, B. Ø. Palsson. 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* 3:121.

65. Lercher, M. J., and C. Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. *Mol. Biol. Evol.* 25:559–567.

66. Ibarra, R. U., J. S. Edwards, and B. O. Palsson. 2002. *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature.* 420:186–189.

67. Vieira-Silva, S., and E. P. C. Rocha. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.* 6:e1000808.

68. Kirschner, D., and S. Marino. 2005. *Mycobacterium tuberculosis* as viewed through a computer. *Trends Microbiol.* 13:206–211.

69. Fong, S. S., and B. Ø. Palsson. 2004. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36:1056–1058.

70. Fong, S. S., J. Y. Marciniak, and B. O. Palsson. 2003. Description and interpretation of adaptive evolution of *Escherichia coli* K-12 MG1655 by using a genome-scale in silico metabolic model. *J. Bacteriol.* 185:6400–6408.

71. Wagner, A., and D. A. Fell. 2001. The small world inside large metabolic networks. *Proc. Biol. Sci.* 268:1803–1810.

72. Barve, A., J. F. M. Rodrigues, and A. Wagner. 2012. Superessential reactions in metabolic networks. *Proc. Natl. Acad. Sci. USA.* 109:E1121–E1130.

73. Ma, H.-W., and A.-P. Zeng. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics.* 19:1423–1430.

74. Hopcroft, J., and R. Tarjan. 1973. Algorithm 447: efficient algorithms for graph manipulation. *Commun. ACM.* 16:372–378.

Biophysical Journal, Volume *113*

Supplemental Information

Constraint and Contingency Pervade the Emergence of Novel Phenotypes in Complex Metabolic Systems

Sayed-Rzgar Hosseini and Andreas Wagner

1. Supplementary Methods:

S1: Genome-scale metabolic networks and their phenotypic representations

Similar to our previous work describing the procedures used here (1), and following common practice in metabolic systems biology (2–4), we represent an organism's metabolic genotype as the set of genomically encoded (enzyme-catalyzed) biochemical reactions proceeding inside the organism. This metabolic genotype specifies a metabolism or metabolic network, a network of chemical reactions encoded by the genotype. A metabolic reaction network enables an organism to extract energy and produce small biomass building blocks, such as amino acids, from extracellular nutrients. Inference of this genotype from genomic and biochemical information has been successful for multiple organisms (5, 6).

Any one metabolic reaction network contains a subset of the "reaction universe" of all biochemical reactions that take place in prokaryotes (See text S2). We have curated a representation of this universe, which comprises 5,906 reactions and is based on current metabolic knowledge (7–10). We represent an organism's metabolic genotype as a binary vector of length 5,906. Each entry of this vector corresponds to a given reaction in the reaction universe, and is equal to one if the corresponding reaction is present in the metabolic network, and zero otherwise. Thus, each genotype can be thought of as a single member of a vast space of all possible metabolic networks, which contains 2^{5906} distinct genotypes.

We define the phenotype of a given metabolic genotype based on its viability in 50 distinct minimal environments that differ only in the carbon source they harbor (See Text S3). We consider that a genotype is *viable* on a given carbon source, if it can produce all essential biomass precursor molecules from the given carbon source, and we use Flux Balance Analysis (FBA, See text S4) to determine viability (11). We represent the phenotype of a given metabolic genotype as a binary vector of length 50. Each entry of this vector corresponds to a given carbon source, and it is equal to one if the genotype is viable on this carbon source, and zero otherwise.

S2: Reaction universe

The reaction universe we curated is a set of metabolic reactions in which each reaction is known to occur in some prokaryotic organisms. For the curation of this universe, we used data from the LIGAND database (7, 8) of the Kyoto Encyclopedia of Genes and Genomes (9). Briefly, the LIGAND database, which is comprised of the REACTION and the COMPOUND databases, provides information on reactions, associated stoichiometric information, chemical compounds involved in a reaction, and the Enzyme Classification (E.C.) identifier of each

reaction. From the REACTION and the COMPOUND databases we excluded (i) all reactions involving polymer metabolites of unspecified numbers of monomers, or general polymerization reactions with uncertain stoichiometry, (ii) reactions involving glycans, due to their complex structure, (iii) reactions with unbalanced stoichiometry, and, (iv) reactions involving complex metabolites without chemical information about their structure (10). Moreover, we do not consider unknown reactions, and we also do not take into account spontaneous reactions, or reactions that depend on external stimuli. The published *E. coli* metabolic model (iAF1260) consists of 1397 non-transport reactions (12). We merged all reactions in the *E. coli* model with the reactions in the KEGG dataset, and retained only the unique (non-duplicate) reactions. This resulted in a universe of reactions consisting of 682 transport, 5,906 non-transport reactions and 5030 metabolites. The reaction universe is available online (https://github.com/rzgar/EMETNET/tree/master/UNIVERSE).

S3: Chemical environments

We consider 50 minimal growth environments, each of which includes oxygen, ammonium, inorganic phosphate, sulfate, sodium, potassium, cobalt, iron (Fe^{2+} and Fe^{3+}), protons, water, molybdate, copper, calcium, chloride, magnesium, manganese, zinc, and a specific carbon source. Importantly, to represent different chemical environments, we vary the carbon source while keeping all other nutrients constant. We consider a metabolic network viable on a given carbon source, if it can synthesize all essential biochemical precursors when this carbon source is provided as the sole carbon source in the minimal medium just described.

We used 50 carbon sources for our analysis of randomly sampled metabolic networks, including the following 27 glycolytic carbon sources: D-glucose, D-glucose 6-phosphate, trehalose, maltose, lactose, D-fructose 6-phosphate, D-fructose, D-mannose, D-mannitol, D-glucose 1-phosphate, D-sorbitol, maltotriose, D-allose, D-ribose, D-xylose, D-gluconate, 5-dehydro-D-gluconate, L-rhamnose, L-fucose, L-arabinose, L-lyxose, D-galactose, melibiose, D-galactonate, N-acetyl-D-glucosamine, N-acetyl-D-mannosamine, N-acetylneuraminate.

In addition, we used the following 20 gluconeogenic carbon sources: pyruvate, L-alanine, L-lactate, D-alanine, D-malate, acetate, L-serine, L-malate, D-serine, glycine, glycolate, L-aspartate, succinate, fumarate, 2-oxoglutarate, D-galacturonate, D-galactarate, D-glucarate, L-galactonate, D-glucoronate. And we used the following three nucleosides as carbon sources: adenosine, deoxyadenosine, inosine.

To study the emergence of novel phenotypes in 55 prokaryotic metabolic networks from the BiGG database (13) (see methods section 2.4 in the main text), we used the following 30 carbon sources on which none of the 55 metabolic networks are predicted to be viable: Biotin, riboflavin, folate, pimelate, urea, carbonic acid, bicarbonate, methanol, trimethylamine, D-

methionine, glycine betaine, gamma-butyrobetaine, choline, L-phenylalanine, L-leucine, L-tyrosine, L-methionine, thiamin, 6-diaminoheptanedioate, (R)-pantothenate, spermidine, taurine, isocytosine, protoheme, nicotinamide adenine dinucleotide, L-fucose 1-phosphate, dimethyl-sulfide, L-carnitine, dimethyl sulfoxide, and 1,5-diaminopentane.

S4: Flux balance analysis

Flux balance analysis (FBA) is a computational method that is widely used for the quantitative analysis and modeling of metabolic networks (11). Based on the stoichiometric coefficients of the metabolites participating in the reactions of a given metabolic network, FBA predicts the metabolic flux through each reaction. Stoichiometric coefficients are stored in a stoichiometric matrix *S*, which is of dimension $m \times n$, where *m* and *n*, denote the number of metabolites and the number of reactions in a metabolic network. FBA constrains the flux through each reaction based on the assumption that a metabolic network is in a steady state where metabolite concentrations do not change, i.e., $Sv = 0$, where *v* is the vector of metabolic fluxes v_i through reaction *i*. The solutions of the equation $Sv = 0$, that is, the null space of matrix S, comprises all flux vectors that are allowable in steady state. The null space is further constrained by physicochemical information regarding the maximum and minimum possible fluxes through each reaction. FBA relies on an optimization procedure called linear programming to identify those among the allowable flux vector(s) that maximize an objective function *Z*. This task can be formulated as finding a flux vector v^* with the property

$$v^* = \max_v Z(v) = \max_v \{ c^T v \mid Sv = 0, a \leq v \leq b \},$$

where the vector *c* contains a set of scalar coefficients representing the maximization criterion, and each entry a_i and b_i of vectors a and b, indicates the minimally and maximally possible flux through reaction *i*. The vector *c* represents the proportions of each small biomass molecule in a cell's biomass. Therefore v^* maximizes the biomass growth flux, that is, the rate at which a metabolic network can produce biomass (11). Here we use FBA to predict qualitatively whether a given metabolic network is viable in a given environment, and we consider a metabolic network viable if it can produce all essential biomass precursors. More precisely, FBA predicts a metabolic network as viable on a given environment, if its biomass flux rate exceeds 0.001 $1/h$. In a free-living bacterium like *E.coli,* there are approximately 60 such molecules including 20 amino acids, DNA, and RNA precursors, lipids and cofactors. We used the biomass composition of the *E. coli* metabolic model iAF1260 to define the vector *c* (12). Moreover, we used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects/coin-or.org/Clp) to solve the linear programming problem of FBA.

The major limitation of FBA is that it neglects regulatory constraints that can arise through suboptimal expression or regulation of enzymes. Newly horizontally transferred genes cannot easily establish regulatory interactions with their host genes, and it may thus take considerable adaptive evolution until they become expressed at a maximal or optimal level (14). Such regulatory constraints would be especially important if we focused on quantitative predictions of biomass growth (15). However, we use FBA solely for qualitative prediction of viability. This focus on qualitative phenotypes is biologically sensible. The reason is that many organisms grow slowly in their native environment (16, 17), implying that regulation for maximal biomass production is far from universal. Moreover, we note that regulatory constraints can easily be broken in evolution, even on the short time scales of laboratory evolution experiments (15, 18, 19).

S5: Generation of random metabolic networks

We here employ a previously described *in silico* process which relies on Markov Chain Monte Carlo (MCMC) random walks to generate metabolic networks that comprise random sets of metabolic reactions that are viable on a given carbon source (10, 20). This procedure can produce metabolic networks that are sampled uniformly from the set of all metabolic networks viable on a given carbon source (10, 20). Briefly, in each step of such a random walk we perform a reaction swap, defined as altering a metabolic network by adding a randomly chosen reaction from the reaction universe, and then deleting a reaction randomly chosen from the set of reactions present in the metabolic network. If the reaction swap disrupts the metabolic network's viability on the given carbon source (as determined by FBA) we reject it, and perform another reaction swap until we find a swap that does not disrupt viability. This procedure also ensures that the total number of reactions remains constant. For the MCMC method to produce random samples of metabolic networks, it is essential to carry out enough reaction swaps to "erase" the random walker's similarity to the initial metabolic network. Previously, it has been shown that 3×10^3 reaction swaps are sufficient for this purpose (10, 20). Each of our random walks starts from *E. coli*'s metabolic network and performs 10^4 reaction swaps before storing the final metabolic network for further analysis. We used 10^4 independent random walks conducted in this way to create 10^4 random metabolic networks viable on each of the 50 carbon sources.

S6: Generation of parental metabolic network pairs

Some of our analyses required us to recombine pairs of "parental" metabolic networks with particular features, such as being viable on a specific carbon source (and only on that carbon source), or having a given genotypic distance (D), defined as the number of reactions differing

between the parents. Generating parents with a given genotypic distance (D) is not straightforward, because the random metabolic networks generated by MCMC sampling generally have genotypic distances sufficiently large ($D \approx 2{,}000$) to be biologically unrealistic for modeling frequently recombining prokaryotic genomes. To create less distant metabolic network pairs, we took an MCMC random walk approach. It revolves around a reaction-swapping random walk starting with a pair of randomly chosen metabolic networks from our sample of 10^4 sampled metabolic networks that are exclusively viable on a given carbon source. In each step of this random walk, we subjected each parental metabolic network to a reaction swap, and we accepted each reaction swap if it (*i*) preserved the original phenotype, and (*ii*) did not increase the genotypic distance of the two metabolic networks after the swap, otherwise we rejected the reaction swap. We continued this procedure until the genotypic distance between the metabolic networks became equal to a desired distance D. We note that this procedure is very time-consuming when applied to the thousands of parents we study here.

Finally, to generate parental metabolic networks with a given number of reactions, we started from a random viable metabolic network generated by MCMC sampling, as described in the text S5. All such metabolic networks have the same number of reactions as *E.coli* (2,079). We then applied a sequence of individual and random reaction deletions, where we required that each deletion preserve viability, until the network had reached the desired size.

S7. Estimation of the metabolic distance between carbon sources

For each pair of carbon sources (C_i, C_j), we calculated metabolic distance with two different approaches, a direct approach that is based on the shortest path between carbon sources in substrate graph (21), and an indirect approach that is based on carbon source-dependent superessentiality of metabolic reactions in metabolic networks (22).

The first approach relies on the substrate graph of a metabolic network, in which vertices correspond to metabolites. Two metabolites are linked via an edge, if the metabolites participate in the same metabolic reactions as either a substrate or a product. From this substrate graph we excluded currency metabolites, which are metabolites that transfer small chemical groups, and are involved in many reactions (23). 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. In addition, we excluded oxidized and reduced

forms of cofactors such as quinone, ubiquinone, glutathione, thioredoxin, flavodoxin and flavin mononucleotide. For all metabolic networks viable on C_i, we measured the shortest path in the substrate graph between C_i and any other $C_j, j \neq i$ using Dijkstra's algorithm (24). Then, we considered the average shortest path between C_i and C_j among metabolic networks viable on C_i as the metabolic distance between C_i and C_j.

In the second approach, we take advantage of the fact that metabolic reactions show varying degrees of essentiality among different metabolic networks that are viable on the same carbon sources. Any one reaction can be essential in one such network and inessential in another, depending on which reactions and pathways are present in the network. One can quantify a reaction's degree of essentiality in randomly sampled viable networks via a "superessentiality index", defined as the fraction of metabolic networks in which the reaction is essential for viability on a given carbon source (22). Highly superessential reactions are essential in most random viable networks, and cannot be by-passed easily by alternative metabolic pathways. We first computed the superessentiality index of each reaction on each carbon source C_i, and assembled this information into a superessentiality vector. Each element of this vector corresponds to one of the 5,906 reactions in the reaction universe, and contains the fraction of random viable metabolic networks in which the reaction is essential for viability on C_i. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources C_i and C_j as a proxy for metabolic distance between the two carbon sources.

S8: Distance measure between carbon sources based on superessential reactions

In the second approach, we take advantage of the fact that metabolic reactions show varying degrees of essentiality among different metabolic networks that are viable on the same carbon sources. Any one reaction can be essential in one such network and inessential in another, depending on which reactions and pathways are present in the network. One can quantify a reaction's degree of essentiality in randomly sampled viable networks via a "superessentiality index", defined as the fraction of metabolic networks in which the reaction is essential for viability on a given carbon source (22). Highly superessential reactions are essential in most random viable networks, and cannot be by-passed easily by alternative metabolic pathways. We first computed the superessentiality index of each reaction on each carbon source C_i, and assembled this information into a superessentiality vector. Each element of this vector corresponds to one of the 5,906 reactions in the reaction universe, and contains the fraction of random viable metabolic networks in which the reaction is essential for viability on C_i. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources C_i and C_j as a proxy for metabolic distance between the two carbon sources.

Previous work showed that highly superessential reactions are more likely to be involved in metabolic innovation (1). We thus also wanted to compute a biochemical distance measure of carbon sources based on this index. To this end, we computed, for each carbon source, the superessentiality index of all reactions belonging to the reaction universe, which yields a superessentiality vector of length 5,906. We then computed the Euclidian distance between the superessentiality vectors for all pairs of carbon sources C_i and C_j as a proxy for the biochemical distance between the two carbon sources. Fig. S27a shows that the number of innovative offspring, which are generated by recombination between parents viable on C_i, and gain viability on a given carbon source C_j is significantly correlated with the Euclidian distance between the superessentiality vectors for (C_i, C_j) (Pearson r = -0.3935, and $P < 10^{-83}$).

S9: Random metabolic networks and erroneous energy generating cycles

A recent study by Fritzemeier et al. showed that most of the published genome-scale metabolic networks include thermodynamically impossible energy-generating cycles (EGCs), which are capable of charging energy metabolites without nutrient consumption (25). It showed that these EGCs can artificially inflate biomass flux by 25% and could be particularly problematic in evolutionary simulations, which involves incorporation of foreign metabolic reactions from other species.

We applied the approach of Fritzemeier et al., to identify EGCs in metabolic networks (25), using 15 different energy dissipation reactions (EDRs) for each of the 15 different types of energy metabolites in the cell. (See https://doi.org/10.1371/journal.pcbi.1005494.s002 for complete information on these reactions). We maximized one energy dissipation reaction flux v_d at a time, while preventing all influx of external nutrients into the model. The problem can be mathematically expressed as follows:

$$\max v_d \text{ subject to:}$$
$$Sv = 0$$
$$\forall i \notin E: v_i^{min} \leq v_i \leq v_i^{max}$$
$$\forall i \in E: v_i = 0$$

where S is the stoichiometric matrix describing a metabolic system, v is the vector of all metabolic fluxes, d is the index of one of the energy dissipation reactions, v^{min} and v^{max} are vectors of lower and upper reaction bounds, and E is the set of indices of all exchange reactions. An optimal value v_d^* for this optimization with $v_d^* > 0$ for at least one of the energy dissipation reactions demonstrates the existence of at least one EGC in the corresponding metabolic network.

Using this approach, we first determined that the initial *E. coli* metabolic network with 2079 reactions (12) from which we started most of our MCMC sampling had no EGCs. However, we found that 97.3% and 97.8% of our randomly sampled metabolic networks viable on

glucose and acetate, respectively, harbored at least one EGC.

To determine whether these EGCs artificially inflated the number of innovative offspring, we sampled EGC-free parental metabolic networks. To do so, we modified our MCMC approach such that each sampled metabolic network not only retained viability in a given environment, but was also EGC-free. To fulfill these goals, we required that each step (reaction swap) in our MCMC sampling preserved viability on a given carbon source, and did not introduce an EGC (checked by the EGCs identification approach described above). Using this approach, we generated 1,000 pairs of EGC-free metabolic networks viable exclusively on glucose, and 1,000 pairs of EGC-free networks viable only on acetate. We then generated 1,000 recombinant offspring from each pair. Recombination between EGC-free metabolisms viable exclusively on glucose resulted in 29,941 innovative offspring, only 7.41% fewer than the corresponding number for EGC-containing metabolisms (32,338). Likewise, we observed 46,941 innovative offspring emerging from EGC-free parental metabolisms viable exclusively on acetate, 5.57% fewer than the corresponding number for EGC-containing metabolisms (49,708). Thus, removing EGCs slightly reduces the incidence of innovation (figure S30). Importantly, the patterns of relative constraints remain almost exactly unchanged (figure S31).

Fritzemeier et al. showed that EGCs could artificially increase the biomass rate of metabolic networks by 25% (25). However, figure S32 indicates that the majority of viable networks we study already have a biomass flux considerably larger than our threshold of viability, so reducing their biomass production rate by 25% will not result in a viability loss for most metabolisms, which is why excluding EGCs does not substantially reduce the emergence of novel phenotypes.

Supporting References:

1. Hosseini, S.-R., O. Martin, and A. Wagner. 2016. Phenotypic innovation through recombination in genome-scale metabolic networks. Proc. R. Soc. B. .
2. Edwards, J.S., R.U. Ibarra, and B.O. Palsson. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat. Biotechnol. 19: 125–30.
3. Edwards, J.S., and B.O. Palsson. 1999. Systems properties of the Haemophilus influenzae Rd metabolic genotype. J. Biol. Chem. 274: 17410–6.
4. Lewis, N.E., H. Nagarajan, and B.O. Palsson. 2012. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. Nat. Rev. Microbiol. 10: 291–305.
5. Feist, A.M., and B.Ø. Palsson. 2008. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nat. Biotechnol. 26: 659–67.
6. McCloskey, D., B.Ø. Palsson, and A.M. Feist. 2013. Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. Mol. Syst. Biol.

9: 661.

7. Goto, S., T. Nishioka, and M. Kanehisa. 2000. LIGAND: chemical database of enzyme reactions. Nucleic Acids Res. 28: 380–2.

8. Goto, S., Y. Okuno, M. Hattori, T. Nishioka, and M. Kanehisa. 2002. LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. 30: 402–4.

9. Kanehisa, M., S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 38: D355-60.

10. Matias Rodrigues, J.F., and A. Wagner. 2009. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS Comput. Biol. 5: e1000613.

11. Orth, J.D., I. Thiele, and B.Ø. Palsson. 2010. What is flux balance analysis? Nat. Biotechnol. 28: 245–8.

12. Feist, A.M., C.S. Henry, J.L. Reed, M. Krummenacker, A.R. Joyce, P.D. Karp, L.J. Broadbelt, V. Hatzimanikatis, and B.Ø. Palsson. 2007. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol. Syst. Biol. 3: 121.

13. King, Z.A., J. Lu, A. Dräger, P. Miller, S. Federowicz, J.A. Lerman, A. Ebrahim, B.O. Palsson, and N.E. Lewis. 2015. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. Nucleic Acids Res. : gkv1049-.

14. Lercher, M.J., and C. Pál. 2008. Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol. Biol. Evol. 25: 559–67.

15. Ibarra, R.U., J.S. Edwards, and B.O. Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 420: 186–9.

16. Vieira-Silva, S., and E.P.C. Rocha. 2010. The systemic imprint of growth and its uses in ecological (meta)genomics. PLoS Genet. 6: e1000808.

17. Kirschner, D., and S. Marino. 2005. Mycobacterium tuberculosis as viewed through a computer. Trends Microbiol. 13: 206–11.

18. Fong, S.S., and B.Ø. Palsson. 2004. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. Nat. Genet. 36: 1056–8.

19. Fong, S.S., J.Y. Marciniak, and B.O. Palsson. 2003. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. J. Bacteriol. 185: 6400–6408.

20. Samal, A., J.F. Matias Rodrigues, J. Jost, O.C. Martin, and A. Wagner. 2010. Genotype networks in metabolic reaction spaces. BMC Syst. Biol. 4: 30.

21. Wagner, A., and D.A. Fell. 2001. The small world inside large metabolic networks. Proc. R. Soc. B Biol. Sci. 268: 1803–1810.

22. Barve, A., J.F.M. Rodrigues, and A. Wagner. 2012. Superessential reactions in metabolic networks. Proc. Natl. Acad. Sci. U. S. A. 109: E1121-30.

23. Ma, H.-W., and A.-P. Zeng. 2003. The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics. 19: 1423–30.

24. Hopcroft, J., and R. Tarjan. 1973. Algorithm 447: efficient algorithms for graph manipulation. Commun. ACM. 16: 372–378.

25. Fritzemeier, C.J., D. Hartleb, B. Szappanos, B. Papp, M.J. Lercher, and G. Fekete. 2017. Erroneous energy-generating cycles in published genome scale metabolic networks: Identification and removal. PLOS Comput. Biol. 13: e1005494.

2. Supplementary Figures

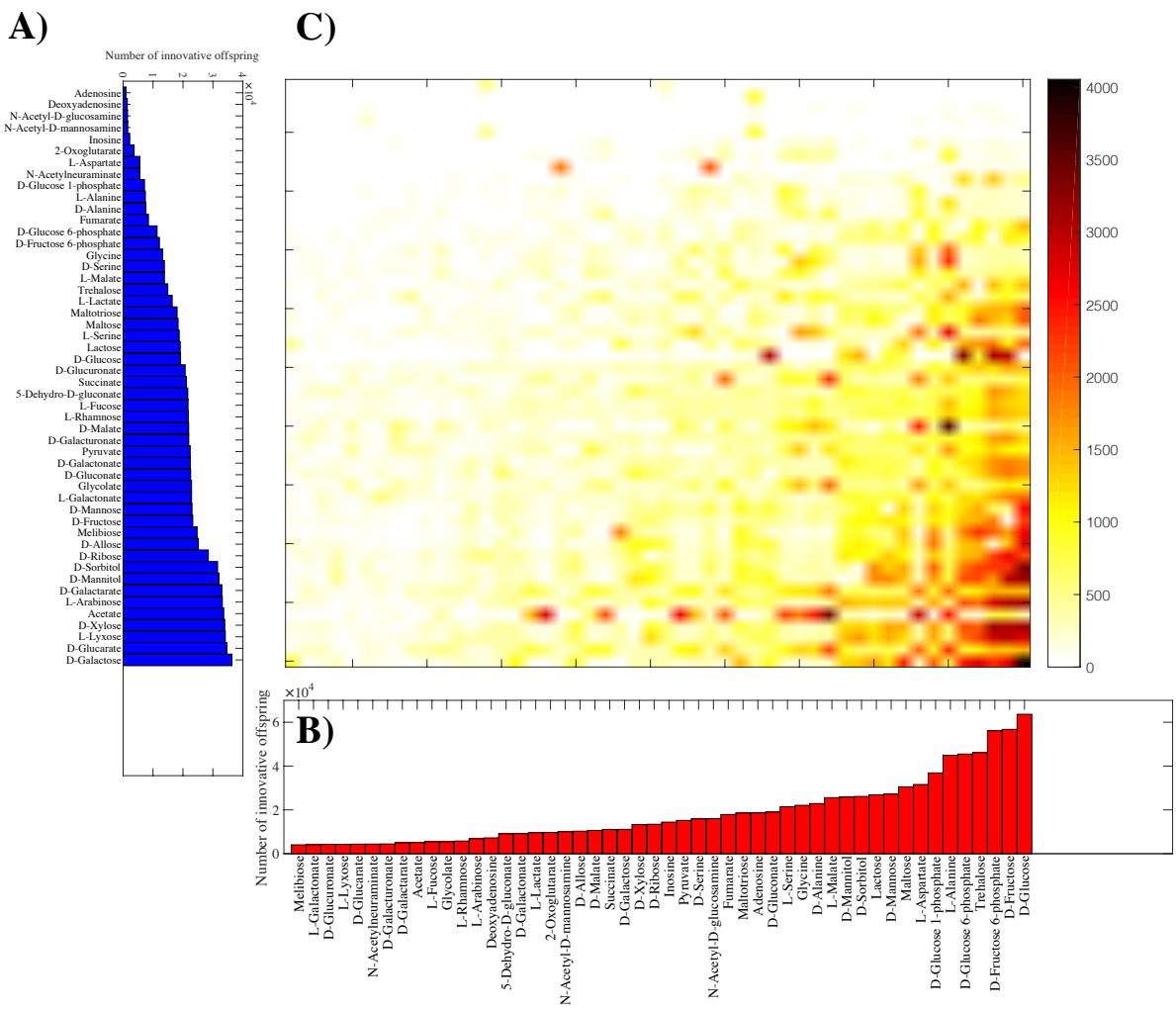


Figure S1: Recombination can create all 50 carbon-use phenotypes considered here (*n* = 20). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 37, ranging from 977 on Adenosine to 356,378 on D-galactose. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 15, ranging from 4,042 on melibiose to 63,634 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same as in the *E.coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 20$ reactions are swapped between parental metabolic networks in a recombination event.

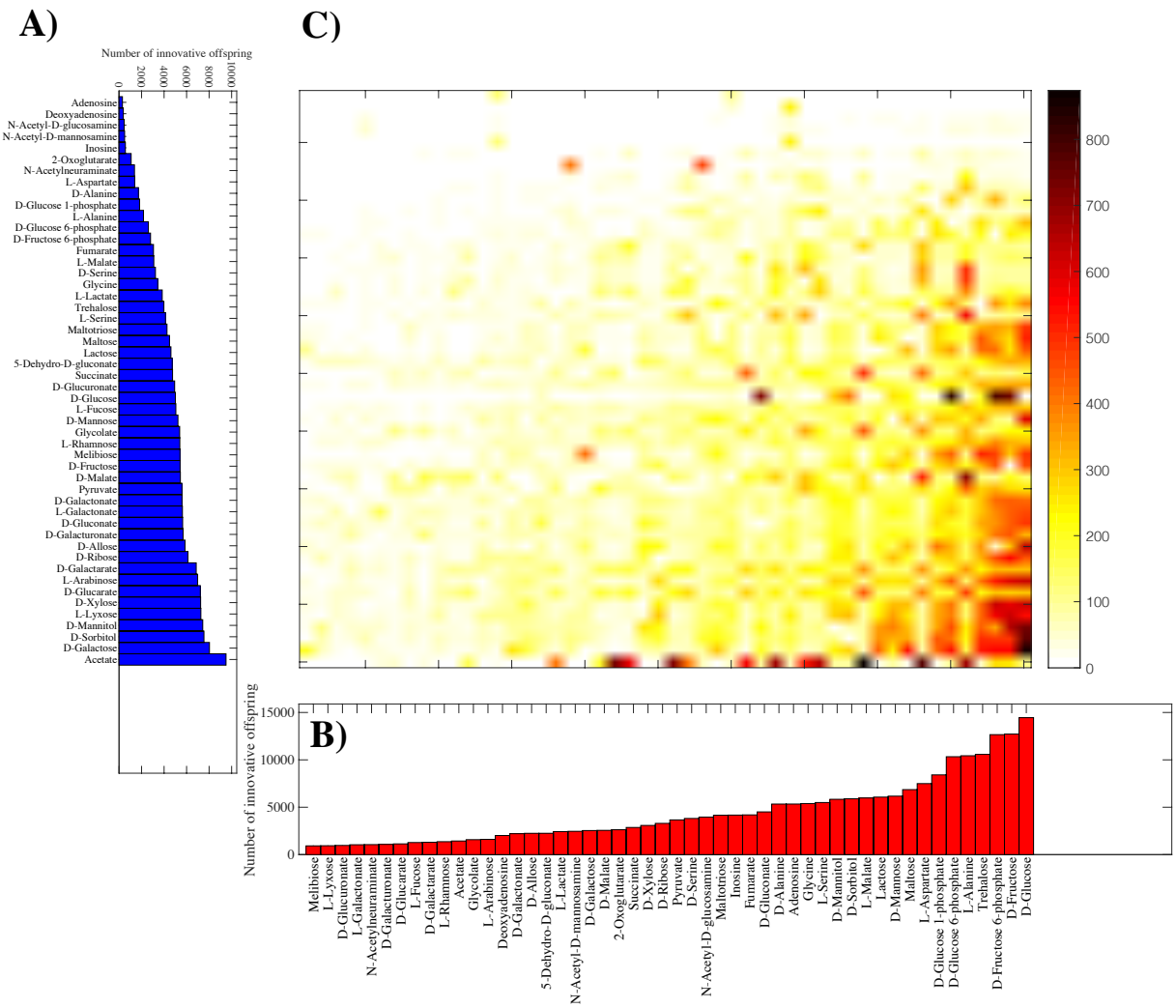


Figure S2: Recombination can create all 50 carbon-use phenotypes considered here (*n* = 30). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 32, ranging from 299 on adenosine to 9,503 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 16, ranging from 923 on melibiose to 14,452 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G|| = 2{,}079$ reactions, the same as in the *E.coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 30$ reactions are swapped between parental metabolic networks in a recombination event.

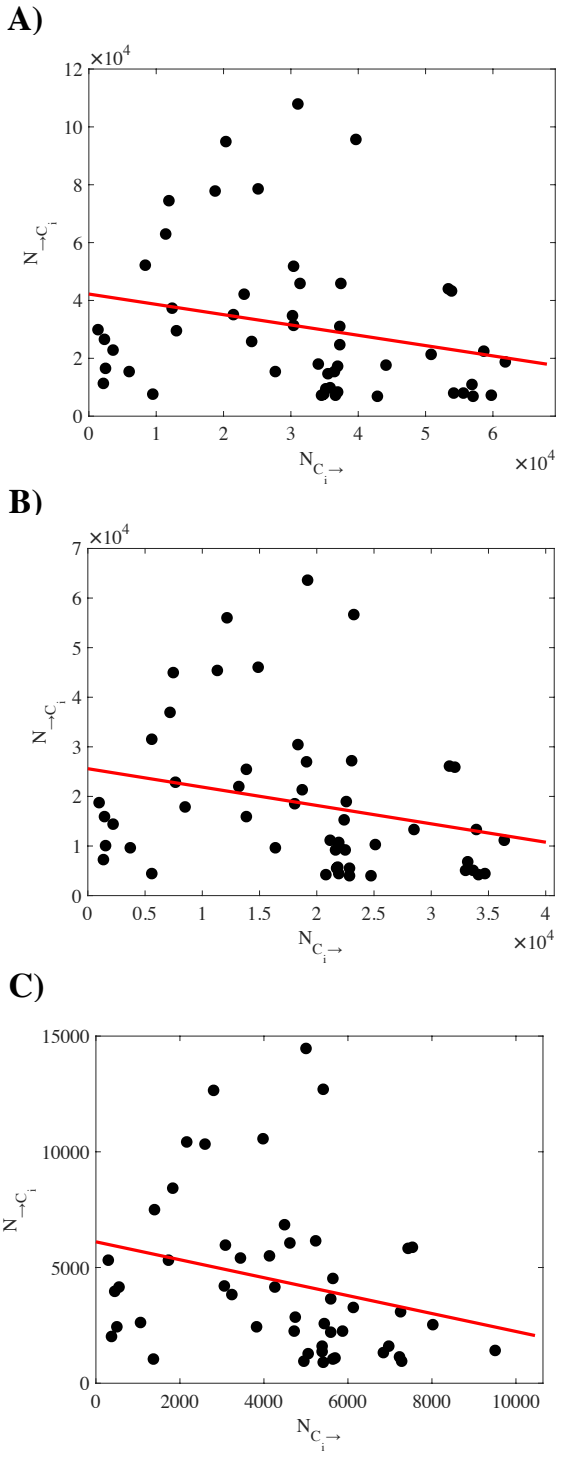
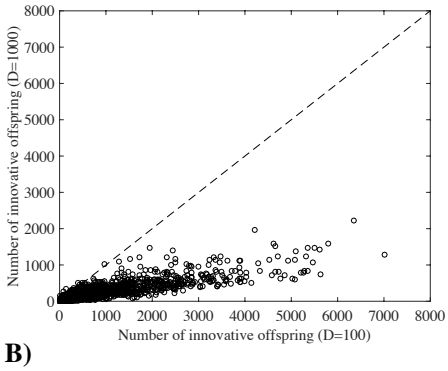
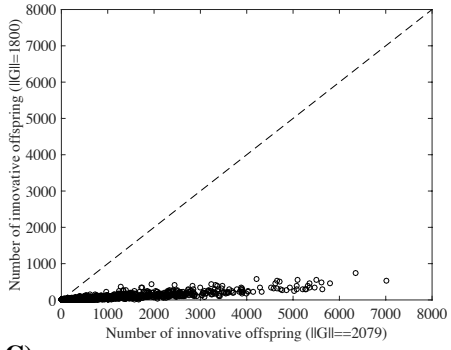


Figure S3: Negative correlation between ($N_{C_i\rightarrow}$) **and** ($N_{\rightarrow C_i}$)**.** Each circle corresponds to a given carbon source C_i. The vertical axis shows ($N_{C_i\rightarrow}$), the number of metabolic innovations emerging from parents viable on carbon source C_i. The horizontal axis shows ($N_{\rightarrow C_i}$), the number of innovations leading to viability on C_i. There is a negative correlation between ($N_{C_i\rightarrow}$) and ($N_{\rightarrow C_i}$), regardless of the number (n) of reactions exchanged: **A)** (n = 10, Pearson r = -0.239, P < 0.093), **B)** (n = 20, Pearson r = -0.248, P < 0.082), **C)** (n = 30, Pearson r = -0.256, P < 0.073). For all analyses the genotypic distance between parents is D = 100.

A)



B)



C)

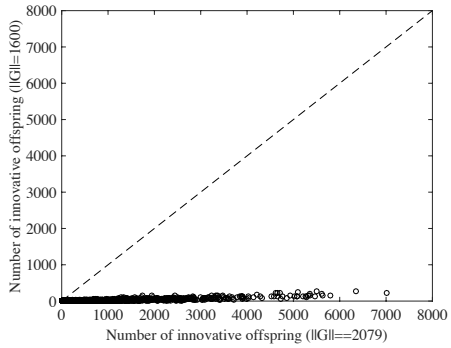


Figure S4: Fewer innovative offspring at higher genotypic distance (*D*) and smaller metabolic network size ||*G*||. Each circle corresponds to a pair of carbon sources (C_i, C_j) and shows the number of innovative offspring gaining viability on C_j, which are generated by recombination between parents viable on carbon source C_i. The horizontal axis specifies the number of innovative offspring where parents have genetic distance *D* = 100, and metabolic network size ||*G*|| = 2,079. The vertical axes provide the same information, but for parents with A) genotypic distance *D* = 1,000, and metabolic network size ||*G*|| = 2,079 reactions, B) genotypic distance *D* = 100, and metabolic network size ||*G*|| = 1,800 reactions, and C) genotypic distance *D* = 100, and metabolic network size ||*G*|| = 1,600 reactions. The dashed diagonal lines correspond to the identity line (*y* = *x*). Note that in all three panels, most or all data lie below this line, indicating that higher parental genotypic distance and lower metabolic network size lead to fewer innovative offspring for almost all carbon source pair.

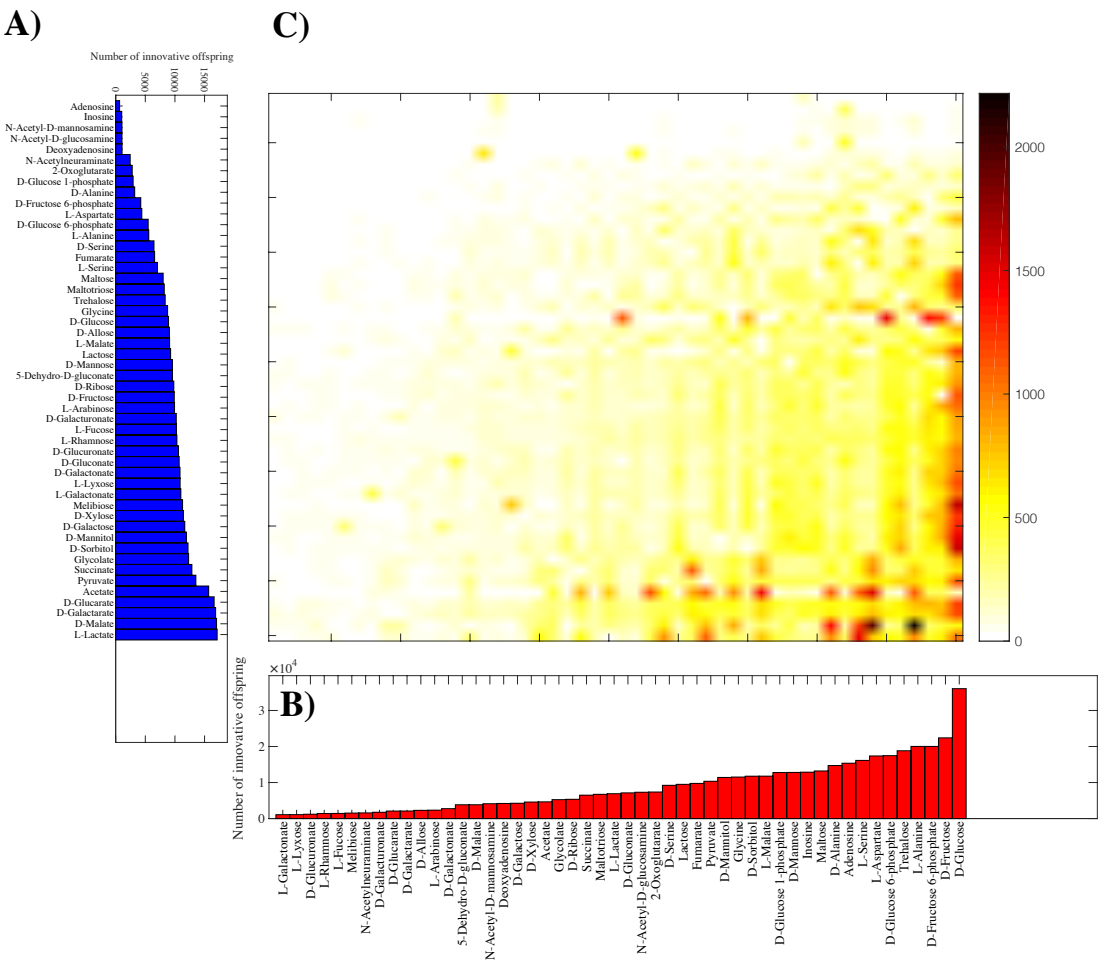


Figure S5: Recombination can create all 50 carbon-use phenotypes considered here (*D* = 1,000). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 25, ranging from 662 on adenosine to 17,132 on L-lactate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 33, ranging from 1081 on L-galactonate to 36,051 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as the *E.coli* metabolic network, and they differ in D = 1,000 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

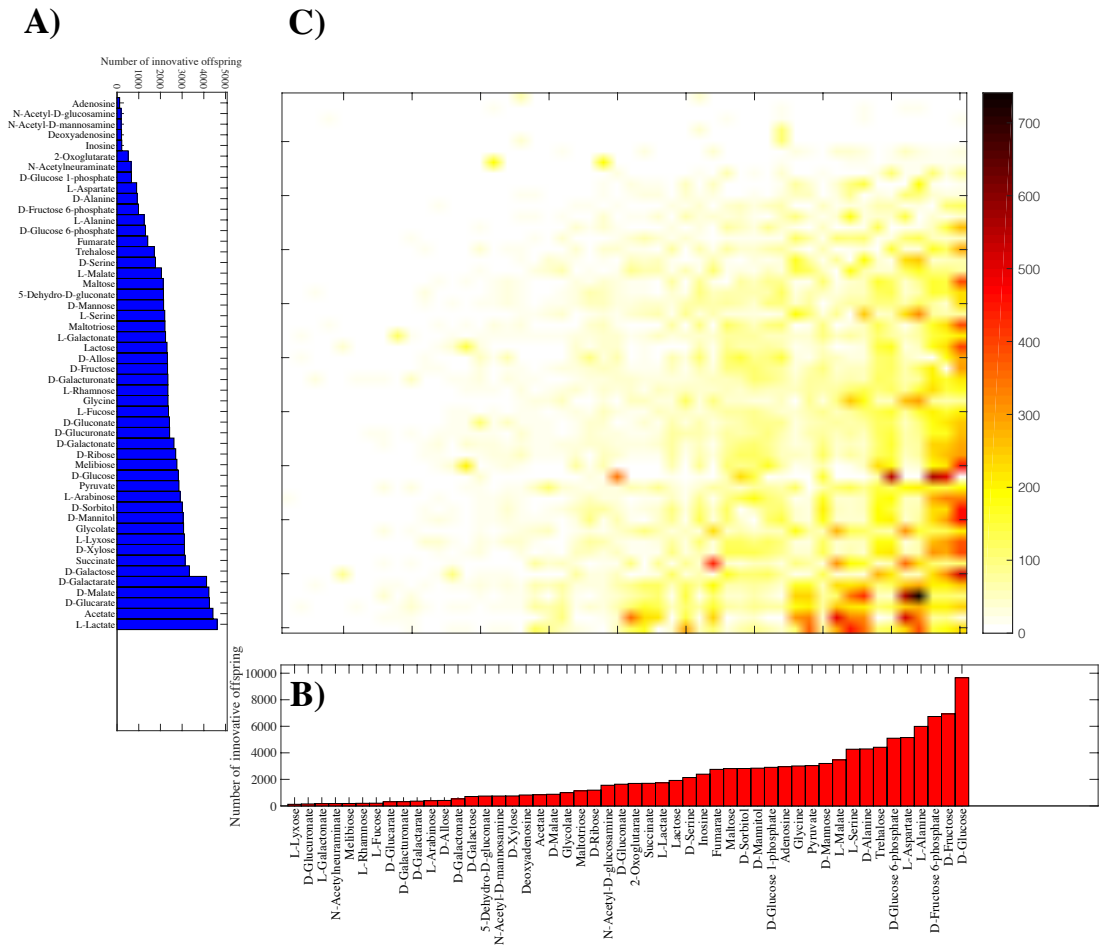


Figure S6: Recombination can create all 50 carbon-use phenotypes considered here (||*G*|| = 1800). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 38, ranging from 120 on adenosine to 4,616 on L-lactate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 79, ranging from 122 on L-lyxose to 9,657 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain ||*G*|| = 1,800 reactions and differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

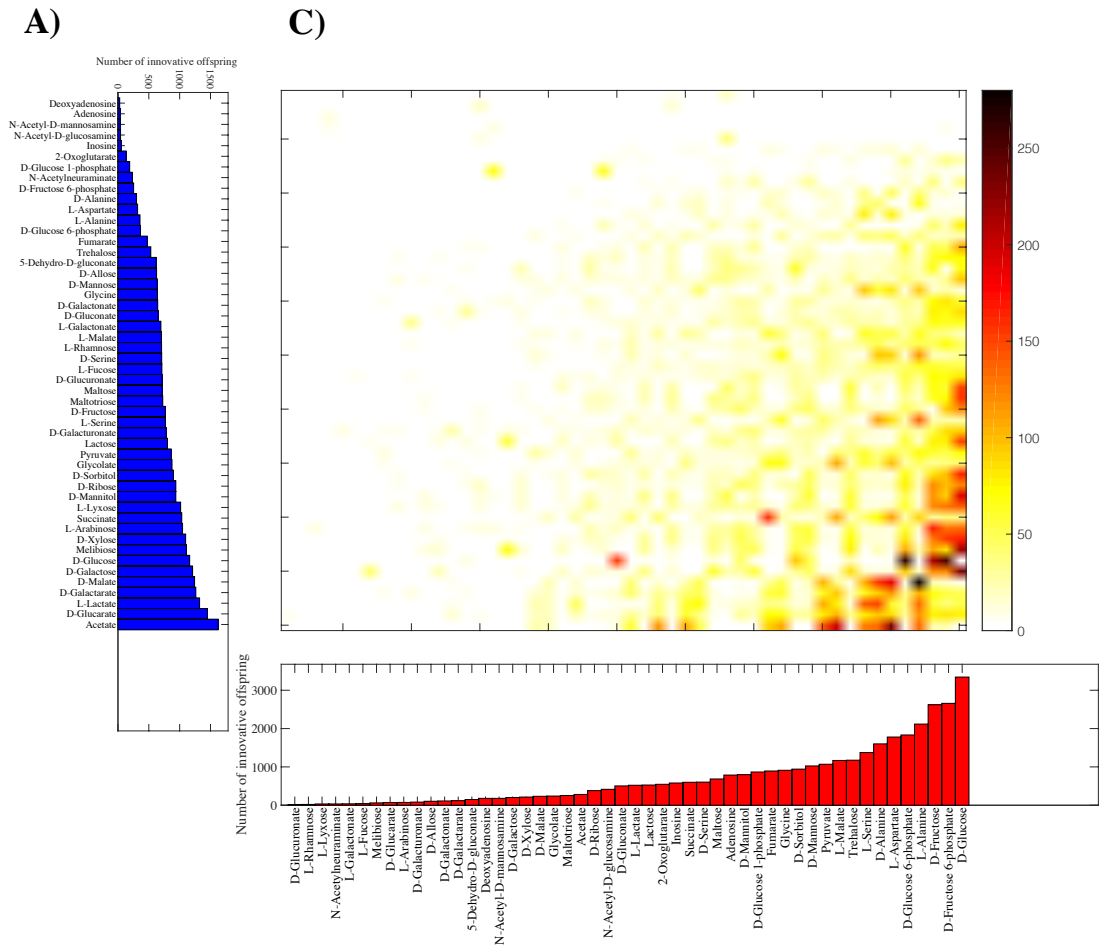


Figure S7: Recombination can create all 50 carbon-use phenotypes considered here (||*G*|| = 1,600). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 58, ranging from 28 on deoxyadenosine to 1,623 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 176, ranging from 19 on D-glucuronate to 3,344 on D-glucose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain ||*G*|| = 1,600 reactions and differ in *D* = 100 reactions. Moreover, *n* = 10 reactions are swapped between parental metabolic networks in a recombination event.

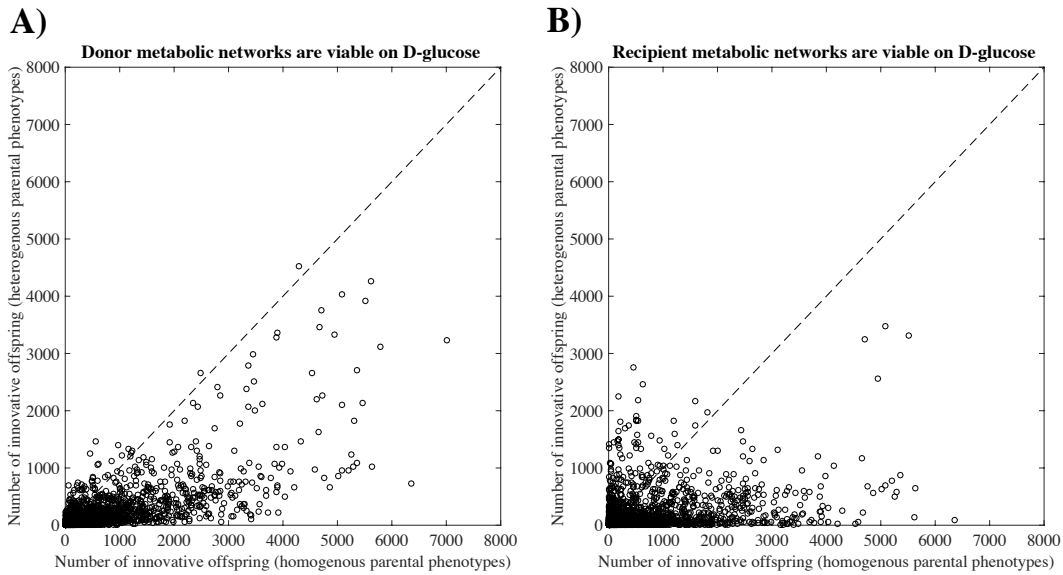


Figure S8: Fewer innovative offspring from phenotypically heterogeneous parents than from phenotypically homogenous parents. Each circle corresponds to a given pair of carbon sources (C_i, C_j) and shows the number of innovative offspring gaining viability on C_j, that are generated by recombination between parents viable on carbon source C_i. The horizontal axis specifies the number of innovative offspring for parents that are viable on the same carbon sources (phenotypically homogeneous parents). The vertical axes show the number of innovative offspring for **A)** parental donors viable on D-glucose and parental recipients viable on C_i, and **B)** parental recipients are viable on D-glucose, and parental donors viable on C_i. In these analyses, all parents have ||*G*|| = 2,079 reactions, the same as the *E.coli* metabolic network, and their genotypic distance (*D*) is constant and equals 100. Note that in both panels, the majority of circles (with few exceptions) are placed below the identity ($y = x$) line, indicating that it is more likely for phenotypically homogenous parents to generate innovative offspring than for phenotypically heterogeneous parents.

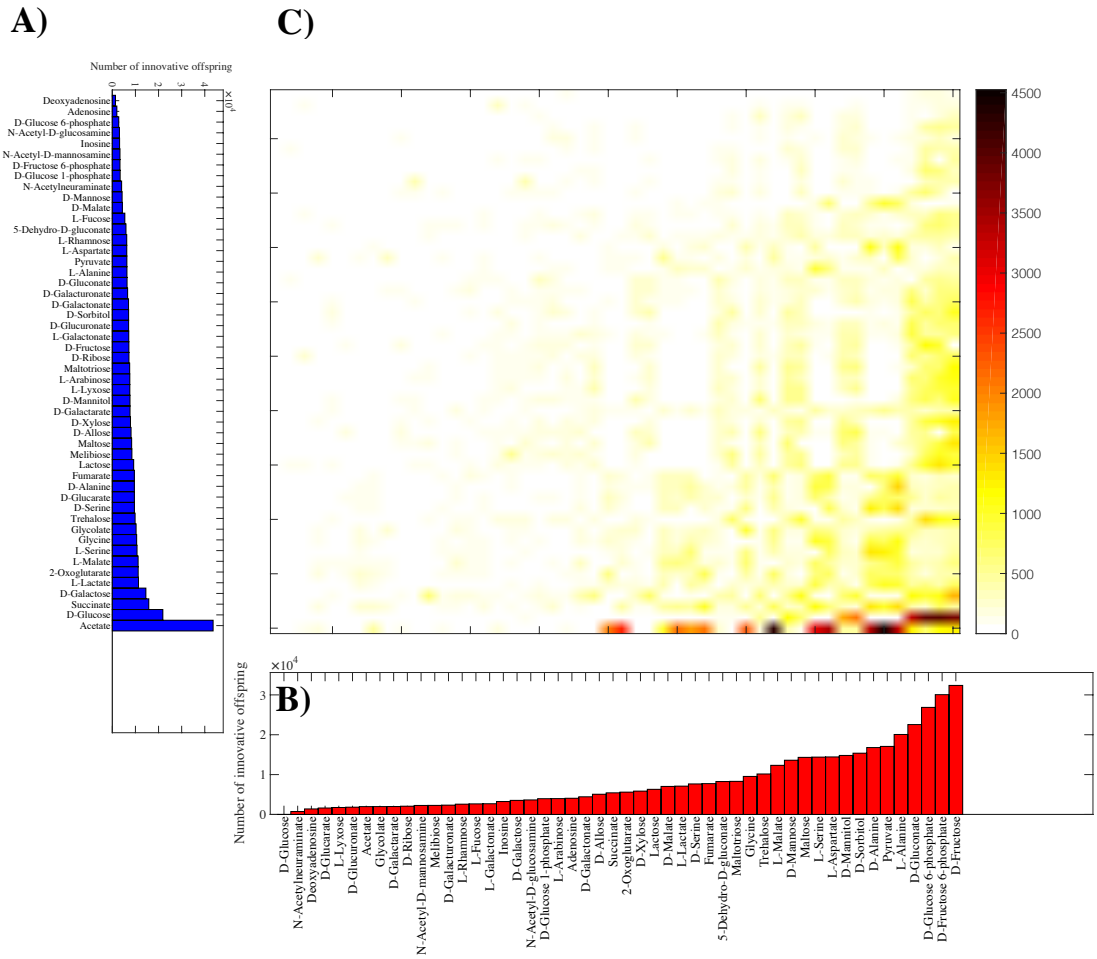


Figure S9: Recombination can create all 50 carbon-use phenotypes considered here (Parents with heterogeneous phenotypes, donors viable only on glucose). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between donor parents viable on glucose and recipient parents that are viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 32, ranging from 1,371 on deoxyadenosine to 43,615 on acetate. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 44, ranging from 729 on N-acetylneuraminate to 32,378 on D-fructose. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between donor parents viable on glucose, and recipient parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

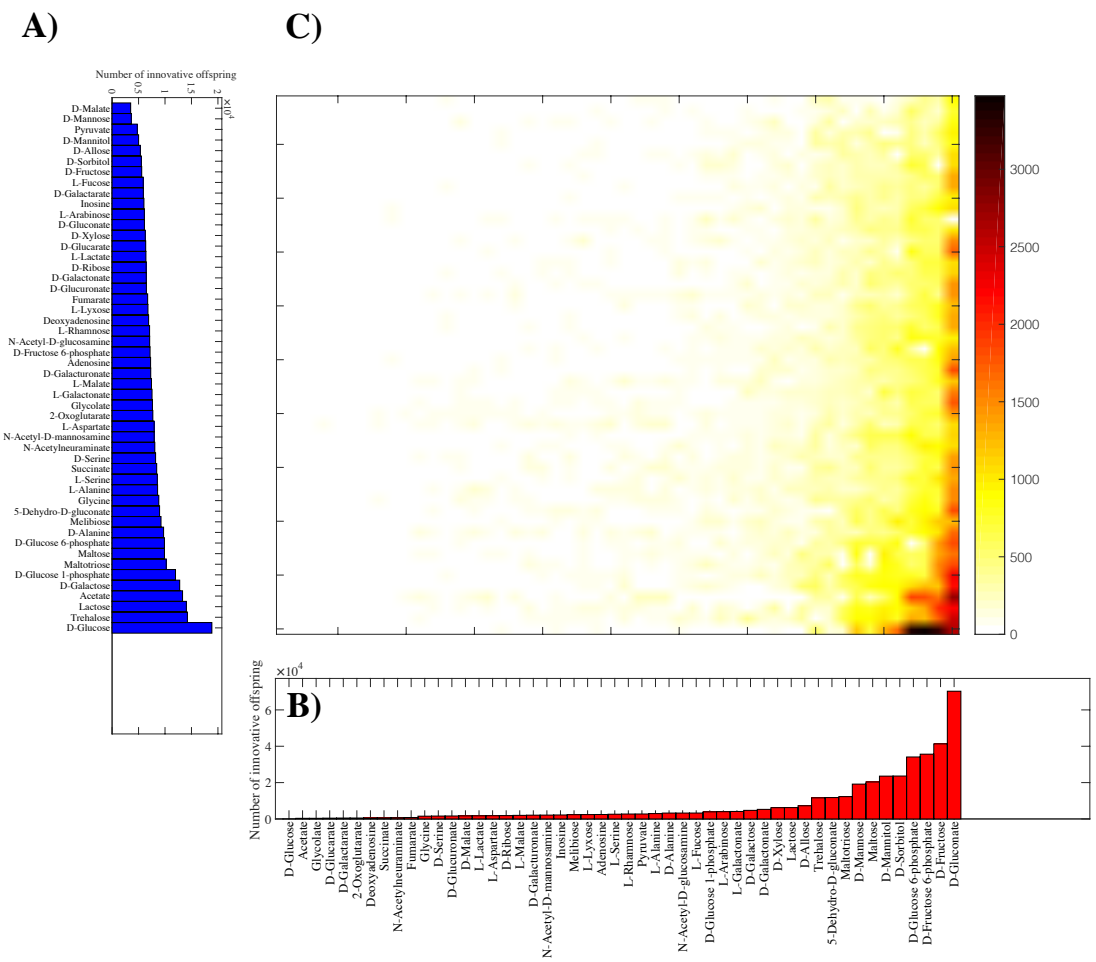
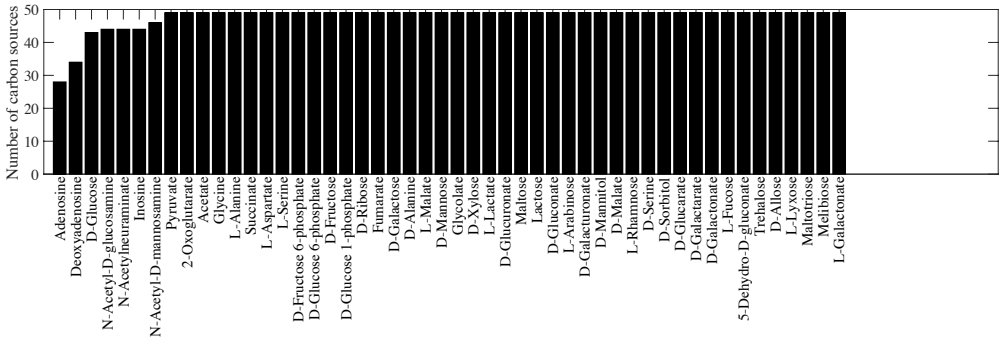
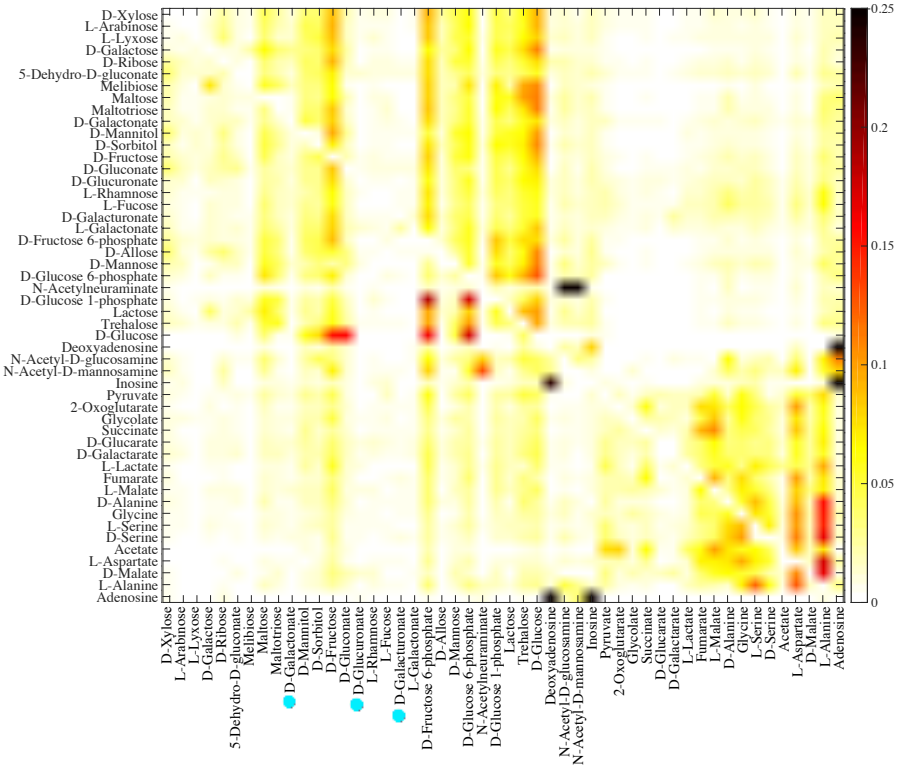


Figure S10: Recombination can create all 50 carbon-use phenotypes considered here (Parents with heterogeneous phenotypes, recipients viable only on glucose). A) The horizontal axis shows the number of innovative recombinant offspring (out of one million offspring) resulting from recombination between recipient parents viable on glucose and donor parents viable exclusively on the carbon source specified on the vertical axis. This number varies by a factor 5, ranging from 3,511 on D-malate to 18,856 on D-glucose. **B)** Number of innovative recombinants (per million offspring) gaining viability on the novel carbon source specified on the *x*-axis. This number varies by a factor 204, ranging from 343 on acetate to 70,292 on D-gluconate. **C)** Number of innovative recombinants (per million offspring, color-coded according to the legend) resulting from recombination between recipient parents viable on glucose, and donor parents viable exclusively on the carbon source specified in panel A, which have gained viability on the novel carbon source specified in panel B. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Xylose
L-Arabinose
L-Lyxose
D-Galactose
D-Ribose
5-Dehydro-D-gluconate
Melibiose
Maltose
Maltotriose
D-Galactonate
D-Mannitol
D-Sorbitol
D-Fructose
D-Gluconate
D-Glucuronate
L-Rhamnose
L-Fucose
D-Galacturonate
L-Galactonate
D-Fructose 6-phosphate
D-Allose
D-Mannose
D-Glucose 6-phosphate
N-Acetylneuraminate
D-Glucose 1-phosphate
Lactose
Trehalose
D-Glucose
Deoxyadenosine
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Inosine
Pyruvate
2-Oxoglutarate
Glycolate
Succinate
D-Glucarate
D-Galactarate
L-Lactate
Fumarate
L-Malate
D-Alanine
Glycine
L-Serine
D-Serine
Acetate
L-Aspartate
D-Malate
L-Alanine
Adenosine
0.25
0.2
0.15
0.1
0.05
0

C)

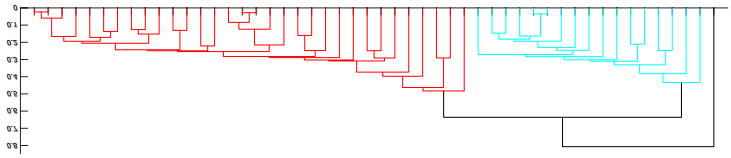
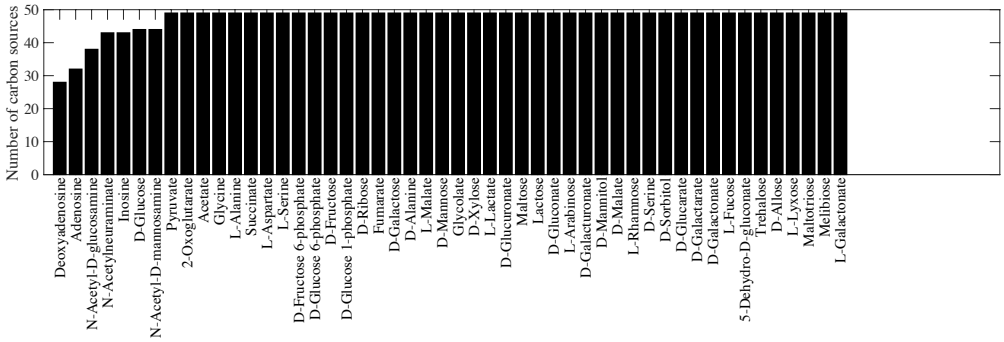
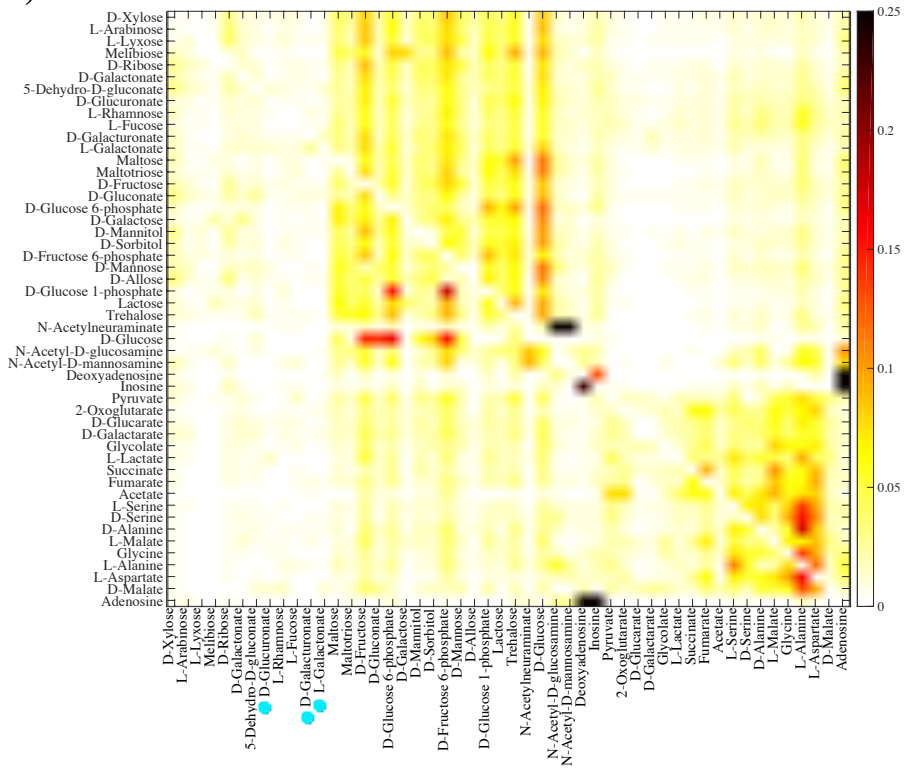


Figure S11: Emergence of innovative offspring can be constrained by parental phenotypes (*n* = 20). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate, (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 20 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
D-Glucose
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Xylose
L-Arabinose
L-Lyxose
Melibiose
D-Ribose
D-Galactonate
5-Dehydro-D-gluconate
D-Glucuronate
L-Rhamnose
L-Fucose
D-Galacturonate
L-Galactonate
Maltose
Maltotriose
D-Fructose
D-Gluconate
D-Glucose 6-phosphate
D-Galactose
D-Mannitol
D-Sorbitol
D-Fructose 6-phosphate
D-Mannose
D-Allose
D-Glucose 1-phosphate
Lactose
Trehalose
N-Acetylneuraminate
D-Glucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
Pyruvate
2-Oxoglutarate
D-Glucarate
D-Galactarate
Glycolate
L-Lactate
Succinate
Fumarate
Acetate
L-Serine
D-Serine
D-Alanine
L-Malate
Glycine
L-Alanine
L-Aspartate
D-Malate
Adenosine
0.25
0.2
0.15
0.1
0.05
0

C)

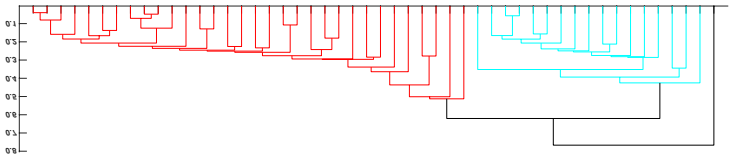
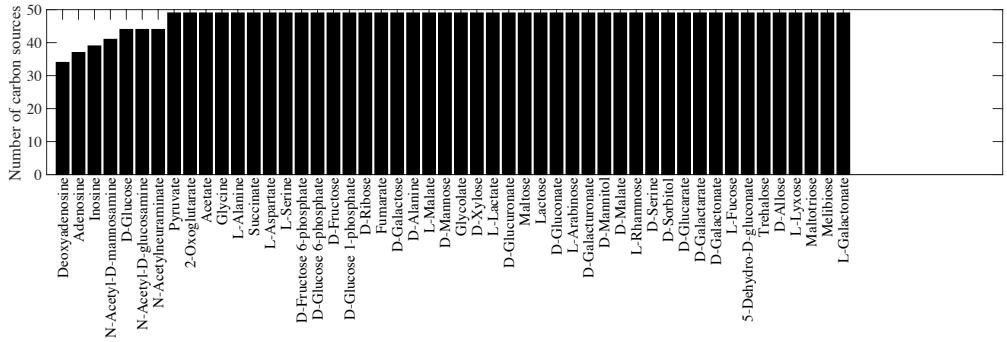
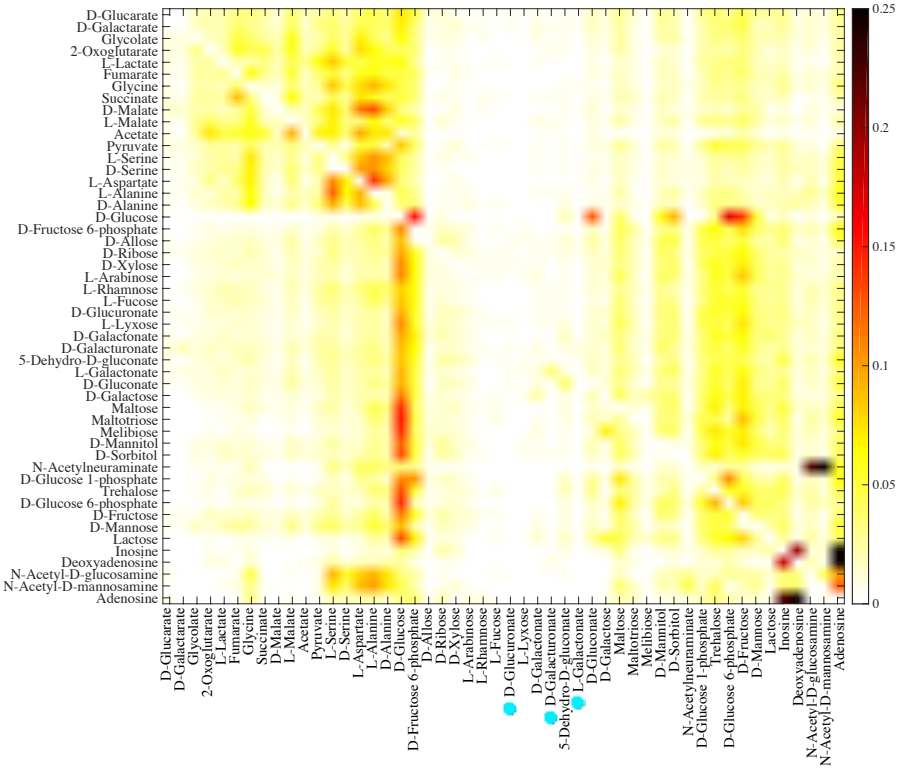


Figure S12: Emergence of innovative offspring can be constrained by parental phenotypes (*n* = 30). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 30 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
50
40
30
20
10
0
Deoxyadenosine
Adenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Glucarate
D-Galactarate
Glycolate
2-Oxoglutarate
L-Lactate
Fumarate
Glycine
Succinate
D-Malate
L-Malate
Acetate
Pyruvate
L-Serine
D-Serine
L-Aspartate
L-Alanine
D-Alanine
D-Glucose
D-Fructose 6-phosphate
D-Allose
D-Ribose
D-Xylose
L-Arabinose
L-Rhamnose
L-Fucose
D-Glucuronate
L-Lyxose
D-Galactonate
D-Galacturonate
5-Dehydro-D-gluconate
L-Galactonate
D-Gluconate
D-Galactose
Maltose
Maltotriose
Melibiose
D-Mannitol
D-Sorbitol
N-Acetylneuraminate
D-Glucose 1-phosphate
Trehalose
D-Glucose 6-phosphate
D-Fructose
D-Mannose
Lactose
Inosine
Deoxyadenosine
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
0.25
0.2
0.15
0.1
0.05
0

C)

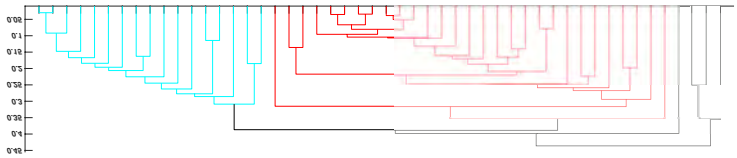
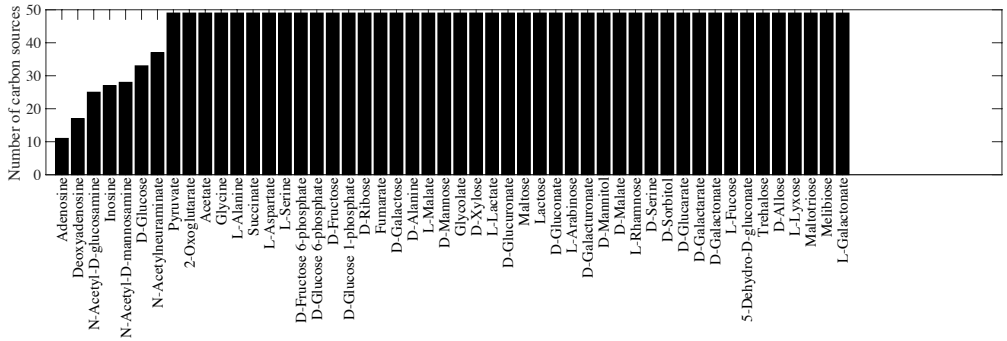
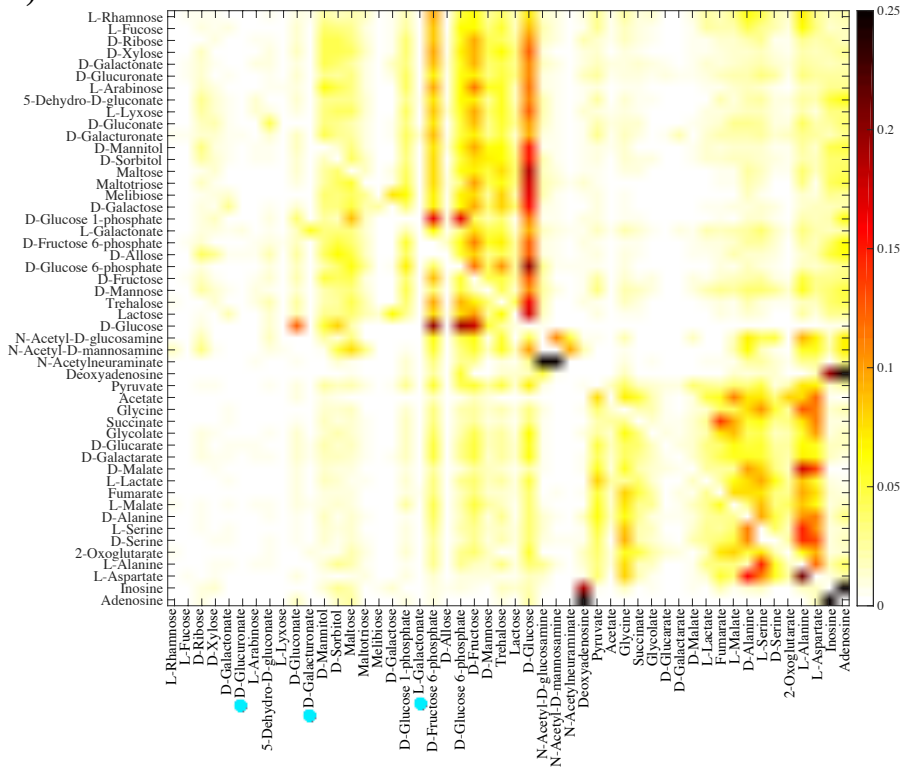


Figure S13: Emergence of innovative offspring can be constrained by parental phenotypes (*D* = 1,000). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 1,000 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


L-Rhamnose
L-Fucose
D-Ribose
D-Xylose
D-Galactonate
D-Glucuronate
L-Arabinose
5-Dehydro-D-gluconate
L-Lyxose
D-Gluconate
D-Galacturonate
D-Mannitol
D-Sorbitol
Maltose
Maltotriose
Melibiose
D-Galactose
D-Glucose 1-phosphate
L-Galactonate
D-Fructose 6-phosphate
D-Allose
D-Glucose 6-phosphate
D-Fructose
D-Mannose
Trehalose
Lactose
D-Glucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
N-Acetylneuraminate
Deoxyadenosine
Pyruvate
Acetate
Glycine
Succinate
Glycolate
D-Glucarate
D-Galactarate
D-Malate
L-Lactate
Fumarate
L-Malate
D-Alanine
L-Serine
D-Serine
2-Oxoglutarate
L-Alanine
L-Aspartate
Inosine
Adenosine
0.25
0.2
0.15
0.1
0.05
0

C)

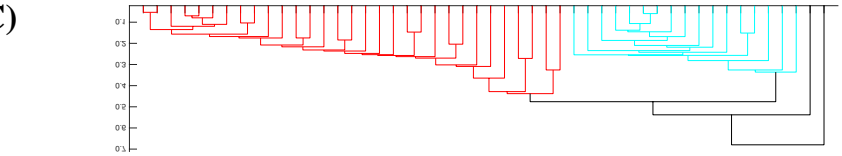
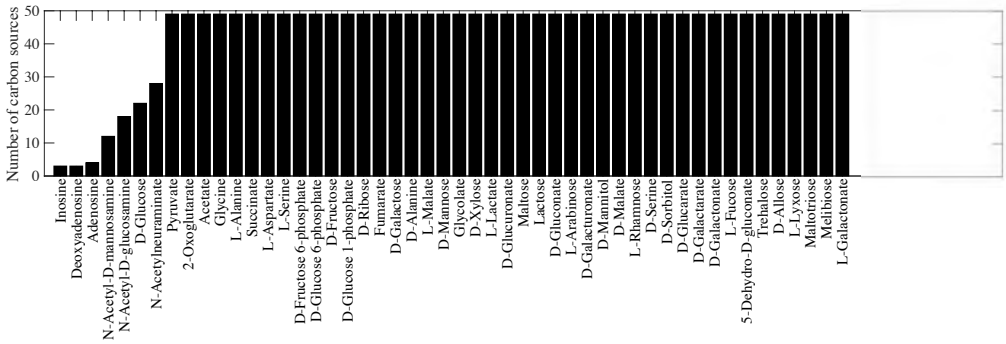
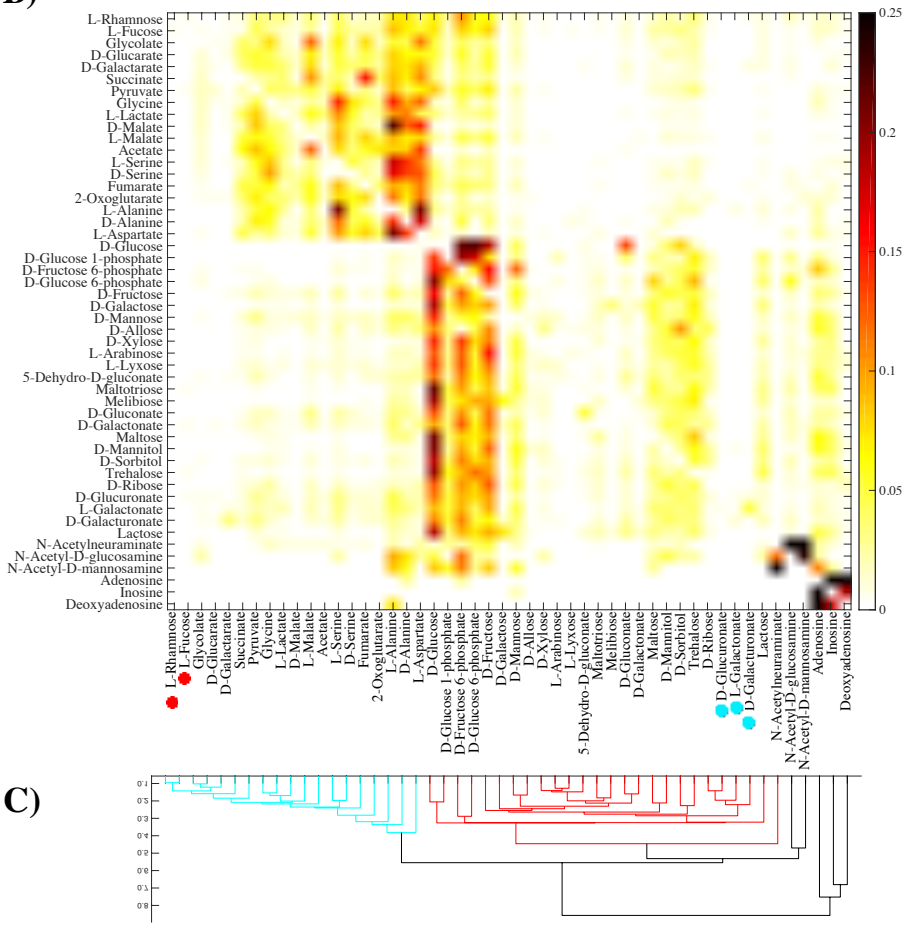


Figure S14: Emergence of innovative offspring can be constrained by parental phenotypes (||*G*|| = 1,800). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain ||*G*|| = 1,800 reactions, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Inosine
Deoxyadenosine
Adenosine
N-Acetyl-D-mannosamine
N-Acetyl-D-glucosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

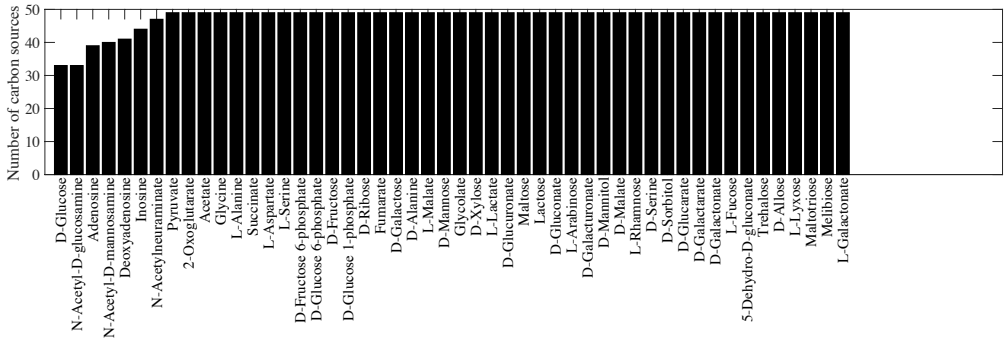
B)


L-Rhamnose
L-Fucose
Glycolate
D-Glucarate
D-Galactarate
Succinate
Pyruvate
Glycine
L-Lactate
D-Malate
L-Malate
Acetate
L-Serine
D-Serine
Fumarate
2-Oxoglutarate
L-Alanine
D-Alanine
L-Aspartate
D-Glucose
D-Glucose 1-phosphate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Galactose
D-Mannose
D-Allose
D-Xylose
L-Arabinose
L-Lyxose
5-Dehydro-D-gluconate
Maltotriose
Melibiose
D-Gluconate
D-Galactonate
Maltose
D-Mannitol
D-Sorbitol
Trehalose
D-Ribose
D-Glucuronate
L-Galactonate
D-Galacturonate
Lactose
N-Acetylneuraminate
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
Inosine
Deoxyadenosine
0.25
0.2
0.15
0.1
0.05
0

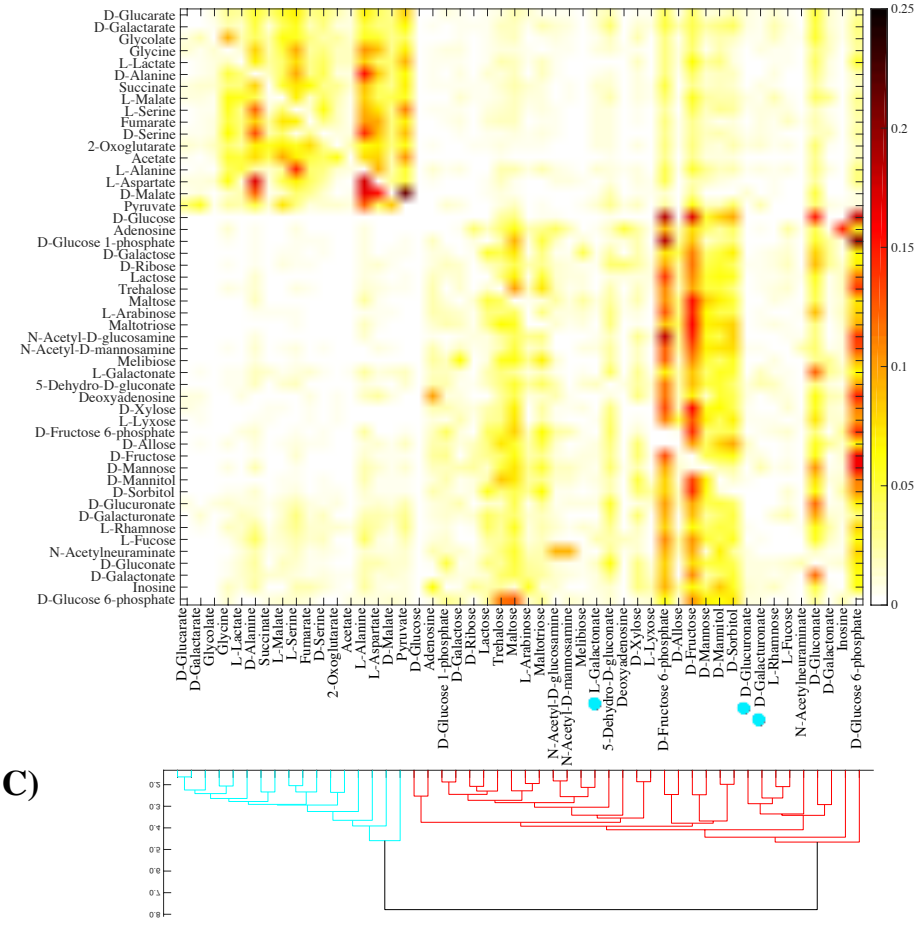
C)

Figure S15: Emergence of innovative offspring can be constrained by parental phenotypes (||*G*|| = 1,600). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between parents viable exclusively on the carbon source specified on the vertical axis, which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). In these analyses, parental metabolic networks contain ||*G*|| = 1,600 reactions, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
D-Glucose
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)

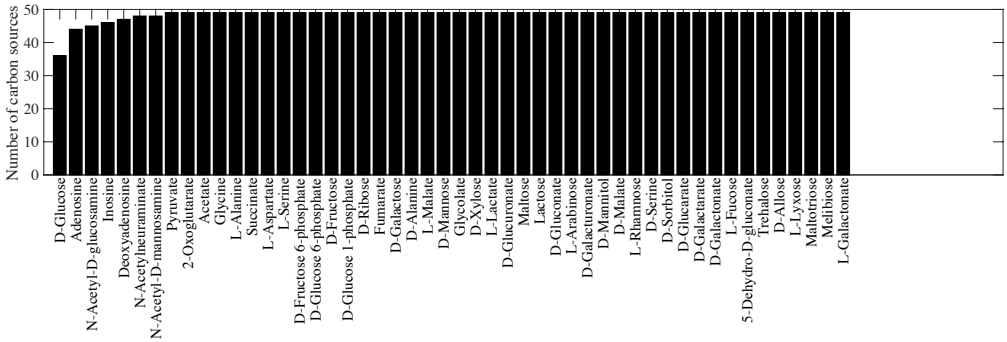

D-Glucarate
D-Galactarate
Glycolate
Glycine
L-Lactate
D-Alanine
Succinate
L-Malate
L-Serine
Fumarate
D-Serine
2-Oxoglutarate
L-Alanine
L-Aspartate
D-Malate
Pyruvate
D-Glucose
Adenosine
D-Glucose 1-phosphate
D-Galactose
D-Ribose
Lactose
Trehalose
Maltose
L-Arabinose
Maltotriose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Melibiose
L-Galactonate
5-Dehydro-D-gluconate
Deoxyadenosine
D-Xylose
L-Lyxose
D-Fructose 6-phosphate
D-Allose
D-Fructose
D-Mannose
D-Mannitol
D-Sorbitol
D-Glucuronate
D-Galacturonate
L-Rhamnose
L-Fucose
N-Acetylneuraminate
D-Gluconate
D-Galactonate
Inosine
D-Glucose 6-phosphate
0.25
0.2
0.15
0.1
0.05
0

C)

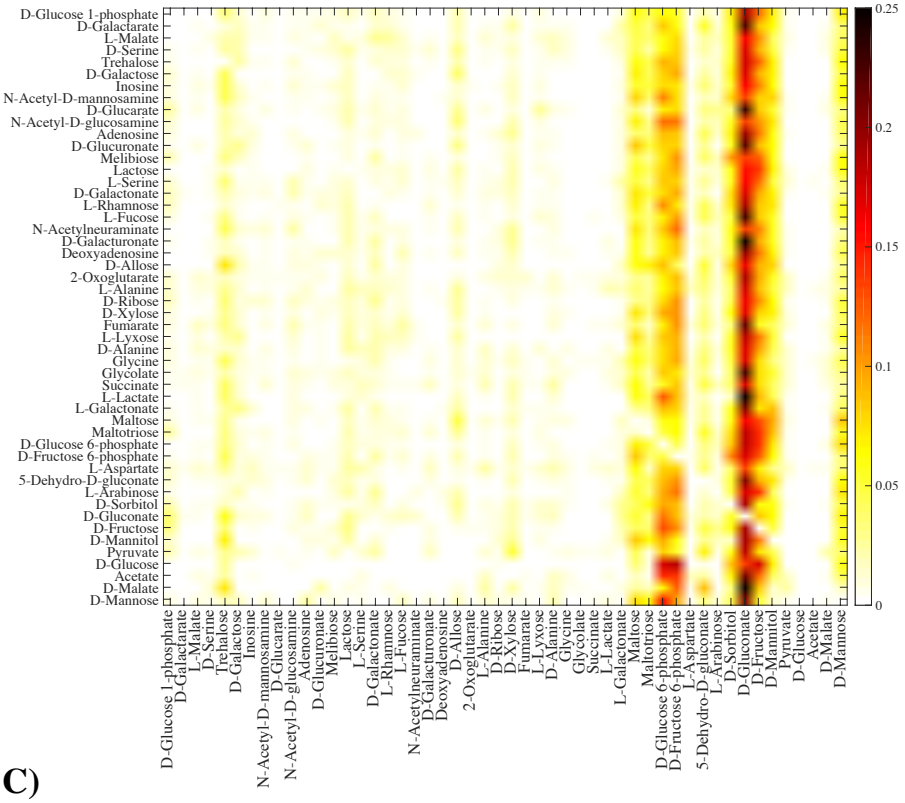
Figure S16: Emergence of innovative offspring can be constrained by parental phenotypes (Parents with heterogeneous phenotypes, donors viable only on glucose). A) The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between donor parents viable on glucose and the recipient parents viable exclusively on the carbon source specified on the vertical axis., which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $\|G\| = 2{,}079$ reactions, the same number as in the *E.coli* metabolic network, and they differ in $D = 100$ reactions. Moreover, $n = 10$ reactions are swapped between parental metabolic networks in a recombination event.

=

A)


Number of carbon sources
0
10
20
30
40
50
D-Glucose
Adenosine
N-Acetyl-D-glucosamine
Inosine
Deoxyadenosine
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Glucose 1-phosphate
D-Galactarate
L-Malate
D-Serine
Trehalose
D-Galactose
Inosine
N-Acetyl-D-mannosamine
D-Glucarate
N-Acetyl-D-glucosamine
Adenosine
D-Glucuronate
Melibiose
Lactose
L-Serine
D-Galactonate
L-Rhamnose
L-Fucose
N-Acetylneuraminate
D-Galacturonate
Deoxyadenosine
D-Allose
2-Oxoglutarate
L-Alanine
D-Ribose
D-Xylose
Fumarate
L-Lyxose
D-Alanine
Glycine
Glycolate
Succinate
L-Lactate
L-Galactonate
Maltose
Maltotriose
D-Glucose 6-phosphate
D-Fructose 6-phosphate
L-Aspartate
5-Dehydro-D-gluconate
L-Arabinose
D-Sorbitol
D-Gluconate
D-Fructose
D-Mannitol
Pyruvate
D-Glucose
Acetate
D-Malate
D-Mannose
0.25
0.2
0.15
0.1
0.05
0

C)

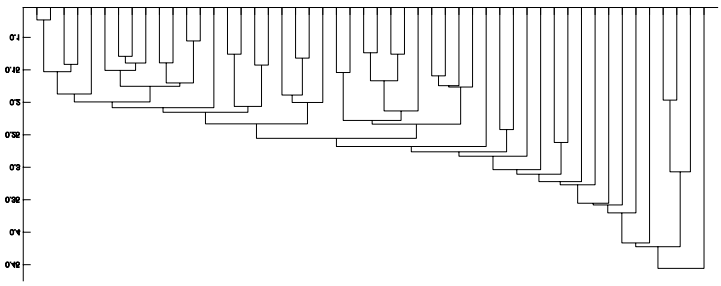
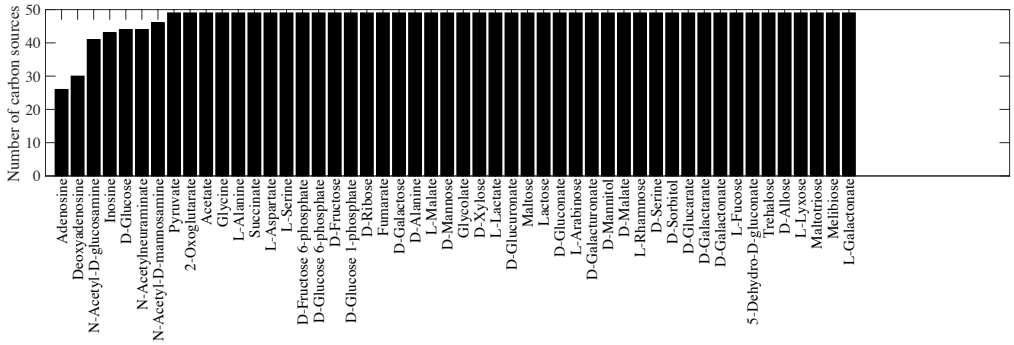
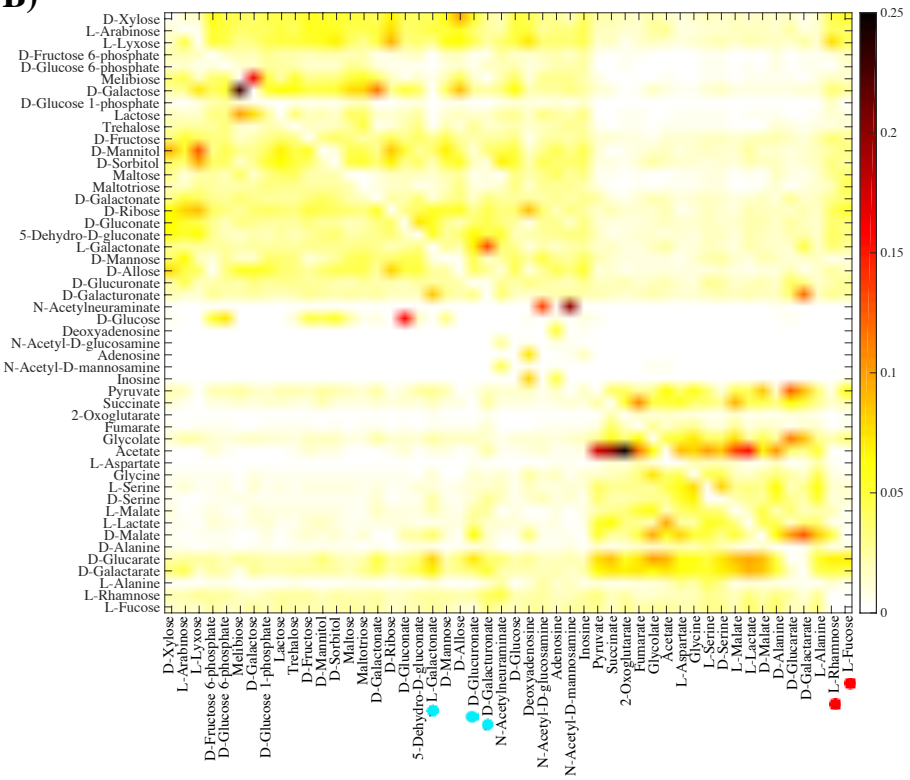


Figure S17: Emergence of innovative offspring can be constrained by parental phenotypes (Parents with heterogeneous phenotypes, recipients viable only on glucose). **A)** The horizontal axis shows the carbon source on which parental metabolisms are viable, and the vertical axis shows the number of novel carbon sources (among the remaining 49 carbon sources) on which at least one innovative offspring results from recombination between parental metabolic networks. **B)** Fraction of innovative recombinants (color-coded according to the legend) resulting from recombination between recipient parents viable on glucose and donor parents viable exclusively on the carbon source specified on the vertical axis., which have gained viability on the novel carbon source specified on the horizontal axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. In this figure, main branches do not reflect glycolytic and gluconeogenic carbon sources as in other figures. In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
D-Glucose
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Xylose
L-Arabinose
L-Lyxose
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Melibiose
D-Galactose
D-Glucose 1-phosphate
Lactose
Trehalose
D-Fructose
D-Mannitol
D-Sorbitol
Maltose
Maltotriose
D-Galactonate
D-Ribose
D-Gluconate
5-Dehydro-D-gluconate
L-Galactonate
D-Mannose
D-Allose
D-Glucuronate
D-Galacturonate
N-Acetylneuraminate
D-Glucose
Deoxyadenosine
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Inosine
Pyruvate
Succinate
2-Oxoglutarate
Fumarate
Glycolate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Malate
L-Lactate
D-Malate
D-Alanine
D-Glucarate
D-Galactarate
L-Alanine
L-Rhamnose
L-Fucose
0.25
0.2
0.15
0.1
0.05
0

C)

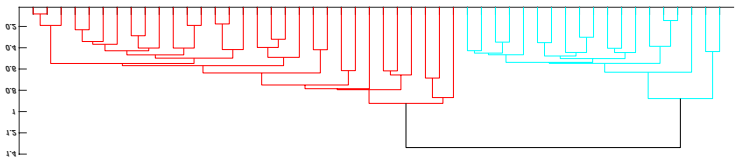
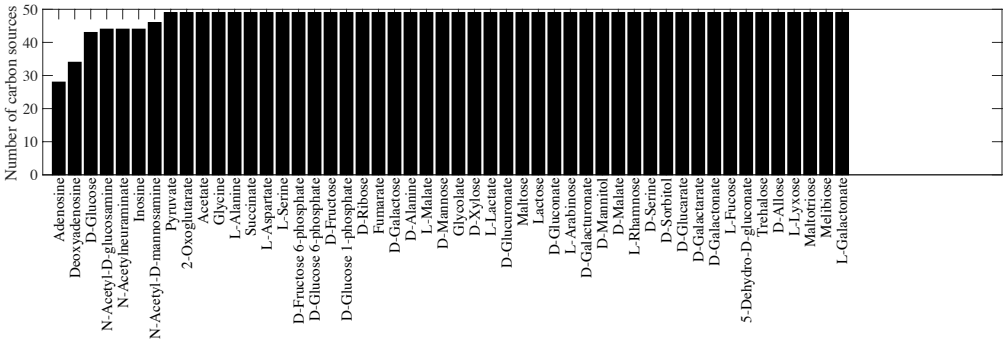
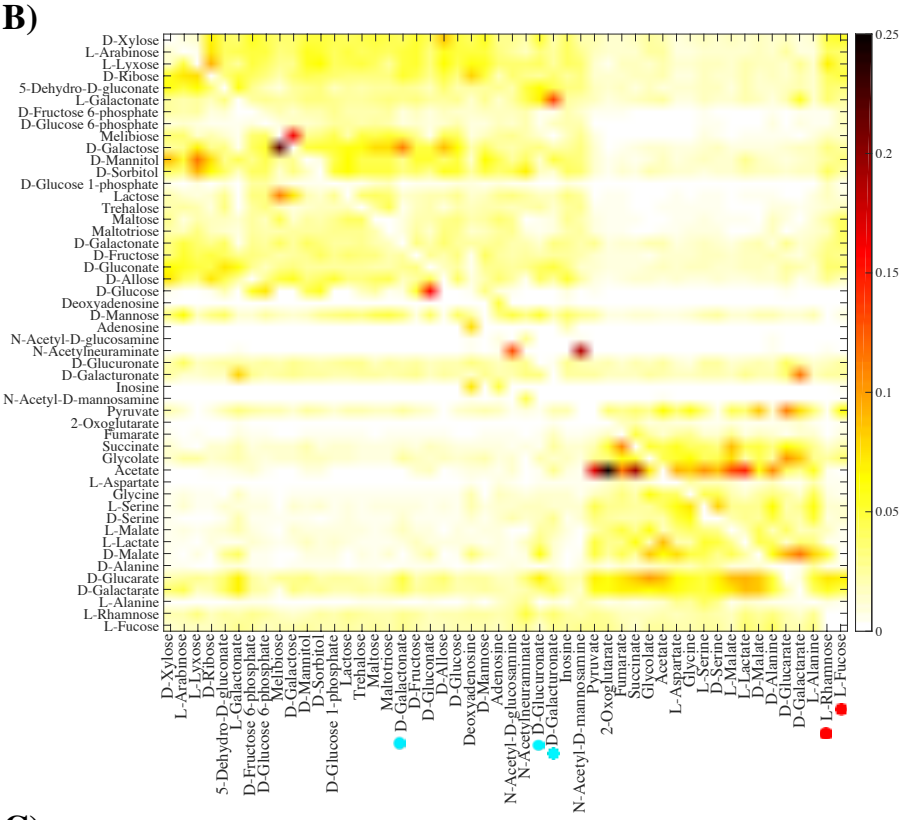


Figure S18: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes. A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis. Recombinants are generated between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, with the exception of the gluconeogenic carbon sources D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), and the glycolytic carbon sources L-rhamnose, and L-fucose (shown by red circles). In these analyses, parental metabolic networks contain $\|G\|$ = 2,079 reactions, the same number as the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks during recombination.

A)


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


0.25
0.2
0.15
0.1
0.05
0
D-Xylose
L-Arabinose
L-Lyxose
D-Ribose
5-Dehydro-D-gluconate
L-Galactonate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Melibiose
D-Galactose
D-Mannitol
D-Sorbitol
D-Glucose 1-phosphate
Lactose
Trehalose
Maltose
Maltotriose
D-Galactonate
D-Fructose
D-Gluconate
D-Allose
D-Glucose
Deoxyadenosine
D-Mannose
Adenosine
N-Acetyl-D-glucosamine
N-Acetylneuraminate
D-Glucuronate
D-Galacturonate
Inosine
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Fumarate
Succinate
Glycolate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Malate
L-Lactate
D-Malate
D-Alanine
D-Glucarate
D-Galactarate
L-Alanine
L-Rhamnose
L-Fucose

C)

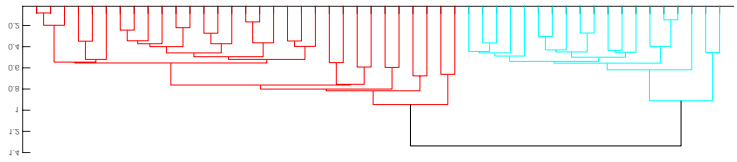
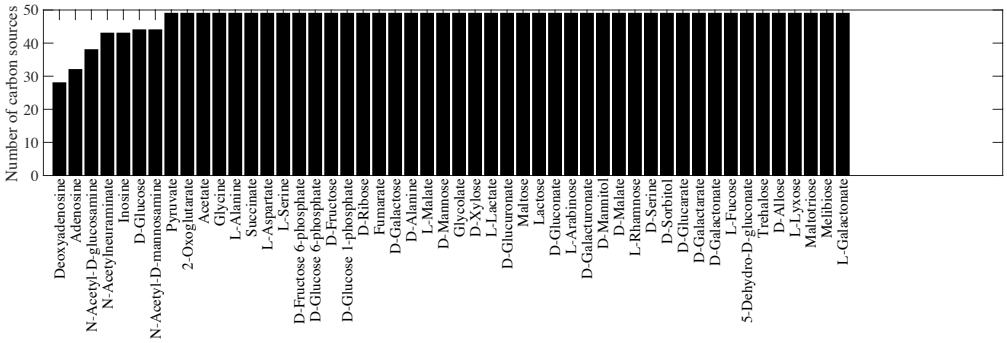
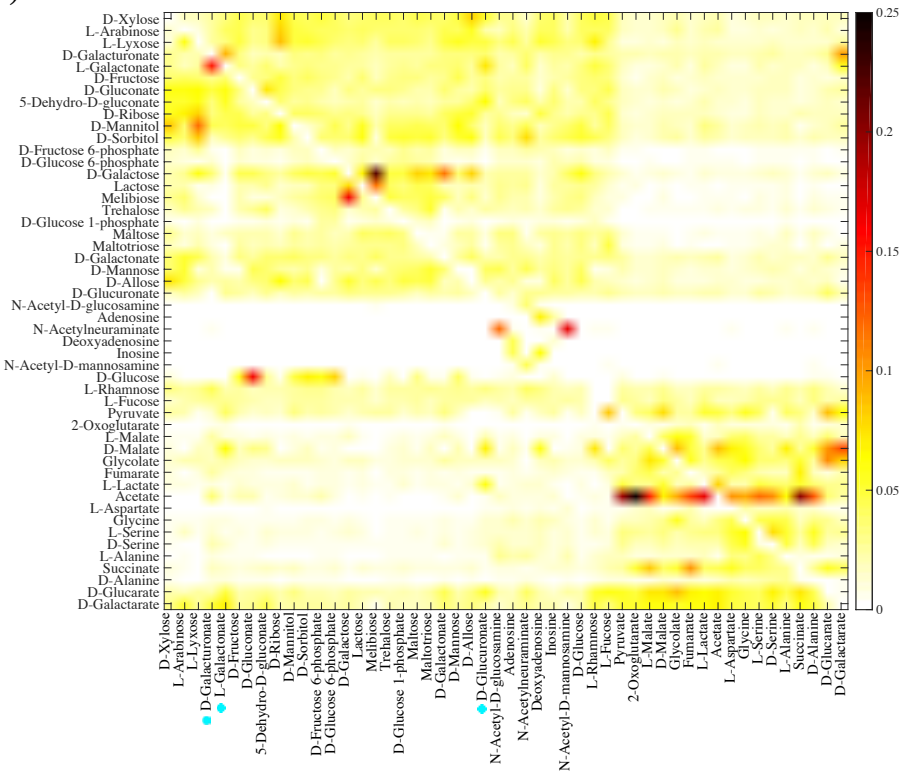


Figure S19: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (n = 20). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 20 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
N-Acetyl-D-glucosamine
N-AcetylNeuraminate
Inosine
D-Glucose
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Xylose
L-Arabinose
L-Lyxose
D-Galacturonate
L-Galactonate
D-Fructose
D-Gluconate
5-Dehydro-D-gluconate
D-Ribose
D-Mannitol
D-Sorbitol
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Galactose
Lactose
Melibiose
Trehalose
D-Glucose 1-phosphate
Maltose
Maltotriose
D-Galactonate
D-Mannose
D-Allose
D-Glucuronate
N-Acetyl-D-glucosamine
Adenosine
N-Acetylneuraminate
Deoxyadenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
L-Rhamnose
L-Fucose
Pyruvate
2-Oxoglutarate
L-Malate
D-Malate
Glycolate
Fumarate
L-Lactate
Acetate
L-Aspartate
Glycine
L-Serine
D-Serine
L-Alanine
Succinate
D-Alanine
D-Glucarate
D-Galactarate
0.25
0.2
0.15
0.1
0.05
0

C)

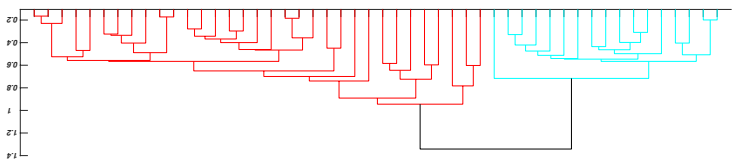
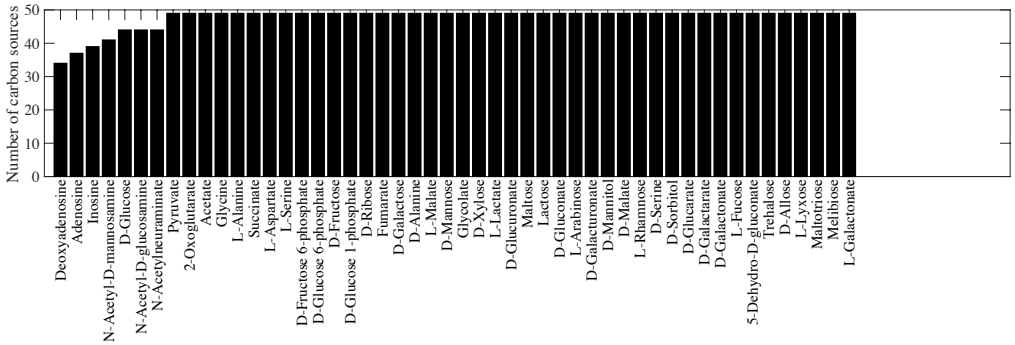
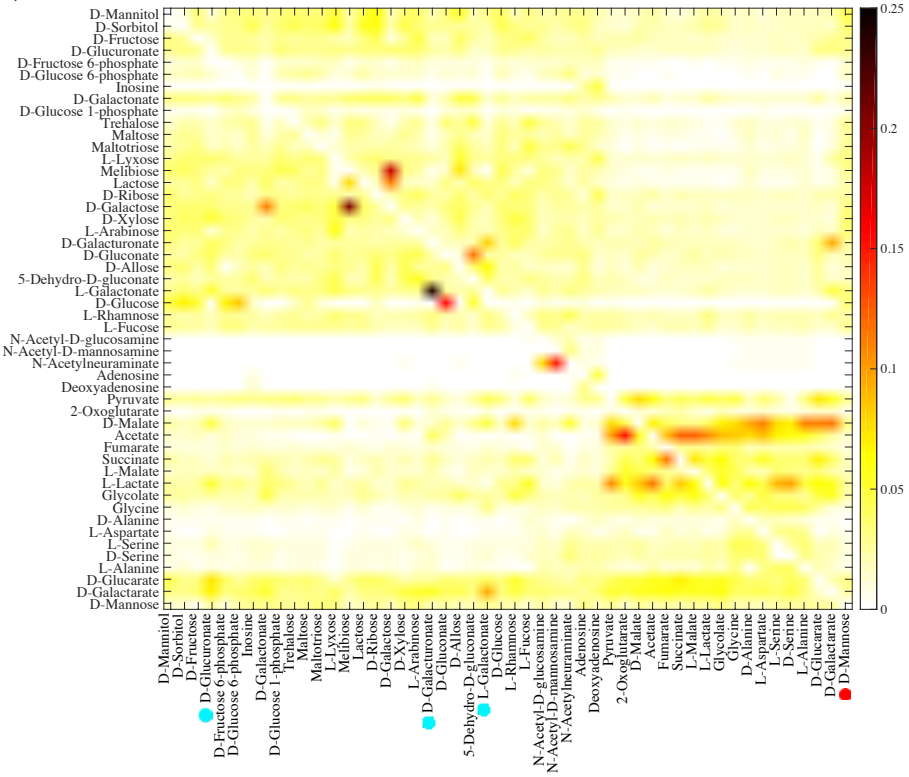


Figure S20: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (n = 30). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 100 reactions. Moreover, n = 30 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Deoxyadenosine
Adenosine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetyl-D-glucosamine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Mannitol
D-Sorbitol
D-Fructose
D-Glucuronate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
Inosine
D-Galactonate
D-Glucose 1-phosphate
Trehalose
Maltose
Maltotriose
L-Lyxose
Melibiose
Lactose
D-Ribose
D-Galactose
D-Xylose
L-Arabinose
D-Galacturonate
D-Gluconate
D-Allose
5-Dehydro-D-gluconate
L-Galactonate
D-Glucose
L-Rhamnose
L-Fucose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
N-Acetylneuraminate
Adenosine
Deoxyadenosine
Pyruvate
2-Oxoglutarate
D-Malate
Acetate
Fumarate
Succinate
L-Malate
L-Lactate
Glycolate
Glycine
D-Alanine
L-Aspartate
L-Serine
D-Serine
L-Alanine
D-Glucarate
D-Galactarate
D-Mannose
0.25
0.2
0.15
0.1
0.05
0

C)

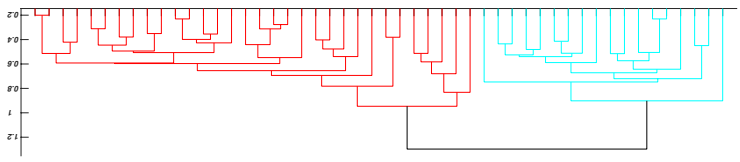
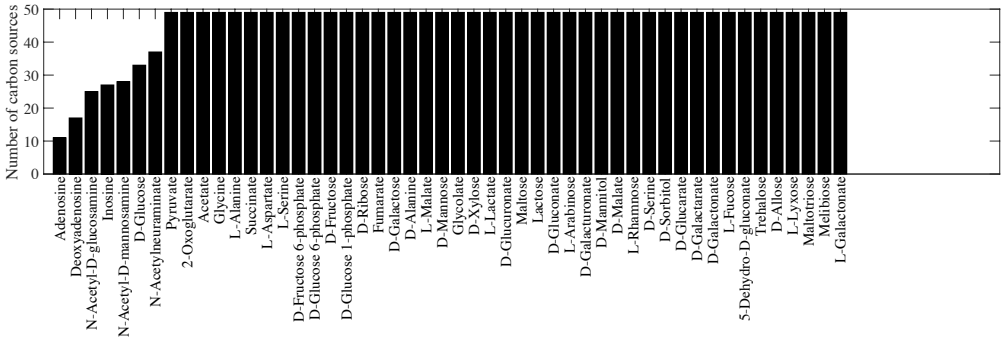
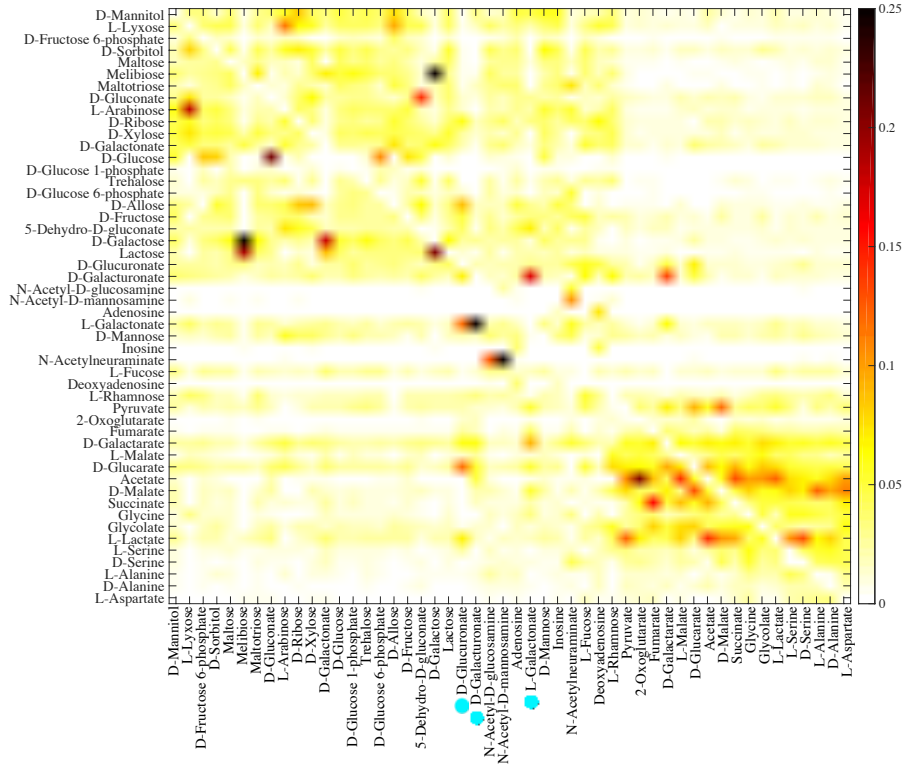


Figure S21: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (D = 1,000). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and D-mannose (shown by red circles), which is a glycolytic carbon source). In these analyses, parental metabolic networks contain $\|G\|$ = 2,079 reactions, the same number as in the *E.coli* metabolic network, and they differ in D = 1,000 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Adenosine
Deoxyadenosine
N-Acetyl-D-glucosamine
Inosine
N-Acetyl-D-mannosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Mannitol
L-Lyxose
D-Fructose 6-phosphate
D-Sorbitol
Maltose
Melibiose
Maltotriose
D-Gluconate
L-Arabinose
D-Ribose
D-Xylose
D-Galactonate
D-Glucose
D-Glucose 1-phosphate
Trehalose
D-Glucose 6-phosphate
D-Allose
D-Fructose
5-Dehydro-D-gluconate
D-Galactose
Lactose
D-Glucuronate
D-Galacturonate
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
L-Galactonate
D-Mannose
Inosine
N-Acetylneuraminate
L-Fucose
Deoxyadenosine
L-Rhamnose
Pyruvate
2-Oxoglutarate
Fumarate
D-Galactarate
L-Malate
D-Glucarate
Acetate
D-Malate
Succinate
Glycine
Glycolate
L-Lactate
L-Serine
D-Serine
L-Alanine
D-Alanine
L-Aspartate
0.25
0.2
0.15
0.1
0.05
0

C)

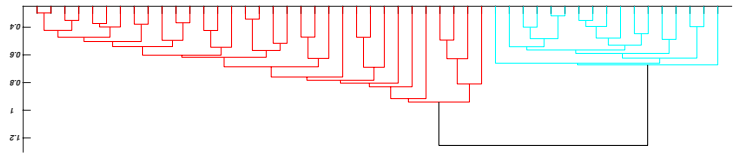
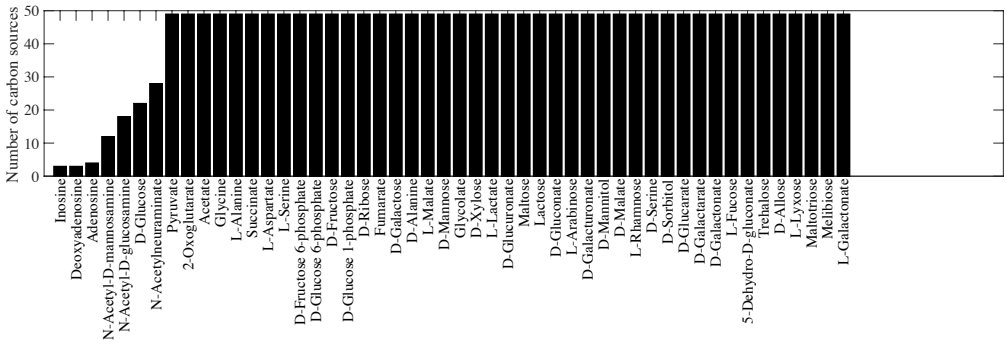
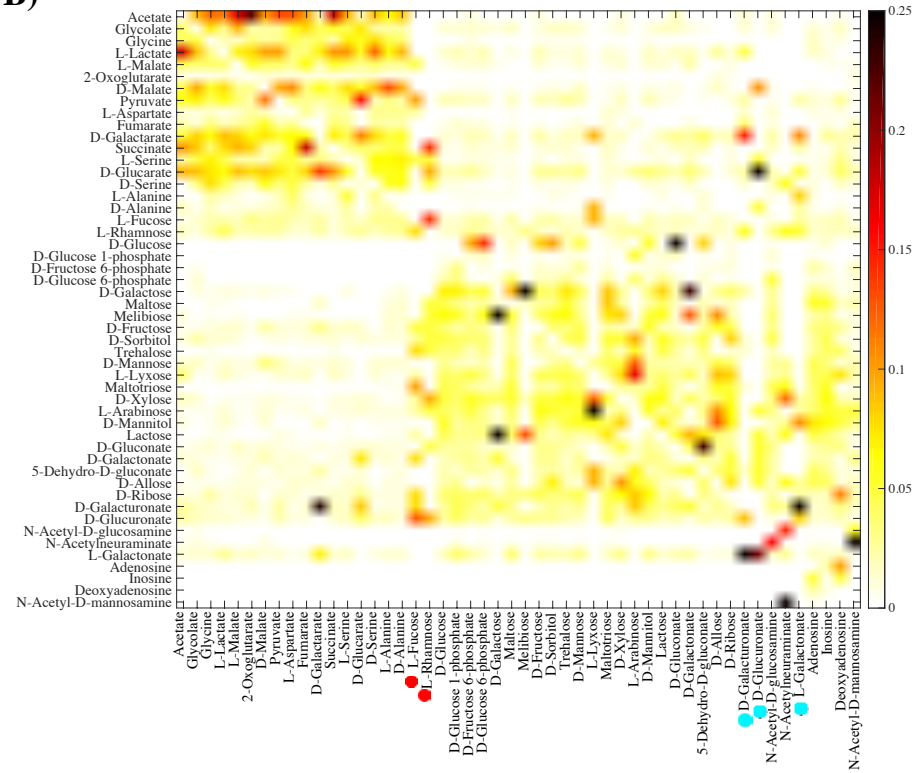

0.4
0.6
0.8
1
1.2

Figure S22: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (||G|| = 1,800). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources). In these analyses, parental metabolic networks contain ||G|| = 1,800 reactions, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
Inosine
Deoxyadenosine
Adenosine
N-Acetyl-D-mannosamine
N-Acetyl-D-glucosamine
D-Glucose
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


Acetate
Glycolate
Glycine
L-Lactate
L-Malate
2-Oxoglutarate
D-Malate
Pyruvate
L-Aspartate
Fumarate
D-Galactarate
Succinate
L-Serine
D-Glucarate
D-Serine
L-Alanine
D-Alanine
L-Fucose
L-Rhamnose
D-Glucose
D-Glucose 1-phosphate
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Galactose
Maltose
Melibiose
D-Fructose
D-Sorbitol
Trehalose
D-Mannose
L-Lyxose
Maltotriose
D-Xylose
L-Arabinose
D-Mannitol
Lactose
D-Gluconate
D-Galactonate
5-Dehydro-D-gluconate
D-Allose
D-Ribose
D-Galacturonate
D-Glucuronate
N-Acetyl-D-glucosamine
N-Acetylneuraminate
L-Galactonate
Adenosine
Inosine
Deoxyadenosine
N-Acetyl-D-mannosamine
0.25
0.2
0.15
0.1
0.05
0

C)

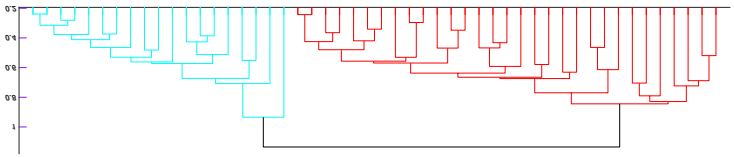
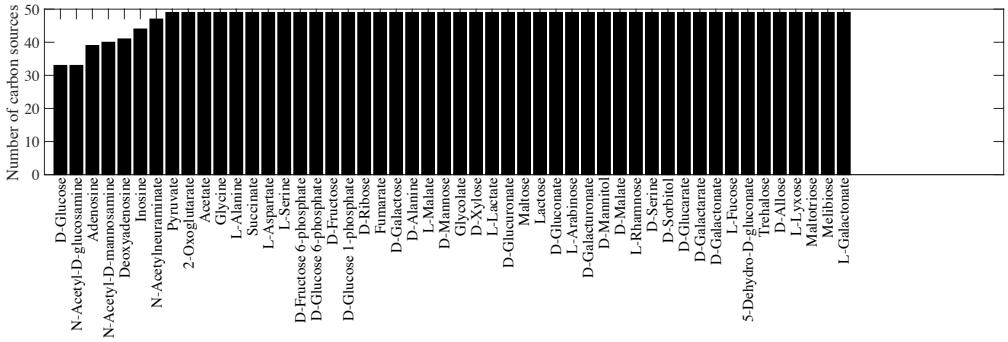
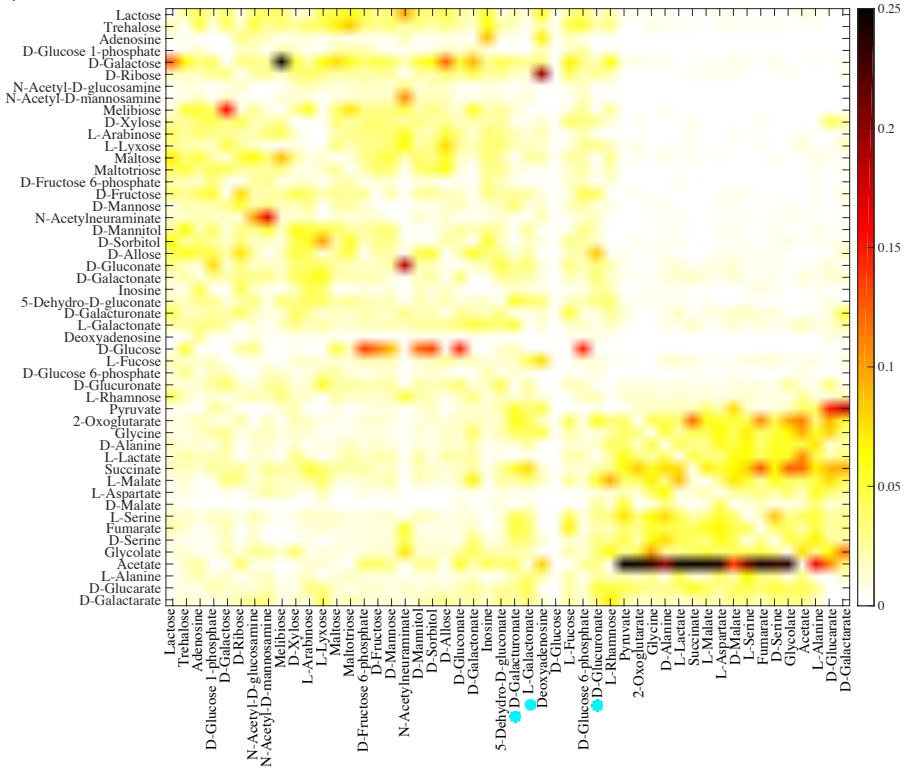


Figure S23: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (||G|| = 1,600). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between parents viable exclusively on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources, and L-rhamnose, and L-fucose (shown by red circles), which are glycolytic carbon sources). . In these analyses, parental metabolic networks contain ||G|| = 1,600 reactions, and they differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
D-Glucose
N-Acetyl-D-glucosamine
Adenosine
N-Acetyl-D-mannosamine
Deoxyadenosine
Inosine
N-Acetylneuraminate
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


Lactose
Trehalose
Adenosine
D-Glucose 1-phosphate
D-Galactose
D-Ribose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Melibiose
D-Xylose
L-Arabinose
L-Lyxose
Maltose
Maltotriose
D-Fructose 6-phosphate
D-Fructose
D-Mannose
N-Acetylneuraminate
D-Mannitol
D-Sorbitol
D-Allose
D-Gluconate
D-Galactonate
Inosine
5-Dehydro-D-gluconate
D-Galacturonate
L-Galactonate
Deoxyadenosine
D-Glucose
L-Fucose
D-Glucose 6-phosphate
D-Glucuronate
L-Rhamnose
Pyruvate
2-Oxoglutarate
Glycine
D-Alanine
L-Lactate
Succinate
L-Malate
L-Aspartate
D-Malate
L-Serine
Fumarate
D-Serine
Glycolate
Acetate
L-Alanine
D-Glucarate
D-Galactarate
0.25
0.2
0.15
0.1
0.05
0

C)

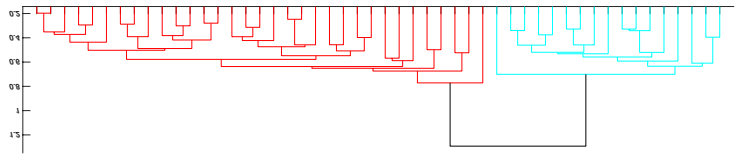
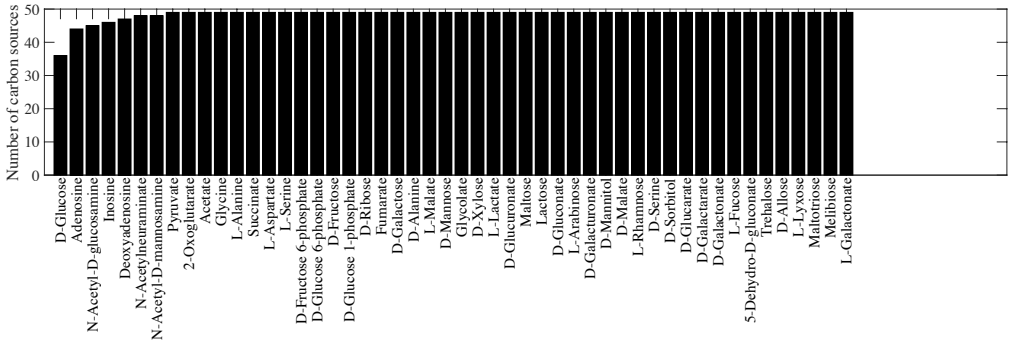
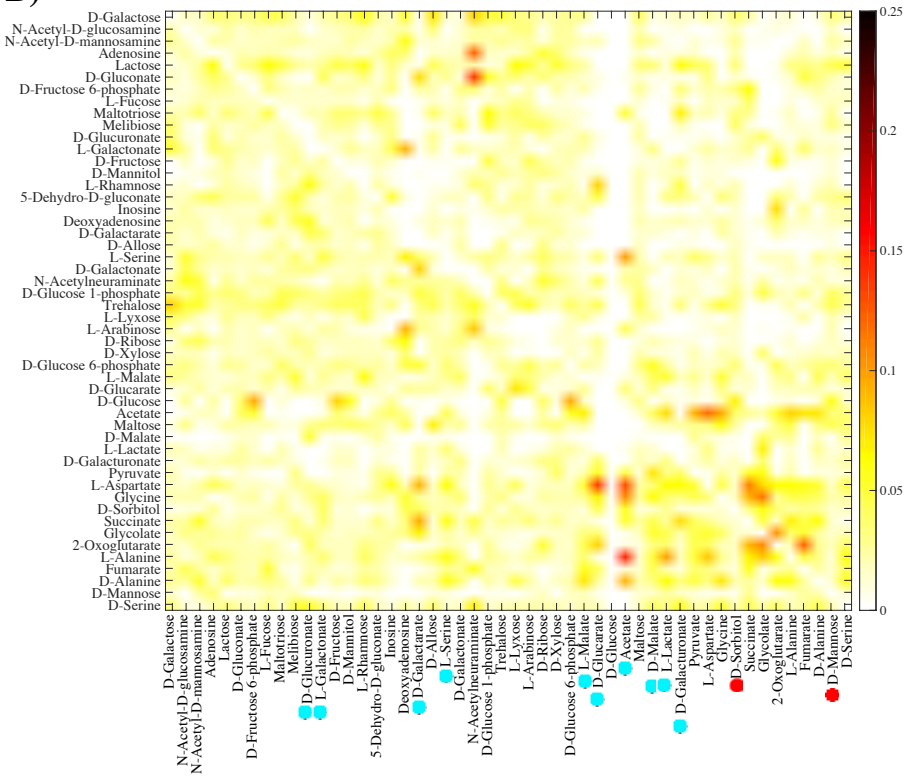


Figure S24: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (Parents with heterogeneous phenotypes, donors viable only on glucose). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between donor parents viable exclusively on glucose and the recipient parents that are exclusively viable on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (except D-galacturonate, L-galactonate, and D-glucoronate (shown by cyan circles), which are gluconeogenic carbon sources.). In these analyses, parental metabolic networks contain $||G||$ = 1,800 reactions, and differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

A)


Number of carbon sources
0
10
20
30
40
50
D-Glucose
Adenosine
N-Acetyl-D-glucosamine
Inosine
Deoxyadenosine
N-Acetylneuraminate
N-Acetyl-D-mannosamine
Pyruvate
2-Oxoglutarate
Acetate
Glycine
L-Alanine
Succinate
L-Aspartate
L-Serine
D-Fructose 6-phosphate
D-Glucose 6-phosphate
D-Fructose
D-Glucose 1-phosphate
D-Ribose
Fumarate
D-Galactose
D-Alanine
L-Malate
D-Mannose
Glycolate
D-Xylose
L-Lactate
D-Glucuronate
Maltose
Lactose
D-Gluconate
L-Arabinose
D-Galacturonate
D-Mannitol
D-Malate
L-Rhamnose
D-Serine
D-Sorbitol
D-Glucarate
D-Galactarate
D-Galactonate
L-Fucose
5-Dehydro-D-gluconate
Trehalose
D-Allose
L-Lyxose
Maltotriose
Melibiose
L-Galactonate

B)


D-Galactose
N-Acetyl-D-glucosamine
N-Acetyl-D-mannosamine
Adenosine
Lactose
D-Gluconate
D-Fructose 6-phosphate
L-Fucose
Maltotriose
Melibiose
D-Glucuronate
L-Galactonate
D-Fructose
D-Mannitol
L-Rhamnose
5-Dehydro-D-gluconate
Inosine
Deoxyadenosine
D-Galactarate
D-Allose
L-Serine
D-Galactonate
N-Acetylneuraminate
D-Glucose 1-phosphate
Trehalose
L-Lyxose
L-Arabinose
D-Ribose
D-Xylose
D-Glucose 6-phosphate
L-Malate
D-Glucarate
D-Glucose
Acetate
Maltose
D-Malate
L-Lactate
D-Galacturonate
Pyruvate
L-Aspartate
Glycine
D-Sorbitol
Succinate
Glycolate
2-Oxoglutarate
L-Alanine
Fumarate
D-Alanine
D-Mannose
D-Serine
0.25
0.2
0.15
0.1
0.05
0

C)

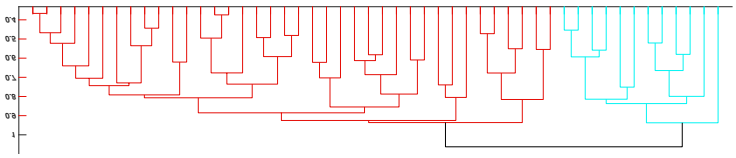


Figure S25: Emergence of innovative offspring is relatively but not absolutely contingent on parental phenotypes (Parents with heterogeneous phenotypes, recipients viable only on glucose). A) The horizontal axis shows the carbon use phenotype C_i of recombinant offspring. The vertical axis shows the number of parental carbon use phenotypes (among 49 possible such phenotypes), from which at least one innovative offspring gained viability on C_i. **B)** Fraction of innovative recombinants (color-coded according to the legend) gaining viability on the novel carbon source specified on the horizontal axis, which are generated from recombination between recipient parents viable exclusively on glucose and donor parents that are exclusively viable on the carbon source specified on the vertical axis. **C)** Dendrogram of carbon sources clustered based on their "innovation distance" defined by the data in panel B. We used UPGMA (unweighted pair group method with arithmetic means) for clustering carbon sources. Branches colored in red (cyan) correspond to glycolytic and gluconeogenic carbon sources, (with 12 exceptions; shown by 10 cyan circles, and 2 red circles.). In these analyses, parental metabolic networks contain $||G||$ = 1,800 reactions, and differ in D = 100 reactions. Moreover, n = 10 reactions are swapped between parental metabolic networks in a recombination event.

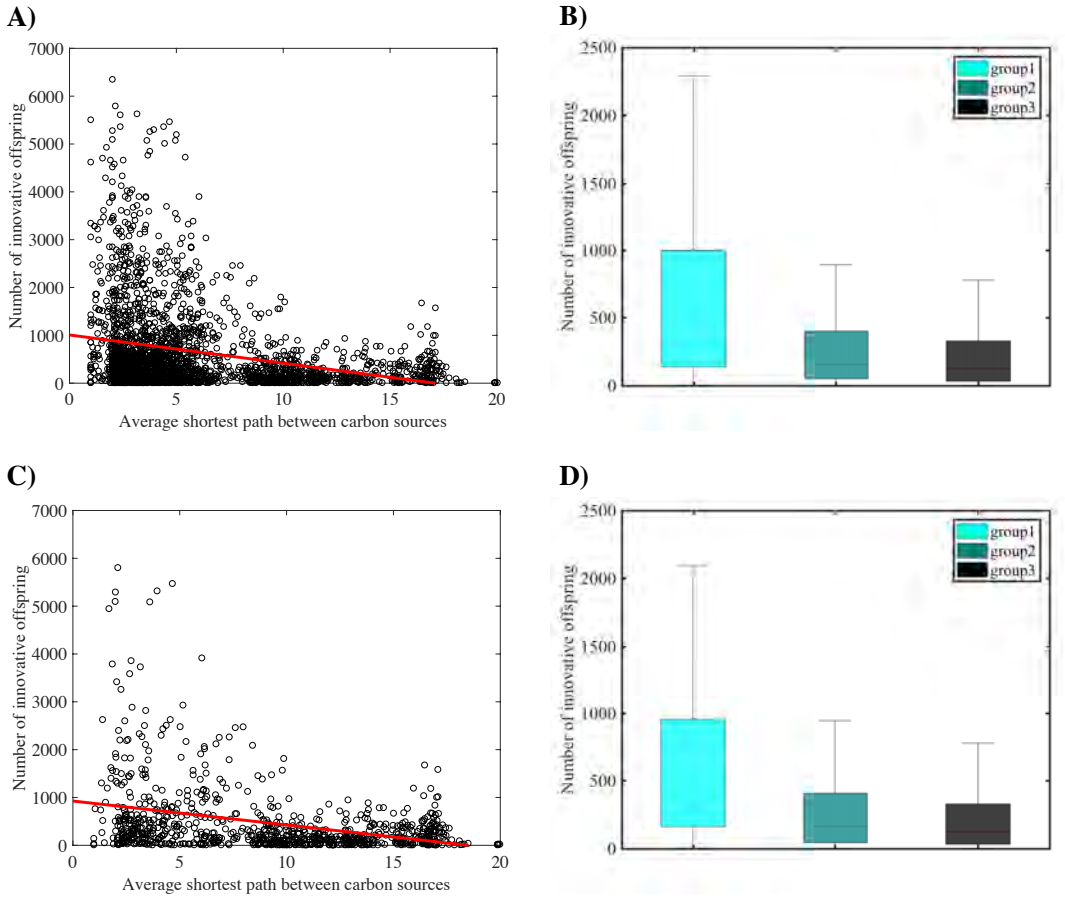


Figure S26: Distance between carbon sources in substrate graphs and relative constraint in the emergence of innovative offspring. In all 4 panels, the vertical axis shows the number of innovative recombinants (per 1 million recombinant offspring) gaining viability on some new carbon source C_j resulting from recombination between parental metabolic networks viable on carbon source C_i. In panels A and C, the horizontal axes show the mean shortest path between carbon source C_i and C_j in the substrate graph (supplementary text S7) of the metabolic networks viable on carbon source C_i. In panel **A)** each circle corresponds to a given pair of carbon sources (C_i, C_j), and data on both axes are significantly correlated (Pearson r=-*0.2722*, and $P<10^{-41}$). In panel **B)** the carbon source pairs (C_i, C_j) are divided into three groups based on their mean shortest path ($||SP\ (i,j)||$) between carbon source C_i and C_j in the substrate graph of metabolic networks viable on carbon source C_i: group 1 $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), group 2 $\{i,j|6 < ||SP\ (i,j)|| \leq 1\}$), and group 3 $\{i,j|\ ||SP\ (i,j)|| > 12\}$. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima.
In panel A, a non-uniform distribution of mean shortest paths ($||SP\ (i,j)||$) between carbon sources is evident on the horizontal axis. To exclude the possibility that the correlation in panel A is significant simply because of a higher number of data points for lower shortest path distances, we repeated the analyses shown in panels A and B by resampling from the 2500 pairs of carbon sources an equal number of pairs in each distance category, i.e., 284 pairs (C_i,

C_j) with $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), 284 pairs ($C_i$, C_j) with $\{i,j|6 < ||SP\ (i,j)|| \leq 12\}$), and 284 pairs ($C_i$, C_j) with $\{i,j|\ ||SP\ (i,j)|| > 12\}$, to create the subsampled data in panels C and D. In panel **C)** each circle corresponds to a given pair of carbon sources (C_i, C_j), and data on both axes are significantly correlated (Pearson r=*-0.3411*, and $P<10^{-24}$). In panel **D)**, analogous to panel B, carbon source pairs (C_i, C_j) are divided into three equally-sized groups based on their mean shortest path ($||SP\ (i,j)||$) between carbon source C_i and C_j in the substrate graph of metabolic networks viable on carbon source C_i: group 1 $\{i,j|1 \leq ||SP\ (i,j)|| \leq 6\}$), group 2 $\{i,j|6 < ||SP\ (i,j)|| \leq 1\}$), and group 3 $\{i,j|\ ||SP\ (i,j)|| > 12\}$. Boxes span the 25-th to 75-th percentile, and whiskers indicate maxima and minima. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E.coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

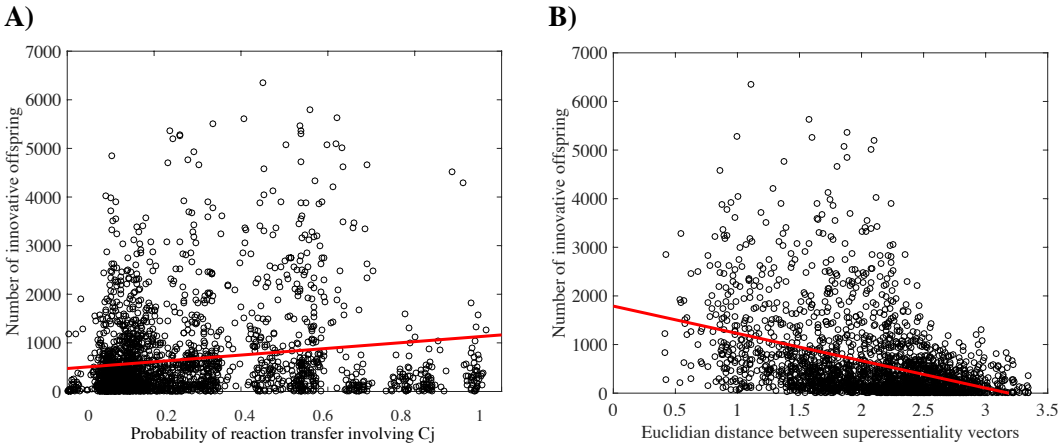


Figure S27: In both panels, each circle corresponds to a given pair of carbon sources (C_i, C_j) and the vertical axis shows the number of innovative recombinants (per 1 million recombinant offspring) gaining viability on some new carbon source C_j resulting from recombination between parental metabolic networks viable on carbon source C_i. The horizontal axes show **A)** the fraction of parental metabolic network pairs viable on carbon source C_i, in which a reaction that can enable viability on carbon source C_j can be transferred from the donor to the recipient metabolic network, and **B)** the Euclidian distance between superessentiality vectors of the corresponding pair of carbon sources, which we use as another proxy for the biochemical distance between carbon sources. In both panels the data plotted against one another are significantly correlated: **A)** Pearson r=*0.163*, and $P<10^{-15}$, and **B)** Pearson r=*-0.3935*, and $P<10^{-83}$. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E.coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

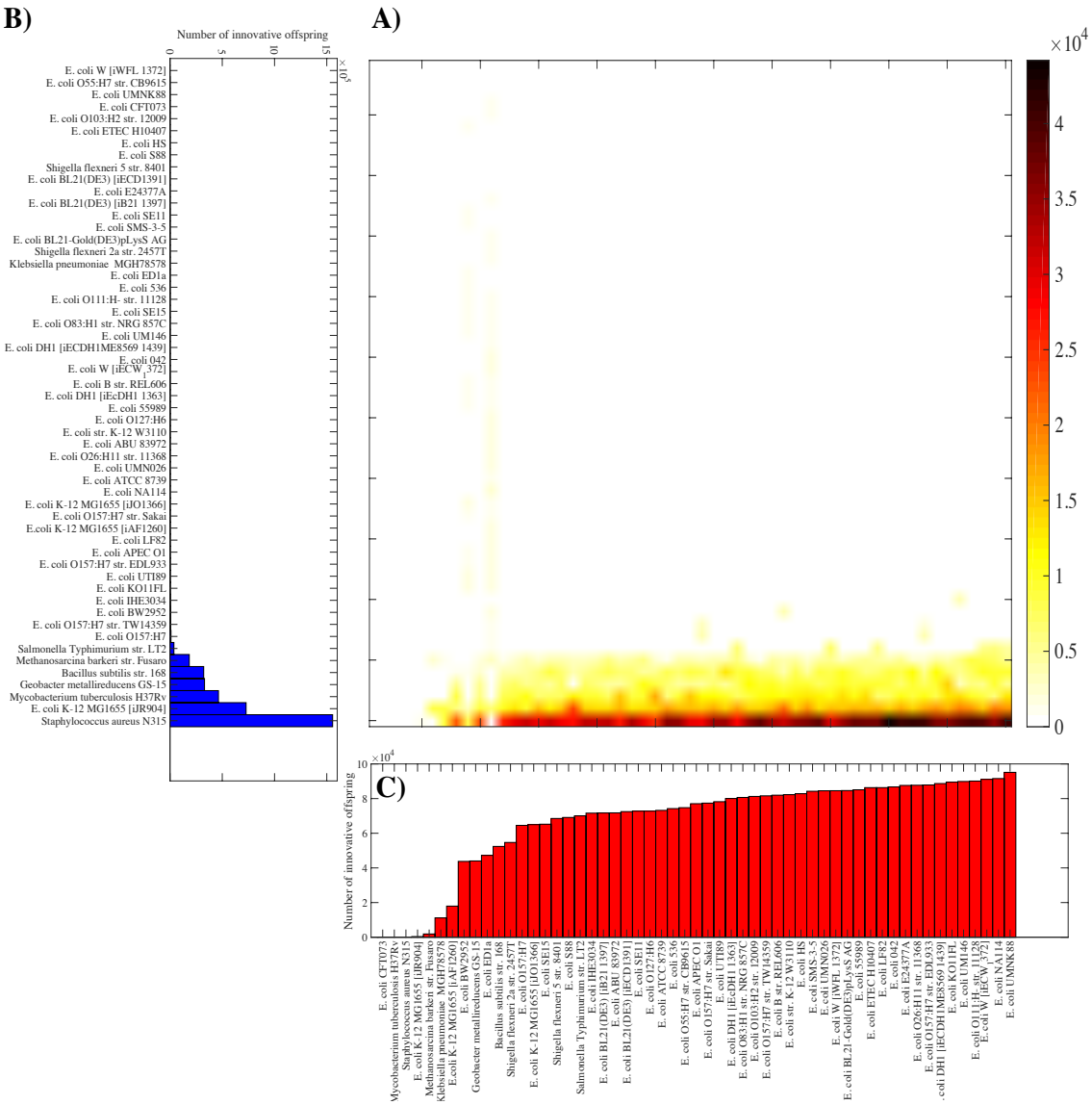


Figure S28: Emergence of innovative offspring is contingent on and constrained by parental genotypes. A) Number of innovative offspring resulting from linkage-based recombination between bacterial DNA donors specified on the vertical axis of panel B, and the corresponding recipient genotypes specified on the horizontal axis of panel C (coded according to the color legend). **B)** Total number of innovative recombinant offspring involving the donor genotype specified on the vertical axis. **C)** Total number of innovative recombinant offspring involving the recipient genotype specified on the horizontal axis.

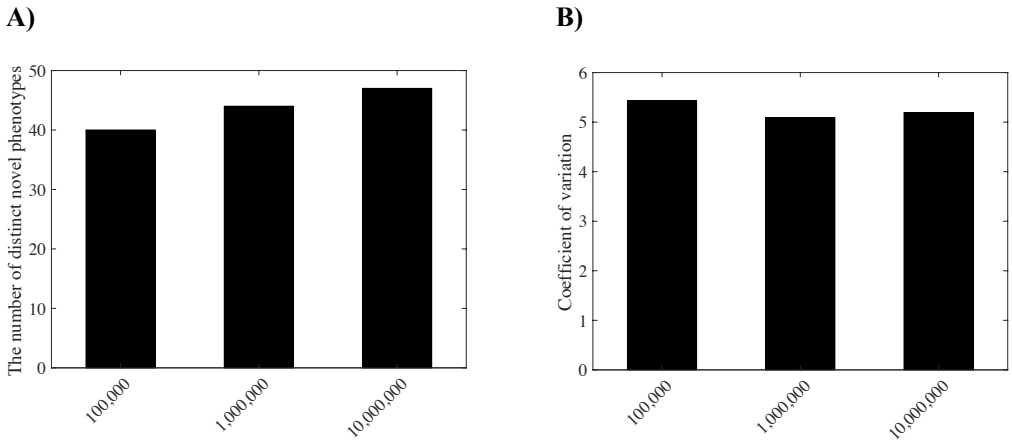


Figure S29: Sample size and its effect on absolute and relative constraints. For this analysis, we used 1,000 parental metabolic networks that are viable exclusively on glucose, and in three different simulations we generated *i)* 100, *ii)* 1,000 and *iii)* 10,000 offspring from each parent, which amounts to *i)* 100,000 *ii)* 1,000,000 and *iii)* 10,000,000 total offspring, as indicated on the horizontal axes. The vertical axes show **A)** the number of distinct novel phenotypes (among a possible total of 49 phenotypes) that emerged in the offspring, and **B)** the coefficient of variation in the number of innovative offspring for different novel carbon usage phenotypes. In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E.coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

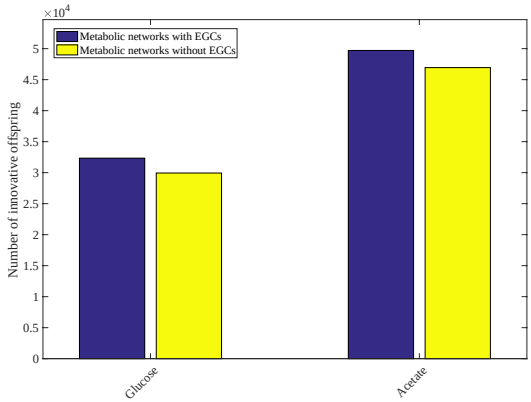
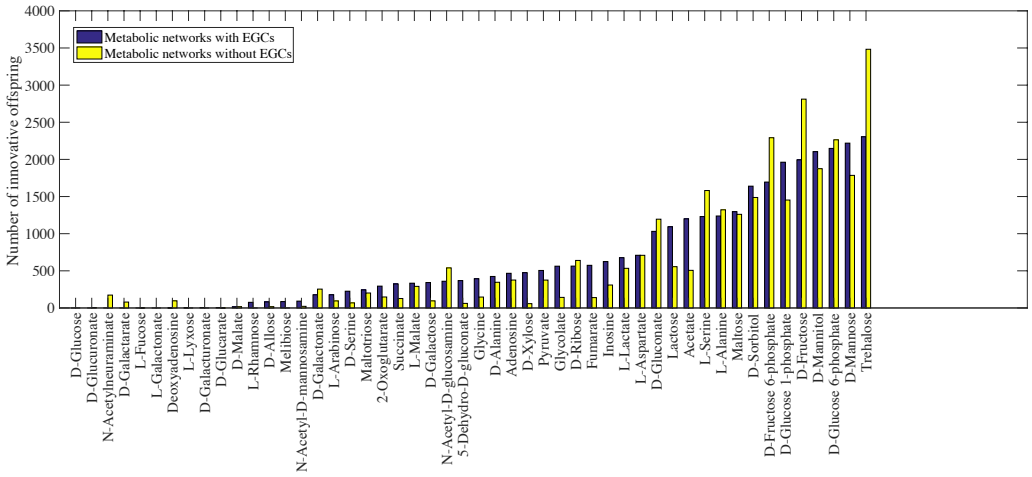


Figure S30: Erroneous energy generating cycles (EGCs) and the emergence of innovative offspring. The number of innovative offspring (per 1 million recombinants) emerging from recombination between parental metabolic networks that contain EGCs (blue) or that do not contain EGCs (yellow), and that are viable exclusively on glucose (left) and acetate (right). In these analyses, parental metabolic networks contain ||G||=2079 reactions, the same as the *E.coli* metabolic network, and they differ in *D=100* reactions. Moreover, *n=10* reactions are swapped between parental metabolic networks during recombination.

A)



B)

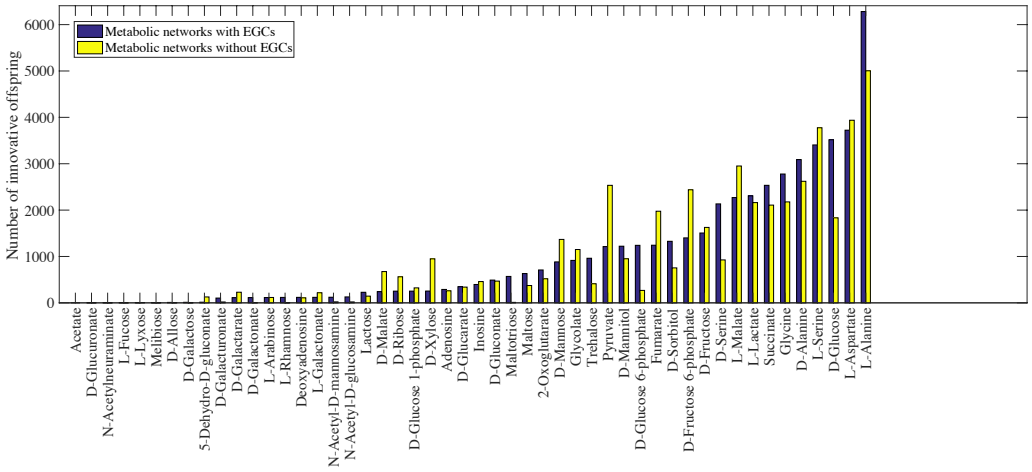


Figure S31: Erroneous energy generating cycles (EGCs) and relative constraints. Horizontal axes show the number of innovative offspring (per 1 million recombinants) emerging from recombination between parental metabolic networks viable exclusively on **A)** glucose and **B)** acetate, where parental metabolisms contain EGCs (blue) or do not contain EGCs (yellow). The ranking of the height of the blue bars and yellow bars in both panels is significantly correlated (panel A: Spearman's $\rho = 0.8913$, and $P < 10^{-18}$; panel B: Spearman's $\rho = 0.9197$, and $P < 10^{-21}$). In these analyses, parental metabolic networks contain $||G||=2079$ reactions, the same as the *E.coli* metabolic network, and they differ in $D=100$ reactions. Moreover, $n=10$ reactions are swapped between parental metabolic networks during recombination.

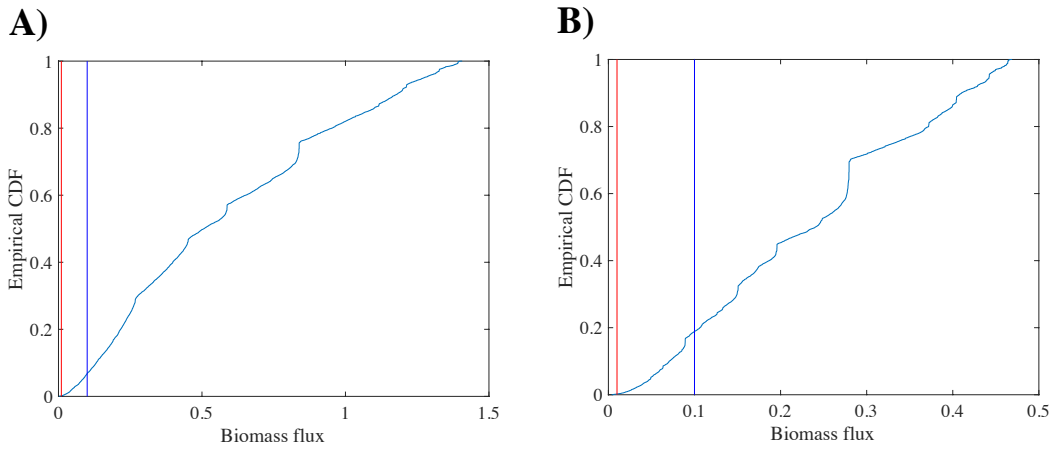


Figure S32: Biomass growth flux of most viable metabolic networks is much greater than our cut-off value for viability. The vertical axes show the empirical cumulative distribution function of the biomass flux among 10,000 MCMC-sampled metabolic networks viable exclusively on **A)** glucose, and **B)** acetate. The vertical red and blue lines show the cut-off value of 0.01 and 0.1 $1/h$. We used 0.001 $1/h$ as the cut-off value for viability.