Research article

The organization of metabolic genotype space facilitates adaptive evolution in nitrogen metabolism

Andreas Wagner[1,2,3], Vardan Andriasyan[4] and Aditya Barve[1,2]

[1]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, 8057, Switzerland
[2]The Swiss Institute of Bioinformatics, Basel, Switzerland
[3]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA
[4]Institute of Molecular Life Sciences, University of Zurich, Zurich, 8057, Switzerland

Received on November 6, 2013; Accepted on February 4, 2014; Published on February 25, 2014

Correspondence should be addressed to Andreas Wagner; Email: andreas.wagner@ieu.uzh.ch

Abstract

A metabolism is a complex chemical reaction system, whose metabolic genotype – the DNA encoding the enzymes catalyzing these reactions – can be compactly represented by its complement of metabolic reactions. Here, we analyze a space of such metabolic genotypes. Specifically, we study nitrogen metabolism and focus on nitrogen utilization phenotypes that are defined through the viability of a metabolism – its ability to synthesize all essential small biomass precursors – on a given combination of sole nitrogen sources. We randomly sample metabolisms with known phenotypes from metabolic genotype space with the aid of a method based on Markov Chain Monte Carlo sampling. We find that metabolisms viable on a given nitrogen source or a combination of nitrogen sources can differ in as much as 80 percent of their reactions, but can form networks of genotypes that are connected to one another through sequences of single reaction changes. The reactions that cannot vary in any one metabolism differ among metabolisms, and include a small core of "absolutely superessential" reactions that are required in all metabolisms we study. Only a small number of reaction changes are needed to reach the genotype network of one metabolic phenotype from the genotype network of another metabolic phenotype. Our observations indicate deep similarities between the genotype spaces of macromolecules, regulatory circuits, and metabolism that can facilitate the origin of novel phenotypes in evolution.

Introduction

Nitrogen is among the top five chemical elements occurring in living systems, comprising of the order of 10 percent of biomass in bacteria, for example (Fagerbakke *et al.* 1996, Heldal *et al.* 1985). Most of this nitrogen occurs in the form of amino acids, but some of it also as RNA and DNA nucleotides, as well as cofactors such as NAD and heme (Neidhardt 1996). The biomass of a free-living heterotrophic organism such as *E. coli* is built from approximately sixty small molecule biomass precursors, of which 48 contain nitrogen (Table S1).

Highly abundant but chemically inert atmospheric molecular nitrogen gas can only be converted into biomass by a select few organisms (Sadava *et al.* 2006). Many other nitrogen sources are less abundant and can limit an organism's rate of growth or reproduction. Organisms can circumvent such limitations by using more than one nitrogen source. For free-living heterotrophic organisms like the bacterium *Escherichia coli*, three nitrogen-containing molecules play an especially important role as nitrogen sources, because the biosynthesis pathways leading to nitrogen-containing biomass precursors require one or more of them. These are ammonia, glutamine, and glutamate. Among them, ammonium supports the fastest growth in *E. coli* and is thus considered a preferred carbon source. Glutamine and glutamate are not only potentially important nitrogen sources; they also serve as precursors for the biosynthesis of amino acids, and of purine and pyrimidine nucleotides (Merrick & Edwards 1995, Neidhardt 1996, Reitzer 2003).

Metabolic generalists like *E. coli* can use dozens of nitrogen sources, including many amino acids, but also compounds such as nitrate and urea (Neidhardt 1996). They can also vary considerably in their ability to use any one nitrogen source. For example, while the proteobacterium *Klebsiella aerogenes* can use histidine (Neidhardt 1996) as a sole nitrogen source, its relative *E. coli* cannot. Some strains of *E.coli* can use agmatine, an intermediate in the degra-

dation of arginine, as a sole nitrogen source, but *Salmonella typhimurium* cannot (Neidhardt 1996). Variation also exists in the biochemical pathways that metabolize or synthesize nitrogen-containing molecules. For example, arginine can be metabolized by multiple different pathways, two of which occur in enteric bacteria (Neidhardt 1996). The first uses arginine decarboxylase to degrade arginine to γ-aminobutyric acid in multiple steps, which then serves as a nitrogen source. In the second pathway arginine is succinylated and then metabolized further to produce succinate and glutamate. Similarly, L-alanine can be synthesized from pyruvate by glutamic-pyruvic transaminase using glutamate as an amino donor, or by transaminase C with valine as an amino donor (Neidhardt 1996). L-γ-glutamic semialdehyde, a precursor to the amino acid proline, can be synthesized from two different compounds, N-acetylglutamic γ-semialdehyde and L-γ-glutamyl phosphate.

Observations like these suggest that the pathways leading from any one nitrogen source to nitrogen -containing biomass precursors are flexible. We aimed to understand the extent of this flexibility, not just at the level of individual pathways, but on the level of the entire complex metabolic reaction network that is needed to synthesize all biomass precursors. More generally, we wanted to understand the basic organizational features of the space of possible metabolisms that can utilize different nitrogen sources. To this end, we used a recently developed approach to study large ensembles of metabolic networks that share the same biosynthetic abilities, but contain an otherwise random complement of biochemical reactions. We next introduce some necessary terminology and sketch the method behind this approach, which has been described in greater detail elsewhere (Rodrigues & Wagner 2009, Samal *et al.* 2010).

Methods

Metabolic genotypes, phenotypes and viability

A metabolism is a complex network of chemical reactions catalyzed by enzymes that are encoded by genes. The metabolic *genotype* of an organism is the part of a genome's DNA sequence that encodes these enzymes. This DNA-level representation of a metabolic genotype is too unwieldy to study qualitative and large-scale differences in the complement of enzyme-catalyzed reactions that specifies a metabolism. A more suitable and more compact representation is based on the observation that our current knowledge of metabolism comprises more than 5,000 enzyme-catalyzed reactions with known stoichiometry that occur in some organism (Goto *et al.* 1998, Kanehisa & Goto 2000). One can write these reactions as a long list, as indicated in Figure 1a. If the genome of an organism, such as that of a human, a fruit fly, or of *E. coli* encodes an enzyme that can catalyze a specific reaction, write "1" next to the reaction, otherwise write "0" (Figure 1a). The result is a representation of a metabolic genotype as a binary vector that completely specifies the reaction complement of a metabolism. This representation also makes clear that any one metabolic genotype is a member of a giant space of genotypes, a metabolic *genotype space* or a space of possible metabolisms. Since the universe of metabolic reactions comprises more than 5000 reactions, this space comprises more than 2^{5000} possible genotypes, many more than could be realized in the history of life on earth. Two metabolisms are *neighbors* in this space if they differ in a single reaction. A metabolism's *neighborhood* comprises all its neighbors. The *genotype distance* of two metabolisms can be defined through a variety of distance measures. We here use the fraction of reactions in which two metabolisms differ (in the representation of Figure 1a) as a distance measure. Specifically, if n_{12} denotes the reactions that two metabolic genotypes G_1 and G_2 have in common, and n_i denotes the number of reactions in genotype G_i, then this distance measure can be written as $1 - (n_{12} / (n_1 + n_2 - n_{12}))$.

The metabolic genotype of any one organism encodes its metabolic *phenotype*. There are many ways to define a metabolic phenotype, but the best-suited for the purpose of this paper is described hereafter. It starts from the observation that the most fundamental task of any one metabolism is to sustain life; that is, to synthesize all major biomass molecules that an organism needs for growth and reproduction, which include all amino acids, nucleotide precursors, lipids, and several co-factors (Feist *et al.* 2007, 2009, Feist & Palsson 2010). An organism whose metabolism is able to do that in a given chemical environment can survive in this environment - we refer to it as *viable*. Clearly, the potential nutrient molecules that occur in a given environment strongly influence whether a metabolism is viable. We will here consider minimal chemical environments that contain a single source of carbon (D-glucose), phosphorus (inorganic phosphate), sulfur (sulfate), oxygen, iron (Fe^{2+}, Fe^{3+}), as well as a single one among multiple possible nitrogen sources. One can write these potential nitrogen sources as a list, as shown in Figure 1b and associate a "1" with a nitrogen source if a metabolism is viable on it, that is, if it can synthesize all nitrogen-containing biomass precursors from it, and a "0" otherwise. In this way, a metabolic (nitrogen utilization) phenotype can be represented as a binary vector that indicates the spectrum of nitrogen

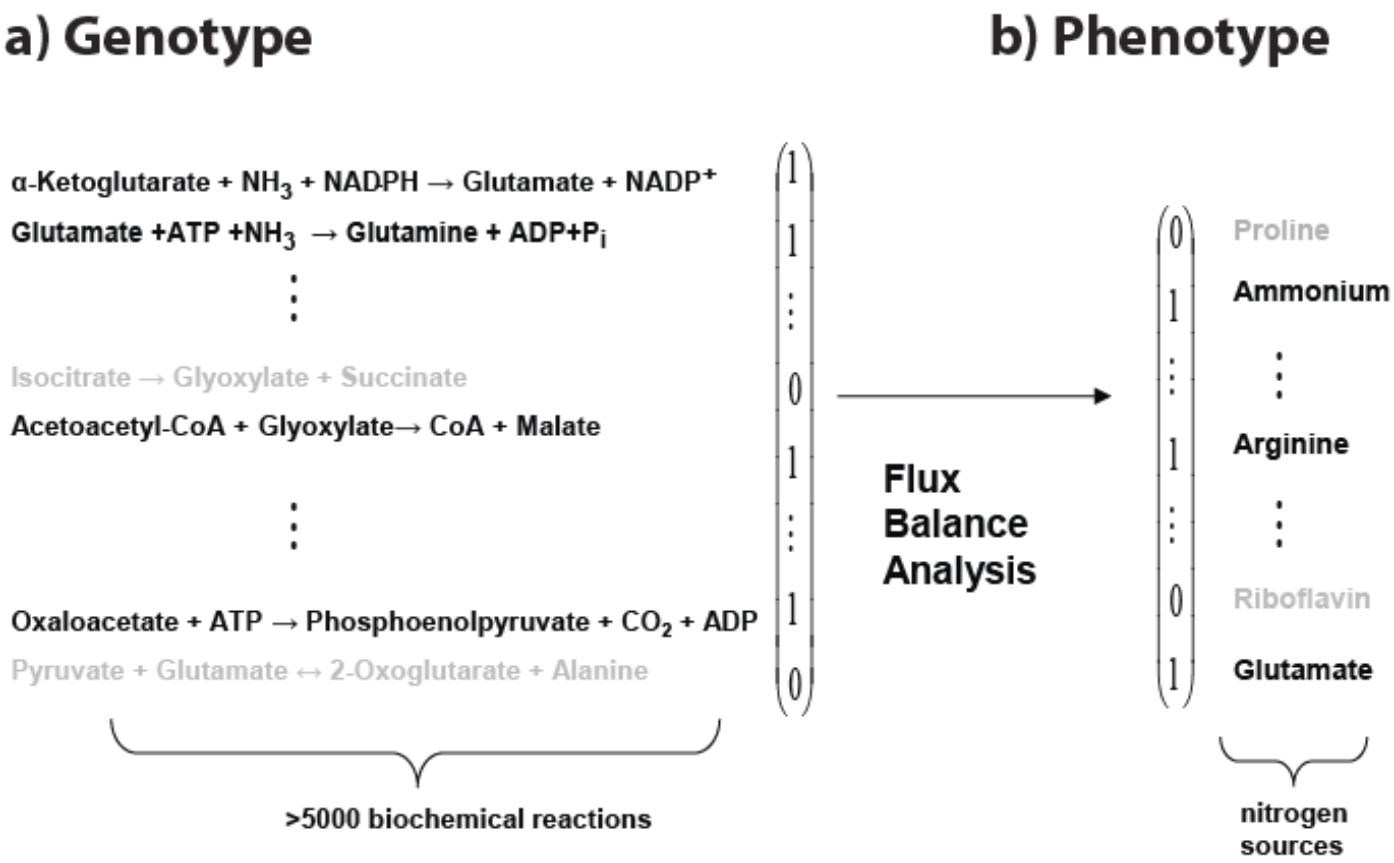

Figure 1. Metabolic genotypes and phenotypes. See text for details.

sources on which a metabolism is viable. In this paper, we consider 50 different nitrogen sources (Table S2).

To characterize those metabolic genotypes within the metabolic genotype space that are viable on a given number of nitrogen sources, we need to study many different metabolic genotypes and their phenotypes. It is possible to determine the metabolic phenotype of any one organism and its metabolic network experimentally on multiple different sources of chemical elements, such as through large scale metabolic phenotyping (Bochner 2009). However, it is currently not yet possible to experimentally manipulate metabolic genotypes on the large scale necessary to create many metabolic networks that are very different from each other. Fortunately, during the last 15 years computational approaches have been developed that allow us to predict metabolic phenotypes (Figure 1b) from qualitative information about metabolic genotypes, such as the stoichiometric equations shown in Figure 1a (Becker *et al.* 2007, Edwards & Palsson 2000, Feist & Palsson 2008, Heinrich & Schuster 1996, Schilling *et al.* 1999). Most notable among such approaches is the constraint-based modeling framework called flux balance analysis (Becker *et al.* 2007, Schilling *et al.* 1999). For a network that operates in a metabolic steady-state, such as would occur in a constant chemical environment under a steady nutrient supply, flux balance analysis predicts the maximal rate of biomass synthesis that a network can achieve in this chemical environment. Importantly, flux balance analysis requires only information about the stoichiometry of a metabolic reaction, and not about its kinetics or the concentrations of the enzyme catalyzing it. For metabolic networks with a well-studied genotype, the predictions of flux balance analysis are in good qualitative agreement with experimental data, for example on the viability of gene deletion mutations (Feist *et al.* 2007, Wang & Zhang 2009). The most important limitation of flux balance analysis is that it can incorporate regulatory constraints on biomass only with difficulty. Aside from the fact that many such constrains are easily broken in laboratory evolution experiments (Fong *et al.* 2005, Fong *et al.* 2003), such constraints are not of central importance for our purpose, because we are concerned mainly with the more fundamental constraints on viability caused by the complete absence of a reaction (enzyme-coding gene) from a metabolic genotype.

In our analysis, we constrained uptake rates of each nitrogen source to a maximum of 10 mmol/gdw/hr, and that of oxygen to a maximum of 20 mmol/gdw/hr. All other nutrients, including glucose as the sole

carbon source in the minimal environment were effectively unconstrained in their uptake rate ($<10^{20}$ mmol/gdw/hr). As we are studying the metabolism of nitrogen sources, choosing s low uptake rate for the nitrogen source makes it the rate-limiting nutrient. This is especially relevant because we define a network as viable if its biomass growth rate (flux) is no less than one percent of the starting *E. coli* network in the same minimal environment. Not having the nitrogen source, but another nutrient as rate-limiting could result in a false estimation of viability. Moreover, our definition of viability also takes into account that many microbes survive in the wild even though they grow slowly (Vieira-Silva *et al.* 2011).

Sampling of random viable metabolisms

In bioengineering, flux balance analysis is often applied to a single metabolism, to help understand the role that individual reactions play in the metabolism and to improve incomplete knowledge about its metabolic genotype (Figure 1a) (Becker *et al.* 2007, Feist *et al.* 2007, 2009, Herrgard *et al.* 2008, Jamshidi & Palsson 2007). In contrast, we will characterize many different metabolisms in metabolic genotype space, as well as their viability on different nitrogen sources. To this end, we employ a variant of Markov Chain Monte Carlo (MCMC) sampling in network space that we have already described previously (Rodrigues & Wagner 2009, 2011, Samal *et al.* 2010). This technique can produce uniform samples of metabolisms with a given phenotype. Briefly, it relies on random walks through genotype space that start with a metabolism of a given number of reactions and a given metabolic phenotype, for example viability on glutamine as a sole nitrogen source. Each step of this random walk consists of a so-called reaction swap, where one reaction chosen at random from the known reaction universe is added to a metabolism, whereas another randomly chosen reaction is deleted from the metabolism. This procedure ensures that each step leaves the number of reactions in the metabolism constant. In addition, each step is required to preserve viability on the chosen nitrogen source. If a step does not fulfill this requirement it is rejected, and another step is tried until one is found that preserves viability. In this way, one can perform long random walks through metabolic genotype space and sample networks at some steps during this random walk.

During a random walk using MCMC sampling, each metabolism in the random walk differs by only a reaction pair from the preceding metabolism. In other words, successive metabolisms in the random walk are "correlated" in their genotype and thus also in their phenotypic properties. As the number of steps between two metabolisms along the random walk increases, this autocorrelation decreases. Earlier work has shown that for metabolisms comprising about 1400 reactions, similar to the number of reactions in the *E. coli* metabolic network and the metabolisms studied here (Feist *et al.* 2007), 3×10^3 steps are sufficient to erase the correlation to the starting metabolism (Barve & Wagner 2013, Rodrigues & Wagner 2009, Samal *et al.* 2010). We thus sampled the first network after 5,000 steps, a number ensuring that the autocorrelation between the starting and the sampled metabolism was minimal. After this "burn-in" period, we sampled a metabolic genotype every 5,000 steps until we had obtained a sample of 1,000 genotypes that are viable on one or more given nitrogen sources, but contain an otherwise random complement of reactions (Rodrigues & Wagner 2009, Samal *et al.* 2010). In other words, our random walks proceeded for at least $5x10^6$ steps, unless otherwise mentioned. We refer to the metabolisms in the samples we thus generated as random viable metabolisms.

Variants of random sampling used in different analyses

Different analyses required us to use different variants of the sampling procedure. To estimate maximal genotype distances of metabolisms viable on a given, single nitrogen source, we started each random walk from the *E. coli* metabolic network (Feist *et al.* 2007), which comprises 1397 metabolic reactions, and required that none of the 5000 viability-preserving steps in the random walk reduce the distance to the starting network. In this way, we generated 1000 metabolisms required to be viable on a sole nitrogen source for each of the 50 nitrogen sources, that is, a total of $5x10^4$ (50 x 1000) metabolisms. We also used these samples to quantify the superessentiality of reactions for each nitrogen source (Figure 3).

To understand if the maximal genotype distances we observed for metabolisms viable on a single nitrogen source changed when metabolisms were required to be viable on multiple nitrogen sources, we needed metabolisms viable on a randomly chosen *n*-tuples (5, 10, 15, 20 and so on) of nitrogen sources. To create them, we first generated a random metabolism viable on all 50 nitrogen sources starting from the *E. coli* metabolism (which itself is viable on all 50 sources) after 5000 viability-preserving steps. We then randomly chose n nitrogen sources and continued the random walk for another 5000 steps while ensuring that the metabolism was viable on these n nitrogen sources. We then generated 100 random metabolisms from this starting metabolism through another 5000 steps of the random walk, with the constraint that none

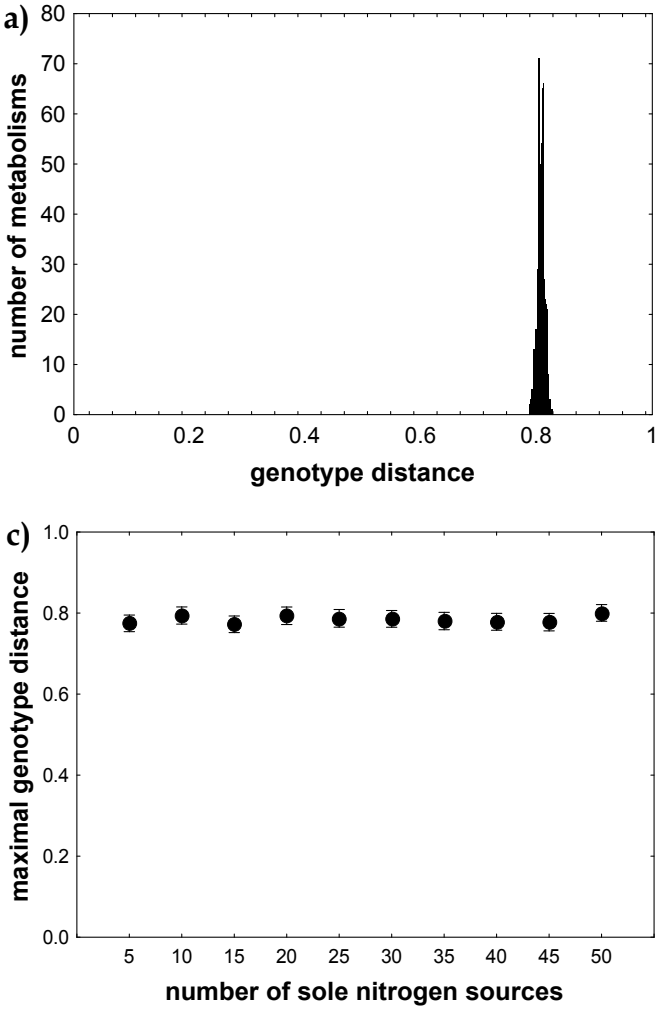

Figure 2. Metabolisms viable on the same sole nitrogen sources can differ in most of their reactions. a) Histogram of genotype distances between the *E. coli* metabolism and 1000 metabolisms that were the endpoints of viability-preserving random walks starting from the *E. coli* metabolism. The metabolisms in this analysis were required to be viable on glutamine as the sole nitrogen source. b) Distribution of the minima (left histogram, black) and maxima (right histogram, grey) of 50 genotype distance distributions obtained as in a), but for all 50 nitrogen sources considered here. Note that all minima and maxima lie within a narrow interval of genotype distance. c) The vertical axis shows means (circles) and three standard deviations (whiskers) of metabolic genotype distances between the start-points and end-points of 1000 viability-preserving random walks that started from metabolisms viable on the number *n* of sole nitrogen sources indicated on the horizontal axis. Each of these 1000 metabolisms were the starting points of a random walk where each step (i) had to preserve viability on the *n*-tuple of nitrogen sources, and (ii) was not allowed to decrease the distance to the starting metabolism (see Methods).

of the 5000 steps reduced the distance from the starting metabolism. We repeated this procedure 9 more times, with a different, randomly chosen n-tuple. In other words, we used this procedure to generate 1000 metabolisms viable on a given number *n* of nitrogen sources (100 metabolisms for each of 10 n-tuples).

As a starting point of our analysis of genotype network closeness, we required metabolisms that were viable on a specific nitrogen source, but not on other nitrogen sources. To generate such metabolisms, we returned to our sample of 1000 metabolisms viable on a sole nitrogen source. All of them were viable on one nitrogen source, but each may also be viable on other nitrogen sources (Barve & Wagner 2013). We chose an arbitrary metabolism among them, that was viable only on the focal nitrogen source (at least one of such metabolisms happened to exist in all of our samples). We used this metabolism as the starting point for random walks in which each reaction-swapping step was required to retain viability only on the focal nitrogen source. That is, if a step created viability on additional nitrogen sources, we discarded it. Through such random walks, we generated 10 random viable metabolisms that were strictly viable only on the focal nitrogen source. We repeated this approach for all 50 nitrogen sources, thus creating a total of 500 metabolisms, in groups of 10, where each group contained metabolisms viable on a specific nitrogen source. To estimate how close two genotype networks of metabolisms viable on two nitrogen sources (termed source 1 and 2) are in genotype space, we chose, with uniform probability, one metabolism G_1 among the ten metabolisms viable on source 1 and another metabolism G_2 among the ten metabolisms viable on source 2. We then used G_1 as a starting point for a phenotype-preserving random walk towards G_2, where each step was required to preserve viability only on the focal nitrogen source, and was not allowed to create viability on a new nitrogen source. After 5000 reaction swaps, we recorded the remaining distance between the random walker and G_2. We note that this distance is an upper bound for the point of closest proximity between genotype networks.

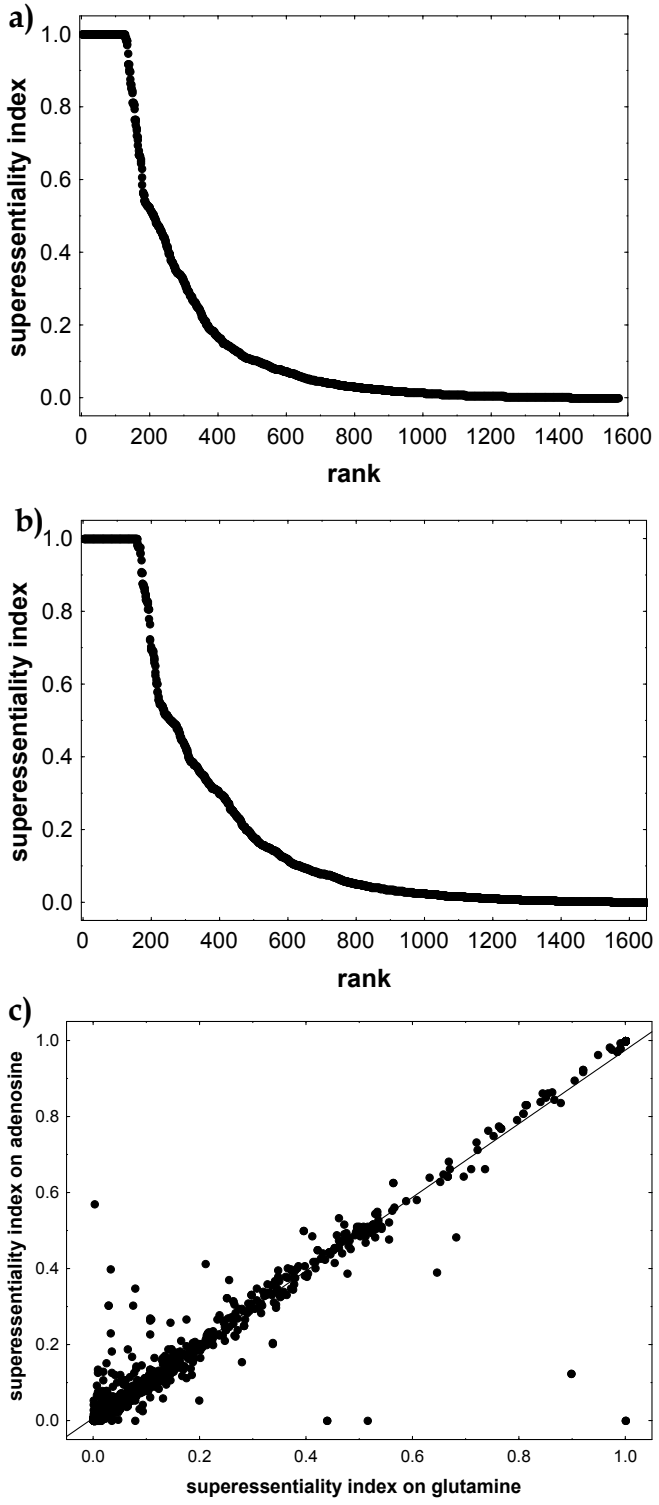

Figure 3. Reaction superessentiality in nitrogen metabolism. Rank plots of superessentiality indices (vertical axis) I_{SE} based on 1000 random metabolisms viable on a) glutamine, b) all 50 nitrogen sources considered here (when each is provided as the sole nitrogen source). c) Superessentiality indices of reactions where I_{SE} >0, for metabolisms viable on glutamine (horizontal axis) or adenosine (vertical axis) as sole nitrogen sources. Data are based on a sample of 1000 random viable metabolic metabolisms generated as described in methods.

We repeated this procedure 100 times, i.e., for 100 randomly chosen pairs of nitrogen sources.

Results and Discussion

Connected networks of viable nitrogen metabolisms extend far through genotype space

We first inquired how different two metabolisms (metabolic genotypes) can be while preserving their viability on a given spectrum of nitrogen sources. To answer this question, we performed the following analysis. Starting from the *E. coli* metabolic network, we performed a random walk of 5000 viability-preserving steps, where none of these steps was allowed to reduce the distance to the starting network. At the end of this walk, we recorded the genotype distance between the random walker and the starting network. We repeated this random walk 1000 times. Figure 2a shows a histogram of the genotype distance from *E. coli*, for 1000 networks viable on glutamine as the sole nitrogen source. The distribution of genotype distances is sharply peaked around a mean of 0.81, with a standard deviation of 0.006, a minimum of 0.79 and a maximum of 0.83. This means that two networks can differ in the vast majority of their reactions – approximately 80 percent – and still retain viability on glutamine as a sole nitrogen source. In addition, the networks that we used in this analysis can be connected in genotype space through long sequences of single reaction changes, none of which eliminates viability. In other words, they form part of a single connected network of genotype networks with the same viability phenotype, a genotype network (Rodrigues & Wagner 2009, Samal *et al.* 2010).

This observation is not a peculiarity of glutamine as a nitrogen source. To show this, we performed 1000 additional random walks for each of the 49 remaining nitrogen sources *N*, such that each random walk had to preserve viability on *N*. The results were 49 further genotype distance distributions like the one shown in Figure 2a. Figure 2b shows a histogram of the minima (black) and the maxima (grey) of all 50 distributions. It demonstrates that these distributions are confined within a narrow interval. Specifically, the smallest minimum genotype distance for all 50 nitrogen sources is 0.78 and the largest maximal genotype distance for all 50 nitrogen sources is 0.83.

Next, we asked whether these observations change fundamentally if we require networks to be viable on multiple different nitrogen sources, when each of these sources is provided as the sole nitrogen source. The answer is shown in Figure 2c, for 1000 end-points of viability-preserving random walks starting from networks viable on different numbers of *sole*

nitrogen sources, as shown on the horizontal axis. Clearly, the large genotype distances we observed for metabolisms required to be viable on only one nitrogen source change little when we consider multiple nitrogen sources. Taken together, these observations mean that metabolisms viable on one or more nitrogen source can differ greatly in the complement of metabolic reactions they harbor. Regardless of their specific nitrogen metabolism phenotype, they form connected networks of metabolic genotypes that range far through genotype space. In other words, they show great internal flexibility in their reaction complements.

Reactions vary widely in their superessentiality

The observation that 20 percent of metabolic reactions cannot change if viability on specific nitrogen sources is to be preserved raises the question of what these unchangeable reactions are, and whether they are the same for each of the 1000 metabolisms we studied during any one random walk. In other words, are some reactions more important than others in this sense? In previous works on carbon metabolism, we had shown that this is indeed the case, and that one can quantify this importance, as described hereafter (Barve *et al.* 2012). In any one metabolism, a reaction can be *essential* to synthesize biomass, that is, its removal will abolish the metabolism's viability. In a random sample of viable metabolisms, a reaction may be essential in some metabolisms but not in others. We call a reaction that is essential in more than one metabolism *superessential* – it is more than just essential. We introduced a *superessentiality index* I_{SE} that denotes the fraction of metabolic networks in which this reaction is essential. This index can range from zero (the reaction is never essential) to one (the reaction is essential in all metabolisms). A reaction with a superessentiality index of one is special, because it cannot be by-passed through an alternative sequence of reactions. We previously showed that assessing superessentiality based on random samples of at least 500 viable networks gives results that are in good agreement with a complementary approach that estimates superessentiality independently of random sampling (Barve *et al.* 2012). We thus proceeded to analyze the distribution of reaction superessentiality in randomly sampled metabolisms.

Figure 3a shows a rank plot of the superessentiality index of those reactions that were essential in at least one metabolism in a sample of 1000 random metabolisms viable on glutamine as a sole nitrogen source. The graph clearly shows that a small number of reactions are essential in all metabolisms – they are absolutely superessential and have a superessentiality index of one. Specifically, there are 126 such reactions, 102 of which involve nitrogen-containing molecules. The vast majority of reactions whose superessentiality index is shown are not always essential and rank from being essential in most metabolisms (left) to being essential only in few metabolisms (right). Figure 3b shows an analogous rank plot, but for metabolisms viable on all 50 nitrogen sources shown here. The overall shape of this plot is very similar, except that the number of absolutely superessential reactions is somewhat larger (157 reactions, 114 of which involve nitrogen-containing compounds). Table S3 shows a list of these reactions.

The absolutely superessential reactions include riboflavin synthase, the last step in the biosynthesis of riboflavin, a component of the cofactors flavin adenine dinucleotide (FAD) and flavin adenine mononucleotide (FMN). Another example is the reaction catalyzed by the enzyme phosphomethylpyrimidine kinase, encoded by the gene thiD (Blattner number b2103), which participates in the biosynthesis of thiamine diphosphate (also known as thiamine pyrophosphate TPP). TPP is an essential cofactor in enzymes such as pyruvate dehydrogenase (Nemeria *et al.* 2010). Yet another example concerns the enzyme histidinol phosphatase, encoded by the gene hisB (Blattner number b2022). Histidinol-phosphatase is essential for the biosynthesis of the amino acid histidine, while the same gene product also catalyzes a further reaction essential for histidine synthesis, that of imidazoleglycerol-phosphate dehydratase. The superessentiality indices of reactions on different nitrogen sources are statistically associated with one another. For example, Figure 3c shows a scatterplot of superessentiality indices of reactions where this index is greater than zero, for metabolisms viable on glutamine (horizontal axis) and adenosine (vertical axis) as sole nitrogen sources (Spearman's $r=0.94$, $P<<10^{-17}$, $n=1480$).

Though most reactions have very similar superessentiality indices for growth on glutamine and adenosine, it is instructive to discuss those outlier reactions whose superessentiality differs greatly on these nitrogen sources. One of them is the reaction catalyzed by the enzyme adenylosuccinate lyase (encoded by the gene purB). It is essential in all metabolisms viable on glutamine, but only in 0.2 percent of metabolisms viable on adenosine. The reaction converts adenylosuccinate to adenosine monophosphate. Whenever glutamine is the sole nitrogen source, this reaction is essential for the synthesis of the DNA precursor deoxyadenosine-triphosphate (dATP) (Baba *et al.* 2006). However, when adenosine is provided as a nutrient, this reaction can be by-passed, because dATP can be synthesized directly from adenosine. Indeed, *E. coli* strains lacking the gene purB are able to grow only when adenosine or adenine is supplied to a minimal

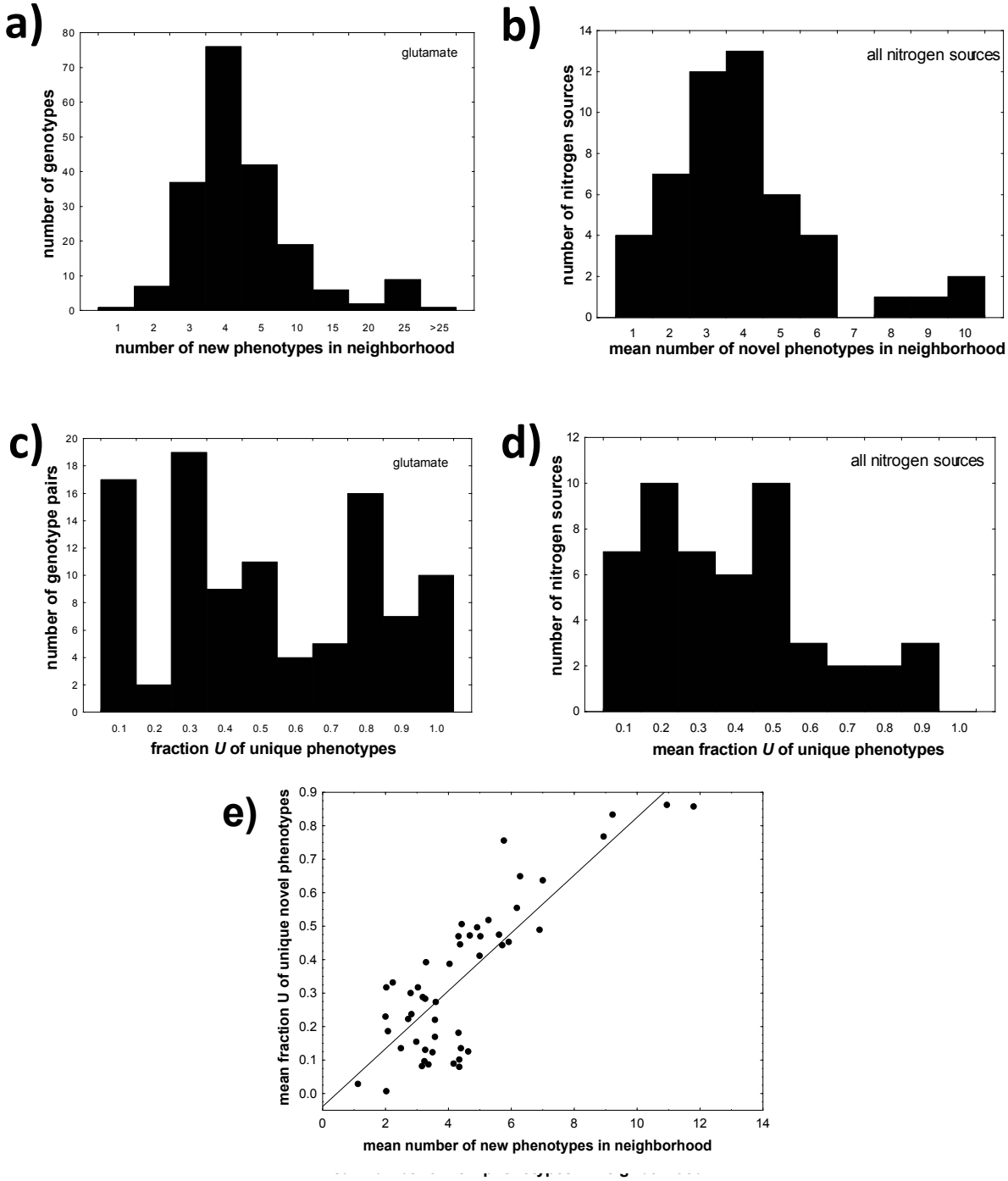

Figure 4. Genotypes contain novel phenotypes in their neighborhoods and some fractions of these novel phenotypes are unique. a) Distribution of the number of novel phenotypes in the neighborhood of 200 genotypes viable on glutamate. b) Distribution of the *mean* number of novel phenotypes in the neighborhood of genotypes for all 50 nitrogen sources considered here. Each data item is based on a sample of 200 genotypes (and each genotype's neighborhoods) for each of 50 nitrogen sources. Thus, the histogram is based on 50 samples of 200 genotypes each. c) For genotypes G_1 and G_2 sampled from the same genotype network, that is, they are viable on the same nitrogen source, and for P_i the set of all phenotypes that are found among genotypes in the neighborhood of G_i, the figure shows the distribution of $U=(|P_1| - |P_1 \cap P_2|)/|P_1|)$. This is the fraction of phenotypes unique to one neighborhood, i.e. without occurring in the other genotype's neighborhood. Specifically, the vertical axis shows the number of genotype pairs whose value of U is shown on the horizontal axis. The data is based on 100 random genotype pairs viable on glutamate. d) Histogram of the mean value of U (horizontal axis) for genotype pairs viable on each of the 50 nitrogen sources considered here. The vertical axis shows the number of nitrogen sources for which genotype pairs have the mean value of U shown on the horizontal axis. The data is based on 100 random genotype pairs (and their neighborhoods) for each of 50 nitrogen sources. Panels b) and d) are based on the mean as a summary statistic, because it is the simplest such statistic for distributions that are not extremely right- or left-skewed.

growth medium (Sun *et al.* 2011). Another example involves citrate synthase, which is essential in 0.3 and 56.9 percent of metabolisms on glutamine and adenosine, respectively. The enzyme citrate synthase is encoded by the gene gltA, which participates in the tricarboxylic acid cycle and produces citrate, which is in turn necessary for the synthesis of important biomass precursors such as 2-oxoglutarate and succinyl-CoA (generated from 2-oxoglutarate) (Noor *et al.* 2010). On glutamine as the sole nitrogen source, enzymes such as glutaminase can convert glutamine to glutamate (Brown *et al.* 2008), which can be further metabolized to 2-oxoglutarate via other enzymes such as aspartate-transaminase (encoded by the gene aspC) (Marcus & Halpern 1969). These biochemical pathways make the enzyme citrate synthase dispensable on glutamine as the sole nitrogen source, because they allow 2-oxoglutate to be directly synthesized from glutamine without the need for citrate synthesis. In contrast, growth on adenosine does not easily allow this bypass and thus renders citrate synthase essential in the majority of metabolisms (56.9 percent).

Different neighborhoods in metabolic genotype space do not contain the same novel phenotypes

In a population of evolving organisms, metabolism would evolve through alteration of an organism's metabolic genotypes. Especially in microbes, such evolution can occur rapidly by adding individual enzyme-coding genes to a genome through horizontal gene transfer, as well as by deleting individual genes. Even in populations that evolve under stabilizing selection for an existing, well-adapted phenotype, genotypic change can occur, because of the flexibility afforded by genotype networks. Such populations will explore the genotype network associated with a well-adapted phenotype, and its member genotypes will also explore local *neighborhoods* around them and around their genotype network. In general, the neighborhood of a genotype is important from an evolutionary perspective, because it comprises all those genotypes – with potentially novel phenotypes – that are easily reached via little genotypic change. Some metabolic genotypes in this neighborhood may have novel metabolic phenotypes, i.e., they may be able to survive on novel combinations of nitrogen sources. Genotype networks would be especially important for evolutionary adaptation, if different neighborhoods contained a different spectrum of novel phenotypes: Because genotype networks allow the exploration of different regions of genotype space, they also allow the exploration of different neighborhoods, and thus the exploration of novel phenotypes that would not be accessible otherwise. We thus wished to find out whether the neighborhood of different genotypes on a genotype network contained different phenotypes. To this end, we carried out the following quantitative analysis.

Consider two arbitrary genotypes G_1 and G_2 that are sampled from the same genotype network, that is, they are viable on the same nitrogen source. Denote as P_1 the set of all phenotypes that are found among genotypes in the neighborhood of G_1, that is, among all those metabolisms that differ in a single reaction from G_1. Define P_2 analogously as the set of all phenotypes found in the neighborhood of G_2. We wished to quantify the fraction of U of phenotypes that are contained in P_1 but not in P_2, i.e., the number of phenotypes that are in this sense unique to P_1. To this end we computed the quantity $U = (|P_1| - |P_1 \cap P_2|)/|P_1|$, where $|X|$ denotes the number of elements in a set. For example, if $|P_1| = |P_2| = 10$ and $|P_1 \cap P_2| = 5$ (5 phenotypes are common to both sets P_1 and P_2), then $U = (10 - 5)/10) = 0.5$; that is, 50 percent of phenotypes are unique to the neighborhood of G_1. More specifically, we computed this quantity for 100 random genotype pairs viable on the same nitrogen source. We obtained these genotype pairs by randomly choosing two different metabolisms from our sample of 1000 metabolisms viable on a given nitrogen source.

Figure 4a shows the distribution of the number of novel phenotypes that occur in the neighborhood of 200 genotypes (100 pairs) viable on glutamate, illustrating that most neighborhoods contain some phenotypes viable on novel combinations of nitrogen sources. Figure 4b shows a histogram summarizing the same data for all 50 nitrogen sources. Specifically, the figure shows the distribution of the *mean* number of novel phenotypes in the neighborhood of genotypes, where each data item is the mean value of U based on a sample of 200 genotypes (and each genotype's neighborhoods) for each of 50 nitrogen source In other words, the histogram is based on 50 samples of 200 genotypes and their neighborhood. The data illustrates that the number of novel phenotypes in a genotype's neighborhood varies broadly between one and ten, depending on the nitrogen source. Figure 4c shows, as an example, the distribution of the fraction U of unique phenotypes for 200 pairs of metabolic genotypes viable on glutamate, and Figure 4d shows the distribution of the mean value of U for genotype pairs viable on each of the 50 nitrogen sources considered here. The panel is again based on 200 genotype pairs for each of the 50 nitrogen sources, i.e., on a total of 50x200 genotype pairs. The panels show that some fraction of novel phenotypes are unique to individual neighborhoods, otherwise U would be equal to zero for all genotype pairs and nitrogen sources. They also demonstrate that U varies broadly among both geno-

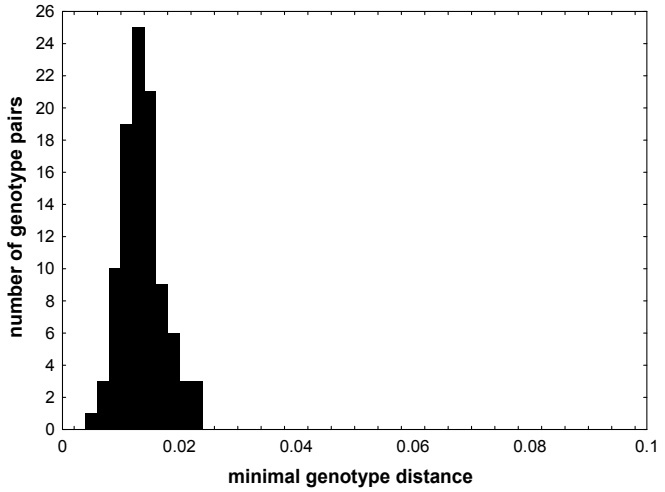

Figure 5. Metabolisms with different nitrogen utilization phenotypes can be very close together in genotype space. The figure shows the distribution of the minimal genotype distance for 100 pairs of metabolisms with different phenotypes, where each member of a pair was required to be viable on a different sole nitrogen source randomly and equiprobably chosen from the 50 nitrogen sources in Table S2 (See Methods for procedures).

types (Figure 4c) and nitrogen sources (Figure 4d). The mean U tends to be lowest for those nitrogen sources where genotypes have, on average, the smallest number of novel phenotypes in their neighborhood (Figure 4e, Spearman's r=0.71, $p=7.35 \times 10^{-9}$).

Some genotypes on two different genotype networks are close to each other in genotype space

Earlier analyses on metabolic and other systems showed that two genotypes with arbitrary different phenotypes can often be found close together in genotype space (Ciliberti *et al.* 2007, Rodrigues & Wagner 2009, Schuster *et al.* 1994). In the context of metabolism, this means that few reaction changes may be necessary for a transition from one phenotype to another, unrelated phenotype. We wished to explore whether this also holds true for our nitrogen utilization phenotypes. In this regard, we conducted an analysis that starts with two metabolic genotypes, G_1 and G_2, each of which is viable only on one nitrogen source, but where these nitrogen sources are different.

We then asked how similar we can make the reaction complement of G_1 to that of G_2 without altering its phenotype. To this end, we carried out reaction-swapping and phenotype-preserving random walks that started from G_1 and approached G_2, i.e., each step in such a random walk was not allowed to increase the distance to the target G_2. After 5000 steps we recorded the distance remaining between G_1 and G_2. We emphasize that our estimates of minimal distances are upper bounds, since our procedure does not exclude the possibility that metabolisms with different phenotypes differ in even fewer reactions. Figure 5 shows the results of this approach for 100 different pairs of metabolisms viable on different nitrogen sources. The figure shows that the minimum genotype distance of networks with different phenotypes is very small, and comprises less than 2% of the total diameter (maximal distance) of genotype space. In other words, metabolism pairs that are viable on different nitrogen sources can share 98% or more of their chemical reactions. Only very few reaction changes are minimally needed to produce one nitrogen utilization phenotype from another such phenotype.

Conclusion

To summarise, our analysis has shown that metabolic genotypes can differ in most of the biochemical reactions they encode, yet share the same nitrogen utilization phenotype. In addition, our Markov chain Monte Carlo approach shows that even very different genotypes with the same phenotype can be transformed into one another through a series of single reaction changes. In other words, such genotypes form large connected networks – genotype networks – that extend far through metabolic genotype space. A second qualitative feature we observed is that different neighborhoods of genotypes on the same genotype network usually do not contain the same novel phenotypes. Together, these properties can facilitate the exploration of novel phenotypes by a population whose metabolism evolves through the addition and deletion of enzyme-coding genes in a genome. Specifically, the individuals in such a population can preserve existing, well adapted phenotypes, while at the same time altering their genotypes in a step-by-step manner, thus exploring different regions and neighborhoods of a genotype network. Because different neighborhoods contain different novel phenotypes, the existence of genotype networks can help in the exploration of novel phenotypes. Any evolutionary search for novel and adaptive phenotypes may be further facilitated by the observation that different genotype networks tend to be highly interwoven and close together in genotype space (Figure 5). These observations are in qualitative agreement with earlier ones on carbon and sulfur metabolism (Rodrigues & Wagner 2009, 2011, Samal *et al.* 2010), and genotype spaces with these features also exist in proteins, RNA, as well as in regulatory circuits, where they help facilitate evolutionary adaptation (Wagner 2011). Although some 80% of chemical reactions in a genotype network may change without affecting nitrogen utilization phenotype, not all reactions are equally changeable. In particular, there is a

small core of super essential reactions that cannot be altered without abolishing viability on any one nitrogen source, at least based on current biochemical knowledge. Reactions like these are potential targets for antimetabolic drugs whose action cannot be easily circumvented through the evolution of alternative metabolic pathways in pathogens targeted by these drugs (Barve *et al.* 2012).

Acknowledgements

We acknowledge support through Swiss National Science Foundation grants 315230-129708, as well as through the YeastX project of SystemsX.ch, and the University Priority Research Program in Systems Biology at the University of Zurich.

Conflicts of Interest

The authors declare no conflicts of interest.

References

Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL & Mori H 2006 Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2** 2006.0008

Barve A, Rodrigues JFM & Wagner A 2012 Superessential reactions in metabolic networks. *Proc Natl Acad Sci U S A* **118** E1121-E1130

Barve A & Wagner A 2013 A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature* **500** 203-206

Becker SA, Feist AM, Mo ML, Hannum G, Palsson BO & Herrgard MJ 2007 Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox. *Nature Prot* **2** 727-738

Bochner BR 2009 Global phenotypic characterization of bacteria. *FEMS Microbiol Rev* **33** 191-205

Brown G, Singer A, Proudfoot M, Skarina T, Kim Y, Chang C, Dementieva I, Kuznetsova E, Gonzalez CF, Joachimiak A, Savchenko A & Yakunin AF 2008 Functional and structural characterization of four glutaminases from Escherichia coli and Bacillus subtilis. *Biochemistry* **47** 5724-5735

Ciliberti S, Martin OC & Wagner A 2007 Innovation and robustness in complex regulatory gene networks. *Proc Natl Acad Sci U S A* **104** 13591-13596

Edwards JS & Palsson BO 2000 The Escherichia coli MG1655 in silico metabolic genotype: Its definition, characteristics, and capabilities. Proc Natl Acad Sci U S A **97** 5528-5533

Fagerbakke KM, Heldal M & Norland S 1996 Content of carbon, nitrogen, oxygen, sulfur and phosphorus in native aquatic and cultured bacteria. *Aquat Microb Ecol* **10** 15-27

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V & Palsson BØ 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

Feist AM, Herrgard MJ, Thiele I, Reed JL & Palsson BO 2009 Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol* **7** 129-143

Feist AM & Palsson BO 2008 The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. *Nat Biotechnol* **26** 659-667

Feist AM & Palsson BO 2010 The biomass objective function. *Curr Opin Microbiol* **13** 344-349

Fong SS, Joyce AR & Palsson BO 2005 Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res* **15** 1365-1372

Fong SS, Marciniak JY & Palsson BO 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

Goto S, Nishioka T & Kanehisa M 1998 LIGAND: chemical database for enzyme reactions. *Bioinformatics* **14** 591-599

Heinrich R & Schuster S 1996 *The regulation of cellular systems*. New York: Chapman and Hall

Heldal M, Norland S & Tumyr O 1985 X-ray microanalytic method for measurement of dry-matter and elemental content of individual bacteria. *Appl Environment Microbiol* **50** 1251-1257

Herrgård MJ, Swainston N, Dobson P, Dunn WB, Arga KY, Arvas M, Blüthgen N, Borger S, Costenoble R, Heinemann M, Hucka M, Le Novère N, Li P, Liebermeister W, Mo ML, Oliveira AP, Petranovic D, Pettifer S, Simeonidis E, Smallbone K, Spasić I, Weichart D, Brent R, Broomhead DS, Westerhoff HV, Kirdar B, Penttilä M, Klipp E, Palsson BØ, Sauer U, Oliver SG, Mendes P, Nielsen J & Kell DB 2008 A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* **26** 1155-1160

Jamshidi N & Palsson BO 2007 Investigating the metabolic capabilities of Mycobacterium tuberculosis H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Systems Biol* **1** 26

Kanehisa M & Goto S 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28** 27-30

Marcus M & Halpern YS 1969 The metabolic pathway

of glutamate in Escherichia coli K-12. *Biochim Biophys Acta* **177** 314-320
Merrick MJ & Edwards RA 1995 Nitrogen control in bacteria. *Microbiologic Rev* **59** 604-622
Neidhardt FC (Ed.) 1996 *Escherichia coli and Salmonella*. Washington, DC: ASM Press.
Nemeria NS, Arjunan P, Chandrasekhar K, Mossad M, Tittmann K, Furey W & Jordan F 2010 Communication between thiamin cofactors in the Escherichia coli pyruvate dehydrogenase complex E1 component active centers: evidence for a "direct pathway" between the 4'-aminopyrimidine N1' atoms. *J Biol Chem* **285** 11197 -11209
Noor E, Eden E, Milo R & Alon U 2010 Central carbon metabolism as a minimal biochemical walk between precursors for biomass and energy. *Mol Cell* **39** 809-820
Reitzer L 2003 Nitrogen assimilation and global regulation in Escherichia coli. *Ann Rev Microbiol* **57** 155-176
Rodrigues JFM & Wagner A 2009 Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput Biol* **5** e1000613
Rodrigues JFM & Wagner A 2011 Genotype networks, innovation, and robustness in sulfur metabolism. *BMC Systems Biol* **5** 39
Sadava D, Heller C, Orians G, Purves W & Hillis D 2006 *Life: The science of biology*. (8th ed.). New York: WH Freeman.
Samal A, Rodrigues JFM, Jost J, Martin OC & Wagner A 2010 Genotype networks in metabolic reaction spaces. *BMC Systems Biol* **4** 30.
Schilling CH, Edwards JS & Palsson BO 1999 Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol Prog* **15** 288-295
Schuster P, Fontana W, Stadler P & Hofacker I 1994 From sequences to shapes and back - a case-study in RNA secondary structures. *Proc Royal Soc Lon Series B* **255** 279-284
Sun YR, Fukamachi T, Saito H & Kobayashi H 2011 ATP requirement for acidic resistance in Escherichia coli. *J Bacteriol* **193** 3072-3077
Vieira-Silva S, Touchon M, Abby SS & Rocha EPC 2011 Investment in rapid growth shapes the evolutionary rates of essential proteins. *Proc Natl Acad Sci U S A* **108** 20030-20035
Wagner A 2011 The molecular origins of evolutionary innovations. *Trends Genet* **27** 397-410
Wang Z & Zhang JZ 2009 Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol Evol* **1** 23-33