Review

Cell PRESS

Feature Review

The molecular origins of evolutionary innovations

Andreas Wagner

University of Zurich Institute of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

The history of life is a history of evolutionary innovations, qualitatively new phenotypic traits that endow their bearers with new, often game-changing abilities. We know many individual examples of innovations and their natural history, but we know little about the fundamental principles of phenotypic variability that permit new phenotypes to arise. Most phenotypic innovations result from changes in three classes of systems: metabolic networks, regulatory circuits, and macromolecules. I here highlight two important features that these classes of systems share. The first is the ubiquity of vast genotype networks – connected sets of genotypes with the same phenotype. The second is the great phenotypic diversity of small neighborhoods around different genotypes in genotype space. I here explain that both features are essential for the phenotypic variability that can bring forth qualitatively new phenotypes. Both features emerge from a common cause, the robustness of phenotypes to perturbations, whose origins are linked to life in changing environments.

A question about origins

Evolutionary innovations can be difficult to define rigorously [1,2], but they are often easy to recognize as qualitatively new and adaptive traits of organisms. In addition, they also frequently provide new platforms upon which further evolutionary change can unfold. Examples include the evolution of eyes, of flowers, and of flight, each of which opened new ecological niches.

We know many examples of innovations, each a fascinating piece of natural history. However, we know few of the principles that explain the ability of living things to innovate through a combination of natural selection and random genetic change. Random change by itself is not sufficient, because it does not necessarily bring forth beneficial phenotypes. For example, random change might not be suitable to improve most man-made, technological systems [3]. Similarly, natural selection alone is not sufficient: As the geneticist Hugo de Vries already noted in 1905, 'natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest' [4]. Any principle of innovation needs to explain how novel, beneficial phenotypes can *originate*. In other words, principles of innovation are principles of phenotypic variability.

Corresponding author: Wagner, A. (andreas.wagner@ieu.uzh.ch).

All macroscopically visible innovations ultimately have a molecular basis. They result from genetic changes that affect the biological function and regulation of biological macromolecules, as well as the interaction networks that such molecules form. Many such molecular changes are innovations in their own right. Examples include the evolution of proteins with new catalytic abilities [5]. Both macroscopic and molecular innovations involve change in three classes of systems. These are genome-scale metabolic networks, regulatory circuits, and macromolecules. Below I will discuss these systems and their involvement in many if not all evolutionary innovations. I will summarize a body of work which suggests that these systems share two important features. The first is the ubiquity of vast genotype networks, connected sets of genotypes with the same phenotype. The second is the great phenotypic diversity of small neighborhoods around different genotypes in genotype space. I will argue that these features jointly permit a principled and systematic understanding of phenotypic variability that can bring forth qualitatively novel phenotypes.

Metabolic network innovations

Metabolic networks are systems of hundreds to thousands of chemical reactions that are catalyzed by enzymes encoded by genes. These networks are responsible for providing cells with energy and multiple molecular building blocks – amino acids, nucleotides, lipids, and others – for cell growth. Innovations involving metabolic networks enable an organism to produce useful secondary metabolites, to detoxify waste products of its metabolism, or to use a novel molecule as a source of energy or chemical elements. The last of these is arguably most fundamental because it allows organisms to survive in novel chemical environments.

Prokaryotes are especially prolific metabolic innovators. Heterotrophic bacteria, for example, have acquired the ability to use a broad spectrum of different molecules as sole carbon sources. To acquire this ability for any one molecule is an innovation in any environment where this molecule is the only available carbon source. It allows the bearer to survive where other organisms would perish. Such metabolic innovations continue to occur to this day. For instance, prokaryotes have acquired the ability to use a wide array of man-made antibiotics as sole carbon sources, including fully synthetic compounds such as

0168-9525/$ – see front matter doi:10.1016/j.tig.2011.06.002 Trends in Genetics, October 2011, Vol. 27, No. 10

ciprofloxacin [6]. They also thrive on many toxic (to us) xenobiotic substances of industrial importance, such as polychlorinated biphenyls [7], chlorobenzenes [8,9], or pentachlorophenol, a synthetic pesticide first produced in 1936 [10,11]. The latter compound, for example, can be digested by the bacterium *Sphingomonas chlorophenolica*. The necessary metabolic pathway (Figure 1a) involves four steps that this organism assembled from enzymes processing naturally occurring chlorinated chemicals, as well as from an enzyme involved in tyrosine metabolism [10]. In other words, the enzymes or reactions themselves are not novel, but their combination is. Such novel reaction combinations are characteristic of metabolic innovations. They are facilitated by horizontal gene transfer, which shuffles existing metabolic genes among organisms [12].

Because horizontal gene transfer is rampant in prokaryotes it is easy to understand why they are prolific metabolic innovators. However, metabolic innovations are not only restricted to microbes. Consider the urea cycle, an innovation of land-living animals. It serves to convert highly toxic ammonia into a less toxic compound that can be excreted through urine. This compound is urea, produced in a chemical cycle of five reactions. The cycle combines four widespread reactions involved in arginine biosynthesis with a reaction catalyzed by arginase, an enzyme involved in arginine degradation [13]. Although the reactions themselves are not new, the combination of these reactions in the urea cycle is novel [13].

Innovation through regulation

Gene regulation changes the expression of a gene or the activity of its product. The products of many genes form regulatory circuits whose members cross-regulate the expression or the activity of each other. Changes in gene regulation are central to many innovations.

One example is the eyespots of butterflies, innovations that serve to deter predators [14–16]. In developing butterfly larvae, eyespots form in a prospective wing region called the eyespot focus, where the transcription factor *Distal-less* is expressed [17]. This protein plays a role in the development of many body structures, including legs and wings [18]. Its early expression in the eyespot focus demarcates the location where the eyespot will later form. *Distal-less* is expressed in all eyespot foci studied to date, and grafts of *Distal-less*-expressing foci to developing wing tissue are sufficient to cause eyespot formation in the recipient tissue [17]. Together with other regulators that are also expressed in eyespot foci, *Distal-less* is thus a prime candidate for a molecule that was involved in the origin of eyespots [19].

Another example is leaf dissection in plants. Some plant leaves are simple in shape, others are highly complex or dissected, consisting of multiple small leaflets (Figure 1b). Leaf dissection is an innovation that can serve many purposes, among them to prevent leaf overheating in hot environments and to increase CO_2 uptake in water [20,21]. The developing leaflets of most flowering plants with complex leaves show a marked increase in the expression of KNOTTED1-like homeobox (KNOX) transcription factors [22] (Figure 1b). This association is causal, as shown in the lamb's cress *Cardamine hirsuta* which has dissected leaves. Reducing the activity of *KNOX* genes severely impairs leaflet formation, whereas an increase in its expression is sufficient to produce additional leaflets [23].

The commonality of these and many other examples of innovation is that a change in the expression of already existing molecules is involved in the innovation. It is no coincidence that the regulatory molecules in both examples are transcriptional regulators. First, transcriptional regulation circuits have prominent roles in patterning plant and animal embryos [18,24]. Second, most regulatory phenomena ultimately affect gene transcription, which can be viewed as the regulatory backbone of life.

Novel macromolecules

Macromolecules are among the smallest system classes in which innovation can occur. Many innovations involve changes in the composition of macromolecules, especially proteins. Some such innovations are based on a single amino acid change. A case in point is the bacterial enzyme L-ribulose-5-phosphate 4-epimerase, where a single mutation at the active site – from histidine to asparagine – gives rise to a new catalytic activity, that of an aldolase joining one molecule of dihydroxyacetone phosphate and glycoaldehyde phosphate [25] (Figure 1c). Other innovations require many amino acid changes. Examples include antifreeze proteins. These proteins protect organisms from temperatures at which their body fluids would otherwise turn to ice [26–28]. Antifreeze proteins have evolved multiple times, and they can evolve rapidly [26,29]. For example, the arctic glaciation which probably has driven antifreeze protein evolution in arctic fish occurred less than 3 million years ago [30]. Antifreeze proteins are representative of much larger classes of innovations that have occurred repeatedly and with sometimes very different solutions to the same problem [31,32].

Towards a systematic understanding of innovation

The three classes of molecular change I discussed are not mutually exclusive. Most innovations involve an entangled combination of them, each with small effects, but jointly transformative. For example, the evolution of new metabolic abilities can involve novel combinations of enzymes created through horizontal gene transfer, changed regulation of existing enzymes, and amino acid changes that create new enzyme functions. It is nonetheless necessary to examine these classes of change separately to identify commonalities that can lead us to a more systematic understanding of innovation.

What specifically should such a systematic understanding of innovation achieve? First and foremost, it has to explain how biological systems can preserve existing phenotypes that are necessary for survival and reproduction, while at the same time exploring the many novel phenotypes necessary to find an adaptive new phenotype. Second, it should offer a unified understanding of innovations at different levels of organization. Third, it should capture the combinatorial nature of phenotypic variation: Innovation often involves new combinations and re-use of existing parts of a system [33,34], such as new combinations of enzymes in metabolic innovations. Fourth, it should account for the multiple origins of many innovations

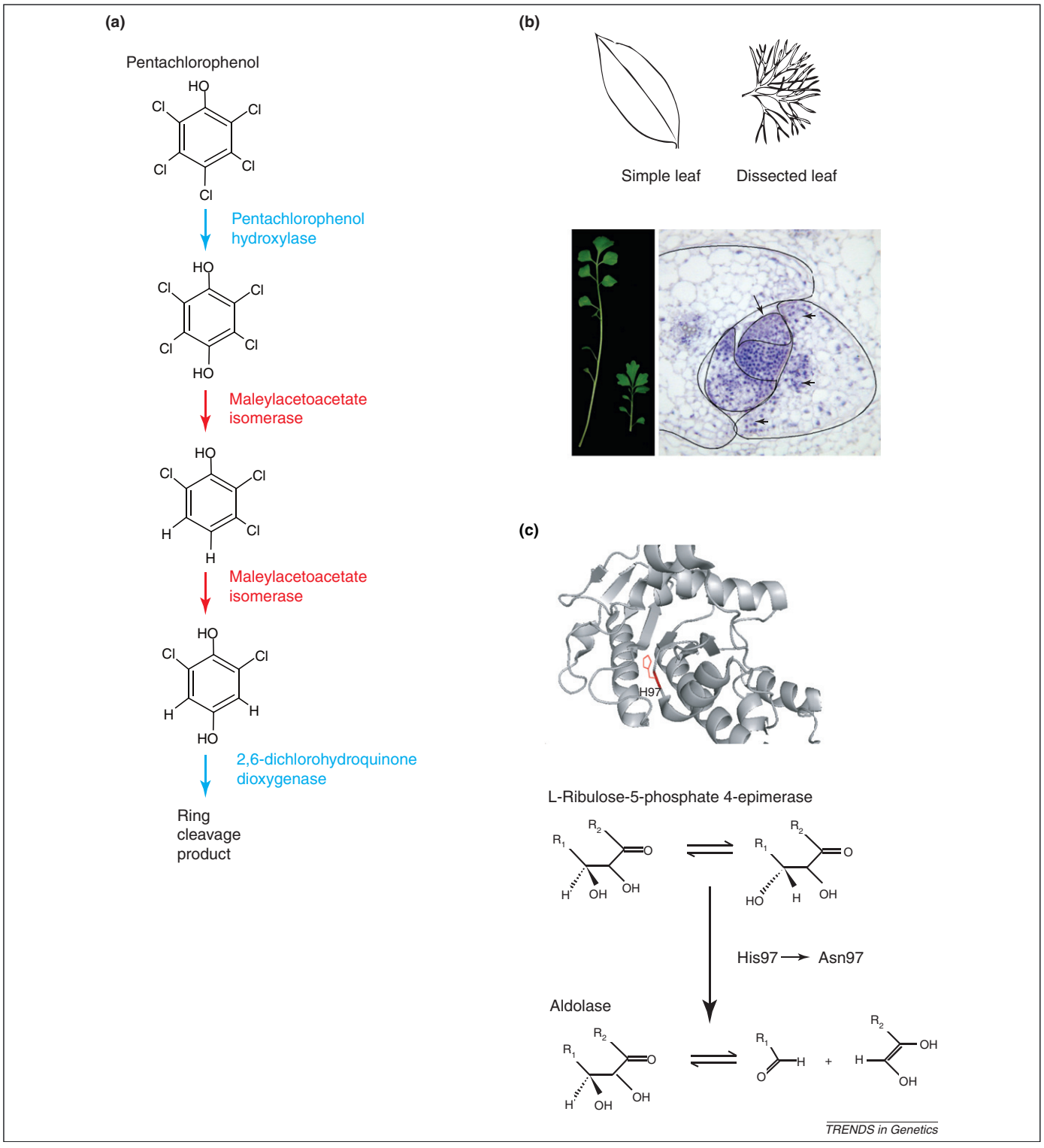

Figure 1. Three example innovations involving change in three different kinds of systems. **(a)** Four enzymatic steps in the degradation of pentachlorophenol. The enzymes marked in blue have probably been recruited to pentachlorophenol degradation from pathways that are involved in the degradation of naturally occurring chlorophenols, such as 2,6-dichlorophenol, which are produced naturally by some fungi and insects. The reactions in red are carried out by maleylacetoacetate isomerase, an enzyme involved in the degradation of phenylalanine and tyrosine in some organisms, including some bacteria, fungi, and humans [10]. **(b)** Upper: a simple and a dissected leaf; lower left: dissected leaf of the lamb's cress *Cardamine hirsuta*; lower right: cross-section of shoot apical meristem (central region enclosed by a black line) together with leaf primordia (enclosed areas surrounding the meristem, one is indicated by an arrow) of *C. hirsuta* [23]. Arrowheads indicate initiating leaflets and regions where KNOX proteins are expressed, as revealed by antibody staining. After Figure 1 of [23], with permission from Nature Publishing Group. **(c)** Upper: one subunit of the homotetrameric L-ribulose-5-phosphate 4-epimerase from *Escherichia coli*. A histidine residue (H97) in the catalytic site is highlighted. The structure is rendered from information in Protein Data Bank file 1K0W [114]. Lower: schematic drawing of the chemical reaction catalyzed by the epimerase shown above, as well as for a mutant with a single histidine to asparagine amino acid change at position 97; after [115]. The mutant catalyzes a new aldolase reaction.

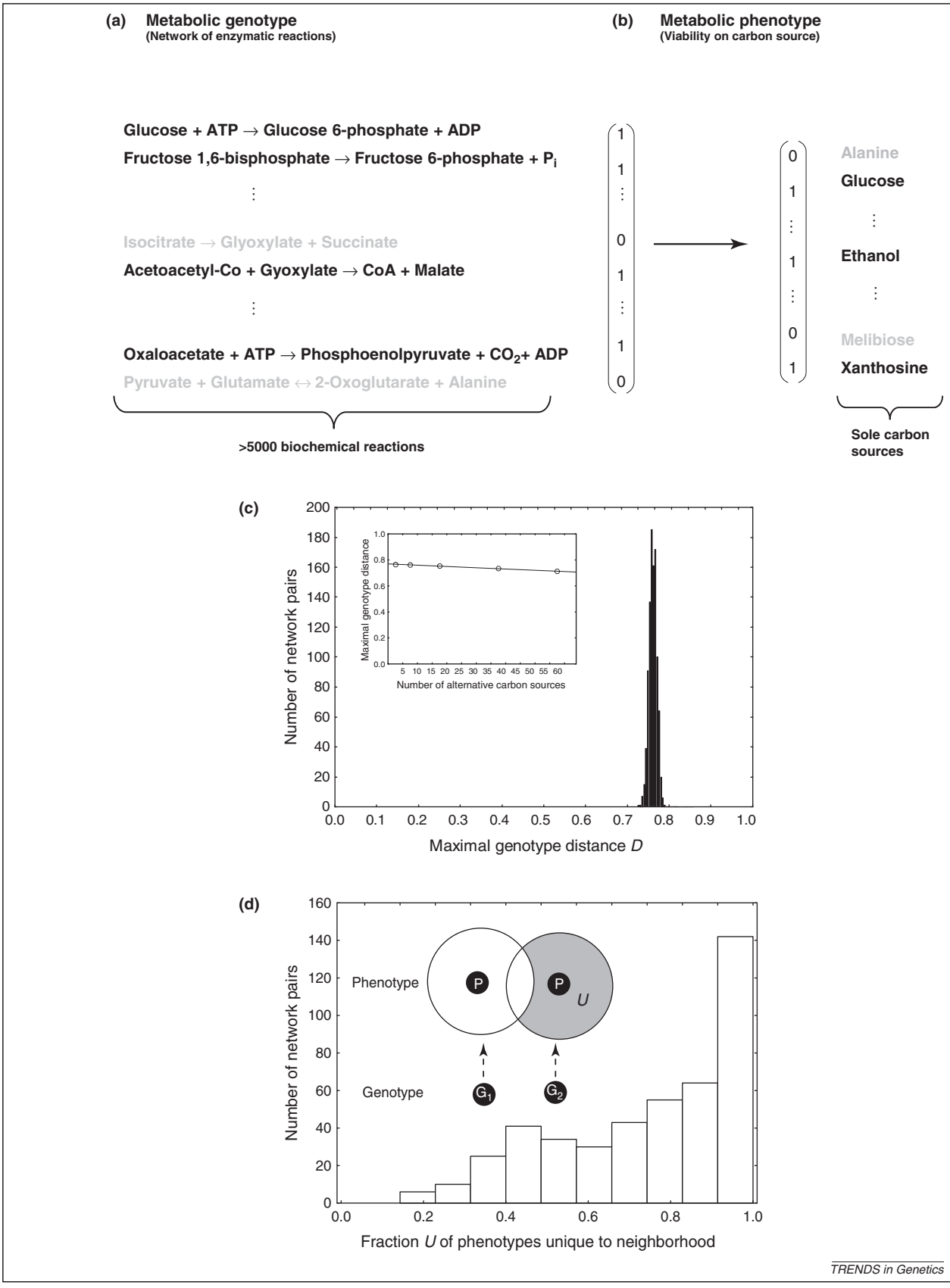

Figure 2. Metabolic genotypes, phenotypes, and genotype networks. Panels **(a)** and **(b)** represent the metabolic genotype and phenotype of a hypothetical metabolic network. **(a)** The metabolic genotype of a genome-scale metabolic network can be represented compactly as a binary string which is as long as the number of known enzyme-catalyzed biochemical reactions (there currently are more than 5000 such reactions.) The string contains a '1' for each reaction that the network can catalyze, and a '0' for all other reactions. **(b)** The entries of a metabolic phenotype string representation correspond to individual carbon sources. The string contains a '1' for every carbon source (black lettering) from which a metabolic network can synthesize all major biomass components. **(c)** Distribution of maximum genotype distance between 1000 pairs of metabolic networks that are the endpoints of random walks in genotype space, where each walk leads away from an initial metabolic network, while preserving the viability of this network on glucose. Inset: maximum genotype distances (vertical axis) between metabolic networks able to sustain life on a given number of carbon sources

[31]. For instance, carbon fixation, the incorporation of inert atmospheric CO2 into biomass, has been achieved through the Calvin–Benson cycle, the reductive citric acid cycle, and the hydroxypropionate cycle in quite different ways [32]. Finally, it should provide insights about the role of the environment and its change in innovation.

To study innovation more systematically than any one case-study would allow it is necessary to represent a system such that every possible phenotype can in principle be studied. The representation of a *genotype space*, the set of all possible genotypes, each with some phenotype, is suitable for this purpose. To understand innovation systematically one needs to understand how genotype change translates into phenotype change. Because genotypes are ultimately DNA sequences, a genotype space is ultimately a space of DNA sequences. However, it is often more expedient to use other, more compact representations of genotypes. I will next revisit the three classes of molecular systems introduced earlier, and discuss the relationship between genotype and phenotype in each. To elucidate this relationship one must examine thousands of genotypes and their phenotypes. Experimental techniques are currently inadequate for this purpose. Computational modeling and comparative analyses of massive amounts of data are thus still essential for this characterization.

The organization of metabolic genotype space

The known 'universe' of chemical reactions in metabolism comprises more than 5000 enzyme-catalyzed reactions [35,36]. The genome of any one organism encodes enzymes for only some of these reactions. We can view this collection of enzyme-coding genes as the *metabolic genotype* of the organism. As a first approximation, we can represent it as a binary string whose length is the number of reactions in the known reaction universe (Figure 2a). This string contains a '1' at some position i if the organism encodes a gene for reaction i (black lettering) and a '0' (grey lettering) if it does not. Reactions catalyzed by more than one enzyme can be represented through one of their enzyme-coding genes.

The set of possible metabolic genotypes forms a vast metabolic genotype space. The metabolic network of any one organism is a point in this space. Two metabolic networks have a genotype distance D in this space, which can be represented as the fraction of reactions in which they differ. The maximal distance in this space – the diameter of the space – corresponds to $D = 1$. Two metabolic networks are *neighbors* in this space if they differ in the presence or absence of a single reaction. Another important notion is that of a network's *neighborhood*, which comprises all the network's neighbors. More generally, the *k-neighborhood* of a metabolic network contains all networks that differ from this network in no more than k reactions.

To characterize the most fundamental metabolic *phenotypes* systematically, one can ask whether a metabolic network can synthesize all biomass molecules life needs in a given chemical environment, such as a chemically minimal environment with a single carbon source. Different metabolic networks can use different molecules as sole carbon sources. These observations motivate a representation of *metabolic phenotypes* as a binary string whose length corresponds to the number of molecules that can serve as a sole carbon source for a particular metabolic network. For any one metabolic genotype this string contains a '1' at position i if the network can synthesize all biomass components whenever carbon source $\underline{i}$ is provided as the only carbon source, in an otherwise minimal environment (Figure 2b). A string with multiple '1's corresponds to a network that is viable – it can synthesize biomass – in multiple minimal environments that differ in the sole carbon source they contain. In this context, a metabolic innovation is a new metabolic genotype viable on a novel carbon source – a novel '1' in the phenotype string. I focus here on innovations involving novel sources of carbon because they are most fundamental, but the framework I discuss also applies to other metabolic innovations [37].

Determining the metabolic phenotype of a single organism – viability on a given spectrum of carbon sources – can be performed experimentally. However, systematically exploring metabolic genotype space requires phenotypic information for many thousands of genotypes, and is currently unfeasible by experiment. Fortunately, it is possible to infer metabolic phenotypes from metabolic genotypes using the computational method of flux balance analysis [38]. Briefly, this method uses information about the stoichiometry of all chemical reactions in a metabolic network, and about available nutrients, to predict the spectrum of biomass molecules that the network can synthesize and how fast it can synthesize them. The predictions of the method are in good agreement with experimental results, with some exceptions, such as when enzyme misregulation prevents growth [39–42].

Starting from a metabolic network with a given phenotype, one can explore metabolic genotype space through random walks in this space. Each step in such a random walk consists of deleting a randomly chosen reaction from a metabolic network, or of adding a randomly chosen reaction from the known reaction universe, as might occur in a horizontal gene transfer event, followed by computation of the new metabolic phenotype of the network. Such an exploration reveals two simple organizational features of metabolic genotype space [37,43,44].

First, metabolic genotypes with the same phenotype form vast interconnected sets that extend far through genotype space. Two metabolic networks with the same phenotype are connected if one can be reached from the other through a series of additions and deletions of reactions, without ever changing the phenotype. Such a network of genotypes – a *genotype network* – can be viewed as a network of metabolic networks, all with the same phenotype. Each metabolic phenotype is associated with one or

(horizontal axis) [43]. The graphs show that most reactions in a metabolic network can be altered while preserving the network phenotype. **(d)** Distribution of the fraction of different and novel metabolic phenotypes that occur in the neighborhood of one but not the other metabolic network, based on pairs of networks chosen at random from the same genotype network. The two circles in the inset stand for the sets of phenotypes in the neighborhoods of two metabolic networks (genotypes), *G1* and *G2*, where phenotypes that would occur in only one but not the other neighborhood have grey shading and are labeled with the letter '*U*' for unique. The graph shows that for most network pairs (*G1,G2*) the majority of phenotypes is unique to one of the two neighborhoods. Data from [43].

more such genotype networks. A genotype network for any typical phenotype contains an astronomical number of genotypes and reaches far through genotype space. Figure 2c illustrates, as an example, that two metabolic networks viable on glucose can differ in more than 70% of their reactions. Networks viable on multiple sole carbon sources can be equally diverse [43] (Figure 2c, inset).

The second feature emerges if one studies the neighborhoods of different genotypes on a genotype network. That is, choose two metabolic genotypes *G1* and *G2* at random from the same genotype network, and examine the (1-)neighborhood of each. In each neighborhood, some networks are inviable, others have the same genotype as *G1* and *G2*, and still other networks are viable on new carbon sources. These neighbors have novel metabolic phenotypes, potential evolutionary innovations. If one compares the two sets of novel phenotypes that occur in each neighborhood one finds that they are very different. Most phenotypes that are found in a neighborhood are unique to that neighborhood, in the sense that they do not occur in the other neighborhood (Figure 2d).

Genotype networks in gene regulatory circuits

To study phenotypic variability in the expression of regulatory genes or the activity of their products, one must study the patterns of molecular activity, the regulatory *phenotypes* that such circuits produce. Because transcription regulatory circuits are central to embryo development and to regulatory innovations, many such circuits are well-studied individually [18,24]. The gene products of such circuits are transcriptional regulators that bind regulatory DNA sequences of other genes and activate or repress their transcription. However, the evolution of such circuits is difficult to study, partly because regulatory DNA sequences evolve very rapidly, partly because they can be spread over vast regions of non-coding DNA, and are thus difficult to characterize [24]. In addition, systematic understanding of novel regulatory phenotypes cannot be achieved by the analysis of a single circuit, and instead requires systematic analysis of thousands of circuits in the *genotype space* of such circuits. For this purpose, computational models of such circuits are currently indispensable [45–49].

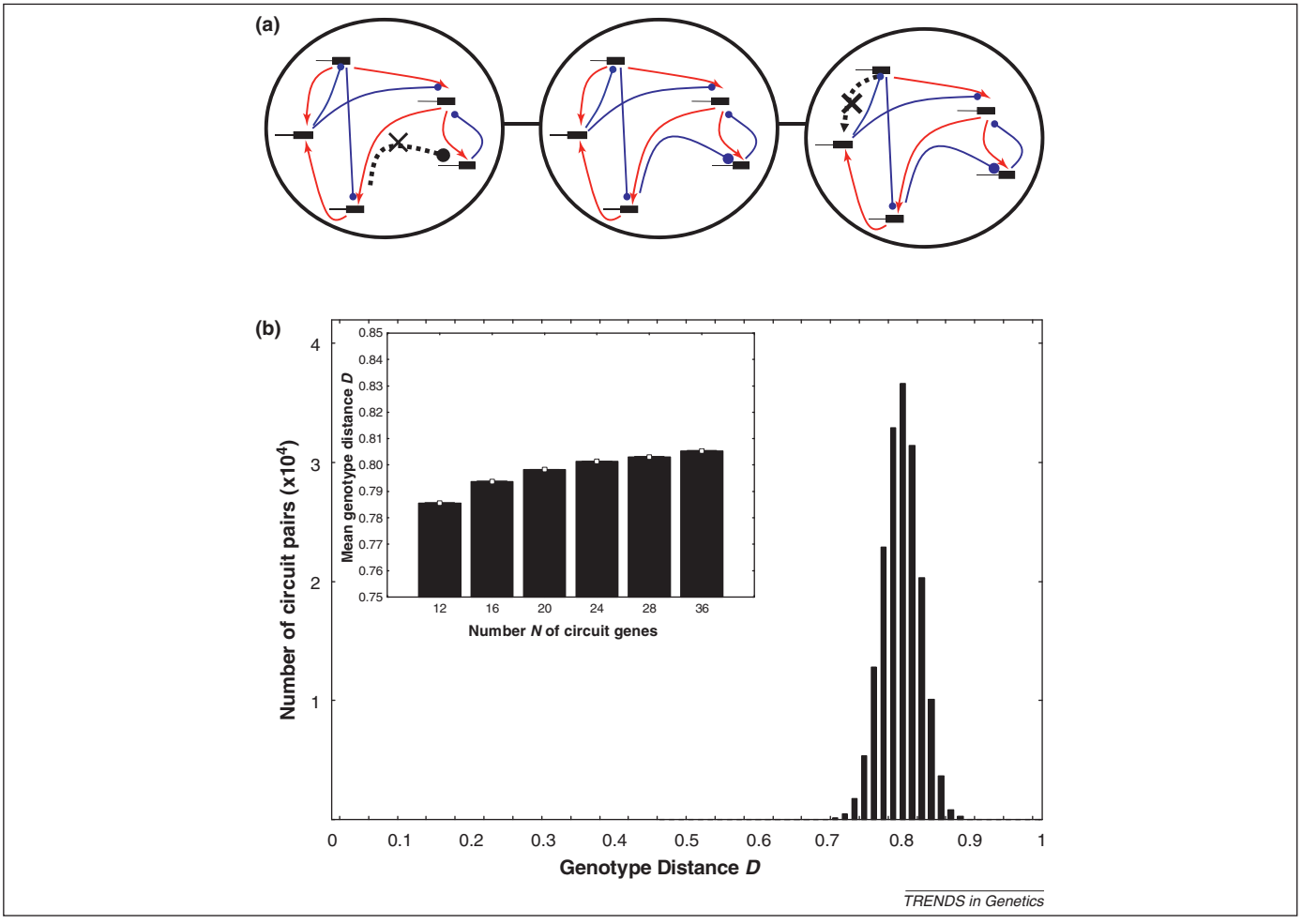

Figure 3. Neighbors and genotype networks in regulatory gene circuits. **(a)** The large central circle shows a hypothetical transcriptional regulation circuit of five genes (black bars) encoding transcriptional regulators. The effect of a regulator on the expression of any other regulator can be activating (red arrows), repressing (blue lines), or absent. The left and right circles each contain one neighbor of the central circuit in the genotype space of these circuits. Neighboring circuits differ from each other in one regulatory interaction (black crosses, dashed lines). Each of the three circuits has a regulatory genotype that is a member of a vast genotype space of circuits. **(b)** The horizontal axis shows genotype distance *D*, the fraction of regulatory interactions that differ between 5×10^5 model regulatory circuit pairs of 24 genes drawn from the same genotype network, that is, they all have the same expression phenotype. The inset shows the mean (and standard deviation) of the same distribution shown for model circuits with different numbers of genes. The graphs show that most regulatory interactions in a circuit can change without changing the phenotype of the circuit. This holds for circuits of various sizes that exceed a minimal complexity [50,51,116]; figures from [50,51].

Computational models that lend themselves to an exploration of a circuit space need to represent the topology of such a circuit – the pattern of activating and inhibiting regulatory interactions – in a systematic way. Figure 3a illustrates such a topology in a hypothetical gene regulatory circuit of five genes. One can think of the pattern of regulatory interactions in a circuit as the *regulatory genotype* of that circuit. Two circuits are neighbors in the genotype space of circuits if they differ in exactly one regulatory interaction. The neighborhood of a circuit comprises all circuits that differ from it in one regulatory interaction. One can represent the distance D between the regulatory genotypes of two circuits as the number or fraction of all regulatory interactions in which they differ. Two circuits are maximally different ($D = 1$) if they have no regulatory interactions in common. The regulatory interactions specified in a regulatory genotype determine the *phenotype* of the circuit. This phenotype reflects the activity or expression level of each gene in any one cell, and these can be represented either continuously or discretely ('on' or 'off'). The latter, discrete representation, although highly simplified, facilitates enumeration and comparison of different circuit phenotypes [49–51].

Just as in the case of metabolic genotype space, one can explore a space of regulatory circuits through random walks in this space. Each step in such a random walk changes one regulatory interaction at a time while preserving the gene expression phenotype of the circuit. This type of analysis reveals an organization of circuit genotype space that is very similar to that of metabolic genotype space.

First, for any one circuit phenotype, there are many circuits with this phenotype. Almost all of these circuits form one vast, connected genotype network that extends far through genotype space. For example, circuits of 20 genes with the same phenotype can differ in more than 75% of their regulatory interactions (Figure 3b). In consequence, there are many solutions to the problem of producing a given gene activity pattern [49–52]. This observation is not specific to any one model of transcription regulatory circuitry. It has also been made for circuits that involve modes of regulation different from transcription [49,53–55]. Supporting empirical evidence for this observation comes from recent analyses on the regulatory circuits driving sex determination, galactose metabolism, and coordinated expression of ribosomal proteins in yeasts, all of which can produce very similar regulatory phenotypes but with different regulatory interactions [56–58].

A second generic feature is that small neighborhoods around different circuits *G1* and *G2* with the same phenotype generally contain very different novel regulatory phenotypes. Many phenotypes in the neighborhood of *G1* are usually unique to that neighborhood, in the sense that they do not occur in the neighborhood of *G2* [51].

Innovation in macromolecules

The genotype of a macromolecule is its amino acid sequence for proteins or its nucleotide sequence for RNA. The phenotype is its folding pattern in 3D space or its biochemical function. The set of all possible genotypes forms a genotype space, a concept that goes back to the late John Maynard Smith who termed it a 'protein space' [59]. Later, computational work showed that genotype networks – connected networks of macromolecules with the same phenotype – exist in simple models of protein folding and of secondary structure phenotypes in RNA [60,61]. Such genotype networks had originally been called neutral networks [60], but evolution along such networks is all but neutral in fitness, and can involve many epistatic mutations with modest individual fitness effects [62,63]. Evidence accumulated since then shows that such networks also exist for real proteins of the same structure and/or function. In consequence, proteins with the same phenotype can be extremely diverse in their genotype [5,64–66].

An example involves the globin fold, a protein phenotype characteristic of oxygen-binding proteins, such as myoglobin and hemoglobin (Figure 4a). The tertiary structures of even distant globin representatives are very similar, but their sequences are highly diverged. For example, a study of six hemoglobins from plants and animals found that as few as 12.4% of amino acid residues were identical between any protein pair. In addition, only four out of 97 amino acids were unchanged in all of these proteins [67]. Phylogenetic information suggests that many known and highly diverse globins stem from a common ancestor, and that amino acid similarities of globins among different animal species reflect the evolutionary relatedness of the species [68,69]. Globins are not unusual in this regard. One of many other examples is the triosephosphate isomerase (TIM) barrel domain, a barrel-like protein structure whose 'planks' are made up of secondary structure elements. The TIM-barrel could derive from a single ancestor [70], but many proteins that harbor its structure have no recognizable sequence similarity to one another. More generally, surveys of multiple proteins reveal that many protein structures can be realized by very different amino acid sequences [5,64–66], although there are exceptions [71].

Thus, protein sequence space is permeated by genotype networks of proteins with similar structure and function. In addition, different neighborhoods in this space generally contain different novel phenotypes. Figure 4b shows results of a pertinent analysis [72]. The vertical axis shows the fraction of novel enzymatic phenotypes that occur in the neighborhood of one but not the other protein in a pair of proteins with the same structure, and with a given genotype distance D (horizontal axis). Even for proteins with moderate genotype distance D, more than 50% of enzyme phenotypes occur in the neighborhood of one protein but not the other. Similar observations exist for RNA structure phenotypes [60,73,74].

Common organizational properties of genotype spaces facilitate exploration of many novel phenotypes

The three system classes examined above are very different, but they share two important features. First, genotypes with the same phenotypes form extended genotype networks that reach far through genotype space. Second, different neighborhoods of genotypes on the same genotype network contain different novel phenotypes. Taken together, both of these features (illustrated in Figure 5a) facilitate the exploration of many novel phenotypes. To visualize this, consider a population of genotypes that

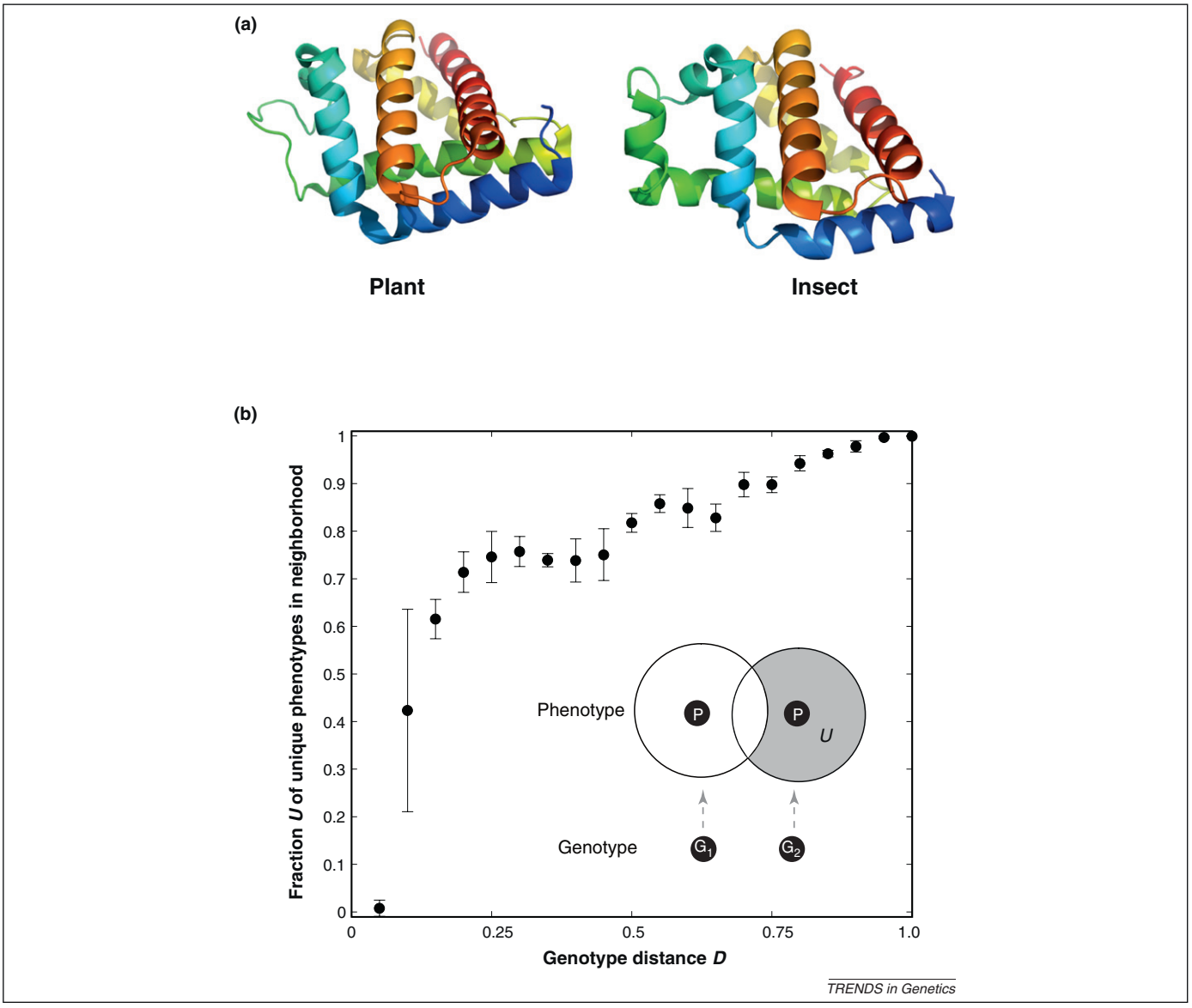

Figure 4. Structural conservation and great functional diversity in different neighborhoods of protein sequence space. **(a)** Two globin molecules with very similar tertiary structures but little amino acid sequence similarity. Left structure: root nodule hemoglobin of the lupine *Lupinus luteus* [117]. Right structure: hemoglobin of the midge *Chironomus thummi thummi* [118]. The root mean square difference in the position of 97 α-carbon atoms in the backbone of these helices is 2.88 Å [67]. Only 15.5% of the amino acids are identical in the seven conserved helices [67]. The globins shown correspond to entries 2LH3 and 1ECO of the protein database PDB [119]. Figure adapted from [95]. **(b)** The horizontal axis shows the mean genotype distance *D* of single-domain protein genotypes *G1* and *G2* with the same structure. This distance *D* is the fraction of amino acids in which the two proteins differ. The vertical axis shows the fraction *U* of proteins with enzymatic functions that occur in a small sequence neighborhood (≤5 point mutations) of one but not the other of the two genotypes. The data are based on 16 574 single-domain proteins of known structure and enzymatic function [72]. The large uncertainty (long error bars) at small distances *D* is caused by the low number of enzyme pairs at low genotype distance *D* [72]. The panel shows that the neighborhoods of even modestly diverged proteins ($D > 0.25$, corresponding to more than 25% amino acid divergence) typically contain proteins with mostly unique new functions, that is, functions that occur in the neighborhood of one but not the other protein genotype.

explores a genotype network through repeated cycles of mutation and selection that preserve an existing phenotype. If the genotype network of this phenotype reaches far through genotype space, individuals in the population can gradually change their genotype – and dramatically so – while preserving their phenotype. They can explore very different regions of genotype space and, through mutations, different neighborhoods of their genotype network. Because these neighborhoods harbor very different novel phenotypes the existence of genotype networks facilitates access to a great diversity of novel phenotypes. Recent experimental work corroborates these ideas, showing that populations of RNA enzymes that are spread out on a genotype network can undergo rapid evolutionary adaptation to a new chemical environment [75].

Figure 5b–d illustrate that both these features are essential by exploring several counterfactual scenarios. First (Figure 5b), if many genotypes were to form the same phenotype, but if these genotypes were isolated from one another, evolving genotypes would remain confined to small regions of this space, and they could no longer access as many different novel phenotypes. They can no longer explore large regions of this space through mutations that leave the phenotype unchanged. Second (Figure 5c), if genotype networks were to be connected, but were highly localized instead of extending far through genotype space,

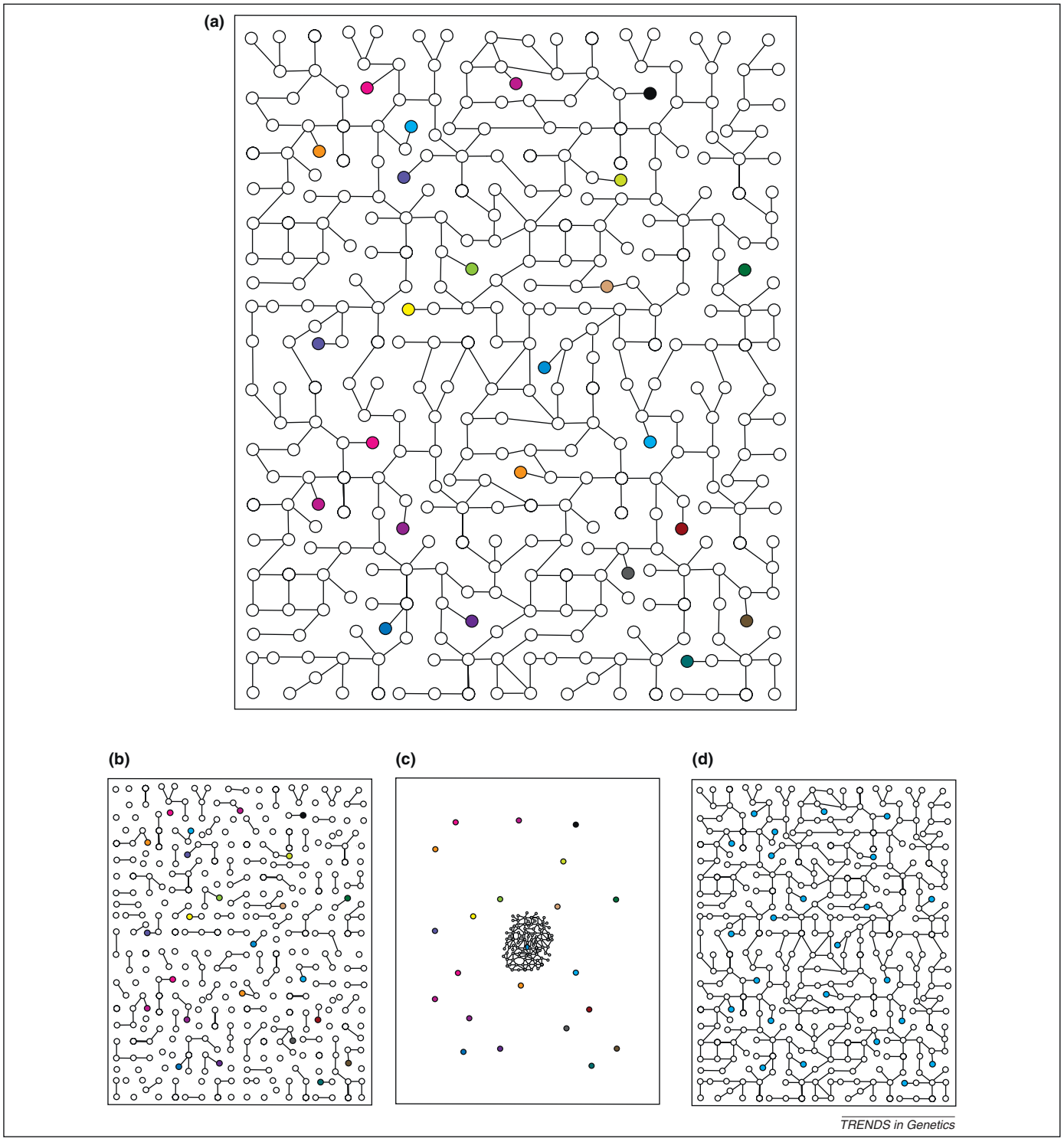

Figure 5. Connected genotype networks facilitate accessibility of many phenotypes. **(a)** The figure schematically represents a set of genotypes (grey circles) in a genotype space (rectangle) that share the same phenotype and form a genotype network; neighboring genotypes are connected by grey lines. Colored circles indicate genotypes that are adjacent to the genotype network, and that have different phenotypes (each color stands for a different phenotype). The figure illustrates that many different novel phenotypes can be accessed from a connected genotype network that spreads far through genotype space. Panels **(b)** through **(d)** show three counterfactual scenarios for genotype space organization, scenarios not typically found in systems studied thus far. **(b)** Most genotypes with the same phenotype are not connected. **(c)** The genotype network does not spread far through genotype space but is highly localized to a small region. **(d)** Neighborhoods of different genotypes are not diverse, but they contain the same novel phenotypes (blue circles). Note that genotype spaces have many dimensions with counterintuitive geometric properties, which a two-dimensional schematic can only represent crudely.

many novel phenotypes occurring elsewhere in genotype space would remain inaccessible. Third (Figure 5d), if the phenotypes in different neighborhoods of a genotype networks were not different but identical, the existence of a genotype network would be irrelevant for evolutionary innovation. Regardless of where a genotype lies on such a network, it would only have access to the same novel phenotypes.

Box 1. Robustness is necessary and sufficient for extended genotype networks

The first step is to show that robustness – a fraction $v > 0$ of any genotype's neighbors with the same phenotype – is necessary for the existence of connected genotype networks. The text below is written with metabolic networks in mind, but the argument would apply to regulatory circuits and molecules as well. Consider a typical phenotype P. It will be adopted by some very large number N_P of genotypes. These phenotypes typically constitute a very small fraction of a vast genotype space [5,44,50,73]. Let us assume that this set of N_P genotypes consists of genotypes chosen at random from genotype space, without requiring that each genotype has many neighbors with the same phenotype. The question is whether many or most of these genotypes would be connected in a genotype network. To address this question, let us examine one such genotype G and its S neighbors. If the reaction universe has S reactions, then each genotype has S neighbors. What is the probability that this genotype is isolated, that is, that none of its neighbors are members of P's genotype set [51]? To answer this question, consider first the probability that a randomly chosen genotype from the $2^S - 1$ genotypes different from G is *not* one of the S neighbors of G. This probability is equal to one minus the number of neighbors of G, divided by the $2^S - 1$ genotypes different from G, that is, it is equal to $1-[S/(2^S-1)]$. Similarly, the probability that a second genotype chosen at random from the now remaining $2^S - 2$ genotypes is *not* a neighbor of G is $1-[S/(2^S-2)]$. The same argument applies for a third, fourth, and further genotypes, until one reaches genotype number $(N_P - 1)$. From these considerations, we can estimate the probability that none of the $(N_P - 1)$ genotypes different from G are neighbors of G as their product. This probability is greater than

$$\left(1-\frac{S}{2^s-N_p+1}\right)^{N_P-1} \approx 1-\frac{S(N_P-1)}{2^s-N_P+1} \quad (1)$$

The ratio $S/[2^S - N_P - 1]$ will be very small, because the numerator is linear in S, whereas the denominator is dominated by the term 2^S, which is exponential in S. For this reason, and because the number N_P of genotypes with phenotype P is typically very small compared to the size 2^S of genotype space, the right hand side of the equation will be extremely close to one. This means that, in the absence of robustness, two genotypes with the same phenotype will typically be isolated from one another. They will not form a genotype network. More generally, a set of random genotypes must contain at least of the order of a fraction $1/S$ of all genotypes in genotype space before a large genotype network arises [111,112]. In a genotype space of 2^S genotypes, this is a gigantic number of genotypes, larger than the genotype sets of most phenotypes [44].

I next show that robustness ($v > 0$) is not only necessary but sufficient for the occurrence of genotype networks. To this end, it is useful to ask how genotype space would be organized if all genotype networks were random networks that shared only this feature. I emphasize that such random networks probably show little resemblance to actual genotype networks. However, they are useful in forming null-hypotheses about genotype space organization. To this end it will be useful to view genotype networks as graphs, mathematical objects that consist of *nodes*, and of *edges* that link these nodes. The nodes in a genotype space graph are genotypes. Two nodes are k-neighbors if they differ in exactly k reactions. An edge connects two genotypes if they are 1-neighbors.

Let us now iteratively construct a random graph in genotype space as follows (Figure I). In the first iteration, connect G to vS of its 1-neighbors at random, with equal probability that each 1-neighbor is chosen. Second, take each of the 1-neighbors of G that are now connected to G, and connect it to vS of its neighbors, most of which will be 2-neighbors of G. Proceed analogously for the 2-neighbors now connected to 1-neighbors (that are themselves connected to G) and connect these 2-neighbors to 3-neighbors, and so forth. By construction, nodes in the resulting graph will be connected to approximately vS of their neighbors. At some iteration step in this graph construction process, genotypes newly added to the graph no longer increase the diameter of the graph, the maximum distance between two nodes. To estimate approximately when this step is reached – how far this graph reaches through genotype space – focus on a k-neighbor G_k of G that is on the graph, and on one of the neighbors that G_k has on the graph. The probability that this neighbor is a $(k+1)$-neighbor of G is $1 - k/S$. The number of newly added nodes that are $(k+1)$-neighbors of G then follows a binomial distribution with parameters vS and $p = 1 - k/S$. The expected number of newly added nodes that are $(k+1)$ neighbors of G is thus $v(S - k)$. It falls below 1.0 if $D > 1 - (1/vS)$, where $D = k/S$ is the random graph diameter, expressed as a fraction of the diameter of genotype space. This inequality provides a lower bound for how far a random genotype network would extend through genotype space. Because S is large, numbering in the thousands for metabolic networks, and $v > 0.1$ for most characterized networks [43,50,60,73], D is close to 1.0. In consequence, one would expect random genotype networks to extend far through genotype space, provided that their genotypes have many neighbors with the same phenotype. Whether all or only some of the genotypes of a particular phenotype lie on one genotype network depends on v and on other system details [113].

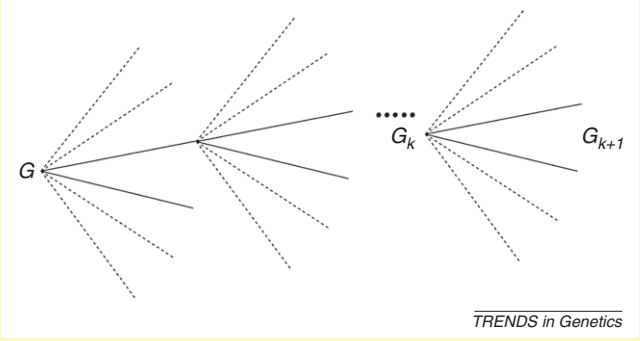

Figure I. Iterative construction of a random graph in genotype space.

The central importance of robustness

Why do the genotype spaces of three very different system classes show a common organization, in particular the existence of extended genotype networks? This common organization emerges from a common underlying cause. Phenotypes in all three systems are to some extent *robust* to genetic change—they are mutationally robust. In the language of genotype spaces, this means that some fraction $v > 0$ of the neighbors of a genotype G have the same phenotype as G itself. Box 1 uses simple mathematics to show that this feature is both necessary and sufficient for the existence of extended genotype networks.

The robustness required for the existence of extended genotype networks exists in all three system classes. For example, many neighbors of a protein G – differing in one amino acid – have the same phenotype P, as shown by random mutagenesis experiments [76–79]. In metabolic networks, both computational and experimental work shows that deletion of many enzyme-coding genes in metabolic networks of free-living organisms does not eliminate viability in any one environment [40,80–83]. In the *Escherichia coli* transcription-regulation network, introducing each of more than 500 new regulatory interactions does little to impair the operation of this network [84].

The second of the two features highlighted, the phenotypic diversity of different neighborhoods of a genotype network, can be easily understood from the following

observation. Consider an arbitrary neighbor of some genotype G and the phenotype P of this neighbor. What is the probability that an arbitrary neighbor of a second genotype G' (either on the same or a different genotype network) will also have phenotype P? Even if phenotypes were to be randomly, uniformly, and independently distributed among all genotypes in a genotype space, this probability would be very small providing there are many more possible phenotypes than a genotype has neighbors with different phenotypes. This is the case for all three system classes [43,44,50,51,85]. Thus, the phenotypic differences between different neighborhoods need no special explanation.

Robustness is widely held to be a key factor influencing evolvability [85–90]. In the framework discussed here it becomes especially clear why robustness is *qualitatively* important for phenotypic variability. Robustness brings forth extended genotype networks that facilitate the exploration of novel phenotype by a population, while allowing the population to preserve existing phenotypes.

Environmental change and the origins of robustness

Because mutational robustness is important for the ability of genetic systems to innovate, we need to ask about the evolutionary origins of robustness in different system classes. Two principal origins are possible. The first is that natural selection directly favors systems with high mutational robustness, and increases robustness over time. Population genetic theory shows that this origin requires populations that are polymorphic in mutational robustness, and this in turn requires large populations or very large mutation rates – because populations that do not fulfill these criteria are generally monomorphic [91,92]. These conditions might not hold for many organisms and systems.

The second candidate origin involves broadly defined change in the environment. For an organism and its metabolic network, such change could affect nutrients in the extracellular environment, whereas in a protein it might also include intracellular fluctuations in ions, metabolites, and various regulators. Mutational robustness in any one environment can arise because most biological systems need to persist in multiple environments. In other words, it can be a by-product of selection for robustness to changing environments [93–95]. In contrast to mutations, which are rare perturbations, environments change constantly for systems at all levels of organization, thus providing ample pressure to increase robustness. And although exceptions exist, mutational robustness and robustness to environmental change are usually positively correlated [94–100].

Metabolic networks are best-suited to illustrate the role of environmental change in promoting mutational robustness. Most reactions in a metabolic network such as that of *E. coli* or yeast are silent in both chemically complex and minimal environments, and even those reactions through which metabolites flow are mostly dispensable [44,80,101–103]. For example, in *E. coli*, more than 70% of reactions do not reduce biomass growth when eliminated [44]. This is not a peculiarity of the *E. coli* metabolic network but is a general property of viable networks with similar numbers of reactions [44]. If *E. coli* lived in only one environment, such as the above glucose minimal environment, it could eliminate most of its chemical reactions without detrimental consequences. The price would be a dramatic loss of robustness, as is observed in endosymbionts that live in highly stable environments [104–106].

As a metabolic generalist [107], the *E. coli* metabolic network can synthesize its biomass from dozens of alternative carbon sources [43,108]. Each requires one or more reactions that are specifically necessary to metabolize this carbon source. If one requires viability on all these carbon source, more than 100 previously dispensable reactions become essential [43]. Many of the remaining dispensable reactions would become indispensable in environments that vary in sources of other elements and of energy. When we examine a metabolic network with multi-environment viability in only one or few of these environments, as laboratory studies typically do, we would see exactly what we see in *E. coli*: many of its reactions are dispensable in one environment. In the language of metabolic genotype space, such a network has many neighbors that preserve viability. It is robust to genetic change *in this environment*. However, any one reaction that is dispensable in this environment might be essential in a different environment. If not, the reaction would eventually disappear – that is, the gene encoding the required enzyme would become eliminated from the genome.

Concluding remarks

Metabolic networks, regulatory circuits, and macromolecules are systems whose phenotypic variability is involved in most if not all evolutionary innovations, from the molecular to the macroscopic level. These systems typically need to be able to sustain life in multiple environments, and this can be a major cause of mutational robustness. Such robustness in turn brings forth genotype networks that extend far through genotype space, and that have phenotypically diverse neighborhoods. Together, these observations satisfy five requirements that a systematic understanding of evolutionary innovation – a theory of innovation – needs to meet.

First and foremost, they explain how biological systems can preserve old phenotypes while exploring many novel phenotypes. Second, they offer a unified understanding of phenotypic variability, and of how systems at different levels of organization can bring forth innovations. Third, the genotype space framework captures the combinatorial nature of innovation because it represents innovations as combinations of chemical reactions (enzymes), regulatory interactions, and amino acids that already exist and are reused for a new purpose [33,34]. Fourth, this framework can readily account for the multiple origins of many innovations [31]. Any two very different genotypes with the same phenotype can be viewed as different solutions to a problem that living systems face. The size and extent of genotype networks shows that most problems have not only more than one but many solutions. Fifth, this framework makes a strong statement about the role of environmental change in phenotypic variability and innovation.

Several other phenomena, such as phenotypic plasticity, gene duplication, and gene cooption, are also important in

evolutionary innovation. I show elsewhere how the framework presented here can help explain their role in innovation [109] and how this framework can also apply to technological innovation [110].

To understand phenotypic variability and how it affects innovations we need to understand genotype spaces. I here highlighted qualitative similarities among different system classes, but these system classes could show even more differences than similarities. We know little about these differences because systematic exploration of genotype spaces – aided by high-throughput genotyping and computation – is only now beginning to develop. And so is our systematic understanding of the ability of life to innovate – its innovability [109].

Acknowledgments

I would like to thank Massimo Pigliucci and Geerat Jacobus Vermeij for their thoughtful comments on the manuscript. I acknowledge support from the Swiss National Science Foundation and from the Yeast X project of SystemsX. Material in this review is a condensed version of a broad theory of innovation presented elsewhere [109].

References

1 Muller, G.B. and Wagner, G.P. (1991) Novelty in evolution-restructuring the concept. *Annu. Rev. Ecol. Syst.* 22, 229–256
2 Pigliucci, M. (2006) What, if anything, is an evolutionary novelty? *Philos. Sci.* 75, 887–898
3 Rechenberg, I. (1973) *Evolutionsstrategie*, Frommann-Holzboog
4 de Vries, H. (1905) *Species and Varieties, Their Origin by Mutation*, The Open Court Publishing Company
5 Todd, A. *et al.* (1999) Evolution of protein function, from a structural perspective. *Curr. Opin. Chem. Biol.* 3, 548–556
6 Dantas, G. *et al.* (2008) Bacteria subsisting on antibiotics. *Science* 320, 100–103
7 Rehmann, L. and Daugulis, A.J. (2008) Enhancement of PCB degradation by *Burkholderia xenovorans* LB400 in biphasic systems by manipulating culture conditions. *Biotechnol. Bioeng.* 99, 521–528
8 van der Meer, J.R. *et al.* (1998) Evolution of a pathway for chlorobenzene metabolism leads to natural attenuation in contaminated groundwater. *Appl. Environ. Microbiol.* 64, 4185–4193
9 van der Meer, J.R. (1997) Evolution of novel metabolic pathways for the degradation of chloroaromatic compounds. *Anton. Leeuw.* 71, 159–178
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 Cline, R.E. *et al.* (1989) Pentachlorophenol measurements in body-fluids of people in log homes and workplaces. *Arch. Environ. Contam. Toxicol.* 18, 475–481
12 Bushman, F. (2002) *Lateral DNA Transfer: Mechanisms and Consequences*, Cold Spring Harbor Press
13 Takiguchi, M. *et al.* (1989) Evolutionary aspects of urea cycle enzyme genes. *Bioessays* 10, 163–166
14 Stevens, M. *et al.* (2008) The anti-predator function of 'eyespots' on camouflaged and conspicuous prey. *Behav. Ecol. Sociobiol.* 62, 1787–1793
15 Stevens, M. *et al.* (2008) Conspicuousness, not eye mimicry, makes 'eyespots' effective antipredator signals. *Behav. Ecol.* 19, 525–531
16 Stevens, M. (2005) The role of eyespots as anti-predator mechanisms, principally demonstrated in the Lepidoptera. *Biol. Rev.* 80, 573–588
17 Brakefield, P.M. *et al.* (1996) Development, plasticity and evolution of butterfly eyespot patterns. *Nature* 384, 236–242
18 Carroll, S.B. *et al.* (2001) *From DNA to Diversity. Molecular Genetics and the Evolution of Animal Design*, Blackwell
19 Keys, D.N. *et al.* (1999) Recruitment of a hedgehog regulatory circuit in butterfly eyespot evolution. *Science* 283, 532–534
20 Gurevitch, J. (1988) Variation in leaf dissection and leaf energy budgets among populations of Achillea from an altitudinal gradient. *Am. J. Bot.* 75, 1298–1306
21 Givnish, T.J. (1987) Comparative studies of leaf form – assessing the relative roles of selective pressures and phylogenetic constraints. *New Phytol.* 106, 131–160
22 Bharathan, G. *et al.* (2002) Homologies in leaf form inferred from *KNOXI* gene expression during development. *Science* 296, 1858–1860
23 Hay, A. and Tsiantis, M. (2006) The genetic basis for differences in leaf form between *Arabidopsis thaliana* and its wild relative *Cardamine hirsuta*. *Nat. Genet.* 38, 942–947
24 Davidson, E.H. and Erwin, D.H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311, 796–800
25 Johnson, A.E. and Tanner, M.E. (1998) Epimerization via carbon-carbon bond cleavage. L-ribulose-5-phosphate 4-epimerase as a masked class II aldolase. *Biochemistry* 37, 5746–5754
26 Cheng, CC-H. (1998) Evolution of the diverse antifreeze proteins. *Curr. Opin. Genet. Dev.* 8, 715–720
27 Fletcher, G.L. *et al.* (2001) Antifreeze proteins of teleost fishes. *Annu. Rev. Physiol.* 63, 359–390
28 Davies, P.L. and Sykes, B.D. (1997) Antifreeze proteins. *Cur. Opin. Struct. Biol.* 7, 828–834
29 Chen, L.B. *et al.* (1997) Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proc. Natl. Acad. Sci. U.S.A.* 94, 3817–3822
30 Shackleton, N.J. *et al.* (1984) Oxygen isotope calibration of the onset of ice-rafting and history of glaciation in the North-Atlantic region. *Nature* 307, 620–623
31 Vermeij, G.J. (2006) Historical contingency and the purported uniqueness of evolutionary innovations. *Proc. Natl. Acad. Sci. U.S.A.* 103, 1804–1809
32 Rothschild, L.J. (2008) The evolution of photosynthesis... again? *Philos. Trans. R. Soc. B: Biol. Sci.* 363, 2787–2801
33 Carroll, S.B. (2008) Evo–devo and an expanding evolutionary synthesis: A genetic theory of morphological evolution. *Cell* 134, 25–36
34 True, J.R. and Carroll, S.B. (2002) Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* 18, 53–80
35 Ogata, H. *et al.* (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 27, 29–34
36 Chang, A. *et al.* (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res.* 37, D588–D592
37 Rodrigues, J.F. and Wagner, A. (2010) Genotype networks in sulfur metabolism. *BMC Syst. Biol.* 5, 39
38 Schilling, C.H. *et al.* (1999) Toward metabolic phenomics: Analysis of genomic data using flux balances. *Biotechnol. Prog.* 15, 288–295
39 Segre, D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117
40 Blank, L.M. *et al.* (2005) Large-scale C-13-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. *Genome Biol.* 6, R49
41 Ibarra, R.U. *et al.* (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 420, 186–189
42 Fong, S.S. and Palsson, B.O. (2004) Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat. Genet.* 36, 1056–1058
43 Rodrigues, J.F. and Wagner, A. (2009) Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* 5, e1000613
44 Samal, A. *et al.* (2010) Genotype networks in metabolic reaction spaces. *BMC Syst. Biol.* 4, 30
45 Jaeger, J. *et al.* (2004) Dynamic control of positional information in the early *Drosophila* embryo. *Nature* 430, 368–371
46 Sanchez, L. *et al.* (2008) Segmenting the fly embryo: logical analysis of the role of the segment polarity cross-regulatory module. *Int. J. Dev. Biol.* 52, 1059–1075
47 Albert, R. and Othmer, H.G. (2003) The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*. *J. Theor. Biol.* 223, 1–18
48 von Dassow, G. *et al.* (2000) The segment polarity network is a robust development module. *Nature* 406, 188–192
49 MacCarthy, T. *et al.* (2003) The evolutionary potential of the *Drosophila* sex determination gene network. *J. Theor. Biol.* 225, 461–468

50 Ciliberti, S. *et al.* (2007) Circuit topology and the evolution of robustness in complex regulatory gene networks. *PLoS Comput. Biol.* 3, e15
51 Ciliberti, S. *et al.* (2007) Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U.S.A.* 104, 13591–13596
52 Martin, O.C. and Wagner, A. (2008) Multifunctionality and robustness trade-offs in model genetic circuits. *Biophys. J.* 94, 2927–2937
53 Giurumescu, C.A. *et al.* (2009) Predicting phenotypic diversity and the underlying quantitative molecular transitions. *PLoS Comput. Biol.* 5, e1000354
54 Wagner, A. (2005) Circuit topology and the evolution of robustness in two-gene circadian oscillators. *Proc. Natl. Acad. Sci. U.S.A.* 102, 11775–11780
55 Nochomovitz, Y.D. and Li, H. (2006) Highly designable phenotypes and mutational buffers emerge from a systematic mapping between network topology and dynamic output. *Proc. Natl. Acad. Sci. U.S.A.* 103, 4180–4185
56 Martchenko, M. *et al.* (2007) Transcriptional rewiring of fungal galactose-metabolism circuitry. *Curr. Biol.* 17, 1007–1013
57 Tsong, A.E. *et al.* (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature* 443, 415–420
58 Tanay, A. *et al.* (2005) Conservation and evolvability in regulatory networks: The evolution of ribosomal regulation in yeast. *Proc. Natl. Acad. Sci. U.S.A.* 102, 7203–7208
59 Maynard-Smith, J. (1970) Natural selection and the concept of a protein space. *Nature* 255, 563–564
60 Schuster, P. *et al.* (1994) From sequences to shapes and back – a case-study in RNA secondary structures. *Proc. R. Soc. Lond. B.* 255, 279–284
61 Lipman, D. and Wilbur, W. (1991) Modeling neutral and selective evolution of protein folding. *Proc. R. Soc. Lond. B* 245, 7–11
62 Kulathinal, R.J. *et al.* (2004) Compensated deleterious mutations in insect genomes. *Science* 306, 1553–1554
63 Kern, A.D. and Kondrashov, F.A. (2004) Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs. *Nat. Genet.* 36, 1207–1212
64 Thornton, J. *et al.* (1999) Protein folds, functions and evolution. *J. Mol. Biol.* 293, 333–342
65 Bastolla, U. *et al.* (2003) Connectivity of neutral networks, overdispersion, and structural conservation in protein evolution. *J. Mol. Evol.* 56, 243–254
66 Rost, B. (2002) Enzyme function less conserved than anticipated. *J. Mol. Biol.* 318, 595–608
67 Aronson, H. *et al.* (1994) Quantification of tertiary structural conservation despite primary sequence drift in the globin fold. *Protein Sci.* 3, 1706–1711
68 Hardison, R.C. (1996) A brief history of hemoglobins: plant, animal, protist, and bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 93, 5675–5679
69 Goodman, M. *et al.* (1988) An evolutionary tree for invertebrate globin sequences. *J. Mol. Evol.* 27, 236–249
70 Copley, R.R. and Bork, P. (2000) Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* 303, 627–640
71 Doolittle, R. (1995) The origins and evolution of eukaryotic proteins. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.* 349, 235–240
72 Ferrada, E. and Wagner, A. (2010) Evolutionary innovation and the organization of protein functions in sequence space. *PLoS ONE* 5, e14172
73 Sumedha *et al.* (2007) New structural variation in evolutionary searches of RNA neutral networks. *Biosystems* 90, 475–485
74 Fontana, W. and Schuster, P. (1998) Shaping space: the possible and the attainable in RNA genotype–phenotype mapping. *J. Theor. Biol.* 194, 491–515
75 Hayden, E. *et al.* (2011) Cryptic genetic variation promotes rapid evolutionary adaptation in an RNA enzyme. *Nature* 474, 92–95
76 Huang, W. *et al.* (1996) Amino acid sequence determinants of beta-lactamase structure and activity. *J. Mol. Biol.* 258, 688–703
77 Rennell, D. *et al.* (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.* 222, 67–87
78 Weatherall, D.J. and Clegg, J.B. (1976) Molecular genetics of human haemoglobin. *Annu. Rev. Genet.* 10, 157–178
79 Kleina, L. and Miller, J. (1990) Genetic studies of the lac repressor. 13. Extensive amino-acid replacements generated by the use of natural and synthetic nonsense suppressors. *J. Mol. Biol.* 212, 295–318
80 Wang, Z. and Zhang, J. (2009) Abundant indispensable redundancies in cellular metabolic networks. *Genome Biol. Evol.* 1, 23–33
81 Stelling, J. *et al.* (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420, 190–193
82 Segre, D. *et al.* (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 15112–15117
83 Edwards, J.S. and Palsson, B.O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* 97, 5528–5533
84 Isalan, M. *et al.* (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature* 452, 840–845
85 Wagner, A. (2008) Robustness and evolvability: a paradox resolved. *Proc. R. Soc. B: Biol. Sci.* 275, 91–100
86 Aharoni, A. *et al.* (2005) The 'evolvability' of promiscuous protein functions. *Nat. Genet.* 37, 73–76
87 Rutherford, S. and Lindquist, S. (1998) Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342
88 Kirschner, M. and Gerhart, J. (1998) Evolvability. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8420–8427
89 Draghi, J. *et al.* (2010) Mutational robustness can facilitate adaptation. *Nature* 463, 353–355
90 Masel, J. and Trotter, M.V. (2010) Robustness and evolvability. *Trends Genet.* 26, 406–414
91 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
92 van Nimwegen, E. *et al.* (1999) Neutral evolution of mutational robustness. *Proc. Natl. Acad. Sci. U.S.A.* 96, 9716–9720
93 Papp, B. *et al.* (2009) A critical view of metabolic network adaptations. *HFSP J.* 3, 24–35
94 Meiklejohn, C. and Hartl, D. (2002) A single mode of canalization. *Trends Ecol. Evol.* 17, 468–473
95 Wagner, A. (2005) *Robustness and Evolvability in Living Systems*, Princeton University Press
96 Ancel, L.W. and Fontana, W. (2000) Plasticity, evolvability, and modularity in RNA. *J. Exp. Zool. Mol. Dev. Evol.* 288, 242–283
97 Szollosi, G.J. and Derenyi, I. (2009) Congruent evolution of genetic and environmental robustness in micro-RNA. *Mol. Biol. Evol.* 26, 867–874
98 Masel, J. and Siegal, M.L. (2009) Robustness: mechanisms and consequences. *Trends Genet.* 25, 395–403
99 Cooper, T.F. *et al.* (2006) Effect of random and hub gene disruptions on environmental and mutational robustness in *Escherichia coli*. *BMC Genomics* 7, 237
100 Milton, C.C. *et al.* (2003) Quantitative trait symmetry independent of Hsp90 buffering: Distinct modes of genetic canalization and developmental stability. *Proc. Natl. Acad. Sci. U.S.A.* 100, 13396–13401
101 Vitkup, D. *et al.* (2006) Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* 7, R39
102 Papp, B. *et al.* (2004) Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429, 661–664
103 Nishikawa, T. *et al.* (2008) Spontaneous reaction silencing in metabolic optimization. *PLoS Comput. Biol.* 4, e1000236
104 Thomas, G.H. *et al.* (2009) A fragile metabolic network adapted for cooperation in the symbiotic bacterium *Buchnera aphidicola*. *BMC Syst. Biol.* 3, 24
105 Pal, C. *et al.* (2006) Chance and necessity in the evolution of minimal metabolic networks. *Nature* 440, 667–670
106 Yus, E. *et al.* (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science* 326, 1263–1268
107 Ebenhoh, O. and Handorf, T. (2009) Functional classification of genome-scale metabolic networks. *EURASIP J. Bioinform. Syst. Biol. 2009* 13 Article ID 570456
108 Reed, J.L. *et al.* (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* 4, R54
109 Wagner, A. (2011) *The Origins of Evolutionary Innovations. A Theory of Transformative Change in Living Systems*, Oxford University Press
110 Raman, K. and Wagner, A. (2010) The evolvability of programmable hardware. *J. R. Soc. Interface* DOI: 10.1098/rsif.2010.0212
111 Reidys, C.M. (2009) Large components in random induced subgraphs of n-cubes. *Discrete Math.* 309, 3113–3124

112 Bollobas, B. *et al.* (1994) On the evolution of random Boolean functions. In *Extremal Problems for Finite Sets* (Frankl, P. *et al.*, eds), pp. 137–156, Janos Bolyai Mathematical Society

113 Reidys, C. *et al.* (1997) Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bull. Math. Biol.* 59, 339–397

114 Luo, Y. *et al.* (2001) The structure of L-ribulose-5-phosphate 4-epimerase: an aldolase-like platform for epimerization. *Biochemistry* 40, 14763–14771

115 O'Brien, P.J. and Herschlag, D. (1999) Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* 6, R91–R105

116 Cotterell, J. and Sharpe, J. (2010) An atlas of gene regulatory networks reveals multiple three-gene mechanisms for interpreting morphogen gradients. *Mol. Syst. Biol.* 6, 425

117 Arutyunyan, E.G. *et al.* (1980) X-ray structure investigation of leg-hemoglobin. VI. Structure of acetate-ferrileghemoglobin at a resolution of 2. 0 Å. *Kristallografiya* 25, 80

118 Steigemann, W. and Weber, E. (1979) Structure of erythrocruorin in different ligand states at 2.1 Ångstroms resolution. *J. Mol. Biol.* 127, 309–338

119 Berman, H. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. B: Biol. Crystallogr.* 58, 899–907