

Structural properties of genotype-phenotype maps

S. E. Ahnert^{1,2}

¹*Sainsbury Laboratory, University of Cambridge, Bateman Street, Cambridge CB2 1LR, UK*

²*Theory of Condensed Matter Group, Cavendish Laboratory,
University of Cambridge, JJ Thomson Avenue, Cambridge CB3 0HE, UK*

The map between genotype and phenotype is fundamental to biology. Biological information is stored and passed on in the form of genotypes, and expressed in the form of phenotypes. A growing body of literature has examined a wide range of genotype-phenotype (GP) maps and has established a number of properties that appear to be shared by many GP maps. These properties are 'structural' in the sense that they are properties of the distribution of phenotypes across the point-mutation network of genotypes. They include: a redundancy of genotypes, meaning that many genotypes map to the same phenotypes, a highly non-uniform distribution of the number of genotypes per phenotype, a high robustness of phenotypes, and the ability to reach a large number of new phenotypes within a small number of mutational steps. A further important property is that the robustness and evolvability of phenotypes is positively correlated. In this review I give an overview of the study of genotype-phenotype maps with particular emphasis on these structural properties, and discuss a model that attempts to explain why these properties arise, as well as some of the fundamental ways in which the structure of genotype-phenotype maps can affect evolutionary outcomes.

I. INTRODUCTION

Reproduction, as a fundamental property of biological systems, depends on the storage, processing, and transfer of biological information. That information is typically stored in the form of sequences, such as DNA, RNA, or amino acid sequences, and is more generally referred to as the *genotype*. In abstract models of biological evolution the genotype can be defined more generally, such as for genetic algorithms [1], where the genotypes are often binary strings. Some have chosen network topologies [2] or Boolean update rules [3] as genotypes. But genotypes are almost always linear and discrete representations of biological information. By contrast, the definition of a *phenotype* is almost impossibly broad. At every level of resolution of biological structure or function any higher-level outcome resulting from a sequence could be viewed as a phenotype. For example, the amino acid to which a given triplet codon maps can already be considered a phenotype. The structure, interactions, and functions of RNA and proteins are phenotypes, as is the function of a metabolic circuit or the state of a biological cell. On a longer time scale the development of a type of tissue or of a whole organism can also be viewed as a phenotype, as even more generally, the behaviour of an organism and its interaction with its ecosystem can be too. This complexity explains the wide range of genotype-phenotype (GP) maps found in the literature. It also explains why many GP maps that are studied are abstract theoretical or computational models, as the space of biological possibilities is often intractably vast, and the hierarchical levels of biological complexity are numerous. An illustration of the difficulties associated with the study of real biological GP maps can be found in the form of the protein folding problem [4], which has occupied biologists for decades, and which represents one of the most immediate connections between genotype and phenotype - the spatial rearrangement of an amino acid sequence into a

protein structure.

It has been known since Mendel [5] that genetic mutations can cause phenotypic changes. Building on Mendel's and Darwin's [6] work, Fisher, Haldane, and Wright produced the modern evolutionary synthesis [7–9] in the early twentieth century, which also introduced the idea of the *fitness landscape* [9]. A fitness landscape relates the space of genotypic variation to survival. It therefore contains a genotype-phenotype map implicitly. But the fitness landscape really consists of two distinct mappings: One is the genotype-phenotype map, and the other is the mapping from phenotypes to fitness values. A seminal contribution to the unravelling of these two mappings was made by Kimura [10], almost fifty years ago, who postulated that many mutations that are important for evolution must be neutral, meaning that they do not affect the fitness of the phenotype. While neutrality refers to selection rather than phenotypic change, and while it is possible for different phenotypes to be equally fit, the ubiquity of neutrality proposed by Kimura made it reasonable to suppose that many mutations are not only neutral in terms of selection, but in fact leave the phenotype entirely unchanged.

This idea was substantially extended by Maynard Smith in 1970 [11], who addressed the apparent contradiction [12] between the vast number of possible amino acid sequences and the tiny fraction these sequences that give rise to the proteins observed in nature. If the evolution of these proteins was driven by natural selection, how would this evolutionary process have found these solutions through random mutations? Maynard Smith's answer was to postulate that "functional proteins must form a continuous network which can be traversed by unit mutational steps without passing through nonfunctional intermediates" [11], which is similar to the definition of what is now commonly referred to as a *neutral network*. Such networks mean that functional proteins occupy connected subsets of genotype space, which makes

their discovery through an evolutionary process, driven by random mutations, feasible. Maynard Smith arguably laid the foundations for the modern study of genotype-phenotype maps in this paper by proposing the concept of a “protein space” of all possible amino acid sequences, in which neighbours are defined by single amino acid substitutions. This construct is almost exactly the same framework that has been used to study genotype-phenotype maps and their properties in RNA, proteins, and many other systems, mainly from the late 1980s onwards. An explicit discussion of genotype-phenotype mappings can be found in R. C. Lewontin’s 1974 work “The Genetic Basis of Evolutionary Change” [13], but Maynard Smith’s earlier work already marks the separation of the fitness landscape from the genotype-phenotype map, as it considers the set of genotypes that map to functional proteins in general. While fitness is very hard to quantify in a biological system, properties of genotype-phenotype maps are much more easily quantifiable. The barrier that remains is the vast size of the space, but the exponential increase in both computational power [14] and the availability of biological data [15] mean that ever more complex genotype-phenotype maps can now be studied quantitatively.

II. THE RNA SECONDARY STRUCTURE GP MAP

While the earliest theoretical work on genotype-phenotype maps [10, 11], and the earliest abstract models of such maps [16] considered protein tertiary structure, one of the most widely studied genotype-phenotype maps is that of RNA secondary structure, which is the specific configuration of base-pair interactions between different parts of a given RNA strand. This is because the folded structure for a given RNA sequence can be predicted computationally to a sufficiently high degree of accuracy [17], which makes large-scale surveys of this genotype-phenotype map possible. Furthermore the phenotype can be characterised precisely in terms of the base pairs that are formed in the folded structure, and can be denoted succinctly by using the ‘dot-bracket’ notation. In this notation a pair of parentheses represents a base pair, while a period represents an unbound base. The genotype-phenotype map therefore can be viewed as the map from an RNA sequence, composed of the four bases A,C,G, and U, such as:

ACGCCUCGGGGA

to a dot-bracket sequence, such as:

.(.(....)).

The structure that corresponds to this sequence is shown in Figure 1. Schuster et al. [18] in 1994 were the first to study global properties of the RNA secondary structure genotype-phenotype map using computational methods.

They established that the number of RNA sequences, or genotypes, vastly outnumbers the number of RNA secondary structures, or phenotypes. In addition they found that the number of sequences per structures followed a generalized form of Zipf’s law, meaning that many structures mapped to only a few sequences each, while a few structures occupied very large proportions of sequence space. Furthermore it was found that these common phenotypes could be reached in only a small number of mutational steps, starting from any random genotype. Lastly this investigation also showed the existence of long neutral paths in genotype space by demonstrating that 21.7% of neutral paths ended on genotypes that had no base in common with the reference sequence used as the starting point. This result implies the presence of extensive neutral networks. It has been proposed [19] that these neutral networks arise as a percolation effect, meaning that a connected network emerges spontaneously for sufficiently high link densities, purely based on the probabilities of individual links. As some phenotypes have a large number of genotypes, such an explanation of neutral networks in terms of percolation would imply that neutral networks result from the Zipfian distribution of the number of genotypes per phenotype. Random graph theory calculations based on this argument yield a theoretical percolation threshold of 0.37 [19] for four letter alphabets (such as in RNA), meaning that a phenotype occupying 37% of genotypes will percolate across the entire genotype space. Very recent work by Greenbury et al. [20] however has established that the genotypes belonging to a given phenotype are in fact much more correlated than random graph theory would suggest, and that neutral networks are therefore even more ubiquitous and pervasive. These results will be discussed in more detail in later sections of this paper.

III. THE HP MODEL GP MAP

The most obvious map between sequence and structure in modern biology, and the one that Maynard Smith addressed in his seminal work [11], is that from an amino acid sequence to a protein structure. The alphabet of amino acid sequences is much larger than that of RNA sequences, as there are twenty amino-acids that appear in protein sequences, compared to only four bases in RNA. The length of amino acid sequences tends to also be considerably larger than the length of RNA sequences that form functional secondary structures in the cell. The former typically consist of several hundred residues, while many examples of the latter are less than 50 bases in length [24]. In addition the folding process in protein tertiary structure relies on a complex network of residue-residue interactions, whereas the folding of RNA secondary structure is primarily governed by hybridization, which is a much simpler mechanism. For these reasons the prediction of protein structures from amino acid sequences is a much greater challenge than the prediction of

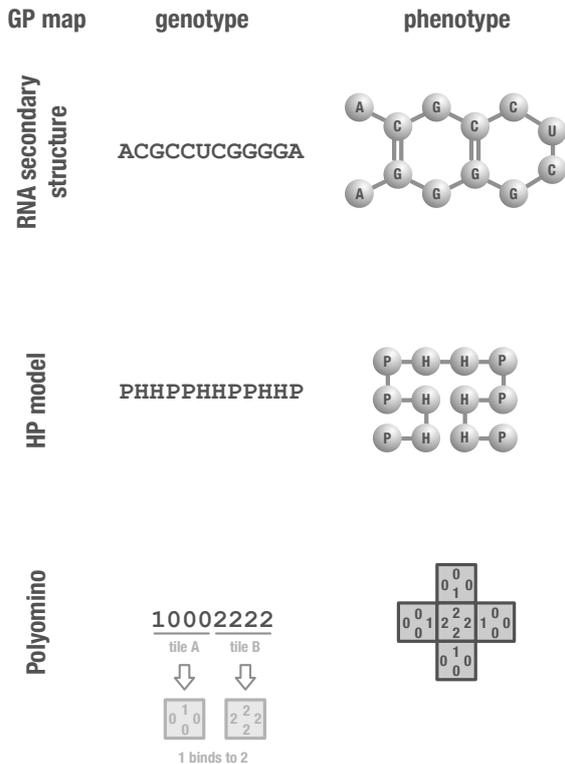


FIG. 1: An illustration of three genotype-phenotype maps. Top: RNA secondary structure [18, 19], which folds from a sequence of the four bases A, C, G, and U. Base pairs are shown as double bonds. The folding process can be predicted computationally with a high level of accuracy. Middle: The HP model [21] of protein folding. The example shown here is a sequence of hydrophobic (H) and polar (P) residues that folds onto a 5x5 two-dimensional lattice. Bottom: The Polyomino model [22, 23] is a self-assembly model in which sequences encode the interfaces of square 2D tiles. These self-assemble on a lattice in a stochastic process. The interfaces (sometimes termed ‘colours’) are written in clockwise order and concatenated for the tiles. In this case interface 1 binds to interface 2, and 0 denotes a neutral interface that does not interact. The two tiles self-assemble into the five-tile cross shape shown on the right.

RNA secondary structure. In 1985 Dill [16] considered an abstract model of protein sequences that focused on a distinction between hydrophilic and hydrophobic residues. The latter would form the core of globular proteins in this model, which led Dill to the conclusion that “the number of accessible conformations in the globular state” was an “exceedingly small fraction of the number accessible to the random coil”. In other words, the phenotypes that correspond to globular proteins occupy a small fraction of the total space of possible phenotypes. Building on this work, Dill and Lau introduced the HP model of protein folding [21], which distinguished hydrophobic (H) and polar (P) residues that fold on a two-dimensional lattice (see Figure 1). This level of abstraction made it possible to study the space of conformations exhaustively

for short chain lengths, and resulted in the work of Lipman and Wilbur [25], published shortly afterwards, who studied the genotype-phenotype map of the HP model extensively. Their findings included the discovery of the large neutral networks, similar to the connected “protein space” Maynard Smith originally envisaged, which led them to conclude that “perhaps the most striking observation is the critical role of neutral mutations in traversing this evolutionary space”, and to raise the question “whether this particular phenomenon is a general property of evolutionary spaces” [25]. A more general version of this question, namely to what extent the structural properties of genotype-phenotype maps are universal in biology, has received increasing attention in the recent literature on genotype-phenotype maps [20, 23, 26, 27], and perhaps the most central question of this review. As will be discussed below, the likely answer to Lipman and Wilbur’s question is positive - that properties such as extensive neutral networks are indeed universal properties of evolving systems. The reason for this may lie in the fundamental way in which biological information is organised [27].

The tractability of the HP model enabled further studies of its genotype-phenotype map, such as the work by Li et al (1996) [28] who found that the number of genotypes per phenotypes in the HP model displayed a highly skewed distribution, similar to the Zipfian distribution found for RNA [18]. This observation was described using the notion of *designability*, meaning that HP model proteins with a large number of genotypes are highly ‘designable’, which makes them more easily accessible in an evolutionary process.

It has been proposed [29] that the neutral networks of different protein phenotypes are largely disconnected from each other, resembling a ‘plum pudding’ [29] in contrast to the RNA genotype-phenotype map, whose interconnected phenotypes could be described with a ‘spaghetti’ metaphor [29]. Recent work [20], discussed in more detail below, has shown however that this strongly depends on the type of HP model used. The compact HP model, which only considers folded structures, exhibits interconnected phenotypes much like RNA.

IV. THE POLYOMINO GP MAP

In RNA secondary structure and protein tertiary structure, the most studied genotype-phenotype maps, the transformation from genotype to phenotype is achieved through a folding process that immediately links the genotype sequence to the final phenotype. Above individual proteins the next level in the hierarchical organisation of biological structures is protein quaternary structure, which describes the binding of multiple proteins to each other, for example in the form of protein complexes [30, 31]. Such structures differ from RNA secondary structure and single proteins because they require several independent subunits to come together and self-

assemble into a larger structure. Recent work by Ahnert et al. introduced a 2D lattice model of self-assembly [22] that can also be used as a genotype-phenotype map to model protein quaternary structure [23, 32]. In this model two-dimensional square tiles can bind to each other, forming larger connected structures called ‘Polyominoes’. In order to use this model as an abstract GP map for protein complexes we can encode the configuration of binding interactions on a set of tiles as a string, or genotype, and define the final assembled ‘Polyomino’ shape on the 2D lattice as the corresponding phenotype (see Figure 1). The Polyomino GP map model has been found to exhibit a rich and complex structure while still being tractable enough to explore genotype-phenotype maps and fitness landscapes exhaustively [23, 32]. Furthermore it exhibits many of the same properties as the RNA and HP model GP maps, such as a large number of genotypes relative to the number of phenotypes, a highly skewed distribution of the number of genotypes per phenotype, and extensive neutral networks [23], among several other properties that will be discussed in more detail below.

V. DELETERIOUS PHENOTYPES IN RNA, THE HP MODEL, AND POLYOMINOES

In all three GP maps discussed so far - RNA secondary structure, the HP model of protein structure, and the Polyomino model - a large proportion of genotype space maps to phenotypes that are in some sense not viable, and therefore deleterious. In RNA this corresponds to the case when the secondary structure without any paired bases has the lowest free energy, meaning that the RNA molecule remains unfolded. This is the case for 85% of RNA sequences of length $L = 12$ [20]. This proportion drops with increasing sequence length, but even for sequences of length $L = 20$, a third of RNA sequences still remain completely unfolded [20].

For the non-compact HP model, which considers all HP sequences of a given length, 98% of sequences of length $L = 24$ do not have a unique ground state structure. The absence of a unique ground state is conventionally regarded as a deleterious phenotype. For the compact HP model, specifically for the sequences on a 5x5 lattice, the fraction of such phenotypes is 82% [20]. Despite this high proportion of deleterious phenotypes the neutral spaces of the 5x5 compact HP model are highly connected to each other, thereby conforming more to the ‘spaghetti’ than the ‘plum pudding’ metaphor [20] discussed earlier. Unlike RNA secondary structure, the HP model retains a high, and possibly constant [33] fraction of unfolded phenotypes with increasing sequence length.

In the Polyomino phenotype certain configurations of interfaces on the square lattice tiles can lead to unbound growth. Furthermore, some configurations are non-deterministic, meaning that different structures are built in a stochastic assembly process, because the same

pair of building blocks can interact in multiple distinct ways. Unbound and non-deterministic (UND) building block sets are categorised as a single deleterious phenotype, similar to the RNA strands that remain unfolded and the HP sequences without a unique ground state. In the Polyomino models most commonly studied the proportion of the UND phenotype is 54% (two tiles, eight interaction types) and 80% (three tiles, eight interaction types).

In all of these three GP maps - RNA, HP 5x5, and Polyominoes - the deleterious phenotype is underrepresented in the neighbourhood of non-deleterious phenotypes, which illustrates that the neutral networks of different phenotypes are connected to each other even when a large proportion of genotype is occupied by the deleterious phenotype [20].

VI. NETWORK GP MAPS

The previous sections have given an overview of genotype-phenotype maps between sequences and biological structures. The definition of a phenotype is of course much broader than this. Biological interactions, such as gene regulatory interactions, metabolic interactions, and signalling networks, offer an alternative way to characterise phenotypes, and are discussed in this section.

A. Gene regulatory network GP maps

The study of genotype-phenotype maps has been extended in recent years to encompass the study of gene regulatory networks. In the model proposed by Ciliberti et al. (2007) [2, 34] an abstract regulatory network consists of genes that are connected by weighted edges, which correspond to activatory or inhibitory transcription factor interactions. Each gene also has an expression level as a function of time, which is updated at discrete timesteps according to a function that depends on the expression levels of the genes that regulate it. In this model the matrix of edge weights is taken to be the genotype, and the long-term equilibrium state of all expression levels is chosen to represent the phenotype. If no such stable state exists the phenotype is regarded as not viable, and discounted. Mutations of the genotype correspond to mutations of the weighted adjacency matrix of the regulatory network, and therefore to the removal or addition of an edge. In this model the neutral networks of the steady-state expression phenotypes are found to extend across genotype space much in the same way they do in RNA, proteins, and the Polyomino model [2]. Furthermore a tradeoff is revealed between the potential of a given regulatory network (i.e. a genotype) to innovate, and its robustness to mutations. This is intuitive, as the neighbours of a given genotype can either predominantly map to the same phenotype as that genotype (making that genotype robust), or to a variety of different phe-

notypes (giving that genotype the ability to innovate), but not both. We will return to this tradeoff and to the concepts of robustness and evolvability (as the ability to innovate is often referred to) in later parts of this review.

Building on this work a more abstract model of the genotype-phenotype maps of genetic regulatory networks was proposed by Payne and Wagner [3, 35, 36]. This model examines three-node Boolean networks, of which there are 104 topologically distinct varieties [37], which can be studied exhaustively. In a Boolean network each node is associated with a Boolean function, which is a bit string of length 2^k where k is the in-degree of that node. The Boolean function specifies the state of the node in response to all possible 2^k binary states of the k nodes connected to it by incoming edges. In order to consider all function sets and topologies one can simply consider all possible Boolean function sets on the fully connected three-node network. Some of these function sets will be independent of the presence of some of the edges, corresponding to less-than-fully connected topologies. For the three-node network the Boolean function set can be written as a binary vector of length $L = 3 \times 2^3 = 24$, which means that there are $2^{24} = 16,777,216$ possible function sets. These function sets are chosen as the genotypes, and the attractors of the network, which represent the long-term gene expression pattern of the regulatory system, can be chosen as the phenotype [35]. Similar to earlier investigations it is found that many genotypes map to few phenotypes, and that the distribution of genotypes per phenotype is skewed. Furthermore the tradeoff between robustness and evolvability at the level of individual genotypes is illustrated.

An alternative perspective on the measurement of network-based genotype-phenotype maps has been introduced by Ibáñez et al. (2014) [38] using the language of network analysis [39–41]. This map is also based on regulatory network model, similar to the work discussed above [2, 3, 34–36]. The genotype is defined as the combination of the network topology and the initial (binary) state of the gene expression. Phenotypes are defined in terms of steady-state dynamics. A difference to previous work is the topology of the gene regulatory networks, which are chosen to be either Watts-Strogatz small-world networks [39] or scale-free networks [40]. As a result the mutations only change the sign of existing network edges (in other words, turning an activatory into an inhibitory interaction, or vice versa) rather than changing the topology of the network. The choice of topologies is informed by the field of network analysis [41], which also provides the authors with the measurements they use to study robustness and evolvability. They apply these network measurements to a ‘pseudo-bipartite graph’ of genotypes and phenotypes, in which genotypes are connected to each other by point-mutation edges (much like in the genotype networks discussed previously), and in which genotypes are also connected to the phenotypes they map to. In this graph, robustness can be characterised in terms of the clustering coefficient of a phenotype node, since this co-

efficient measures the density of triangles formed by that phenotype and two genotypes. Evolvability can be measured by considering the so-called ‘one-mode projection’ of the genotype-phenotype edges onto the phenotypes, which produces a network of phenotypes connected by weighted edges. Conventionally the one-mode projection of a bipartite network produces a network of one node type with weighted edges that represent shared associations. For example, the bipartite network of authors and co-authored journal articles, projected onto authors, results in a weighted network of authors in which the edge weight is the number of articles these two authors have written together. Since two different phenotypes do not share genotypes, the projection here is more involved: Two phenotypes are connected if a path of length three exists from one to the other, and leading through two genotypes that are one mutation step apart. The edge weight is the number of such paths. Evolvability can then be established in terms of the emergence of a giant component in this phenotype network, which suggests that many phenotypes can be accessed through one-mutation steps [38].

B. Metabolic and signalling network GP maps

Metabolic networks aim to describe the entire interaction network of metabolites, chemical reactions, and related regulatory interactions that take place in a cell, and are another system for which genotype-phenotype map models have been developed. In [42] the authors examine 1,000 mutated versions of the metabolic network of *E. coli*, regarding each such network as a genotype, and use flux-balance analysis to determine the metabolic viability of this network. Flux-balance analysis is an efficient mathematical modeling technique for metabolic networks [43]. This notion of viability encompasses a number of possible phenotype definitions, such as the ability of the network to produce all biochemical precursors from glucose in an aerobic minimal medium, or the ability to sustain the metabolism given a range of external carbon sources. The outcomes of this model suggest that the neutral networks of the same metabolic phenotypes extend across genotype space, but also that from any given phenotype many other phenotypes can be accessed in just a few mutational steps [42]. These results characterise the robustness and evolvability of metabolic networks. A similar analysis has been undertaken for the TOR signalling circuit in *S. cerevisiae* [44], which examines 69,120 alternative topologies of the signalling circuit. For each topology an optimisation procedure is used to find the set of kinetic parameters that best reproduce reference signalling behaviour. The phenotypes are defined in terms of clustered trajectories in the space of the concentrations. This analysis too finds neutral networks that extend across genotype space, as well as the heavily skewed distribution of genotypes per phenotype found in RNA, the HP model, and Polyominoes. It also

shows that it is the phenotypes with the largest neutral networks that are both the most robust and the most evolvable [44].

C. Multi-level GP maps

As is evident from the wide range of examples discussed so far, genotype-phenotype maps can be defined at many levels. An approach that has sought to capture the hierarchy of phenotypes and thus multiple levels of the relationship between genotype and phenotype is the toyLIFE model [45]. This model encompasses gene expression, protein folding, gene regulation, and metabolic reactions. Like in other systems, this model too displays many more genotypes than phenotypes and a highly skewed distribution of genotypes per phenotype. Many aspects of the toyLIFE model remain to be explored, due to its high complexity. Its potential may lie particularly in addressing research questions that connect different levels of the genotype-phenotype map.

VII. DIRECT MEASUREMENTS OF BIOLOGICAL GP MAPS

All genotype-phenotype maps discussed so far are to a greater or lesser extent abstractions of real biological systems. The RNA secondary structure GP map is the closest to a real biological genotype-phenotype map, as small RNAs fulfill a wide range of functions in the cell, and the map has been studied in the context of this naturally occurring variety [24]. Nevertheless the phenotype representation in terms of the base pairs only captures part of the complex three-dimensional structure of folded RNAs. A level of abstraction is inevitable if the goal is to study general properties of genotype-phenotype maps, which at the level of protein tertiary structure remain intractable due to the enormous complexity of the folded amino acid chain, and at the level of interaction networks remains challenging due to the difficulty of reliably measuring and verifying the large number of interactions between proteins, genes, DNA, and RNA in the cell. There have however been attempts to measure 'real' GP maps of protein domains [46] and of transcription factor binding sequences [47] to see whether they, at least within the limited view that existing sequence data affords us, share the same properties as the more abstract, comprehensive GP map models. The genotype-phenotype map of single-domain protein sequences uses the enzymatic function classification as the phenotype [46]. Because this is a functional classification (in terms of the enzyme commission (EC) classification [48]), and not the domain structure itself, there is an additional mapping between structure and function that complicates the genotype-phenotype map. Several different structures can fulfill the same function - a phenomenon that the authors of [46] refer to as *structural promiscuity*. While 86% of functions

are only carried out by one structure, some functions (such as DNA-directed DNA polymerase) can be carried out by up to 14 different structures. In this GP map the sets of enzymatic functions in the neighbourhoods of two genotypes diverge rapidly with increasing distance between the genotypes. This is even true when considering genotypes that map to the same protein structure. The multitude of functions exhibited by some protein families (which group together proteins of the same structure) is the reason for this rapid growth in functional diversity with increasing genotype distance, or in other words "Functional neighbourhood diversity emerges from the multifunctionality of structures." [46]. Another prominent example of an attempt to directly measure the properties of a real biological genotype-phenotype map is the work by Payne and Wagner [47] on the transcription factor binding sites for 104 transcription factors in mouse, and 89 in yeast. By examining the space of all binding sites of eight nucleotides in terms of an enrichment score [49] that represents the binding affinity, one can construct the genotype network and observe where the sequences that bind to a given TF fall on this network. In almost all cases the majority of bound sites for a particular TF form a single connected component, and for a majority of TFs all bound sites are part of a single connected genotype network. The binding site sequences are significantly more robust than expected by chance, and also significantly more evolvable (measured by considering how many mutations were required to reach a binding sequence for a different TF). Both robustness and evolvability increase with the size of the binding site repertoire. These findings mirror the results regarding robustness and evolvability found in more abstract model GP maps of RNA secondary structure, the HP model, Polyominoes, and network phenotypes.

The observation that many biological phenotypes are highly robust has been also made on a more general level, supporting the more detailed investigations of genotype-phenotype maps discussed so far. For example, robustness may explain why similar gene expression patterns are observed across a wide range of species, even when the cis-regulatory regions are highly divergent [50]. A similar result was found across a range of *Drosophila* species [51].

VIII. STRUCTURAL PROPERTIES OF GENOTYPE-PHENOTYPE MAPS

The previous sections provide an overview of research into the properties of genotype-phenotype maps, and show that certain characteristics of the distribution of phenotypes over the space of genotypes have been observed across many different phenotypes and modelling approaches. In this section I explore these 'structural' properties in more detail, particularly in the context of relatively tractable genotype-phenotype map models that can be studied on a global level, such as the RNA sec-

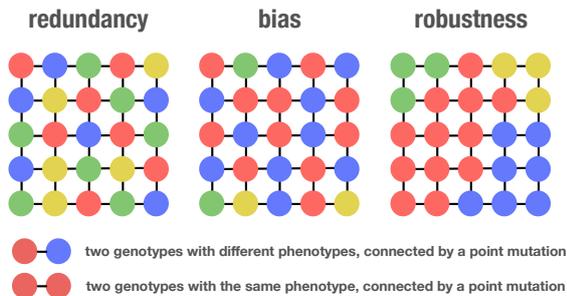


FIG. 2: The three basic structural properties of genotype-phenotype maps, illustrated on a simple network of genotypes (nodes) that map to particular phenotypes (colours), and are connected by single-point mutations (edges). The redundancy (left) of genotype-phenotype maps means that there is a much larger number of genotypes than phenotypes. The bias (middle) of a genotype-phenotype map is the extent to which the distribution of the number of genotypes per phenotypes is non-uniform. In many real genotype-phenotype maps this distribution is approximately Zipfian. The presence of robustness (right) is one way of describing the fact that the different genotypes of a given phenotype are often close to each other in genotype space. As we discuss in the main text, the presence of large and robust neutral networks does not simply follow from bias.

ondary structure, HP model and Polyomino GP maps. The first three - redundancy, bias, and robustness - are illustrated in Figure 2.

A. Redundancy

All observed genotype-phenotype maps exhibit one fundamental property that is a prerequisite for all the other structural properties discussed below. This property is *redundancy*, in the sense that there are many more genotypes than phenotypes. Redundancy is a necessary (but not sufficient) condition for neutral networks to exist. Without redundancy, evolutionary processes would never find viable phenotypes in the vast space of possible sequences. The work of Kimura [10] and Maynard Smith [11] already implied redundancy by postulating the ubiquity and necessity of neutral (or near-neutral) mutations, and early studies of the HP model GP map [25] and RNA secondary structure map [18] established the presence of neutral networks, which too implied redundancy. However the first comprehensive study of an entire genotype-phenotype map, and thus the first explicit calculation of redundancy is arguably the introduction of the concept of the “designability” of protein structures by Li et al. [28] in the context of the HP model. A recent overview of redundancies across RNA, HP and Polyomino models is given in [23]. It reveals that, while redundancy is present in all maps, it scales differently with sequence length. In RNA, we have 57 phenotypes and 1.7×10^7 genotypes for sequences of length $L = 12$, while we have 431 phenotypes and 1.1×10^9 genotypes for $L = 15$, and 11218

phenotypes and 1.1×10^{12} genotypes for $L = 20$. An estimate of the number N_p of RNA phenotypes for a given L is given by $N_p = 1.4848 \times L^{-\frac{3}{2}} (1.8488)^L$ [18], which grows much slower with L than the number of genotypes, $N_g = 4^L$. In the Polyomino model, in which the sequence alphabet is larger, the genotype space grows more rapidly relative to the number of phenotypes (13 phenotypes and 1.7×10^7 genotypes for two tiles and eight colours, 147 phenotypes and 6.9×10^{10} genotypes for three tiles and eight colours, and around 2237 phenotypes and 1.8×10^{19} genotypes for four tiles and 16 colours). In the HP model there is a very large difference in redundancy between compact and non-compact structures. If only compact structures are considered as phenotypes, such as in [28], then the number of phenotypes is significantly smaller. For example all HP sequences of length 25, folded onto a 5x5 grid (sometimes described as the HP5x5 model) yield 549 phenotypes for the $2^{25} \approx 3.36 \times 10^7$ genotypes [20], whereas the equivalent non-compact map (referred to as HP25) yields 107,336 folded phenotypes - in other words around 200 times as many - for the same number of genotypes [23]. However, even in the HP25 GP map genotypes still outnumber phenotypes by a factor of 300.

B. Bias

Redundancy is the foundation for the next most fundamental property of genotype-phenotype maps, which is the skewed distribution of the number of genotypes per phenotypes, also described as *bias* [23]. This more subtle property does not immediately follow from the presence of neutral networks, as in principle the neutral networks of different phenotypes could all be similar in size. Lipman and Wilbur [25] observed this bias in one of the first comprehensive studies of the HP model GP map, as did Schuster et al. [18] in their work on RNA secondary structure, who also pointed out the Zipfian character of the distribution that has since been confirmed in other studies [26, 28], and for other genotype-phenotype maps [23]. The Zipf-like distribution could therefore be regarded as the distinctive hallmark of genotype-phenotype bias. Another way to quantify bias, in the form of a simple number, is to consider the fraction of phenotypes with the largest number of genotypes that cover 95% of genotype space. For example in $L = 20$ RNA, the top 10% of the most frequent phenotypes cover 95% of genotypes, and in three-tile polyominoes this fraction is 16% [23]. But the relationship between bias and the presence of neutral networks is more complex. As will be discussed in the next section, bias alone cannot explain the presence and extent of neutral networks.

C. Robustness

Robustness is typically measured in terms of the fraction of possible mutations that leave a phenotype un-

changed. This measurement can be applied at the genotypic level or at the phenotypic level [52]. In the latter case the robustness of a given phenotype can be calculated as the average of genotypic robustness over the genotypes that map to that phenotype [52]. Robustness as a local measurement of neutrality is a key property of neutral networks. A comparison of the number of genotypes that map to a phenotype - in other words the *frequency* of a phenotype - and its robustness shows that a phenotype's robustness ρ_p scales as the logarithm of its frequency f_p in RNA secondary structure [20, 23, 53, 54], Polyominoes [20, 23] and the HP model [20], as illustrated schematically in Figure 3. Formally we therefore have:

$$\rho_p \approx a + b \log f_p$$

where a and b are positive constants. If genotypes of the same phenotype were uncorrelated (while still following the same distribution in terms of bias), the robustness of a phenotype would simply be equal to its frequency [20]. The robustness of the most frequent phenotypes in the RNA, HP, and Polyomino maps lies between 0.5 and 1, which means that the logarithmic increase in the robustness makes the phenotypes far more robust than they would be according to the linear increase of the null model. Across these GP maps phenotypes are therefore significantly more robust than one would expect based on bias alone [20]. This also becomes clear if one considers the size of the largest neutral component of a given phenotype, and the number of its neutral components. In the null model one sees the emergence of a giant component around a phenotype frequency of 0.37, which corresponds to the percolation threshold found by Schuster [19]). By comparison the observations for the RNA, HP, and Polyomino GP maps show that the sizes of the largest neutral components scale logarithmically with phenotype frequency, making them much larger than the null model expectation below the percolation threshold. Furthermore, at the percolation threshold the number of distinct neutral components in the null model, which grows linearly with phenotype frequency up to that point, drops rapidly, as expected. This again contrasts with the three observed GP maps, in which most phenotypes are divided into less than a hundred (and in many cases less than ten) components [20], rather than the many thousand components observed in the null model below the percolation threshold. These findings again show that genotypes are far more correlated than one would expect based solely on GP map bias. The consistency of the logarithmic relationship between robustness and frequency, across three entirely different GP maps and many orders of magnitude of phenotype frequency, strongly suggests that it is a fundamental property of genotype-phenotype maps. This logarithmic scaling has been discussed in the literature [53, 54], and an analytical derivation of this relationship has been proposed, based on the scaling of the neutral space size as a result of the number of paired and unpaired bases in the sequence. We will return to this later on, in section IX.

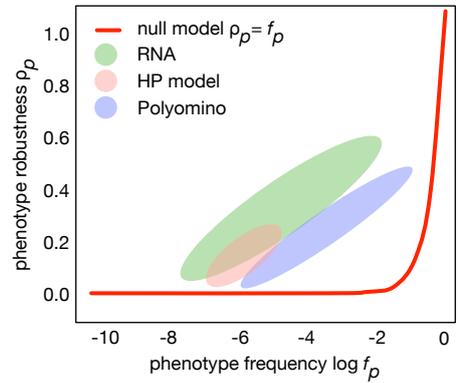


FIG. 3: Schematic illustration of the relationship between phenotypic robustness and phenotypic frequency for the RNA, HP, and Polyomino GP maps. In these maps, robustness scales logarithmically with frequency. If the phenotypes were randomly distributed (according to the biased distribution), we would expect to see the null model $\rho_p = f_p$ (red line). The three shaded areas show the location of the vast majority of values of ρ_p versus f_p for the three maps. For the full data and background, see [20].

D. Evolvability

Robustness is often contrasted with another basic property of phenotypes, evolvability, which attempts to measure the ability of a phenotype to adapt under evolutionary pressure. Formally evolvability is often quantified in terms of the variety of phenotypes that lie within a certain mutation distance of a genotype or phenotype, such as Wagner's definitions of genotypic and phenotypic evolvability [52]. The former counts the number of different phenotypes in the 1-mutation neighbourhood of a genotype, while the latter does the same but for the entire 1-mutation neighbourhood of a phenotype. At the genotypic level, robustness and evolvability are opposed to each other. A given genotype can either be surrounded by neighbours of the same phenotype, making it robust, or surrounded by a variety of other phenotypes, making it evolvable. Comparisons of the genotypic robustness and evolvability therefore consistently show a negative correlation between the two, as observed in regulatory network GP maps, [2], RNA [23, 52] and Polyominoes [23]. This tradeoff is illustrated in Figure 4A. Phenotypic evolvability however is much more complex. The variety of phenotypic neighbours of a given phenotype depends on the structure of the neutral networks of phenotypes in genotype space. Wagner [52] showed that phenotypic robustness and evolvability is positively correlated in the RNA secondary structure GP map - a finding that has also been reproduced in Polyominoes [23]. Phenotypes can therefore benefit from both robustness and evolvability despite the fact that these two properties appear to, on the surface, oppose to each other. Figure 4B illustrates this relationship between the two quantities.

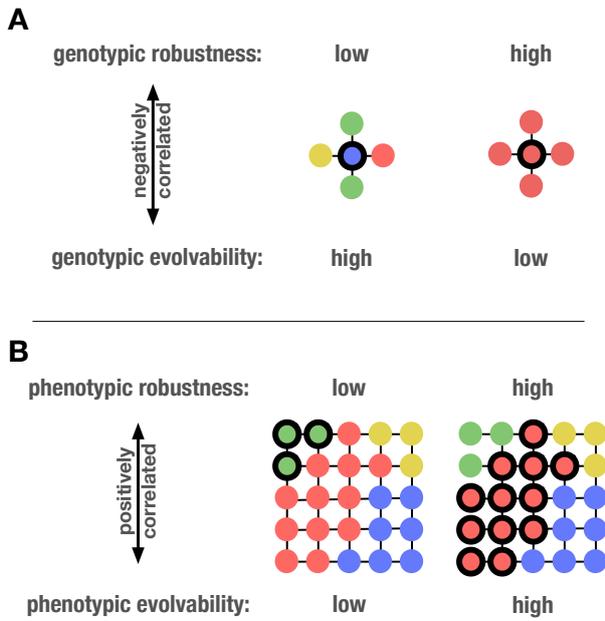


FIG. 4: Illustration of the relationship between robustness and evolvability, which can be defined for a genotype or a phenotype. At the genotypic level (A) a single genotype (highlighted by a black ring) can either be evolvable, meaning that it is a single mutation away from genotypes of many different phenotypes (A, left), or it can be robust, meaning that it is surrounded by genotypes that map to the same phenotype as itself (A, right). A genotype therefore faces a tradeoff between these two quantities, and cannot be evolvable and robust at the same time. By contrast, at the phenotypic level (B) these quantities are positively correlated, which means that a phenotype (shown as the set of genotypes highlighted by black rings) can be both evolvable and robust (B, right). For phenotypes the evolvability is defined as the total number of different phenotypes that lie within the point-mutation neighbourhood of a phenotype, and robustness is defined as the average fraction of genotypic neighbours that leave the phenotype unchanged, taken across all genotypes of that phenotype.

E. Shape space covering

Shape space covering is a way of measuring the accessibility of phenotypes in genotype space, typically by showing the fraction of phenotypes that can, on average, be accessed from a genotype, as a function of the number of mutational steps. In the RNA, HP, and Polyomino GP maps the majority of phenotypes has been shown to lie in close proximity of any genotype [18, 23, 26]. This is because of the high-dimensional nature of GP maps.

IX. SEQUENCE CONSTRAINTS AND GP MAP PROPERTIES

The structural properties discussed above are remarkably consistent across different genotype-phenotype

maps. This suggests that they are in some way fundamental to biological systems. A possible explanation for the universality of these properties is offered in [27], based on the fact that biological sequences are often divided into evolutionarily constrained and (relatively) unconstrained parts. In DNA this distinction comes in the form of exons versus introns, and genes versus intergenic sequences. In RNA this distinction can be made in terms of base pairs (which are more constrained) and loops (less constrained) [54]. In Polyominoes one can distinguish interactive faces and neutral (or effectively neutral) faces. Furthermore the level of constraint of a given sequence element can be changed by mutating the sequence. In DNA a mutation can alter the position of the start and stop codons that determine the boundaries between coding and non-coding DNA. In RNA (or Polyominoes) a mutation can dissolve or establish a base pair (or an interaction between tiles), reducing or increasing the level of constraint at that sequence position. The approach in [27] uses a very simple genotype-phenotype map, in which the binary genotypes are subdivided into a single ‘coding’ and a single ‘non-coding’ part by a simple ‘stop codon’ sequence. Any point mutation of the coding part of the sequence (including the stop codon) changes the phenotype, whereas the non-coding part is entirely free to mutate. This GP map can be used to study the effect of constrained and unconstrained sequences on the GP map structure. It is named the ‘Fibonacci GP map’ because it is closely related to the Fibonacci code in computer science, and because many of its analytical properties can be described in terms of Fibonacci numbers. The Fibonacci GP map exhibits all of the structural genotype-phenotype map properties discussed above - redundancy, bias, the logarithmic scaling of phenotype robustness as a function of phenotype frequency, the negative correlation of genotypic robustness and genotypic evolvability, the positive correlation of phenotypic robustness and phenotypic evolvability, and shape space covering. This implies that the structure of genotype-phenotype maps is heavily influenced by the fact that unconstrained subsequences of the genotypes result in phenotypes that occupy whole subspaces of genotype space, allowing them to form extensive neutral networks that combine robustness and evolvability. The logarithmic scaling relationship between phenotypic robustness and frequency is a particularly strong indicator that unconstrained genotypic sequences play an important role in determining the structure of genotype-phenotype maps. This aspect of the Fibonacci GP map is closely related to the analytical derivation of the same logarithmic relationship in RNA secondary structure [54], which is specifically based on the constraints imposed on RNA sequences by base-pairing interactions. The Fibonacci GP map also shows that the possibility of mutations that change the level of sequence constraint leads to much greater evolvability, and is likely to be one of the reasons why phenotypic robustness and evolvability are positively correlated. It is in fact the analytical term that accounts for mutations of

the stop codon that turns the correlation between phenotypic robustness and evolvability from a negative one into a positive one. This is because mutations of the stop codon shift the boundary between the constrained and unconstrained parts of the sequence, enabling access to a much wider variety of different phenotypes.

X. EVOLUTIONARY IMPLICATIONS OF GP MAP STRUCTURE

While selection remains the driver of phenotypic change, it has become increasingly clear how much the road map - in other words, the genotype-phenotype map - matters. Characteristics such as the skewed distribution of neutral network sizes and their shape in genotype space can strongly determine evolutionary outcomes. On an abstract level this insight is already present in Iwasa's concept of free fitness [55]. But in terms of concrete genotype-phenotype maps evidence has accumulated only more recently, for RNA secondary structure [56] and regulatory network GP maps [57], that the neutral network size strongly determines how likely a phe-

notype is to arise in evolution. In the context of RNA structures it was also shown more recently that fitter phenotypes remain undiscovered if their neutral network is too small [58]. The most striking result in this respect is that the functional RNA secondary structures that have arisen naturally in the course of biological evolution closely mirror the distribution one would expect based on the neutral network size [24]. This provides strong evidence that, rather than being the result of strong selective pressures and evolutionary optimisation, functional non-coding RNAs represent the most accessible solutions to a given evolutionary challenge. More generally speaking this result implies that the accessibility of phenotypes may be just as important a determinant of evolutionary outcomes as the pressures of natural selection.

XI. ACKNOWLEDGMENTS

SEA was supported by The Royal Society and the Gatsby Foundation.

-
- [1] M. Mitchell, *An introduction to genetic algorithms* (MIT Press, 1996).
 - [2] S. Ciliberti, O. C. Martin, and A. Wagner, *PNAS* **104**, 13591 (2007).
 - [3] J. L. Payne and A. Wagner, *PLOS Computational Biology* **9**, (2013).
 - [4] C. M. Dobson, A. Šali, and M. Karplus, *Angewandte Chemie International Edition* (1998).
 - [5] G. Mendel, *Verhandlungen des naturforschenden Vereines in Brunn* **4**, 3 (1866).
 - [6] C. Darwin, *On the Origin of Species* (John Murray, London, 1859).
 - [7] R. A. Fisher, *Transactions of the Royal Society of Edinburgh* **52**, 399 (1919).
 - [8] J. B. S. Haldane, *The causes of evolution* (Princeton University Press, 1990).
 - [9] S. Wright, *Proceedings of the Sixth International Congress on Genetics* pp. 356–366 (1932).
 - [10] M. Kimura, *Nature* **217**, 624 (1968).
 - [11] J. Maynard Smith, *Nature* **225**, 563 (1970).
 - [12] F. B. Salisbury, *Nature* **224**, 342 (1969).
 - [13] R. C. Lewontin et al., *The genetic basis of evolutionary change*, vol. 560 (Columbia University Press New York, 1974).
 - [14] R. R. Schaller, *Spectrum, IEEE* **34**, 52 (1997).
 - [15] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, *Nucleic Acids Research* **37**, D26 (2009).
 - [16] K. A. K. Dill, *Biochemistry (Washington)* **24**, 1501 (1985).
 - [17] I. L. Hofacker, *Nucleic Acids Research* **31**, 3429 (2003).
 - [18] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker, *Proceedings of The Royal Society B - Biological Sciences* **255**, 279 (1994).
 - [19] P. Schuster, *Reports on Progress in Physics* **69**, 1419 (2006).
 - [20] S. F. Greenbury, S. Schaper, S. E. Ahnert, and A. A. Louis, *PLOS Computational Biology* **12**, e1004773 (2016).
 - [21] K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
 - [22] S. E. Ahnert, I. G. Johnston, T. M. A. Fink, J. P. K. Doye, and A. A. Louis, *Physical Review E* **82**, 026117 (2010).
 - [23] S. F. Greenbury, I. G. Johnston, A. A. Louis, and S. E. Ahnert, *Journal of Royal Society Interface* **11**, 20140249 (2014).
 - [24] K. Dingle, S. Schaper, and A. A. Louis, *Interface focus* **5**, 20150053 (2015).
 - [25] D. J. Lipman and W. J. Wilbur, *Proceedings of The Royal Society B - Biological Sciences* **245**, 7 (1991).
 - [26] E. Ferrada and A. Wagner, *Biophysical Journal* **102**, 1916 (2012).
 - [27] S. F. Greenbury and S. E. Ahnert, *Journal of The Royal Society Interface* **12**, 20150724 (2015).
 - [28] H. Li, R. Helling, C. Tang, and N. Wingreen, *Science* **273**, 666 (1996).
 - [29] R. A. Goldstein, *Current Opinion in Structural Biology* **18**, 170 (2008).
 - [30] E. D. Levy, J. B. Pereira-Leal, C. Chothia, and S. A. Teichmann, *PLOS Computational Biology* **2**, e155 (2006).
 - [31] S. E. Ahnert, J. A. Marsh, H. Hernandez, C. V. Robinson, and S. A. Teichmann, *Science* **350**, aaa2245 (2015).
 - [32] I. G. Johnston, S. E. Ahnert, J. P. K. Doye, and A. A. Louis, *Physical Review E* **83**, 066105 (2011).
 - [33] R. D. Schram and H. Schiessel, *Journal of Physics A: Mathematical and Theoretical* **46**, 485001 (2013).
 - [34] S. Ciliberti, O. C. Martin, and A. Wagner, *PLOS Com-*

- putational Biology **3**, e15 (2007).
- [35] J. L. Payne, J. H. Moore, and A. Wagner, *Artificial Life* **20**, 111 (2014).
- [36] J. L. Payne and A. Wagner, *BMC Systems Biology* **8**, 64 (2014).
- [37] J. L. Payne and A. Wagner, *Scientific reports* pp. 1–10 (2015).
- [38] E. Ibáñez-Marcelo and T. Alarcón, *Journal of Theoretical Biology* **356**, 144 (2014).
- [39] D. Watts and S. Strogatz, *Nature* **393**, 440 (1998).
- [40] A. Barabasi and R. Albert, *Science* **286**, 509 (1999).
- [41] M. Newman, *Networks: An Introduction* (OUP Oxford, 2009).
- [42] J. F. Matias Rodrigues and A. Wagner, *PLOS Computational Biology* **5**, e1000613 (2009).
- [43] J. D. Orth, I. Thiele, and B. Ø. Palsson, *Nature biotechnology* **28**, 245 (2010).
- [44] K. Raman and A. Wagner, *Molecular Biosystems* **7**, 1081 (2011).
- [45] C. F. Arias, P. Catalán, S. Manrubia, and J. A. Cuesta, *Scientific reports* **4**, 7549 (2014).
- [46] E. Ferrada and A. Wagner, *PLOS ONE* **5**, (2010).
- [47] J. L. Payne and A. Wagner, *Science* **343**, 875 (2014).
- [48] A. Bairoch, *Nucleic acids research* **28**, 304 (2000).
- [49] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, et al., *Science* **324**, 1720 (2009).
- [50] M. T. Weirauch and T. R. Hughes, *Trends In Genetics* **26**, 66 (2010).
- [51] E. E. Hare, B. K. Peterson, V. N. Iyer, R. Meier, and M. B. Eisen, *PLOS Genetics* **4**, e1000106 (2008).
- [52] A. Wagner, *Proceedings of The Royal Society B - Biological Sciences* **275**, 91 (2008).
- [53] T. Jörg, O. C. Martin, and A. Wagner, *BMC bioinformatics* **9**, 464 (2008).
- [54] J. Aguirre, J. M. Buldú, M. Stich, and S. C. Manrubia, *PLOS ONE* **6**, e26324 (2011).
- [55] Y. Iwasa, *Journal of Theoretical Biology* **135**, 265 (1988).
- [56] M. C. Cowperthwaite, E. P. Economo, W. R. Harcombe, E. L. Miller, and L. A. Meyers, *PLOS Computational Biology* **4**, e1000110 (2008).
- [57] E. Borenstein and D. C. Krakauer, *PLOS Computational Biology* **4**, e1000202 (2008).
- [58] S. Schaper and A. A. Louis, *PLOS ONE* **9**, e86635 (2014).