THE UNIVERSITY OF CHICAGO

BIASED FORMS MOST BEAUTIFUL: THE STRUCTURE OF A MOLECULAR GENOTYPE-PHENOTYPE MAP AND ITS INFLUENCE ON PHENOTYPIC DIVERSITY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES

AND THE PRITZKER SCHOOL OF MEDICINE

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ECOLOGY AND EVOLUTION

BY

SANTIAGO HERRERA-ÁLVAREZ

CHICAGO, ILLINOIS

MARCH 2025

Dedicatoria

A mis padres y mi hermana, quienes con su amor y apoyo incondicional me impulsaron a perseguir mi sueño de ser un científico, alimentaron mi curiosidad y asombro por la naturaleza, y me enseñaron que la gentileza y rigurosidad científica pueden ir de la mano.

Dedication

To my parents and my sister, who with their unconditional love and support pushed me to pursue my dream of becoming a scientist, nurtured my curiosity and wonder for nature, and taught me that kindness and scientific rigor can go hand in hand.

Table of Contents

List of Figures

List of Tables

Acknowledgements

This work represents the culmination of one of the most transformative experiences of my life. I can hardly express how much I have grown as a person and as a scientist in the past 6.5 years—the scientist in me has learned to embrace kindness and empathy, while my inner child has been taught to follow his passion and curiosity with scientific rigor. These experiences, which are now a part of me, were possible due to the support and warmth of my lab. The Thornton lab felt like home from the beginning: I still remember everyone's friendliness when I first stepped into the lab for my rotation and I was struck by everyone's eagerness to help and give feedback on everyone's project, and, most importantly, by everyone's profound human values. For this, I also want to thank Joe Thornton, my advisor. Much of our lab culture reflects Joe's conviction that science is a social endeavor—that the way we do our science is a direct reflection of our values as a society. His scientific rigor has also profoundly shaped my view and approach to science—as every good craftmanship, I learned that doing science requires lots of patience, persistence and self-criticism. Finally, his emphasis on writing made me also realize the profound impact of story-telling—that every scientific discovery is a story awaiting to be written. Because of this, I am better scientist.

A PhD is hardly a one-person's effort and mine is definitely not the exception. I am especially grateful to Jaeda Patton with whom I developed a close and long-lasting collaboration. The huge scientific project we undertook together felt like an adventure, and I am grateful that our interests, knowledge, skills and ideas were always a complement to each other's—I learned a lot from this process and from her, and I believe that working together made this project better and more fun. I am also grateful to all past and present members of the lab, each of which has taught me something special and who also became my friends. Thank you for your friendship,

your thoughtful discussions and all your deep feedback—I will always be inspired by your talent and grateful for your advice. I also want to thank the rest of the members in my thesis committee—Marcus Kronforst, John Novembre and Dave Jablonski—who inspired me and pushed me to be a better scientist.

I would also like to thank my friends in Colombia and the friends I made in Chicago. Being surrounded by these people during all these years has been a gift. I am grateful to all my Latin American friends and, in particular, to my small Colombian family in Chicago. Sharing coffee, conversations, and laughter in Spanish—having a little piece of home here—has made this journey all the more special.

Finally, I want to thank my parents and my sister. Nothing I write here will ever make justice to how grateful I am to them. I could not be here writing these words—feeling a somewhat selfish sense of accomplishment—if it were not for them. Thank you for your continuous and unconditional love and support; they have been my fuel throughout these years. Thank you for never let me lose sight of the big picture and of what really matters in life: happiness, kindness and passion. You are my deepest source of gratitude and inspiration and I hope to honor you as I continue pursuing my dreams and passion for science.

Chapter 1

Introduction

1.1 The tension between production and sorting of variation

Evolutionary change follows a two-step process: the production of variation through mutations and the subsequent sorting of this variation by selection and drift. Mutations generate phenotypic variation when genetic changes are translated into phenotypic variants via biological mechanisms such as biochemistry, cell biology, development, and physiology. Once these phenotypic variants arise, they are subject to selection and drift, determining whether they persist or are lost within populations based on fitness and other population-level dynamics. At the dawn of the 20th century, however, the Modern Synthesis reshaped the study of evolutionary change by emphasizing the sorting of variation while de-emphasizing its production (1, 2). The rediscovery of Mendel's laws of inheritance positioned genes and their alleles as the central focus of evolutionary biology, granting them the status of being both necessary and sufficient to explain phenotypic change (1, 3). This shift effectively excluded the biological mechanisms underlying phenotype production from the explanatory framework of evolutionary theory. Consequently, evolutionary change became synonymous with changes in allele frequencies, with the assumption that populations already possessed abundant genetic and phenotypic variation, so a new evolutionary theory needed to explain how that variation was sorted, not how it was produced (1–3).

The emphasis on the sorting of variation in the new evolutionary theory significantly influenced the research agenda in evolutionary biology but left major biological phenomena unexplained. For instance, it offered no causal explanation for homology—the shared structural

features in related lineages—or for covariation, where phenotypic changes in one trait are often accompanied by correlated changes in others (1, 4, 5). Similarly, the sparseness and structure of phenotypic variation in nature—the observation that many conceivable phenotypes never materialize, and existing ones are confined to specific lineages—remained unaddressed (1, 6). Changes in allele frequencies alone proved inadequate to account for these phenomena.

Advances in molecular and developmental biology during the latter half of the 20th century prompted the reintegration of the production of variation into the explanatory framework of evolutionary theory (2, 7–9). This led to the emergence of a central heuristic object in evolutionary biology: the genotype-phenotype (GP) map (10–12). By mapping all possible phenotypes onto all possible genotypes connected in a network of mutations, the GP map captures the capacity of a biological system to produce and access phenotypic variation and provides a mechanistic framework for examining the role of the production of variation in evolutionary outcomes. Mutations transform one genotype into another, and the GP map determines the phenotypic consequences of these changes. If biological mechanisms favor the generation of certain phenotypic variants over others—or completely preclude the production of some—then the production of variation could exert a profound influence on the trajectory of evolution (2, 13–16).

Despite the conceptual significance of the GP map, the intellectual inertia of the Modern Synthesis created a persistent tension between the roles of selection and the GP map as causal factors in shaping evolution and phenotypic diversity. At one extreme, phenotypic evolution is attributed solely to adaptation through selection; at the other, it is viewed as predominantly driven by production biases. In reality, phenotypic evolution is almost certainly the result of both processes (2, 17–19), but disentangling their relative roles remains fundamentally unsolved.

1.2 The causal role of the GP map in evolutionary biology

A key assumption for selection to be the dominant force in evolution is that any genotype should, in principle, be capable of producing any phenotype through mutation. The Modern Synthesis implicitly ascribed a specific structure to the GP map: to enable selection to act without constraint, the GP map must be both isotropic and homogeneous (20, 21). Isotropy implies that mutations affect all phenotypes equally, with no inherent bias favoring the production of certain phenotypic variants over others. Homogeneity suggests that all genotypes produce an identical distribution of phenotypic variation through mutation. Under this structure, the GP map exerts no influence on evolutionary outcomes. However, an anisotropic GP map would preferentially generate certain phenotypes, making them more likely to evolve on average. Similarly, a heterogeneous GP map would result in genotypes differing in their mutational propensities, leading to lineage-specific evolutionary biases. Understanding the structure of the GP map is therefore critical for disentangling its causal influence from that of selection in shaping phenotypic diversity.

Production bias strongly suggests that the GP map can, in principle, influence the outcomes of evolution. Evidence from diverse developmental, morphological and molecular phenotypes has shown that biological systems do not produce phenotypic variants uniformly (22–24), and that these biases are often congruent with the patterns of variation observed in nature (24–26). However, there have been three important limitations to establish whether the GP map has caused patterns of phenotypic variation in nature. First, a correlation between production bias and observed variation is necessary but not sufficient to establish causality. Second, we lack a clear understanding of how different structural features of the GP map would

cause evolutionary outcomes as no study to date has examined biases in the production and access of variation simultaneously in the same system. Third, production biases in present-day lineages may not reflect the biases that existed when the phenotypes evolved.

My thesis research addresses three key questions on the nature and causal role of the GP map in evolution: 1) What is the extent of anisotropy and heterogeneity of an empirical GP map? 2) How did the structure of an ancient GP map influence the phenotypic diversity of present-day lineages? And 3) What are the genetic determinants of the GP map's structure? Answering these questions is fundamental to understanding the causal role of the GP map in evolution. However, despite numerous studies addressing some of these questions, we still lack clear answers to them. Below, I explain why these knowledge gaps persist and lay out the approaches I developed to address these questions.

1.3 The structure of empirical GP maps

Characterizing the structure of the GP map—the extent of anisotropy and heterogeneity—requires measuring the phenotypes of all possible genotypes, a task that is nearly impossible due to the astronomical size of genotype spaces. For instance, a protein of 100 amino acids has 20^{100} possible genotypes, equivalent to approximately 1.26×10^{130}. This number vastly exceeds the number of atoms in the universe, making exhaustive characterization unfeasible. Additionally, even if the genotype space were experimentally tractable, the number of possible phenotypes to evaluate is virtually infinite.

Characterizing GP maps therefore inherently requires reducing the problem's dimensionality by limiting the size of the genotype space, the phenotype space, or both. Studies of GP maps can be grouped into four broad categories. First, morphospaces are quantitative

representations of phenotypic variation across multiple trait dimensions and reveal how lineages occupy the phenotype space—most regions are unoccupied and lineages are clumped (27–30). However, morphospaces are not strict GP maps because they completely lack an explicit genotype space and therefore can only indirectly suggest production biases without excluding selection. Second, organismal GP maps examine a small subset of the genotype space—often just a few genotypes—and map their developmental effects onto a single phenotype (31–33). While they capture production mechanisms, they cannot reveal production biases because studying biases requires comparing the relative production of different phenotypes. Third, molecular GP maps investigate a larger portion of the genotype space but similarly focus on a single biochemical phenotype (34–36). These maps illustrate the mutational accessibility of genotypes producing the same phenotype but do not reveal production biases. Lastly, computational GP maps explore broader genotype and phenotype spaces, suggesting production biases (22, 23), but they only exist for a few systems and their generalizability remains an open question. Overall, the structure of the GP map—and its role in causing evolutionary outcomes—remains poorly understood, as no biological system has been comprehensively mapped from genotypes to phenotypes within a defined scope.

I address this knowledge gap in the first chapter of my thesis by experimentally characterizing the structure of a complete molecular GP map. I combine multi-phenotype deep mutational scanning with ancestral sequence reconstruction to comprehensively chart the DNA specificity of every possible genotype at the DNA-binding interface of a transcription factor and directly assess how this ancestral GP map shaped the functional diversification of the protein family. I did these experiments and analyses in close collaboration with Jaeda Patton, whom will be co-first author in the publication arising from this chapter.

1.4 The genetic architecture of the GP map's structure

The structure of the GP map is a direct consequence of the association between genotypes and phenotypes—the number and distribution of phenotypes encoded across the connected network of genotypes. This association is in turn determined by the GP map's genetic architecture—the set of causal rules by which individual genetic states contribute to a phenotype, as well as the epistatic contribution of every possible pair of states and higher-order combination. By determining how genotypes encode phenotypes, the genetic architecture directly shapes the extent of anisotropy and heterogeneity in the map.

However, we still lack a clear understanding of the genetic determinants of the GP map's structure for two reasons. First, no biological system has been fully characterized with a comprehensive GP map, which prevents us from uncovering its structure. Even for partial GP maps, the extent of anisotropy and heterogeneity has never been assessed simultaneously within the same system. Second, traditional models of genotype-to-phenotype relationships often rely on simplifying assumptions about the genetic architecture. These models typically assign phenotypes to genotypes at random (37) and treat epistasis as noise rather than as a meaningful mutational effect (38, 39). GP maps simulated under these models provide little information about the mechanistic links between genetic architecture and the structure of the GP map because many different architectures produce the same pattern of bias (40–42). Uncovering the underlying genetic causes of the GP map's structure is critical to understanding when and how would the GP map affect evolutionary outcomes.

I address this knowledge gap in the second chapter of my thesis by directly decomposing the genetic architecture of a complete experimental GP map of a transcription factor. I use a

statistical framework of genetic effects to evaluate how different genetic architectures, varying in the extent of epistasis, simultaneously shape the extent of anisotropy and heterogeneity in the GP map, and show how different structures affect evolutionary outcomes in predictable ways.

1.5 The hierarchical structure of the GP map

Besides their astronomical size, GP maps represent a series of complex transformations across multiple levels of biological organization. An ideal GP map would capture the entire chain of biological processes, mapping the phenotype of every genotype at molecular, cellular, and organismal levels. While constructing such a comprehensive map is unfeasible, the available data from molecular, developmental, and organismal systems provide a foundation for understanding how phenotypic variation is transformed across levels. In the final chapter of my thesis, I propose a perspective on integrating fundamental concepts and experimental approaches from evolutionary biochemistry and developmental evolution into a broader and more comprehensive research framework. I argue that all biological systems share fundamental ways in which they can change, revealing general principles of how phenotypic variation arises and propagates across all levels of biological organization. Furthermore, I identify specific questions where combining experimental approaches from both fields could move us forward towards building multi-scale GP maps.

Chapter 2

Bias in an ancient genotype-phenotype map caused the functional diversity of the steroid hormone receptor family

2.1 Summary

Biological systems may be biased in the phenotypes they can access by mutation, but the extent of these biases and their causal role in the evolution of phenotypic diversity remains unclear. There are three major challenges: it is difficult to isolate the effect of bias in the genotype-phenotype (GP) map from that of natural selection in producing natural diversity, the map is so vast and complex that a direct characterization has been impossible, and most extant phenotypes evolved long ago in species whose GP maps cannot be recovered. Here we use multi-phenotype deep mutational scanning to experimentally characterize the complete GP maps of two reconstructed ancestral steroid receptor proteins from an ancient phylogenetic interval during which a new phenotype—specific binding of a new DNA response element—evolved. We measured all possible DNA specificity phenotypes encoded by all possible amino acid combinations at sites in the protein's DNA binding interface. We found that the ancestral GP maps are structured by very strong global and local biases—unequal propensity to encode the different phenotypes and extreme heterogeneity in the phenotypes accessible around each genotype. Distinct biases in each ancestral map steered evolution toward the lineage-specific phenotypic outcomes that occurred during history. Our findings establish that ancient biases in the GP relationship were causal factors in the historical evolution of a protein family and shaped phenotypic diversity among present-day descendant lineages.

2.2 Introduction

Countless conceivable lifeforms have evolved rarely or never, and those that exist are mostly restricted to specific lineages (27–29, 43). No flying vertebrates have two pairs of wings, for example, and no turtles or frogs fly. What explains the biased distribution of phenotypes in nature? Classical explanations focus on the influence of selection (44, 45), but it is possible that the propensities of biological systems to produce phenotypic variation could also shape evolutionary outcomes. A phenotype can become fixed in an evolving population only if it is first generated by mutation. If biological systems are more likely to produce some phenotypes than others (9, 11, 13–16, 46), and if these propensities change over time as lineages diverge (47, 48), then some phenotypes will be more likely to evolve in some taxa than in others.

The importance of phenotype production as a cause of evolutionary outcomes is unclear, because it is difficult to disentangle its influence from that of selection in producing patterns of variation observed in nature (17, 18, 46, 49, 50). Ideally, we would isolate the phenotype production process by directly characterizing the complete genotype-phenotype (GP) map, which maps all possible combinations of mutations to the phenotypes they encode. Although the space of genotypes and phenotypes is vast, we reasoned that recent technical advances make this goal tractable for proteins and their biochemical phenotypes. Deep mutational scanning (DMS) allows huge libraries of protein variants to be characterized experimentally (51). The scope of genetic variation to be measured for a protein's GP map can be defined as all combinations of all possible amino acid states at the sequence sites that determine the phenotype of interest (34, 35). The scope of phenotypic variation mapped should also be comprehensive, because understanding why evolution turned out as it did requires knowledge of the capacity to produce not only the phenotypes that evolved historically but also those that did not. Although most DMS studies

have addressed only one or a few phenotypes that exist in extant proteins (36), a complete set of possible phenotypes can in principle be assessed in a multi-phenotype DMS by measuring, for example, binding or catalysis of all possible substrates in a biologically relevant class (52–54). By mapping all possible phenotypes onto all possible genotypes connected in a network of all possible mutations, the total capacity of the system to produce and access phenotypic variation could be characterized.

The phenotypes of extant lineages evolved long ago, so we would ideally characterize the GP maps that existed during history. This goal can be also accomplished for protein systems by performing DMS on reconstructed ancestral proteins (55) in a phylogenetic time series (56, 57). These ancestral GP maps would reveal how biases imposed by the phenotype production process may have changed over time and whether these biases are congruent with the historical trajectories of phenotypic evolution. Here, we apply this approach to assess how phenotype production shaped the functional diversification of the steroid hormone receptor protein family. We use multi-phenotype DMS to experimentally characterize GP maps of the binding interface of two reconstructed ancestral steroid hormone receptor DNA binding domains (SR DBDs), assess how the maps may have shaped the historical diversification of SR specificity for DNA response elements, and understand the mechanisms that changed key features of the maps across evolutionary time.

2.3 Results

2.3.1 Two complete ancestral GP maps

SRs are a family of transcription factors that regulate physiological and reproductive biology in bilaterian animals. Most bilaterian taxa have a single SR, which specifically binds to inverted

palindromes of the motif AGGTCA, called the estrogen response element (ERE; Fig. 2.1a). In chordates, a gene duplication of the ancestral SR (AncSR1) produced two major SR classes, which have different DNA specificity phenotypes: chordate estrogen receptors (ERs) retain the ancestral ERE specificity, but a novel specificity for a palindrome of AGAACA, called the steroid response element (SRE), evolved in the lineage leading to AncSR2, the common ancestor of the chordate ketosteroid receptors (kSRs; Fig. 2.1a, b) (58). Specificity for DNA is determined primarily by the amino acid sequence of a recognition helix (RH) that binds in the DNA major groove (59, 60). AncSR1 and AncSR2 DBDs differ by 34 amino acid replacements, but experiments on the reconstructed proteins established that three amino acid changes in the RH were the primary cause of the evolution of SRE specificity (58).

To understand how phenotype production may have shaped the evolution of SR-DBD specificity, we characterized combinatorially complete GP maps of the DBD-response element (RE) interface at the key ancestral timepoints AncSR1 and AncSR2. The scope of genotypes is all possible $20^4 = 160{,}000$ amino acid variants at four variable sites in the recognition helix—the three that changed between AncSR1 and AncSR2, plus one other that varies in the broader nuclear receptor family (Fig. 2.1c). The scope of specificity phenotypes consists of all $4^2 = 16$ possible RE sequences that can be produced by all combinations of nucleotides at the two base positions that vary between ERE and SRE. These two maps of the recognition helix-RE interface can be thought of as submaps within the much larger GP map of the entire DBD, which are connected by the 31 other "background" substitutions that occurred between the AncSR1 and AncSR2 proteins (Fig. 2.1b).

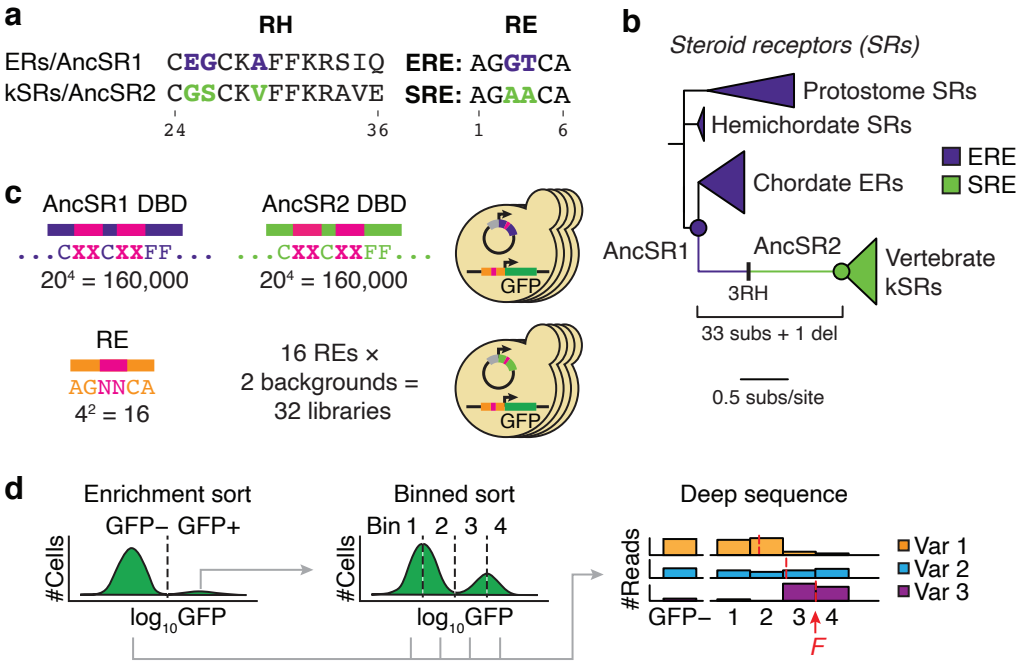


Figure 2.1. Characterizing ancestral GP maps using multi-phenotype DMS. a, Amino acid sequence of the recognition helix (RH) in extant and ancestral steroid receptor (SR) proteins and the sequence of the RE they bind to. Colored residues are responsible for differences in protein-RE specificity. **b**, Phylogeny of SRs. Each clade of proteins is colored by the RE sequence it recognizes. In chordates, a historical transition from ERE to SRE specificity occurred along the branch between AncSR1 (the common ancestor of all chordate SRs) and AncSR2 (the common ancestor of vertebrate kSRs). The number of historical sequence changes along the AncSR1-AncSR2 branch is shown; three of these in the recognition helix (RH) caused the specificity switch (58). **c**, **d**, DMS experiment to assay effects of RH genotype on binding to variable REs. **c**, We built combinatorial libraries of all combinations of 20 amino acid states at four variable sites in the RH (pink Xs), using the rest of the AncSR1 and AncSR2 DBDs as backgrounds (top left). These were transformed into 16 *S. cerevisiae* strains, each containing one of the 16 possible RE motifs (pink Ns, bottom left) genomically integrated upstream of a GFP reporter gene (right). **d**, We assayed binding of DBD-RE complexes using FACS coupled with deep sequencing. For each library, we performed an initial enrichment sort to select for GFP+ cells. We then grew up the selected cells, pooled them across the 32 libraries, and resorted them into four fluorescence bins in triplicate (binned sort). Sorted cells were deep sequenced to estimate the mean $\log_{10}$GFP (F) of each combination of protein and RE genotypes.

We engineered two protein libraries, each containing all 160,000 variants of the recognition helix in the background of either the AncSR1 or AncSR2 DBD, along with16 yeast strains, each containing a GFP reporter driven by one of the REs (Fig. 2.1c, Fig. A1.1a–e). We transformed each RE strain separately with the two protein libraries, with barcodes to mark the

strain and the ancestral background, for a total of 5.12 million protein-DNA complexes. We used an initial round of fluorescence-activated cell sorting to enrich the yeast libraries for GFP-positive cells, pooled the enriched libraries, sorted cells in three replicates by their fluorescence, and sequenced the sorted bins (Fig. 2.1d, Fig. A1.1f, g). Using this strategy, we obtained empirical fluorescence estimates for the majority of complexes with good replicability ($r^2 = 0.92$ across replicates, excluding complexes at the lower bound of fluorescence; Fig. A1.2). Fluorescence of the remaining complexes was predicted using a generalized linear model trained on the experimental data (Fig. A1.3a–d) (61, 62).

Each protein variant was assigned a DNA specificity phenotype based on these experiments. A protein variant is classified as specific if it is functional in complex with only one RE, promiscuous if it is functional on multiple REs, or nonfunctional if it is not functional on any RE. We defined functional complexes as those having fluorescence at least as great as the wild-type complex in each background (*i.e.* EGKA:ERE for the AncSR1 library and GSKV:SRE for AncSR2) (Fig. A1.3e–g).

2.3.2 Global bias in the AncSR1 GP map

The probability that a phenotype will evolve equals the probability that it will be produced by mutation times the probability that, once produced, it will be fixed. The GP map would have no effect on evolutionary outcomes if and only if it had two properties: isotropy—encoding all phenotypes with equal probability—and homogeneity—producing the same distribution of phenotypes from all starting genotypes in the map (20, 21, 63, 64). If the map is anisotropic, then phenotypes more likely to be produced would be more likely to evolve; if the map is heterogeneous, then the probability that each phenotype will be produced—and hence evolve—

would change as lineages diverge from each other across the map.

We assessed the isotropy of the AncSR1 GP map by characterizing the frequency distribution of DNA specificity phenotypes encoded by all functional protein variants. Only 107 out of 160,000 total genotypes in the library were functional (0.07%). Of these, the majority (91) were specific for a single RE. We calculated the bias (*B)* of this global phenotype distribution, defined as 1 minus the Shannon entropy (base 16); *B* can range from 0 when specificity for all 16 REs is encoded with equal frequency to 1 when only a single phenotype is encoded. We found that the distribution is strongly anisotropic ($B = 0.42$). Two specificity phenotypes—ERE and SRE—together account for >60% of all specific genotypes, and only five others can be produced at all; nine phenotypes are not encoded by any protein variant (Fig. 2.2a).

We refer to this anisotropy as global bias in the GP map (64). Global bias in the AncSR1 map imposes hard limits on phenotypic evolution—the majority of conceivable phenotypes could never evolve in this map, even if they conferred strong fitness advantages. The global bias is also congruent with evolutionary history—the phenotypes that evolved historically in the two lineages descending from AncSR1 are also the most frequently encoded.

2.3.3 Local bias in the AncSR1 GP map

We next assessed the homogeneity of the AncSR1 GP map using Maynard-Smith's classic network model of sequence space (65). Each functional protein variant is a node with its experimentally defined phenotype. Nodes are connected by edges if their amino acid sequences can be interconverted by a single nucleotide change given the standard genetic code. Nonfunctional variants are excluded from the network, based on the assumption that they will be removed quickly from evolving populations by purifying selection.

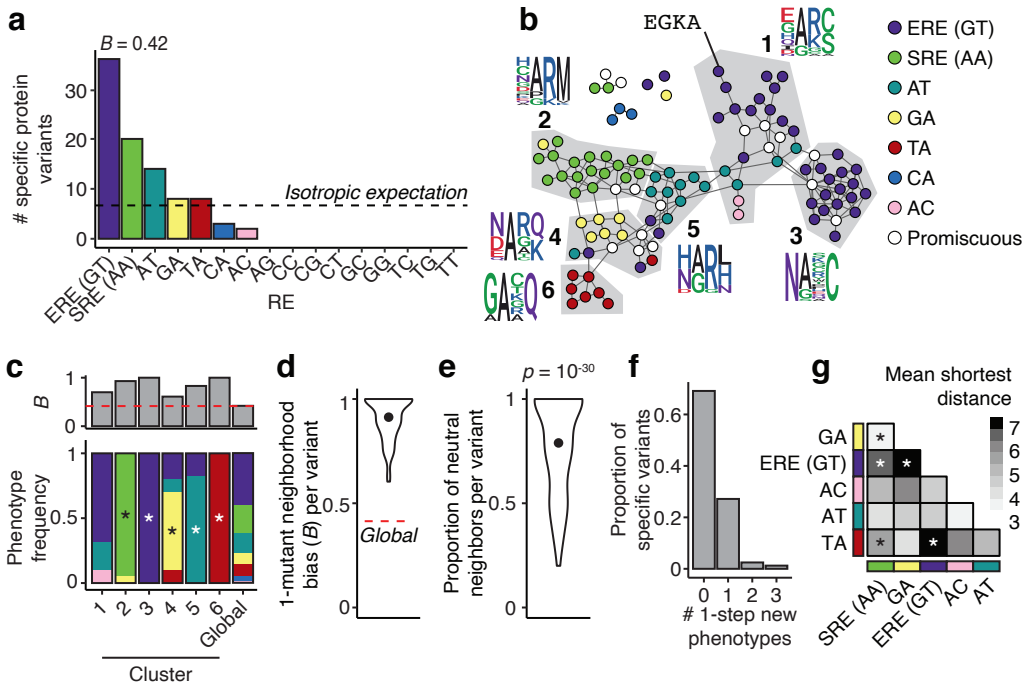


Figure 2.2. Global and local bias in the AncSR1 GP map. **a**, Global production distribution in the AncSR1 GP map. Bars represent the number of protein variants that bind specifically to each RE. The dashed line shows the expected frequencies if the distribution (base 16) of the distribution. **b**, Sequence space network of the AncSR1 GP map. Nodes were unbiased. *B,* phenotype bias, calculated as one minus the entropy represent functional protein variants, colored by their RE specificity; white nodes, promiscuous genotypes. Edges connect protein variants that can be interconverted by a single nucleotide change. Genotype clusters (1–6, ordered by decreasing size) identified by a community structure detection algorithm are shown in gray. Sequence logos show amino acid frequencies at the variable RH sites in each cluster. **c**, Bottom: Frequencies of specificity phenotypes within each genotype cluster; the global production distribution is shown for comparison. Asterisks, phenotypes significantly enriched within a cluster relative to the global production distribution (Fisher's exact test, $p < 0.05$ after Bonferroni correction). Top: strength of phenotype bias (B) in each cluster. Red line, B of global production distribution. **d**, Distribution of phenotype bias (B) of the 1-mutant neighborhood of every RE-specific protein variant in the main network component. Dot shows the mean. Dashed red line, global phenotype bias. **e**, Proportion of neutral neighbors per RE-specific protein variant in the main component of the AncSR1 map. Dot shows the mean. *P-value,* probability that the mean would be at least as great as observed if phenotypes were randomly reassigned in the main component ($n = 91$). **f**, Distribution of the number of new phenotypes accessible within one mutation, across all RE-specific variants in the AncSR1 main component. **g**, Mean distance between pairs of phenotypes in the AncSR1 main component. The color of each cell shows the mean of the length of the most direct path from every genotype encoding one phenotype to every genotype encoding the other. Bonferroni corrected p-values for a two-sided permutation test where phenotype associations were shuffled within the main component: * $p < 0.001$.

We found that the distribution of phenotypes in AncSR1 sequence space is strongly

heterogeneous. Although the majority of functional genotypes (91%) and phenotypes (6 of 7) are mutually connected in a single main network component, each phenotype tends to be sequestered in a local region (Fig. 2.2b). Using a community structure detection algorithm (66), we found that the main network component can be partitioned into six clusters of genotypes that have dense connectivity within clusters and weak connectivity between (Fig. 2.2b). The phenotype bias *B* within every single cluster is higher than the global bias of the map, and 5 of 6 clusters are significantly enriched for a single specificity phenotype, which differ among all 5 clusters (Fig. 2.2c). The clumpy distribution of phenotypes in sequence space arises from the simple fact that similar genotypes, which are connected to each other in sequence space, are likely to encode similar phenotypes (Fig. 2.2b, logos).

This heterogeneity creates local bias (64): the propensity to produce phenotypes depends strongly on the particular genotype occupied at the time. The one-mutant neighborhood around every genotype has extremely high bias (mean $B = 0.91$; Fig. 2.2d), indicating that individual genotypes can access much less phenotypic variation than is encoded across genotype space as a whole. Most mutations are phenotypically neutral (79% of edges; Fig. 2.2e), and most genotypes can directly access at most one new phenotype (Fig. 2.2f). The historical starting genotype (EGKA), for example, has access to only one functional neighbor, which also has ERE specificity. Another consequence of heterogeneity is that phenotypes, aggregated over the genotypes that encode them, are differentially accessible to each other, with substantial variation in the number of mutations required to transform each phenotype into the others (Fig. 2.2g). For example, SRE-specific protein genotypes are directly accessible from nodes encoding specificity for AT and GA, but they are multiple substitutions away from all ERE-specific genotypes (Fig. 2.2b).

The GP map of AncSR1 is therefore both anisotropic and heterogeneous, and these properties impose global and local biases on the production of phenotypes. Global bias favors production of the historical phenotypes ERE and SRE and entirely prevents the production of most conceivable phenotypes. Local bias further restricts the number of accessible phenotypes from each particular genotype, favoring conservation over the evolution of new phenotypes, including from the historical genotype EGKA.

2.3.4 Biases in the GP map affect evolutionary outcomes

To characterize the potential influence of the AncSR1 GP map on the outcomes of evolution, we modeled evolution on the network of functional amino acid genotypes as a discrete-time Markov chain from every possible starting genotype given a variable trajectory length. Each time-step in a trajectory is an amino acid substitution, the probability of which is weighted by the number of single-nucleotide mutations that can mediate it; the relative probability of evolving a given phenotype at the end of the trajectory is the sum of the probabilities of evolving all genotypes that encode it. This model, which corresponds to neutral molecular evolution in which all functional genotypes have equal fitness (65, 67), represents a null scenario: the fixation process imposes no biases on evolutionary outcomes except to prevent the loss of function via purifying selection, thus allowing us to isolate the influence of biases imposed by the GP map on evolutionary outcomes.

We first computed the equilibrium distribution of phenotypic outcomes after an infinite number of substitutions. This represents the limiting case at which the distribution of outcomes is insensitive to the starting genotype and does not change with additional substitutions. The equilibrium outcome distribution is well correlated with the global production distribution (Fig.

2.3a, $r^2 = 0.82$), reflecting the constraints imposed by the global production bias. However, there are differences: the equilibrium distribution is more biased (B= 0.46), and whereas ERE and SRE specificity are the two most frequently encoded phenotypes, ERE and AT specificity are the most likely equilibrium outcomes (Fig. 2.3a). This difference arises because most AT-specific genotypes are located centrally within the network, while SRE-specific genotypes are in a more peripheral cluster (Fig. 2.2b) and are therefore less likely to be occupied. The heterogeneous connectivity of the GP network and global production bias therefore affect evolutionary outcomes, even over infinitely long timescales.

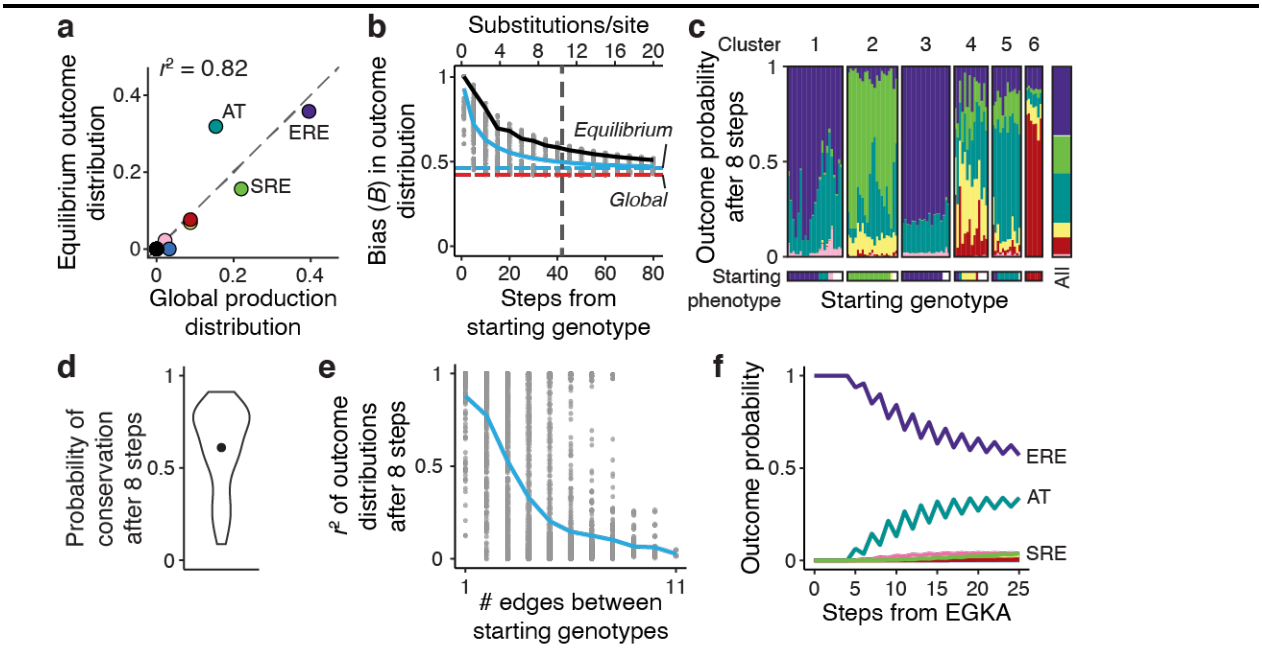


Figure 2.3. The AncSR1 GP map biases evolutionary outcomes towards phenotype conservation. a, Comparison between the global production distribution and the long-term equilibrium distribution of phenotypic outcomes in the AncSR1 main network. Each dot shows the frequency of one specificity phenotype in the two distributions. Black dot at the origin represents nine phenotypes not encoded in the map. Dashed gray line, $y = x$. Squared Pearson's correlation coefficient is shown. **b**, Strength of bias (B) in evolutionary outcomes as a function of the length of evolutionary trajectories. Each gray dot shows the B of the outcome distribution for trajectories of a given number of substitutions starting from one node on the main network component. Solid blue and black lines show the mean across all starting genotypes and from EGKA, respectively. Dashed horizontal red and cyan lines show B of the global production distribution and the equilibrium distribution, respectively. Vertical dashed line shows the number of substitutions required for mean B to reach within 0.05 units of the equilibrium value. The

secondary *x*-axis (above) shows the trajectory length as substitutions per site. **c**, Distribution of evolutionary outcomes after 8 substitution steps from every starting genotype in the AncSR1 main network component, organized by the cluster of the starting genotype (top). Bottom bar shows the phenotype of each starting genotype. Bars at right show the average outcome distribution for all starting genotypes. **d**, Distribution of the probability of phenotype conservation after 8 substitution steps across all specific starting genotypes in the AncSR1 main network component. Dot shows the mean. **e**, Evolutionary outcomes become less similar as starting genotypes diverge from each other. Each dot shows the similarity of the distributions of phenotypic outcomes (Pearson's r^2) of 8-step trajectories starting from a pair of genotypes, versus the number of network edges between the pair. Blue line, mean similarity across all pairs of starting genotypes. **f**, Probability of evolving each specificity phenotype starting from EGKA as a function of the number of substitutions.

On finite timescales, local bias strongly affects evolutionary outcomes. After 3 substitutions, for example—the shortest path between the historical ancestral and derived genotypes—the outcome distributions are very strongly biased (mean $B = 0.8$ across starting genotypes, Fig. 2.3b), because most genotypes can reach only a few new specificity phenotypes by a path of this length (Fig. A1.4). The bias in outcomes gradually decays as trajectories get longer, but it takes 42 substitutions (10.5 per site) for the mean bias to decrease to within 0.05 units of the equilibrium (Fig. 2.3b, vertical dashed line). By comparison, the maximum root-to-tip branch length in the steroid receptor DBD phylogeny (Fig. 2.1b), which spans over 500 million years of evolution, is just 2.2 substitutions per site. The phenotypes likely to evolve on phylogenetically relevant timescales are therefore strongly affected by local bias in the GP map.

Another consequence of local bias is that outcomes are strongly contingent on the genetic starting point. Consider a trajectory length of 8 substitutions—long enough for new phenotypes to become accessible from most starting points, but not so long that the influence of local bias is lost. At this timescale, genotypes differ dramatically in the distribution of phenotypes that evolve from them (Fig. 2.3c). Much of this variation is explained by the genotype cluster to which the starting node belongs (Fig. 2.3c), because evolutionary trajectories rarely jump between weakly

connected clusters and clusters are strongly enriched for individual phenotypes. Even at this timescale, the direction of phenotypic evolution on average favors conservation of the starting phenotype (Fig. 2.3d), but when new phenotypes evolve, these too differ strongly among starting genotype (Fig. 2.3c).

A final consequence of local bias is that as lineages diverge from each other across the map, the distributions of phenotypic outcomes likely to evolve from them become increasingly dissimilar. The correlation between the distributions of phenotypic outcomes after eight-step evolutionary trajectories from pairs of starting genotypes depends strongly on the distance between those genotypes in the network. For pairs of genotypes that are one substitution apart, the average r^2 is 0.88, but this correlation drops to 0.50 when the genotypes are three steps apart and is entirely lost at 11 steps ($r^2 = 0.02$, the maximum distance on the network) (Fig. 2.3e). Biases in the outcomes of phenotypic evolution therefore become distinct among lineages as they traverse the GP map.

2.3.5 The AncSR1 GP map favored historical conservation of ERE specificity

Local and global bias have a particularly strong and long-lasting impact on the outcomes of evolutionary trajectories that begin from the historical genotype of the recognition helix in AncSR1 (EGKA). It takes 80 substitutions for the bias in phenotypic outcomes from this starting point to decay to within 0.05 units of equilibrium, almost double the average across genotypes (Fig. 2.3b, blue vs. black solid lines). It takes at least 5 substitutions for any new specificity phenotype to be accessed, and even after 8 substitutions the probability of conserving ERE specificity is still 0.90 (Fig. 2.3f). The AncSR1 GP map heavily favors phenotypic conservation from the historical starting genotype across phylogenetically relevant timescales. Bias imposed

by the GP map is therefore congruent with the long-term historical conservation of ERE specificity in the lineages that descend from AncSR1 and lead to modern-day estrogen receptors.

The historical outcome in AncSR1's other descendant lineage—the acquisition of SRE specificity in the kSR clade—was very unlikely on phylogenetic timescales. SRE-specific genotypes are distant from EGKA (Fig. 2.2b), so the probability of evolving SRE specificity after eight substitutions is only 0.0008 (Fig. 2.3f), despite the fact that this is the second-most frequently encoded specificity phenotype in the network overall. Strong local bias around EGKA therefore overrides the global bias towards SRE specificity, making the historical outcome in the kSR clade extremely unlikely.

2.3.6 Evolution of a different GP map in AncSR2

Given that local bias made SRE specificity unlikely to evolve from the ancestral genotype in the AncSR1 map, how could this phenotype have historically evolved in the kSRs? We reasoned that the GP map must have changed along the branch leading to AncSR2 when SRE specificity was acquired. Previous experiments showed that the background substitutions that occurred outside the recognition helix during this interval had a nonspecific permissive effect on both ERE and SRE activation, allowing the protein to tolerate the historical substitutions and other mutations in the RH (Fig. 2.1b) (56, 58). We predicted that the background substitutions had a similarly permissive effect across all REs, increasing the number of functional genotypes in the map and the number of phenotypes they encode, including SRE specificity and others.

To assess this hypothesis, we characterized the GP map of the RH sites in AncSR2 and compared it to the map in the AncSR1 background. As predicted, the number of functional genotypes and phenotypes both massively increased (Fig. 2.4a, b). There are 2,407 functional

protein genotypes in the AncSR2 map, an increase of >20-fold over the AncSR1 background. Fourteen of the 16 possible specificity phenotypes are now encoded in the map, twice as many as in AncSR1 (Fig. 2.2a, 2.4a). The background substitutions therefore dramatically expanded the functional genetic and phenotypic variation that can be produced within the recognition helix.

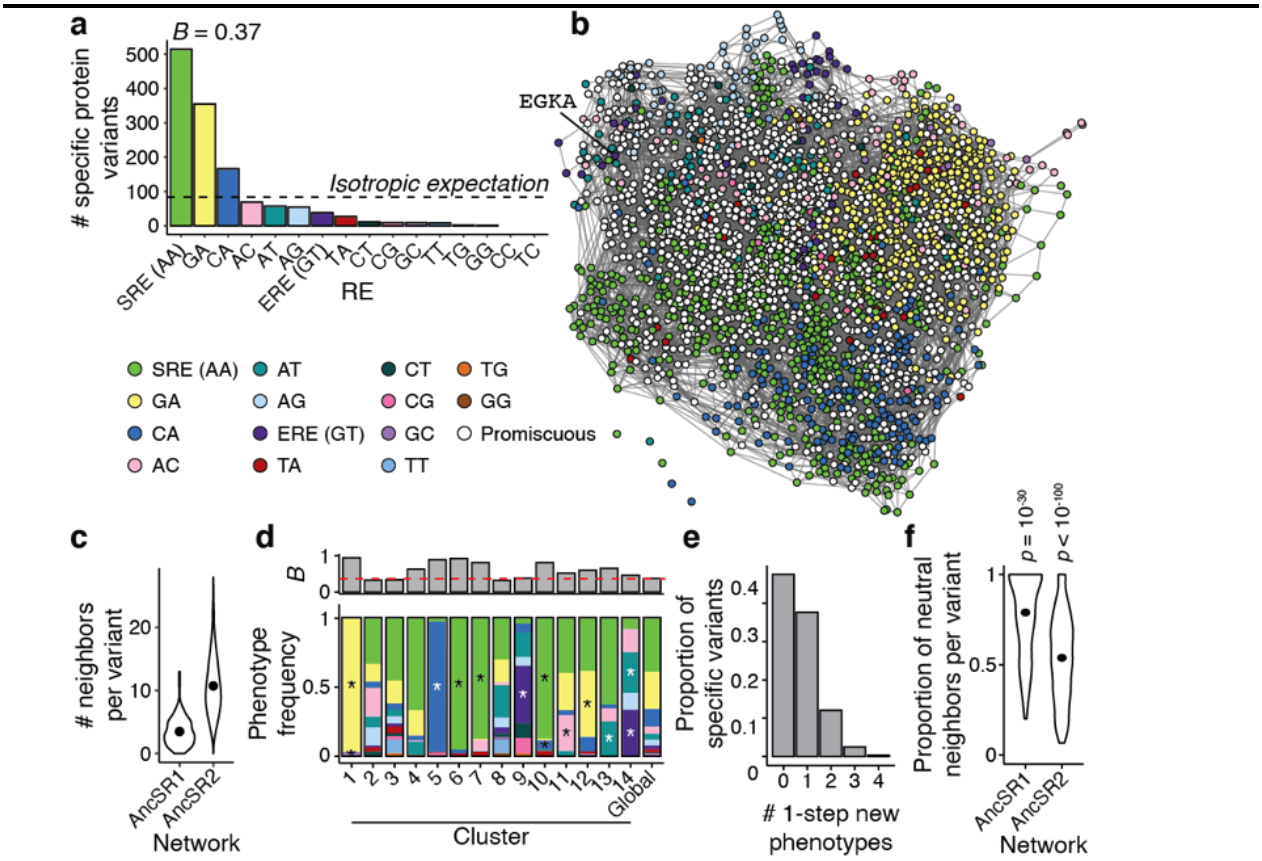


Figure 2.4. Global and local bias and connectivity changed in the AncSR2 GP map. a, Global production distribution and global *B* of the AncSR2 GP map. **b**, Sequence space network of the AncSR2 GP map. **c**, Number of one-step neighbors per protein variant in each network. Dots show the mean of each distribution. **d**, Bottom: Frequencies of specificity phenotypes within each genotype cluster (1–14, ordered by decreasing size); the global production distribution is shown for comparison. Only the 14 largest clusters, which contain >90% of genotypes, are shown. Asterisks, phenotypes significantly enriched within a cluster relative to the global production distribution (Fisher's exact test, $p < 0.05$ after Bonferroni correction). Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. **e**, Distribution of the number of new phenotypes accessible within one mutation, across all RE-specific protein variants in the AncSR2 main component. **f**, Proportion of neutral neighbors per RE-specific variant in the main network component of the AncSR1 and AncSR2 maps. Dots show the mean. *p-value,* probability that the mean would be at least as great as observed if

phenotypes were randomly reassigned in the main component of each map (AncSR1 n = 91, AncSR2 n = 2,402).

Connectivity between genotypes in the map increased, reducing local bias and facilitating access to new phenotypes. In the AncSR2 network, all but five of the 2,407 functional nodes are connected in a single main component (Fig. 2.4b), and the mean number of edges per node is 10.7, a three-fold increase compared to the AncSR1 network (Fig. 2.4c). Genotype clusters are still present, but bias within clusters is weaker than in the AncSR1 map (Fig. 2.2c, 2.4d). As a consequence, genotypes have more access to new phenotypes: >50% of genotypes in the AncSR2 map can access between 1 and 4 new phenotypes within a single mutation (Fig. 2.4e, compare to Fig. 2.2e), because genotypes are typically connected to far more non-neutral neighbors (Fig. 2.4f).

Finally, the global production distribution of phenotypes also changed across this interval. In the AncSR2 map, SRE became the most frequently encoded phenotype (39% of specific variants), and ERE's rank declined from first to seventh (encoding just 3% of specific variants) (Fig. 2.2a, 2.4a). The background substitutions therefore realigned the global phenotype bias from favoring the ancestral specificity to producing the derived specificity.

2.3.7 The AncSR2 GP map favored evolution of SRE specificity

These changes in the AncSR2 GP map dramatically altered the likely phenotypic outcomes of evolution. At long-term equilibrium using our Markov model and the AncGR2 map, the most likely evolutionary outcome is now SRE specificity, with a probability close to 40% (Fig. 2.5a, compared to <20% in the AncSR1 map). At moderate timescales as well, SRE specificity is the most likely outcome across the majority of starting genotypes (Fig. 2.5b). The probability of

evolving new phenotypes overall is considerably higher in the AncSR2 network compared to AncSR1 (mean probability of conservation after 8 steps 0.47 in AncSR2 but 0.61 in AncSR1, Fig. 2.3d, 2.5c).

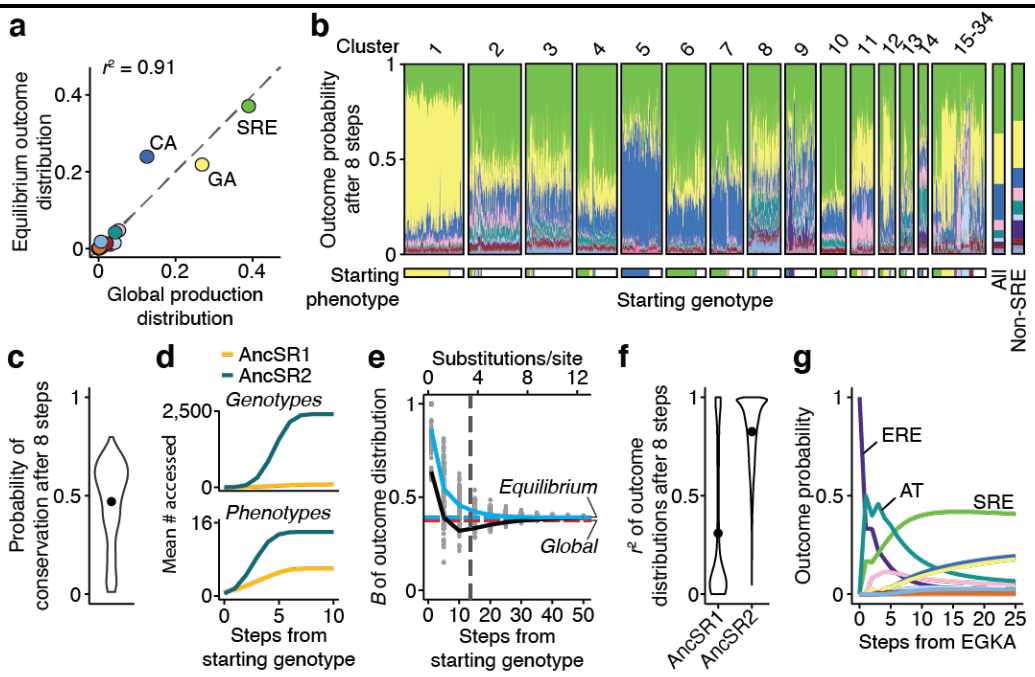


Figure 2.5. The AncSR2 GP map biases evolutionary outcomes towards SRE specificity. a, Comparison between the global production distribution and the long-term equilibrium distribution of phenotypic outcomes in the AncSR2 main network. Dashed gray line, $y = x$. **b**, Distribution of evolutionary outcomes after 8 substitution steps from every starting genotype in the AncSR2 main network component, organized by the cluster of the starting genotype (top). Bottom bar shows the phenotype of each starting genotype. Bars at right show the average outcome distribution for all starting genotypes and all non-SRE-specific starting genotypes, respectively. **c**, Distribution of the probability of phenotype conservation after 8 substitution steps across all specific starting genotypes in the AncSR2 main network component. Dot shows the mean. **d**, Number of genotypes (top) and phenotypes (bottom) accessible as a function of the length of evolutionary trajectories. Lines show the mean across all starting genotypes in each network. Gold, AncSR1 network; teal, AncSR2 network. **e**, Strength of bias (B) in evolutionary outcomes as a function of the length of evolutionary trajectories. Lines and colors are the same as in Fig. 3b. **f**, Distribution of the similarity in outcome distributions (Pearson's r^2) for 8-step trajectories starting from all pairs of genotypes in the AncSR1 and AncSR2 main networks. Dots show means. **g**, Probability of evolving each specificity phenotype starting from EGKA as a function of the number of substitutions.

These changes in evolutionary outcomes arise because of the increased connectivity of

the AncSR2 network and the shift in global production distribution. From any starting point, the increase in functional nodes and connectivity allows access to far more genotypes and new phenotypes (Fig. 2.5d). As a result, the influence of local bias is lost faster, and trajectories more rapidly converge on the equilibrium distribution (Fig. 2.5e), which more closely resembles the production distribution than in the AncSR1 background (Fig. 2.5a). Evolutionary outcomes are also more similar across pairs of starting points than they were in the AncSR1 map (Fig. 2.5f). Combined with the shift in the global production distribution, this causes SRE specificity—which was already the second-most likely outcome in the AncSR1 map—to become the most likely outcome from a majority of starting points in the AncSR2 background.

From the historical RH genotype EGKA (the AncSR2 protein with the RH states reverted to their ancestral states), the likely outcomes of phenotypic evolution are dramatically different than in the AncSR1 map. EGKA is much less mutationally isolated in the AncSR2 network, so the probability of conserving ERE specificity after 8 substitutions drops from 0.9 in the AncSR1 map (Fig. 2.3f) to 0.07 in the AncSR2 map (Fig. 2.5g). The probability of evolving new specificity phenotypes on moderate timescales increases accordingly: after just three steps, two new phenotypes—including SRE specificity—are more likely than conservation of ERE. By six steps, SRE specificity becomes the most likely of all phenotypic outcomes.

The background substitutions that occurred along the branch to AncSR2 therefore changed the GP map of the RH in a way that dramatically changed the probable phenotypic outcomes of evolution. This map strongly favors phenotypic diversification, and it makes the particular phenotype that historically evolved in the kSR lineage the most likely of all possible outcomes.

2.3.8 Simple biophysical mechanisms changed the GP map

Finally, we sought insight into the biophysical mechanisms that changed the GP map of the recognition helix between AncSR1 and AncSR2. Although our experiments provide a functional rather than biophysical readout, different biophysical mechanisms predict different patterns of functional change between the AncSR1 and AncSR2 maps. We therefore analyzed the change in fluorescence of each protein-DNA complex variant between the two backgrounds to identify potential biophysical mechanisms and considered them in light of existing crystal structures. We found evidence for two major mechanisms.

First, the background substitutions between AncSR1 and AncSR2 appear to have caused a universal increase in affinity across all protein-DNA complexes. Previous experiments and crystal structures showed that the background substitutions improve nonspecific DNA contacts and binding cooperativity to both ERE and SRE (56, 58); we therefore hypothesized that affinity increased universally for all amino acid variants across all 16 REs. To test this hypothesis, we fit a simple model in which fluorescence in each ancestral background is a function of a complex's affinity, and affinity is scaled by a constant factor in AncSR2 relative to AncSR1. The model fits the data very well ($r^2 = 0.95$; Fig. 2.6a), with an estimated 70-fold universal improvement in affinity in the AncSR2 background. This apparent increase in affinity explains the vast increase in functional genotypes and specificity phenotypes between AncSR1 and AncSR2, because many protein-DNA complexes that had weak affinity in the AncSR1 background—and were therefore nonfunctional—bind strongly enough in AncSR2 to produce functional levels of fluorescence. The number of promiscuous protein variants also increases, because many variants cross the threshold for functionality on multiple REs (Fig. A1.5a). A universal improvement in affinity explains not only the increased size but also the greater connectivity of the AncSR2 network: the

background substitutions do not qualitatively change the amino acid determinants of binding but instead make them less stringent (Fig. 2.6b), so many of the newly functional nodes in AncSR2 are close neighbors of those that were already functional in AncSR1, with an average gain of 11 new neighbors per node (Fig. 2.6c).

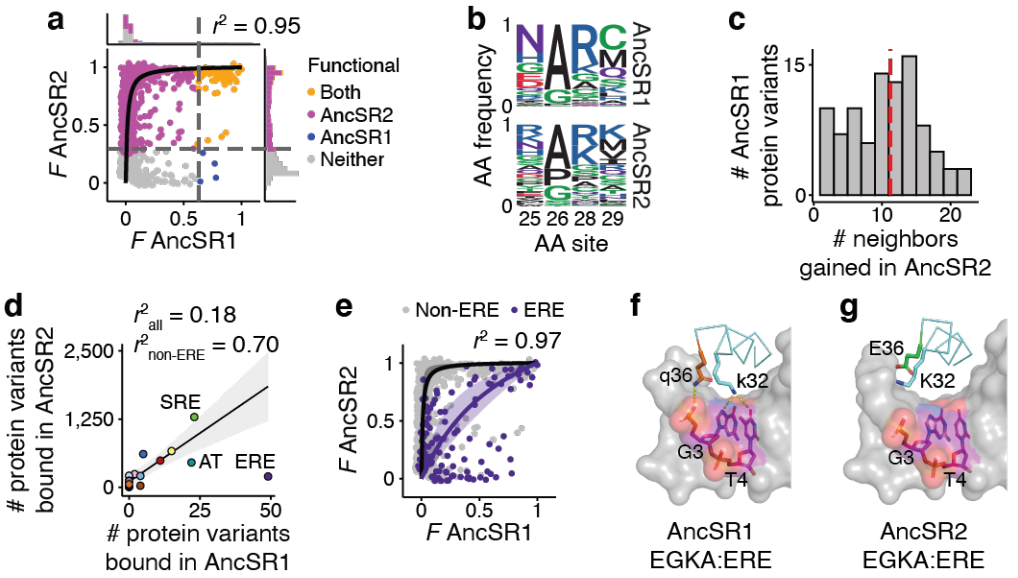


Figure 2.6. Nonspecific effects of background substitutions on DBD-RE affinity. a, Fluorescence of each complex in the AncSR1 vs. AncSR2 background, scaled between the upper and lower bounds for each background. Curve shows best-fit model assuming that the affinity of every complex in the AncSR2 background is related to its affinity in the AncSR1 background by the same scaling factor. Shaded region around the curve (barely visible) shows bootstrapped 95% confidence interval (CI). The Pearson's r^2 between the data and model predictions is shown (n = 2,627). Histograms show distribution of F in each background. Dashed lines show the fluorescence of the wild type complex in each background (AncSR1-EGKA:ERE or AncSR2-GSKV:SRE). Colors indicate the backgrounds in which each genotype is functional. **b**, Amino acid frequencies at the variable RH sites across all functional protein variants in the AncSR1 and AncSR2 maps. **c**, Distribution of the number of neighbors gained in the AncSR2 background across all functional protein variants in the AncSR1 background that remain functional in the AncSR2 background. Dashed line, mean. **d**, Correlation between the number of protein variants bound per RE in each background. Black line, linear fit to all REs except ERE; shaded region, 95% CI. **e**, Same as **a**, but fitting a model in which the background substitutions affect affinity of all variants for ERE by one scaling factor and for all other REs by a different scaling factor. Purple, observed fluorescence and best-fit model predictions for ERE complexes; gray, for non-ERE complexes. **f**, Crystal structure of the AncSR1-EGKA protein in complex with ERE (PDB 4OLN). The RH backbone is shown as a ribbon, with key side chains shown as sticks. The gray surface shows ERE, with variable bases and backbone as sticks. In this complex, glutamine (q) at site 36 forms a hydrogen bond (yellow dashed line) with the DNA backbone, and lysine (k) at site 32 forms two hydrogen bonds to the ERE-specific bases G and T. **g**, Same as **f**, but with the

AncSR2-EGKA crystal structure (PDB 4OND). Substitution to glutamic acid (E) at site 36 abolishes the ancestral hydrogen bond to the DNA backbone and results instead in electrostatic repulsion from the backbone. This deforms the recognition helix, abolishing the hydrogen bonds between K32 and the G and T bases. In **F** and **G**, lowercase letters represent ancestral amino acid states, and uppercase derived.

The second apparent mechanism is that the background substitutions negatively affect specific binding to ERE, shifting the global production bias away from ERE and leaving SRE as the most-encoded phenotype in the AncSR2 background. A universal affinity increase predicts that the number of variants with every specificity phenotype should increase proportionally across the AncSR1-AncSR2 interval; this pattern holds, but ERE is an outlier, with far fewer variants than would be expected given the pattern for other phenotypes (Fig. 2.6d). Moreover, ERE complexes exhibit notably lower fluorescence in the AncSR2 background than predicted by a universal increase in affinity (Fig. A1.5b). We estimated the effect of the background substitutions on ERE affinity by incorporating a background-by-ERE interaction term into our affinity-fluorescence model; adding this parameter improves the fit to the data ($r^2 = 0.97$), with the background substitutions improving ERE affinity by an estimated 2.3-fold, compared to 99-fold for all other REs (Fig. 2.6e). The extent of the relative reduction in fluorescence differs among protein variants, however, suggesting additional specific interactions between background substitutions and amino acids in the recognition helix (Fig. A1.5c). Crystal structures of the EGKA:ERE complex (58) suggest a possible structural basis for the global reduction in ERE affinity: one of the background substitutions (q36E) deforms the protein backbone of the recognition helix, abolishing two hydrogen bonds that are formed between a conserved residue and bases in the ERE (Fig. 2.6f, g). Corroborating this mechanism, the background substitutions also shift the global bias away from AT specificity (Fig. 2.6d), and this is the only other RE that can form these hydrogen bonds.

The structure of the GP map therefore changed between AncSR1 and AncSR2 via two simple biophysical mechanisms. By increasing all proteins' affinity for all REs, while also impairing their affinity for ERE, the background substitutions reduced local bias and changed the direction of global bias, facilitating the evolution of many new genotypes and phenotypes and shifting the protein's global propensity away from conserving ERE specificity to evolving the new specificity for SRE.

2.3.9 Robustness to assumptions

To assess whether our conclusions are sensitive to assumptions that we made in our analysis, we reanalyzed our experimental data under different models and assumptions. First, we applied different thresholds to classify genotypes as functional or nonfunctional, included promiscuous genotypes when characterizing global production distributions, and characterized these distributions using only genotypes with experimentally measured phenotypes. In every case, we observed similar forms of bias in both the AncSR1 and AncSR2 GP maps to those reported above (Fig. A1.6).

Second, instead of treating the protein as an evolutionary unit independent of the RE, we repeated our analyses using an alternative sequence space network in which the protein and RE coevolve as a complex. In this model, evolution may occur via single-step amino acid mutations in the protein or nucleotide mutations in the RE. Our main conclusions again hold: global and local biases impact phenotypic evolution over long and short timescales, favoring ERE conservation in the AncSR1 map and evolution of SRE specificity in AncSR2 (Fig. A1.7).

Finally, we addressed uncertainty about the ancestral sequences. AncSR1 and AncSR2 DBD reconstructions have very high confidence, containing just five and zero ambiguously

reconstructed sites, respectively (57). Experimental data from a prior single-mutant DMS study show that the effects of mutations in the RH are virtually identical when they are introduced into the AncSR1 background or into an alternative reconstruction of AncSR1 that incorporates all plausible alternative amino acids at the ambiguously reconstructed sites ($r^2 > 0.99$; Fig. A1.8) (57). The very limited uncertainty about the AncSR1 ancestral sequence is therefore likely to have little or no effect on our conclusions.

2.4 Discussion

2.4.1 The GP map is a cause of phenotypic evolution

Our data establish that global and local biases in the two ancestral GP maps we studied were causal factors in the historical lineage-specific evolution of DNA specificity. Establishing causality in a multifactorial framework requires 1) evidence that a putative cause increases the probability of the outcome(s) of interest, and 2) evidence for a specific mechanism by which the cause affects the outcome's probability (68). Concerning the first requirement, our experiments show that biases in the AncSR1 map increased the probability that ERE specificity would be evolutionarily conserved, and biases in the AncSR2 map increased the probability that SRE specificity would be acquired. The second requirement is satisfied by a simple axiom of population genetics: the probability that a phenotype will evolve is the product of its probability of production and its probability of fixation under the influence of selection and drift. If biases in the GP map increase the production probability, then evolutionary outcomes will in turn be biased.

A cause must also precede its effect. The biases that favored the conservation of ERE specificity in the AncSR1 map are ancestral to the ER lineage in which that outcome occurred

(Fig. 2.1b). This map persisted unchanged for hundreds of millions of years of phenotype conservation, because zero amino acid changes anywhere in the DBD occurred along the descendant branches leading from AncSR1 to ERa in the ancestor of all bony vertebrates. Even most present-day ERa DBDs contain zero or at most a single substitution relative to AncSR1 (Fig. A1.9). As for the acquisition of SRE specificity in the AncSR2 lineage, the global bias that favors production of SRE specificity as the second-most encoded phenotype was already present in the AncSR1 map. Further, the massive increase in connectivity of the AncSR2 map, which dramatically increased the propensity for new phenotypes to evolve, must have been acquired before SRE specificity actually evolved, because the recognition helix substitutions that conferred SRE specificity during history cannot be tolerated unless the background substitutions that nonspecifically increased DNA affinity occurred first (58). Our experiments do not resolve whether the third major property of the AncSR2 map—a shift in global bias away from ERE specificity that further enhanced the propensity to encode SRE specificity—occurred before or after this phenotype was historically acquired.

We do not argue that selection played no historical role in the evolution of specificity. It seems likely that purifying selection favored conservation of ERE specificity in the chordate ERs, and positive selection could have contributed to fixation of SRE specificity in the AncSR2 lineage. If so, however, selection would have further increased the probability of outcomes that were already favored by the biases imposed by the GP map.

Our data show that the GP map's influence is strong enough to override the influence of selection in many cases. For example, some global biases we observed are absolute—there are 9 specificity phenotypes that cannot be encoded at all in the AncSR1 GP map, and two cannot be encoded in AncSR2; these phenotypes could never evolve, no matter how large a fitness benefit

they might confer. Local bias is also absolute in many cases: from every starting point, the vast majority of phenotypes are impossible to produce directly by mutation, and most require many substitutions before they become accessible. Selection would therefore be powerless to fix these phenotypes over short or medium timescales. The GP map limited evolution to a small subset of possible phenotypes; history, further influenced by selection and chance, played out within this set.

Features similar to those we observed in the steroid receptor GP map affect biological systems and their evolution across levels of organization. Global bias is apparent in other molecular (23, 69) and developmental systems (70, 71, 22, 72), and the resulting biases are often congruent with natural patterns of diversity (25, 26, 73, 74). Local bias also appears to be widespread, because most random mutations are phenotypically neutral if they are tolerated (72, 75–78), and long-term phenotype conservation is widespread in the fossil record (79). When new phenotypes are acquired, identical perturbations often yield different phenotypes in different lineages (80–83), and convergent evolution becomes less likely among distantly related lineages (84). As lineages evolve across their GP maps, their biology inevitably changes, imposing new biases on the production and future evolution of genotypes and phenotypes. It therefore seems likely that anisotropy and heterogeneity are near-universal characteristics of GP maps (14, 20, 21), and that the biases these properties create have shaped large-scale patterns of phenotype conservation and lineage-specific evolutionary change across the tree of life.

2.5 Methods

2.5.1 RE reporter strains

To measure binding of SR DBD to the 16 RE variants, we adapted a yeast GFP reporter system

previously developed to measure binding to ERE and SRE, where GFP expression is well correlated with DNA affinity over a range of at least 2 M^{-2} (r^2 = 0.74) (57). We engineered 16 yeast strains, each of which reports on binding of the DBD to one RE. We modified the yeast strain CM997 (YPS1000 MATa ho::KMX)(85) to replace the *KMX* gene at the *HO* locus with a construct containing yeast-enhanced GFP downstream of a minimal *CYC1* promoter with an array of four palindromic RE sites (tcaAGNNCAcagTGNNCTtga), each separated by a 19-nt sequence, along with a *HygR* gene. To ensure a consistent dynamic range of fluorescence across strains, we made changes to two RE strains in the nucleotide sequences flanking the palindromes at sites that do not affect specificity (59, 60) (see next section). These constructs were transformed into yeast using the lithium acetate method(86) and selected for resistance to hygromycin and susceptibility to G418; integration was confirmed by Sanger sequencing.

To validate this reporter system, we measured fluorescence of each strain in the presence and absence of a DBD variant with universally high affinity to all REs (AncSR1+11P+GGKA) (53, 58). We used a low-copy yeast vector (pDBD) to express this DBD variant as a C-terminal fusion with an SV40 nuclear localization signal and a *S. cerevisiae* Gal4 activation domain (Gal4AD) under control of a pGAL1 promoter. We transformed this construct into each yeast strain using the lithium acetate method followed by G418 selection (50 μg/mL). Single colonies were inoculated in YPD+G418 and transferred to YPGal+G418 media for 6 hours to induce DBD expression. GFP fluorescence was measured on a BD LSRFortessa flow cytometer using a 488 nm laser with 505 nm long pass and 525/50 nm band pass filter. We used as the metric of fluorescence $\log_{10}(\text{GFP}/\text{FSC-A}^{1.5})$, which normalizes fluorescence to cell volume. All 16 strains showed DBD-dependent fluorescence across a similar dynamic range (Fig. A1.1a–c).

2.5.2 Dynamic range correction for RE reporter strains

14 of 16 strains showed DBD-dependent fluorescence across a similar dynamic range, with two fluorescence peaks in the presence of DBD—a GFP-negative peak (GFP–) corresponding to autofluorescence from cells that had spontaneously lost the DBD plasmid, and a GFP-positive peak (GFP+) from cells that retained the plasmid and expressed GFP—and a single GFP– peak in the absence of DBD (Fig. A1.1a, b). With the CC RE strain, however, the GFP+ peak was absent (Fig. A1.1a); with the GA RE, the GFP– peak was right-shifted, indicating high basal fluorescence even in the absence of DBD (Fig. A1.1a, b). We repeated the transformation and obtained the same results. We hypothesized that inserting the CC and GA RE sites may have introduced cryptic yeast TF binding sites or caused chromatin remodeling that resulted in constitutive GFP repression and activation, respectively. The flank and spacer (FS) sequences surrounding the RE half sites do not affect SR binding specificity (60, 87, 88), so we reasoned that introducing mutations into these regions might preserve the ability of the strains to act as reporters for RE binding while disrupting any recognition sites for endogenous yeast proteins. We experimentally identified a set of FS mutations for the CC strain (aacAGCCCAaaaTGGGCTgtt) and another for the GA strain (ccaAGGACAatcTGTCCTtgg) that result in DBD-dependent GFP expression similar to the other RE strains (Fig. A1.1c). We therefore used these two FS-modified strains along with the original strains for the remaining 14 RE variants for DMS experiments.

2.5.3 AncSR1 and AncSR2 combinatorial library construction

We used as the wild-type protein sequences the maximum *a posteriori* AncSR1 and AncSR2 DBD sequences inferred from a maximum likelihood phylogeny of nuclear receptors(57).

We optimized codon usage for yeast and cloned the ancestral DBDs into the pDBD2.1 expression vector, which is modified from the pDBD vector(56, 57) to express GFP at a level within the dynamic range of fluorescence for the wild type AncSR1:ERE and AncSR2:SRE complexes. A bidirectional pGAL1/GAL10 promoter simultaneously drives DBD and mCherry expression, which allowed us to monitor plasmid retention in yeast (Fig. A1.1d).

Combinatorial mutant libraries were created by synthesizing oligos (IDT) with degenerate NNS codons to encode all 20 amino acids and a stop codon at four recognition helix sites of each ancestral protein (Fig. A1.1e). To distinguish sequencing reads coming from different RE strains, 16 synonymously barcoded versions of the library were designed for each background (Fig. A1.1e, Table A1.1). Each barcode (REBC) differed by at least three nucleotides to ensure accurate read assignment despite sequencing errors. The oligos were cloned into the pDBD2.1 vector using the BsaI-HF Golden Gate Assembly kit (NEB), transformed into Invitrogen ElectroMAX DH5α-E E. coli, and maxiprepped (Supplementary Methods). Transformation yields exceeded 1.08×10^7 cfu per barcoded library, providing 56-fold coverage of the amino acid library size (Supplementary Table 2). Assemblies were validated by Sanger sequencing of independent transformants and PCR of the plasmid libraries to confirm the correct insert size.

Maxiprepped libraries (GenElute HP, Sigma-Aldrich) were transformed into the yeast reporter strains using an optimized yeast electroporation protocol. Transformation yields exceeded 10^7 cfu per library (50-fold coverage), estimated by dilution plating (Table A1.2). Yeast libraries were flash-frozen in liquid N_2 in 200 OD_{600}-mL aliquots with 25% glycerol and stored at –80°C. Multiple transformant rates estimated from Sanger sequencing of individual colonies(89) were estimated to result in 0.03% or fewer cells with multiple plasmid copies at

time of sorting.

2.5.4 Cell sorting

We used fluorescence-activated cell sorting (FACS) to separate cells based on their GFP expression. We performed two rounds of sorting: an initial "enrichment sort" to enrich for GFP+ variants in the full libraries, and a second, higher resolution "binned sort" on the enriched libraries to generate quantitative fluorescence estimates for each variant. Enrichment sorting was performed in batches of 8 libraries. Two glycerol stocks per library were thawed on ice, after which cells were recovered for 2 hours in 400 mL YPD+chloramphenicol (chlor) per library at 30°C and 225 rpm. After recovery, G418 was added to the culture and a sample of cells was taken for dilution plating. We recovered a minimum of 1.6×10^7 cfu per library (82-fold coverage). After 15 hours of overnight growth, libraries were washed once in PBS, resuspended to OD_{600} 0.25 in 50 mL YPGal+G418, and grown for 6 hours to induce DBD expression. Cells were then spun down, washed once in PBS, resuspended in 5 mL PBS, and kept on ice for sorting.

Sorting was performed at the University of Chicago Cytometry and Antibody Technology Facility on a BD FACSAria Fusion machine. We used a 488 nm laser with 495 nm long pass filter and 515/20 nm band pass filter for GFP detection, and a 561 nm laser with 595 nm long pass filter and 610/20 nm band pass filter for mCherry detection. After gating on homogeneous single cells and mCherry expression, we sorted cells into GFP– and GFP+ populations (Fig. A1.1f). To normalize fluorescence to cell volume, GFP gates were drawn to have a slope of 1.5 on a log(FSC-A)-log(GFP) plot. We sorted 2.5×10^7 cells per library in the enrichment stage (129-fold coverage, Table A1.2).

Enriched cells from different libraries were pooled by GFP bin and grown in either 700 mL (GFP+) or 2 L (GFP–) of YPD+G418+chlor. Cultures were grown overnight at 225 rpm and 22–30°C, depending on the ratio of cells to media, until they were at least OD_{600} 3 but not yet saturated. 200 OD_{600}-mL 25% glycerol stocks were then made for both the GFP+ and GFP– cultures. 10 OD_{600}-mL of the GFP– culture was used for plasmid extraction using a previously described protocol(51).

The binned sort was performed to yield three replicates per library. For each replicate, two 200 OD_{600}-mL glycerol stocks of GFP+ cells per enrichment sort batch were thawed on ice, recovered in 400 mL YPD+chlor for 2 hours, and sampled for dilution plating. After adding G418, cultures were grown overnight, achieving a recovery rate at least 4X the number of GFP+ cells collected during the enrichment sort (Table A1.3). Overnight cultures were pooled proportionally to the GFP+ cell counts from the enrichment sort, yielding a total of 100 OD_{600}-mL. The pooled cells were washed with PBS, induced for DBD expression in 400 mL YPGal+G418 for 6 hours, washed again, resuspended in 40 mL PBS, and kept on ice for sorting. Binned sorting followed the enrichment sort protocol but used four GFP bins instead of two (Extended Data Fig. 1g), with ~1.6×10^8 cells collected per replicate. The number of sorted cells and recovered reads was consistent across libraries and replicates (Table A1.4).

2.5.5 Deep sequencing

After sorting, cells were grown in 100 mL YPD+G418+chlor per 10^7 sorted cells, or at least 100 mL per bin. Cultures were grown overnight to at least OD_{600} 3.0 but not yet saturated, and 50 OD_{600}-mL was collected per 10^7 sorted cells for plasmid extraction.

Sequencing libraries were constructed from plasmids extracted from the enrichment sort GFP– population and the four binned sort populations using two rounds of amplification. In the first round, the RH scanning and REBC regions of the DBD were amplified with primers that added a 6-nt barcode for bin and replicate identification (BRBC)(90). For every 10 OD_{600}-mL of yeast used for plasmid extraction, 3 μL of plasmid template was used in a 10 μL Q5 PCR reaction (NEB). AncSR1- and AncSR2-specific primers were mixed proportionally to background-specific cell counts (estimated from flow cytometry) to minimize amplification bias. To introduce nucleotide diversity for improved cluster identification during Illumina sequencing, eight unique forward and reverse primer pairs were used per bin and background to encode frameshift diversity and attach read 1 primer sequences in both directions. PCR conditions included 52°C annealing for 13 cycles. Reactions were then pooled by bin/replicate and purified using the Zymo DNA Clean & Concentrator Kit. In the second round, half of the first-round product was amplified with primers to add Illumina P5 and P7 adapter sequences. PCR was performed in 50 μL Q5 reactions (NEB) per 10 μL round 1 product reaction at 68.4°C annealing for 12 cycles. The final product was size-selected on a 2% agarose gel, excised, purified using the Qiagen Gel Extraction Kit, and re-purified with the Zymo DNA Clean & Concentrator Kit.

Final sequencing library concentrations were quantified by Qubit. Libraries were pooled according to the number of cells sorted per bin/replicate, and 1.8 pM dilutions were prepared according to Illumina's standard protocol. Replicate 1 of the binned sort libraries was sequenced on a NextSeq High Output run. The remaining replicates were sequenced on a NovaSeq S1 run at the University of Chicago Genomics Facility. We used standard read primers and 86 cycles for read 1 and 80 cycles for read 2. This enabled us to bidirectionally sequence the region containing the variable RH codons and REBC.

Replicate 1 of the binned sort libraries was sequenced on a NextSeq High Output run. The remaining replicates (2, 3 and 4) were sequenced on a NovaSeq S1 run at the University of Chicago Genomics Facility. We obtained 1.27×10^8 and 2.1×10^9 read pairs for the NextSeq and NovaSeq runs, respectively.

2.5.6 Mean fluorescence estimation, data cleaning and validation

Sequencing reads were processed using a custom pipeline. We used *sickle* v1.33 (91) to filter reads based on their quality: we kept reads with a Phred score ≥ 30 and a minimum length of 79 nucleotides. We then used *PEAR* v0.9.6 (92) to merge the trimmed paired-end reads (minimum assembly length 100 nucleotides). Finally, we used Biopython toolkit v1.79 (93) to demultiplex the assembled reads by DBD background, REBC, and BRBC. We only considered reads that mapped exactly to the DBD background and allowed reads with at most one mismatch in the REBC and one in the BRBC.

The mean fluorescence for protein:RE complexes observed in the binned sort data was estimated as previously described(57). We first estimated the proportion of cells of each complex g in each bin b ($c_{g,b}$) from the proportion of reads in b that mapped to g. The mean fluorescence estimate F_g for each complex was then estimated by taking the weighted mean fluorescence across bins (mean fluorescence of each bin was measured during sorting), with weights $c_{g,b} / \sum_b c_{g,b}$.

We applied several filtering and correction steps to reduce global measurement error and normalize fluorescence estimates between replicates. First, complexes with fewer than 27 reads per replicate were removed to ensure >95% had a standard error (SE) of ≤ 0.1 (5% of the assay range; Extended Data Fig. 2a). Second, complexes observed in only one replicate were excluded.

Third, batch effects were corrected by fitting I-splines to normalize fluorescence between replicates (Fig. A1.2b). Finally, SE was recalculated and complexes with SE > 0.1 were removed (Fig. A1.2c). The final dataset had a mean pairwise Pearson's $r^2 = 0.55$ across replicates. The poor correlation arises primarily because the vast majority of complexes are at the lower fluorescence bound, so r^2 is dominated by measurement noise; for variants with fluorescence above the lower bound (roughly $F \geq -4.0$), r^2 improved to 0.92. Altogether, we obtained fluorescence estimates for 628,732 AncSR1 and 658,475 AncSR2 variants, covering 24.6% and 25.7% of possible variants, respectively (excluding nonsense variants).

Many variants were observed at high read depth in the GFP– bin of the enrichment sort but not in the binned sort. We assigned these a null phenotype (lower-bound fluorescence) using a statistical procedure based on read depth (see next section), resulting in 859,171 AncSR1 and 638,762 AncSR2 protein:RE null complexes (FDR = 0.1; Fig. A1.2d). This increased the total phenotyped variants to 1,487,903 in AncSR1 and 1,297,237 in AncSR2, covering 58% and 51% of all possible variants, respectively.

To evaluate the accuracy of the sort-seq fluorescence values, we measured the fluorescence of 5 isogenic variants by flow cytometry, which were also spiked into the DMS libraries prior to the binned sort. We found a high correlation between the fluorescence estimates from flow cytometry and sorting (Pearson's $r^2 = 0.87$, Fig. A1.2e). We additionally compared the fluorescence estimates of the same variants that were contained in the DMS libraries and again observed a strong correlation with flow cytometry measurements (Pearson's $r^2 = 0.97$, Fig. A1.2e).

To evaluate whether the REBC mutations affected fluorescence, we constructed AncSR1 and AncSR2 "mini-libraries" consisting of each of the 16 REBCs engineered into the respective

wild-type protein variant. These were transformed via electroporation into the ERE or SRE reporter strain, respectively, at 1:16 the scale of the full libraries, and spiked into the full-scale libraries before sorting. The fluorescence of the mini-library variants did not differ significantly by REBC (p = 0.98 AncSR1, p = 0.99 AncSR2, one-way ANOVA), indicating that fluorescence estimates are directly comparable between libraries with different REBC mutations.

2.5.7 Statistical classification of null variants from enrichment sort GFP– bin

In addition to quantitative fluorescence estimates from the binned sort dataset, we reasoned that we could assign variants a null phenotype (*i.e.* baseline-level fluorescence) if they were observed in the enrichment sort GFP– bin but were not observed in the binned sort data, since this implies that they did not express GFP during the enrichment sort. A complication is that variants observed at low frequency in the enrichment sort GFP– bin may simply have not been sorted to sufficient depth to be detected in the binned sort; these variants cannot be confidently classified as null. We therefore set out to test whether variants were sorted to sufficient depth in the enrichment sort to be confidently assigned a null phenotype if they did not appear in the binned sort dataset.

We reasoned that if a variant was sorted to a high enough depth in the enrichment sort to be detectable as fluorescent in the binned sort, then if it did not appear in the binned sort it should be classified as null. We therefore estimated the probability p_m of failing to detect a variant m as significantly fluorescent in the binned sort data; we take this to be the probability of capturing fewer than k cells of variant m in the enrichment sort GFP+ bin, where k is the minimum number of enrichment sort GFP+ cells required to detect a fluorescent variant in the binned sort. This can be calculated as a binomial sampling probability:

$$p_m = Pr\left(c_m^{d+} < k \right) = F_{Binom}\left(k; c_m^d, f \right) \quad (1)$$

where c_m^{d+} is the number of cells of variant m sorted into the enrichment sort GFP+ bin, c_m^d is the total number of cells of variant m sorted in the enrichment sort, f is the minimum fraction of cells in the enrichment sort GFP+ bin for a variant to be detected as fluorescent in the binned sort, and F_{Binom} is the binomial cumulative distribution function. If p_m is low then it is likely that variant m was sorted to sufficiently high depth in the enrichment sort to be detected as fluorescent in the binned sort; p_m can therefore be used as a p-value for classifying variants as null if they are not observed in the binned sort.

We first estimated c_m^d for variants observed in the enrichment sort GFP– sequencing data. To do this, we estimated both the number of cells sorted into the enrichment sort GFP– bin, c_m^{d-}, and the number of cells sorted into the enrichment sort GFP+ bin, c_m^{d+}. For variant m in library l, we took c_m^{d-} to be the fraction of enrichment sort GFP– reads from library l that map to variant m, multiplied by the total number of GFP– cells sorted for library l:

$$c_m^{d-} = \frac{r_m^{d-}}{r_l^{d-}} \times c_l^{d-} \quad (2)$$

where r_m^{d-} is the number of enrichment sort GFP– reads for variant m, r_l^{d-} is the total number of enrichment sort GFP– reads for library l, and c_l^{d-} is the number of GFP– cells sorted for library l (Table A1.2). Since we did not sequence the enrichment sort GFP+ bin directly, to estimate c_m^{d+} we assumed that the fraction of cells of variant m in library l in the binned sort data was proportional to the fraction of cells of variant m in library l in the enrichment sort GFP+ bin:

$$c_m^{d+} = \frac{c_m^b}{c_l^b} \times c_l^{d+} \quad (3)$$

where c_m^b is the number of inferred binned sort cells for variant *m* (summed across all sort bins), c_l^b is the total number of inferred binned sort cells for all variants in library *l* (summed across all sort bins), and c_l^{d+} is the number of GFP+ cells sorted for library *l* in the enrichment sort (Table A1.2). The sum of c_m^{d-} and c_m^{d+} is our estimate of c_m^d.

Next, we estimated the minimum number (*k*) and fraction (*f*) of GFP+ cells in the enrichment sort required for a variant to be detected as fluorescent in the binned sort. Variants with lower fluorescence are less likely to be detected in the binned sort since they have a lower fraction of GFP+ cells, so we set out to define a class of minimally fluorescent variants with which to estimate of *k* and *f*. To do this, we classified variants observed in the binned sort as active (*i.e.,* significantly more fluorescent than null) or null by computing the fraction of nonsense variants of similar read depth and protein background with higher fluorescence than the test variant; this was used as a *p*-value for classifying variants as active (FDR = 0.1). Minimally fluorescent variants were defined as those with fluorescence within ±0.05 units of that of the least-fluorescent active variant. We then used the median c_m^{d+} of minimally fluorescent variants, taken as a weighted average across read depth bins, to obtain an estimate of $k = 9$; the median number of cells in binned sort bins 3 and 4 (roughly equivalent to the enrichment sort GFP+ population) for minimally fluorescent variants, averaged across read depth bins, was calculated to obtain an estimate of $f = 0.23$.

Using our estimated parameter values, we were able to classify an additional 859,171 AncSR1 and 638,762 AncSR2 variants as null (FDR = 0.1; Fig. A1.2d). This brought the total number of phenotyped variants to 1,487,903 in AncSR1 and 1,297,237 in AncSR2, corresponding to 58% and 51% of all possible variants, respectively.

2.5.8 Generalized linear model to predict fluorescence of missing variants

A remaining 42% of AncSR1 and 49% of AncSR2 variants were either unobserved or had insufficient read depth to be confidently assigned a phenotype. To predict the fluorescence of these variants, we used a type of generalized linear modeling approach called reference-free analysis (RFA)(61, 62) to predict fluorescence based on the effects of sequence states inferred from empirically phenotyped variants. RFA is an unbiased method for inferring the phenotypic effects of genetic states and their interactions (*i.e.,* epistasis) in a combinatorial genotype space. Each genotype *g* of length *n* is represented as a vector of genetic states at each site $\left(g_1, g_2, \ldots, g_n \right)^T$. RFA relates the genetic states in *g* to a latent phenotype *s*, also referred to as the genetic score, through a linear combination of main and epistatic effects:

$$s(g) = e_0 + \sum_{i}^{n} e_i(g_i) + \sum_{i<j}^{n} e_{i,j}(g_i, g_j) + \ldots + \varepsilon \quad (4)$$

Here, e_0 is the global mean genetic score across all variants, $e_i(g_i)$ is the first-order (average) effect of state g_i at site *i*, and $e_{i,j}(g_i,g_j)$ is the pairwise epistatic effect of states g_i and g_j at sites *i* and *j*; the model can be extended to include higher order epistatic interactions between sites. The genetic score is related to phenotype through a sigmoid link function:

$$F(g) = L + \frac{U - L}{1 + e^{-s(g)}} + \varepsilon \quad (5)$$

where *F(g)* is the empirically measured phenotype of *g*, *L* and *U* are global parameters representing lower and upper phenotype bounds; and ε is experimental noise, assumed to be normally distributed. The sigmoid link function accounts for nonspecific epistasis arising from biological and/or experimental bounds on the dynamic range of *F*.

We estimated separate RFA models for each ancestral DBD background. The model estimates the effects of all amino acid states at the 4 variable sites in the DBD and all nucleotide

states at the 2 variable sites in the RE; it contains intramolecular interactions up to 3rd order amino acid interactions in the DBD and 2nd order nucleotide interactions in the RE, and intermolecular interactions up to 3rd order DBD-by-2nd order RE interactions. All variants with empirical fluorescence estimates from either the binned or enrichment sorts were used in training; for variants classified as null from the enrichment sorts, we used the mean fluorescence of nonsense variants from the same background. Models were fit in R using *glmnet* v4.1-6(94). Because *glmnet* does not allow for estimation of parameters in the link function, we first fit an unregularized RFA model with up to 2nd order DBD-by-2nd order RE effects using nonlinear least squares regression to estimate the global U and L parameters. We then used these estimates to specify a link function for *glmnet*, which we used to fit the full model using L2-regularized regression. 10-fold cross validation (CV) was used to select the L2 penalty that minimized prediction error (Fig. A1.3a). Our final models fit the data for active variants with $R^2 = 0.96$ (AncSR1) for and $R^2 = 0.99$ (AncSR2); for all variants, $R^2 = 0.31$ (AncSR1) and $R^2 = 0.88$ (AncSR2), because most variation in fluorescence for null variants is caused by measurement error (Fig. A1.3b, c).

We used the RFA models to predict fluorescence for the missing variants in our dataset and classified these variants as null or active. The final dataset with combined empirical and predicted fluorescence estimates for AncSR1 contained 460 active protein variants, of which 114 (25%) were from model predictions; for AncSR2, there were 7,601 active variants, of which 1,838 (24%) were from model predictions.

We used the model to correct for the observation that the GA strain has systematically lower fluorescence than expected given previous measurements of affinity for this RE and a panel of protein variants (Fig. A1.3c) (58), presumably because the FS mutations we introduced

reduce affinity (see 2.5.2: *Dynamic range correction for RE reporter strains*). We estimated the magnitude of this effect by fitting the log-affinity-fluorescence relationship for these variants, including an effect of the RE, using the equation

$$F = L + \frac{U - L}{1 + e^{-\frac{\log(K_A) + d}{a}}} \quad (6)$$

where d is the difference in $log(K_A)$ for GA-bound variants and a determines the steepness of the curve. The best-fit estimate is that the FS mutations reduce the K_A of GA across DBD variants by 0.95 ± SE 0.12 $\log_{10}(\mu M^{-2})$. The genetic score of a complex in the RFA model is linearly related to its affinity, so we used this estimate to adjust the genetic score of all variants on GA and used the fluorescence predicted by the model after this adjustment (Fig. A1.3d). The resulting transformation increased the number of inferred active GA RE variants from 5 to 75 in the AncSR1 background and from 449 to 1,172 in the AncSR2 background. It also increased the fluorescence of the wild type AncSR2 protein on GA to $F = –4.28$, which is closer to the measured fluorescence of AncSR2 wild type on SRE, as predicted by (53, 58).

2.5.9 Classification of functional complexes

We classified complexes as functional if their fluorescence was not significantly lower than the wild type complex, *i.e.* EGKA:ERE in the AncSR1 background and GSKV:SRE in the AncSR2 background. Complexes inferred as null from the enrichment sort were classified as nonfunctional. For complexes observed in the binned sort, we used a *t*-test to account for measurement error. For complexes with predicted fluorescence from the RFA models, we performed a nonparametric bootstrap test using the distribution of model residuals concatenated over the ten cross-validation fits to account for model prediction error. We concatenated residuals across the 10 RFA cross-validation models, then sampled residuals from within an

interval of ±0.1 fluorescence units from the inferred fluorescence of each complex (Fig. A1.3e). *p*-values were calculated as the proportion of bootstrap samples ($n = 250$) with fluorescence greater than or equal to that of the wild type complex. (Fig. A1.3e). For both tests, we used a Benjamini-Hochberg FDR threshold of 0.25 to classify variants as nonfunctional if they were significantly less fluorescent than the wild type complex (Fig. A1.3f). The low stringency of the FDR threshold was chosen to reduce the false positive rate for calling variants functional. The majority of complexes classified as functional in both backgrounds had fluorescence estimates obtained from the binned sort experiment (59.3% AncSR1, 75.4% AncSR2; Fig. A1.3g).

2.5.10 Genotype networks

Following Maynard Smith's sequence space formalism (65), we built protein genotype networks consisting of all functional RH variants in each DBD background. RH genotypes are connected by an edge if they differ by a single amino acid mutation that can be produced via a single nucleotide mutation given the standard genetic code. Genotype networks for joint protein-DNA models follow a similar logic. We considered two functional protein-RE complexes as mutational neighbors if they differed by a single amino acid in the protein *or* a single nucleotide in the RE. For example, the genotypes EGKA:GT and EGKV:GT are neighbors by mutations in the protein, and the genotypes AAAI:AA and AAAI:TA are neighbors due to mutations in the RE. In this network, promiscuous genotypes (such as AAAI) are represented as two separate nodes, one for each complex.

We used the R package *igraph* v1.5.1(95) to build and analyze the genotype networks, and the software *gephi* v0.10.1(96) for network visualization. To identify clusters of densely

connected genotypes within the networks, we used the cluster_edge_betweenness function from the R *igraph* package.

2.5.11 Model of evolution on GP maps

We modeled evolution on the genotype networks as an origin-fixation process under a strong selection-weak mutation regime (97, 98) To isolate the effect of the GP map's structure on evolution, we considered a scenario in which all functional genotypes have equal fitness, so the fixation probability is affected only by drift, and nonfunctional variants are removed by purifying selection. The relative probability $P(i,j)$ of substitution from protein genotype i to genotype j is therefore equal to the amino acid mutation rate μ_{ij}, normalized over all single-step neighbors of i in the network. We assumed that there are no biases in the nucleotide mutation process (*e.g.* transition vs. transversion rate), so μ_{ij} is affected only by unequal mutational access between amino acids imposed by the genetic code. To incorporate this effect, we scaled μ_{ij} by the number of possible nucleotide mutations that can change any nucleotide sequence that encodes i to any nucleotide sequence that encodes j:

$$\mu_{ij} = \eta_{ij}^{o^*} \times \prod_{o \neq o^*} c_o \tag{7}$$

where o indexes the amino acid position, o^* is the position at which the amino acid change occurs, $\eta_{ij}^{o^*}$ is the number of possible single nucleotide changes that can produce the state in j from the state in i at site o^*, and c_o is the number of possible codons for the invariant amino acid state at site o.

We used these transition probabilities to specify a discrete time Markov model for each ancestral genotype network, where each step is a single amino acid substitution. Genotypes that

are more than one nucleotide change apart cannot access each other in a single time step, and the probability of staying in the same genotype across a single step in the Markov chain is also zero. We only considered functional genotypes within the main component of each network (the largest connected component). With this model, we computed the probability distribution $\pi_{(k)}$ of evolving all possible genotypes after k substitution steps given any specified set of starting genotypes:

$$\pi_{(k)} = \pi_{(0)} \times P^k \tag{8}$$

where P is the transition matrix with entries $P(i, j)$, $k > 0$, and $\pi_{(0)}$ is the vector of the probability distribution of genotypes at time step $k = 0$. Setting a single element i of $\pi_{(0)}$ to 1 and all others to zero corresponds to evolution from a single starting genotype; setting all elements of $\pi_{(0)}$ to $1/n$, where n is the number of functional genotypes in the network, averages over all possible starting genotypes. We calculated the relative probability of evolving a given specificity phenotype at time step k by summing over all elements of $\pi_{(k)}$ that encode that specificity and normalizing by the total probability across all specific protein genotypes.

2.5.10 Effects of background substitutions

To estimate the effect of the background substitutions between AncSR1 and AncSR2 on binding affinity, we first considered a model where the background substitutions have a universal nonspecific effect on affinity across all RH and RE genotypes. We assumed that fluorescence is proportional to the fraction of protein bound to DNA. If a complex g has dissociation constant $K_d(g)$ in the AncSR1 background, then its AncSR1 fluorescence (normalized to scale between 0 and 1) is:

$$F(g)_{AncSR1} = \frac{1}{1 + \frac{K_d(g)}{[RE]}} \tag{9}$$

where $[RE]$ is the concentration of RE. If the background substitutions scale $K_d(g)$ by a factor α, then fluorescence in the AncSR2 background is

$$F(g)_{AncSR2} = \frac{1}{1 + \alpha\left(\frac{K_d(g)}{[RE]}\right)} \tag{10}$$

Rearranging these equations gives an expression for fluorescence in the AncSR2 background as a function of fluorescence in the AncSR1 background and α:

$$F(g)_{AncSR2} = \frac{1}{1 + \alpha\left(\frac{1 - F(g)_{AncSR1}}{F(g)_{AncSR1}}\right)} \tag{11}$$

We fit this model to the AncSR1 and AncSR2 fluorescence data using orthogonal regression, which accounts for measurement error in both backgrounds. We used only complexes that had fluorescence measurements from the binned sort in both backgrounds, and whose fluorescence was significantly greater than that of nonsense variants in either background (n = 2,627). Fluorescence was normalized in each background to scale between the upper and lower bounds inferred from the RFA models. Confidence intervals (CI) were constructed by bootstrapping the data and refitting the model. The effect of the background substitutions was estimated to be α = 0.014 (95% CI: 0.010–0.014), corresponding to a 70-fold increase in affinity (95% CI: 70–99).

We next considered a model where the background substitutions have a different effect on ERE affinity than they do on other REs. We modified the model such that α_1 represents the ERE-specific effect of the background substitutions and α_2 the effect on the other 15 REs. We fit

this model as before and obtained parameter estimates of α_1= 0.43 (95% CI: 0.19–0.76) and α_2 = 0.010 (95% CI: 0.0028–0.010), corresponding to fold-increases in affinity of 2.3 (95% CI: 1.3–5.2) on ERE and 99 (95% CI: 99–361) on other REs.

Chapter 3

Epistasis shapes the genotype-phenotype map via structural integration

3.1 Summary

The influence of the genotype-phenotype (GP) map in evolution depends on its structure. A GP map in which every genotype could produce every phenotype upon mutation would be evolutionarily inconsequential. But if the map biased the production of some phenotypes on average, or genotypes had differential access to phenotypes, then the map's structure would influence evolution. The structure of the GP map is determined by its genetic architecture, but how this architecture shapes the map's structure—and the range of structures that can arise from it—remains poorly understood. Here, we characterize the genetic architecture of a transcription factor's GP map from a deep mutational scanning experiment charting the production of every possible DNA specificity phenotype encoded by all steroid hormone receptor (SR) protein genotypes. By treating the production and accessibility of DNA-specificity phenotypes as two separate structural properties, we reveal how epistasis shapes the SR GP map. Epistasis increases the diversity of phenotypes encoded in the map, but also confines them to particular regions of protein genotype space. As a result, variation in epistasis drives correlated changes in both structural properties, outlining a limited space of possible GP map configurations, each with predictable evolutionary consequences. Our findings provide a mechanistic framework to study the structure of the GP map and suggest that the SR's GP map—and likely that of other systems, too—is a permanent causal factor in evolution.

3.2 Introduction

The relationship between genotypes and phenotypes (the GP map) is a central object in evolutionary biology, yet its role in shaping patterns of phenotypic diversity remains debated (10, 9, 99, 24, 20). Whether the GP map could influence evolutionary outcomes depends on its structural properties (100), particularly those arising from the distribution of phenotypes across the connected network of all possible genotypes. If the GP map were isotropic—where all mutations generate all possible phenotypes with equal probability—and homogeneous—where all genotypes produce identical distributions of phenotypes—it would impose no bias on evolution, leaving selection as the sole determinant of phenotypic outcomes (21, 20). However, an anisotropic map would bias the production in favor of certain phenotypes, making them more likely to evolve. And a heterogeneous map would make some phenotypes more accessible through mutations from specific genotypes, biasing evolution toward lineage-specific outcomes. The structure of a GP map—and its potential to influence evolution—could be therefore characterized along two axes: the extent of bias in the production of variation (the isotropy-anisotropy axis) and the extent of bias in the accessibility of variation (the homogeneity-heterogeneity axis) (Fig 3.1).

Biases in the production and access to variation are direct consequences of the association between genotypes and phenotypes. This association is in turn determined by the GP map's genetic architecture—the set of causal rules by which individual genetic states contribute to a phenotype, as well as the epistatic contribution of every possible pair of states and higher-order combination. By determining how genotypes encode phenotypes, the genetic architecture defines how different sets of genetic determinants shape the extent of anisotropy and heterogeneity in the map. However, to address the complexity and vastness of GP maps,

traditional models of genetic architecture have made simplifying assumptions about the underlying causes of the GP relationship. For example, Fisher's geometric model assumes that mutations affect all phenotypes equally (37), the NK model treats epistasis between *K* random loci as noise (38), and quantitative genetics attributes trait variation solely to the additive effects of numerous small-effect loci (39). While these models capture macroscopic patterns of the map, they overlook detailed genetic and biological processes. As a consequence, infinitely many possible genetic architectures produce the same GP bias (40–42, 101), obscuring the mechanistic link between genetic architecture and the structure of the GP map.

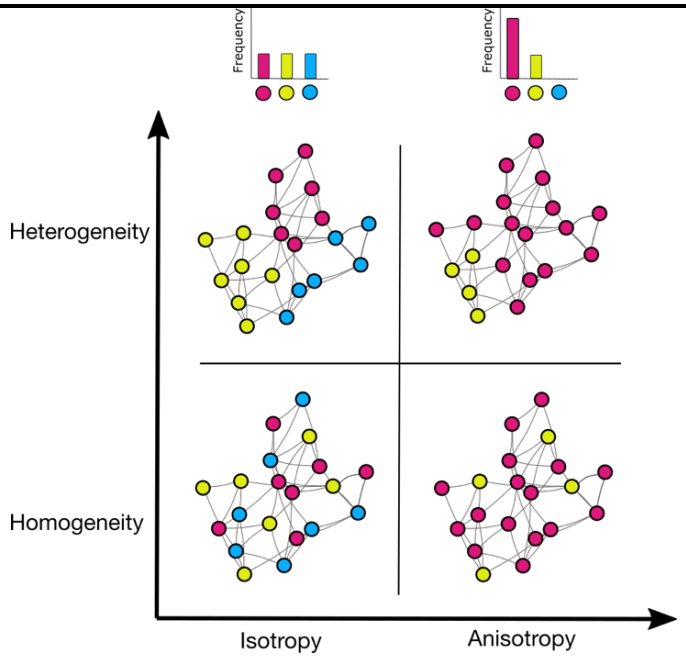


Figure 3.1. Space of conceivable GP map configurations. A hypothetical GP map with three conceivable phenotypes (colors) is shown. Top barplots, frequency distribution of phenotypes produced across all possible genotypes. Networks, distribution of encoded phenotypes across genotypes (nodes). The x-axis captures the extent of bias in the production of variation and the y-axis captures bias in the access to variation from individual genotypes. Changes on both structural properties of the GP map can define a space of conceivable structures. Only when the GP map is both isotropic and homogeneous (lower left), it has no influence on phenotypic evolution.

Determining the extent and causes of bias in a GP map also requires mapping the production of every possible phenotype, as biases can only be studied by comparing the relative production of different phenotypes. However, most experimental studies have examined only one or a few phenotypes (36), while comprehensive mappings of genotypes and phenotypes exist only in computational GP maps. Circumstantial evidence from both partial and computational GP maps suggests that epistasis could shape the structure of the map, but its effects remain debated (102, 103). In computational GP maps, epistasis reduces the set of conceivable phenotypes to a small subset of possible ones because many genotypes map to the same phenotype (23, 104). Yet, simulations with quantitative genetic models indicate that epistasis increases phenotypic variation from new mutations (105), and experimental studies in molecular systems show that it enables the production of intermediate phenotypes (106). Similarly, the role of epistasis in the access to variation is debated. Computational studies of molecular, regulatory, and metabolic GP maps suggest that epistasis restricts access to variation due to landscape ruggedness and degeneracy (77, 78, 107–109). Yet, experimental studies in molecular systems reveal that epistasis can facilitate access to novel phenotypes by opening mutational paths between genotypes with distinct phenotypes (35, 110).

Overall, the influence of the genetic architecture—particularly epistasis—on shaping the GP map's structure remains poorly understood. This is due to the absence of complete empirical GP maps, the oversimplified assumptions of current models of genetic architecture, and the lack of studies characterizing the effects of epistasis on both the production and accessibility of variation within the same system. How does epistasis influence the diversity of phenotypes encoded in the map? How does it shape the distribution of phenotypes across the genotype

network? Is the structure of the GP map constrained by its underlying genetic architecture? Answering these questions is essential for understanding the GP map's role in evolution.

Here, we address this knowledge gap by assessing how epistatic interactions shape the structure of a complete experimental GP map: a combinatorial deep mutational scanning assay of all possible genotypes at four sites critical for DNA binding in a reconstructed ancestral steroid hormone receptor, and all possible DNA specificity phenotypes that can be encoded by them. Previously, we characterized the structure of this map by quantifying the production frequency of every possible DNA specificity phenotype and their distribution across the protein genotype network. Now, we apply a statistical framework of genetic effects to decompose the genetic architecture of the map, and to evaluate how different architectures, varying in the extent of epistasis, simultaneously affect the production and access of variation in this map. This fine-scale analysis allowed us to link the genetic architecture to the macroscopic structure of the GP map, revealing how it shapes the possible configurations of the map and the influence of the map's structure on evolutionary outcomes.

3.3 Results

3.3.1 Experimental data

Steroid hormone receptors (SRs) are a class of ligand-activated transcription factors that regulate chordate development, behavior, and reproduction. These receptors originated from a single ancestral gene that underwent duplication, giving rise to two main clades with distinct DNA specificities. Estrogen receptors (ERs) retained the ancestral DNA specificity for the estrogen response element (ERE)—an inverted palindrome with the motif AGGTCA (58, 59). In contrast, ketosteroid receptors (kSRs), which include receptors for androgens, progestogens, glucocorticoids, and mineralocorticoids, evolved a new specificity for the steroid response

element (SRE, AGAACA) (58, 60). Experiments on reconstructed ancestral SR DNA-binding domains (DBDs) revealed that this functional shift in DNA specificity was driven by three substitutions in the DBD's binding interface—these changes occurred between the last common ancestor of the two subfamilies (AncSR1) and the most recent common ancestor of kSRs (AncSR2) (58).

The data we analyze here come from the GP map for AncSR2 we generated in Chapter 2. Briefly, the space of genotypes consists of all possible 160,000 genotypes at four critical sites in the protein's binding interface—the three that changed between AncSR1 and AncSR2, plus one other that varies in the broader nuclear receptor family. The space of phenotypes consists of all possible 16 RE sequences that can be produced by combinations of the two nucleotide sites that differ between ERE and SRE. The experiment charted the DNA specificity phenotype of every protein genotype, by transforming a library of protein variants into yeast and measuring GFP fluorescence driven from every RE sequence via a FACS-based sort-seq assay.

The experiment also characterized the macroscopic structure of the GP map by assessing its two structural properties. Along the production axis, the map shows anisotropy. Genetic variation in the map produces 14 of the 16 possible specificity phenotypes. Furthermore, the frequency distribution of DNA specificity phenotypes encoded across all protein variants is highly nonuniform. Along the accessibility axis, the map shows heterogeneity. Specificity phenotypes are unevenly distributed across the space of functional protein variants. Each genotype can access only a subset of the possible phenotypes by single mutations, and these sets differ between genotypes.

3.3.2 Intermolecular epistasis is a key component of the GP map's genetic architecture

To characterize the genetic architecture of the GP map, we implemented a statistical approach based on reference-free analysis (RFA) on the experimental data (61, 110). In RFA, each model term is an explicitly defined genetic effect, including additive effects and epistatic interactions, and each term is encoded relative to the global functional average. This allows us to express the portion of the phenotypic variation in the GP map that is attributable to any particular set of genetic determinants. We used the experimentally measured fluorescence as the phenotype and encoded each protein-DNA complex in the model as a genotype with 6 genetic states (4 amino acid states and 2 nucleotide states); each model term can correspond to the additive effect of an amino acid or nucleotide state, a pairwise interaction or higher-order interactions between states. To identify the best description of the genetic architecture, we built seven RFA truncated models (Table 3.1) with varying order of epistatic interactions and two different forms of epistasis: intramolecular epistasis (interaction between states in the same molecule) and intermolecular epistasis (interaction between states across the protein-DNA interface). We fitted each truncated

Table 3.1. Truncated RFA models. Models vary in the form of epistasis (intra or intermolecular) and the order of epistatic interactions. Interactions, terms included in each model; *A*: Amino acid state at a protein site; *N*: Nucleotide state at a DNA site; symbol ":" is an interaction between a site/state combination. Order, highest order of interaction in the model.

Model	Interactions	Type of epistasis	Order	Number of terms
1	A + N	NA	1	88
2	A + N + A:A + N:N	intramolecular	2	2,504
3	A + N + A:A + N:N + A:A:A	intramolecular	3	34,504
4	A + N + A:N	intermolecular	2	728
5	A + N + A:N + A:A + N:N	Intra + inter	2	3,144
6	A + N + A:A + N:N + A:A:A + A:N + A:A:N + A:N:N	Intra + inter	3	55,624
7	A + N + A:A + N:N + A:A:A + A:N + A:A:N + A:N:N + A:A:N:N + A:A:A:N + A:A:A:N:N	Intra + inter	5	862,024

model by least-squares regression and L1 regularization to reduce overfitting (Fig A2.1a-b); we then used 10-fold cross validation to compute the fraction of phenotypic variance explained by each model as an out-of-sample R^2, which measures how well a model inferred from a random subset of the data can predict the phenotypes of unsampled variants. We evaluated the performance between models as the increase in out-of-sample R^2 for active protein-DNA complexes—those whose fluorescence is in the dynamic range of the assay, providing quantitative variation in fluorescence.

We found that intermolecular epistasis is both necessary and sufficient to capture the phenotypic variation present in the GP map. Models with intermolecular epistasis explain between 15-75% of the phenotypic variance for active complexes (Fig 3.2a, blue regions). In contrast, models without intermolecular interactions have virtually zero explanatory power (Fig 3.2a, yellow and green regions). Adding only the simplest family of intermolecular terms—the interaction between one amino acid and one nucleotide—to a completely additive model is

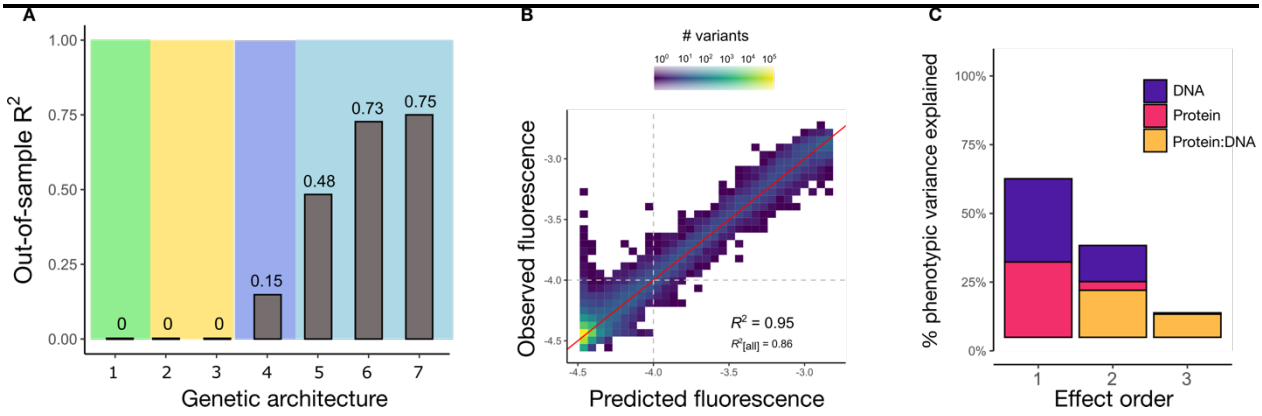


Figure 3.2. Genetic architecture of the SR GP map. (A) Phenotypic variance explained by truncated models of genetic architecture evaluated by 10-fold cross validation (CV). Bars, average out-of-sample R^2 across 10 CV models. Colored regions group models by the form of epistasis included; green: no epistasis, yellow: intramolecular epistasis, dark blue: intermolecular epistasis, light blue: intra + intermolecular epistasis (see Table1 for details). **(B)** Prediction accuracy of the best-fit model (model 6 in **A**). Red line, y = x line. Gray dashed lines, fluorescence cutoff for active variants—those in the dynamic range of the assay. R^2 for active (top) and all (bottom) variants and is shown. **(C)** Fraction of phenotypic variance explained by

order and form of interaction in the best-fit model. Bars, fraction of total phenotypic variance explained by terms of a given order. Colors, fraction of total phenotypic variance explained by terms that include only protein sites, only DNA sites or both.

sufficient to notably increase the phenotypic variance explained by the model (Fig 3.2a, compare model 4 vs 1; Table 3.1). We also detected interactions between forms of epistasis. For example, pairwise intramolecular effects—interactions between two amino acids or two nucleotides—are not sufficient to increase the phenotypic variance explained by a completely additive model (Fig 3.2a, compare model 2 vs 1; Table 3.1), except when they occur in the background of pairwise intermolecular interactions (Fig 3.2a, compare model 5 vs 4; Table 3.1). Intermolecular epistasis is therefore also necessary to unmask the phenotypic effects of intramolecular interactions.

We next focused on the models that incorporate intermolecular epistasis to identify the best-fit model of genetic architecture. These models vary in the order of epistasis, from pairwise (models 4 and 5) to higher-order interactions (models 6 and 7). We found that the genetic architecture is best described by model 6, which incorporates up to three-way epistatic interactions (Fig 3.2a; Table 3.1); increasing the order of interactions only marginally increases the phenotypic variance explained. The genetic architecture implied by the best-fit model is sparse. With just 7,627 of the 55,624 possible terms (14%), the model achieves very high prediction accuracy ($R^2 = 0.95$; Fig 3.2b; Fig A2.1c-d). Of all the phenotypic variance explained by the model, about a half (43%) is attributed to epistasis. As expected, intermolecular interactions are the most important form of epistasis, accounting for 60% of the variance due to epistasis and for a quarter (25%) of the total phenotypic variance (Fig 3.2c). Taken together, these results show that epistasis is pervasive in the GP map and that intermolecular epistasis, in particular, is a fundamental component of the genetic architecture.

3.3.3 Epistasis reduces anisotropy in the GP map and increases diversity

Having characterized the genetic architecture of the GP map at a microscopic scale, we next sought to understand how this architecture shapes the macroscopic structure of the GP map—the production and accessibility of DNA specificity phenotypes. Changes in DNA specificity require amino acid replacements to interact with nucleotide states differentially, which involves epistasis across the protein-DNA interface (58, 111, 112). We therefore hypothesized that the prominence of intermolecular epistasis in the genetic architecture is likely to be linked to both structural properties of the map. The best-fit model contains two families of intermolecular epistatic terms: pairwise interactions between one amino acid and one nucleotide (main specificity terms) and three-way interactions involving two amino acids and one nucleotide or one amino acid and two nucleotides (higher-order specificity terms, Table 3.1). We thus hereafter refer to the best-fit model as the S_2 model.

To assess the effects of sequentially removing each of these families of epistatic interactions, we characterized the structure of the GP map with two truncated models of genetic architecture. Model S includes additive terms and main specificity terms only, while model A includes only additive terms, removing both specificity families. We fitted both models to the experimental data and used the estimated coefficients to predict the GP relationships. For each model, we re-classified functional protein-DNA complexes as those whose predicted fluorescence was not significantly worse than the experimentally measured AncSR2 wild-type complex (GSKV:SRE), accounting for model prediction error. Each protein variant was categorized as specific (functional with a single RE), promiscuous (functional with multiple REs), or nonfunctional. We then compared the structural properties of the inferred maps for models S and A to the experimentally characterized GP map and its genetic architecture (S_2).

We first evaluated how these alternative genetic architectures influence the production property of the GP map, specifically the diversity and frequency distribution of specificity phenotypes encoded across all protein variants. Consistent with the genetic mechanisms underlying protein-DNA specificity switches, removing families of specificity terms leads to a dramatic reduction in the number of encoded phenotypes in the map (Fig 3.3a). However, the overall number of bound DNA elements—both specific and promiscuous—remains nearly identical (Fig. A2.2a). Intermolecular epistasis therefore only affects the production of specificity phenotypes.

The reduction in phenotypic diversity is a simple consequence of the genetic mechanisms underlying protein-DNA specificity. Specific binding occurs when the combined contributions of amino acids and nucleotides in a protein-DNA complex push the system over a functional threshold, resulting in activation for a single RE. Under an additive genetic architecture, context-dependent binding is absent: genetic states only affect affinity and the effects on affinity depend solely on the sum of the main amino acid and nucleotide contributions, not on any specific combination of states (110). As a result, specificity only arises when the contribution of a single pair of nucleotide states is sufficient to exceed the functional threshold. In this system, the SRE element is the only one whose independent nucleotide contributions are capable of pushing protein sequences above the threshold (Fig. A2.2b), making SRE the only specificity that can be encoded (Fig 3.3b, left). Genetic changes cannot encode alternative specificities because any increase in affinity beyond the functional threshold inevitably leads to promiscuous binding. Introducing epistasis across the protein-DNA interface decouples the phenotypic effects on affinity from specificity, thereby increasing the phenotypic diversity of encoded DNA specificities in the GP map (Fig 3.3a).

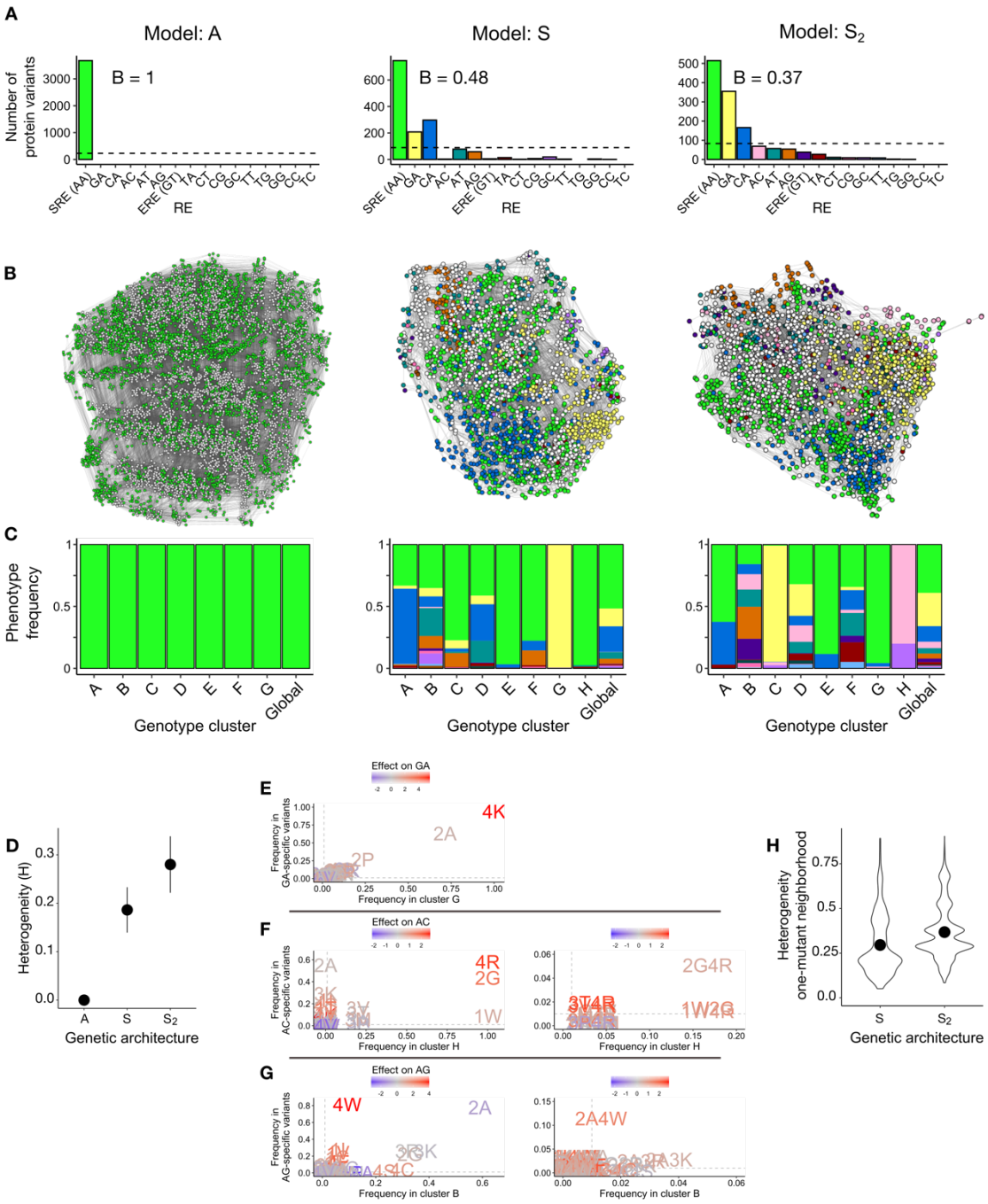


Figure 3.3. Epistasis shapes the structural properties of the GP map. (A) Global production distribution and global bias *B* for each model. Bars represent the number of functional protein variants that bind specifically to each RE. The dashed line shows the expected frequencies if the distribution were unbiased. Global bias *B* is calculated as 1 minus the Shannon entropy (base 16) of the distribution. **(B)** Genotype networks of functional protein genotypes. Nodes represent functional protein genotypes, connected by edges if a single nucleotide mutation can produce the amino acid change. Nodes are colored by their predicted specificity phenotype under each model of genetic architecture. White nodes, promiscuous genotypes. For model S_2, genotypes are colored according to their experimentally measured phenotype, or predicted phenotype—for

unmeasured genotypes—based on the model. **(C)** Frequencies of specificity phenotypes within each genotype cluster for each model; the global production distribution is shown for comparison. **(D)** Heterogeneity (*H*) of the GP networks. Point, average heterogeneity across genotype clusters. Error bars, standard error of the mean *H*. Heterogeneity (*H*) is calculated as the Kullback-Leibler divergence between each cluster's phenotype distribution and the expected global production distribution of the map. (**E-F)** Frequency of amino acid states per site (left) and pairwise combinations (right) in each cluster and across all the protein genotypes encoding a given DNA specificity phenotype. Amino acid states and combinations are colored by their genetic contribution to each phenotype. Dashed lines, 1% frequency. **(H)** Distribution of heterogeneity (*H*) of the single-mutant neighborhood around every specific genotype in each epistatic GP network. Point, average local heterogeneity across specific genotypes.

To directly quantify the effect of the genetic architecture—specifically of intermolecular epistasis—on the production property of the GP map, we computed the extent of bias in the global production distribution of DNA specificities (*B*) as 1 minus the Shannon entropy (base 16). *B* ranges from 0 in a completely isotropic map—if all 16 phenotypes are produced with equal frequency—to 1 if only a single phenotype can be produced. The additive map is maximally anisotropic because only one DNA specificity is encoded (Fig 3.3a, left). Adding main specificity terms dramatically increases the diversity of encoded phenotypes from 1 to 14 DNA specificities, decreasing *B* by 52% (Fig 3.3a, center). Including higher-order specificity terms further modulates the phenotype frequencies, reducing *B* by an additional 23% (Fig 3.3a, right).

Since model S_2 also contains pairwise intramolecular interactions (Table 3.1, see model 6), we directly tested whether this form of epistasis could also shape the production property of the map. We excluded the effects of specificity terms by fitting a truncated model including additive terms and only pairwise intramolecular interactions and predicted the GP relationships. We found that the predicted GP map remains maximally anisotropic (Fig A2.2c). The observed decrease in the global production bias of the map is entirely attributable to specificity terms. Together, these data indicate that epistasis across the protein-DNA interface is the cause of

phenotypic diversity in functional specificity in the GP map—intermolecular epistasis reduces anisotropy, giving rise to a more uniform range of encoded possible phenotypes.

3.3.4 Epistasis creates heterogeneity and distorts access to variation

We next assessed the effect of the alternative genetic architectures on the accessibility property of the GP map, specifically the distribution of encoded specificity phenotypes across protein genotype space. We built a genotype network of protein variants for each GP map, where nodes represent protein genotypes and edges connect two sequences that can be directly interconverted by a single nucleotide change given the standard genetic code. To assess the distribution of phenotypes in the network, we assigned to each node the phenotype (DNA specificity or promiscuous) predicted by the model of genetic architecture (Fig 3.3b). Nonfunctional variants were excluded from the network because they are expected to be removed from an evolving population under purifying selection (65). Since these networks sometimes comprise multiple disconnected components, we only considered the largest component, which we refer to simply as the network.

In a homogenous GP map, all genotypes have uniform access to all phenotypes encoded in the map. As a result, the frequency distribution of phenotypes encoded in every region of the genotype network should match the global production distribution. To test this, we used a community detection algorithm (113) to identify genotype clusters—groups of highly connected genotypes but with fewer connections among clusters—and computed the frequency distribution of specificity phenotypes encoded within each cluster (Fig 3.3c). We then computed the extent of heterogeneity (H) as the normalized Kullback-Leibler divergence between each cluster's phenotype distribution and the expected global production distribution of the map. H ranges from 0 in a completely homogenous map to 1 if the phenotype distribution of every genotype cluster

maximally differs from the global production distribution. We found that complete homogeneity—when every region of the genotype network produces the same expected distribution—only occurs in a fully additive GP map. In contrast, genetic architectures with epistasis produce heterogeneous GP maps (Fig 3.3c-d, mean H across clusters > 0).

Homogeneity in the additive architecture arises from the absence of context-dependent binding—specificity does not rely on particular combinations of states, allowing many different protein genotypes to become specific for the SRE element (Fig. 3.3b, Fig. A2.3a). In contrast, context-dependent binding in the epistatic architectures creates heterogeneity. To quantify this, we computed the genetic contributions of individual amino acid states and pairs of states in protein variants to every encoded phenotype in the map, summing their inferred coefficients across both nucleotide states in the RE. Heterogeneity arises because states (or combinations of states) with strong contributions to a phenotype tend to be concentrated in specific clusters. For example, in the S network, GA specificity is predominantly localized to cluster G (Fig. 3.3c), where all protein genotypes share a lysine (K) at site 4 of the DNA-binding interface—a residue found in all GA-specific variants and with the strongest contribution to GA specificity (Fig. 3.3e). Similarly, in the S_2 network, AC specificity is largely confined to cluster H (Fig. 3.3c) because genotypes in this cluster share an arginine (R) at site 4 and a glycine (G) at site 2, both of which are common in AC-specific genotypes and contribute strongly to AC specificity (Fig. 3.3f, left). Furthermore, the epistatic interaction between these residues positively contributes to the phenotype, further driving its confinement to cluster H (Fig. 3.3f, right).

Epistasis also redistributes phenotypes within the network. For example, AG specificity is equally frequent in the S and S_2 maps (frequency = 0.04; Fig. 3.3a), but while it is distributed across three clusters in the S network, it becomes sequestered to a single cluster in the S_2

network (Fig. 3.3c). In the S network, amino acid states contributing to AG specificity are present at moderate frequencies in clusters B, C, and F, spreading AG specificity across these clusters (Fig. A2.3b). However, pairwise amino acid interactions in the S_2 network concentrate this phenotype in cluster B. Alanine (A) at site 2 is the most frequent state in cluster B but individually has a negative contribution to AG specificity. This changes when it interacts with tryptophan (W) at site 4—the state with the strongest positive contribution to AG specificity—turning alanine's overall contribution strongly positive (Fig. 3.3g).

At a finer scale, heterogeneity strongly affects the phenotypes that can be accessed by individual genotypes. For both epistatic GP maps (S and S_2), the distribution of accessible phenotypes in the single-mutant neighborhood of every RE-specific genotype significantly differs from the global production distribution (Fig 3.3h, mean $H > 0$). Furthermore, local heterogeneity increases with the number of families of specificity terms in the genetic architecture (mean $H = 0.29$ in S map vs. 0.37 in S_2 map; Fig 3.3h). We tested whether the higher local heterogeneity in the S_2 map could arise from differences in connectivity (number of neighbors per genotype) and neutrality (fraction of neighbors with the same phenotype) relative to the S map, because these features directly affect access to genotypes with novel phenotypes. Connectivity in both GP maps, however, is very similar (mean 12 edges per node in S and 10.7 in S_2), and neutrality is identical (mean 0.53 in both maps; Fig A2.4). The increase in local heterogeneity in the S_2 map is therefore driven by higher-order specificity terms.

Epistasis across the protein-DNA interface is the cause of heterogeneity in the GP map. While epistasis facilitates the production of a higher and more uniform diversity of phenotypes on average, it also sequesters phenotypes to different regions of the network and creates local distortions in the propensity of individual genotypes to access phenotypic variation. When

epistasis is present in the map, the access of variation from individual genotypes depends upon the genotype's location within the network.

3.3.5 Structural integration shapes the space of possible GP map structures

Our results so far show that epistasis across the protein-DNA interface shapes both structural properties of the GP map. The models examined, however, comprise substantial, discontinuous changes to the genetic architecture—the complete addition or removal of families of terms—rather than changes in the magnitude of the effect of those terms. To simulate quantitative variation in the genetic architecture, we titrated the amount of epistasis into the S_2 model and assessed the impact of these changes on both structural properties of the GP map as we did previously. For both specificity families in the model, we successively scaled the terms by a constant and re-inferred the GP map with each new set of model coefficients. After each scaling procedure, we computed the amount of epistasis in the map as the epistatic variance (ε_{GP})—the total phenotypic variance in the map attributed to both families of specificity terms.

We evaluated the change in global bias (B) and heterogeneity (H) over a range of values of ε_{GP} spanning more than eight orders of magnitude—at the lower end, ε_{GP} accounts for less than 0.001% of the phenotypic variance in the map, and at the higher end it accounts for more than 95% of the phenotypic variance. We found that both structural properties change in a monotonic fashion with ε_{GP}: B always decreases while H always increases with epistatic variance (R^2 for B = 0.95 and R^2 for H = 0.97; Fig 3.4a-b). Quantitative variation in epistasis, however, affects the access of variation more strongly than the production of variation in the map: H changes 1.5 times faster with the amount of epistasis than does B. As a result, H reaches its maximum value when ε_{GP} accounts for ~30% of the phenotypic variance, while B reaches its

minimum value when ε_{GP} accounts for ~90% of the phenotypic variance. Changes in the genetic architecture therefore produce asymmetric variation to the structure of the GP map.

To understand how epistasis shapes the possible configurations of the GP map's structure, we characterized the degree to which the two structural properties covary. Using the sigmoidal relationships between ε_{GP} and each property, we predicted GP map structures across a range of ε_{GP} values. We then projected the predicted structures onto a bidimensional space, with one axis representing variation in *B* (the isotropy-anisotropy axis) and the other representing variation in *H* (the homogeneity-heterogeneity axis). This analysis revealed a fundamental trade-off between production and access of variation (Fig 3.4c). High ε_{GP} produces maps with greater and more uniform phenotypic diversity, which also tend to be heterogeneous. In contrast, low ε_{GP} yields less diverse, more anisotropic maps that are typically more homogeneous. And no amount of epistatic variance can generate a map that is simultaneously isotropic and homogeneous (*i.e.*, $B \approx 0$ and $H \approx 0$). Changes in the genetic architecture therefore drive correlated variation in both structural properties, as both are intrinsically linked to epistasis. We call this phenomenon structural integration.

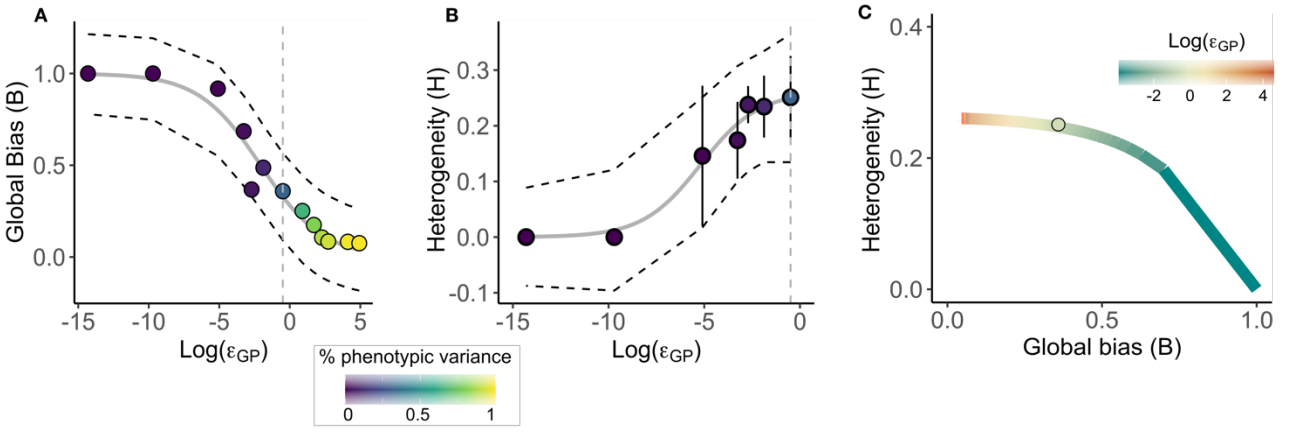


Figure 3.4. Epistasis shapes the space of possible GP map structures. Effect of epistatic variance (ε_{GP}) on global bias **(A)** and heterogeneity. **(B)**. Gray line, best sigmoidal fit. Black dashed lines, 95% confidence interval (CI) of the fit. Vertical dashed line, reference epistatic

variance in the best-fit model of genetic architecture. Points, value of *B* or *H* computed after scaling the model coefficients of the specificity terms by a constant and re-inferring the GP relationships. Points are colored by the percent of the total phenotypic variance captured by ε_{GP}. Error bars, standard error of the mean *H* across genotype clusters. **(C)** Covariation between global bias and heterogeneity as predicted from sigmoidal fits in (A) and (B). Line shows the predicted structure of the GP map (combinations of *B* and *H*) as a function of the amount of epistatic variance (color). Point, structure of the experimental GP map.

These observations have two major implications for evolution and the genetic architecture of the GP map. First, due to structural integration, evolution of the genetic architecture can only produce a limited set of possible configurations for the SR GP map. Variation over a wide range of epistatic variance results in anticorrelated and asymmetric changes in the map's structural properties. Consequently, even substantial changes to the genetic architecture are unlikely to generate all conceivable GP map configurations. Second, the SR GP map will always influence phenotypic evolution. The pattern of structural integration precludes the generation of an evolutionarily inert map—one that is both isotropic and homogeneous. Even if the genetic architecture evolved to reduce bias in one property, it would increase the bias of the other.

3.3.6 The GP map's structure predicts its effects on evolutionary outcomes

Finally, we investigated the evolutionary consequences of variation in GP map structure. Structural integration between production and access of variation suggests that changes in genetic architecture modulate the relative influence of these properties on evolutionary outcomes. In maps with minimal epistasis, access to variation is largely homogeneous, making phenotypic evolution more dependent on the anisotropy of the global production distribution. Conversely, in highly epistatic maps, production tends toward isotropy, shifting the influence to

the heterogeneous access to variation, where phenotypic outcomes should be more strongly determined by the starting genotype.

To test these predictions, we modeled evolutionary trajectories on GP maps with different genetic architectures. We used a discrete-time Markov chain to compute the probability that each possible DNA specificity phenotype would evolve from every starting genotype in the network given trajectories of variable length; the model allows a protein genotype to change into any directly connected genotype, with a probability proportional to the number of single-nucleotide substitutions that can mediate it. After a given trajectory, the likelihood of evolving any phenotype is proportional to the sum of the probabilities of the genotypes encoding that phenotype. This model corresponds to neutral molecular evolution with purifying selection against nonfunctional variants (65, 67), allowing us to directly test the impact of alternative GP map structures on evolutionary outcomes.

We modeled evolution on the experimental GP map (S_2 model) and the map produced by the S model. The experimental GP map is 55% more heterogeneous than the S map, while the S map is 30% more anisotropic than the experimental map (Fig. 3.3). Phenotypic evolution should therefore be more strongly influenced by the heterogeneous access to variation in the experimental map, whereas it should be more shaped by the global production distribution in the S map. Evolutionary modeling confirms these predictions. First, the distribution of phenotypic outcomes across genotypes more closely resembles the global production distribution in the S map. We computed the probability distribution of evolutionary outcomes from every starting genotype over an evolutionary timescale of k substitutions. Then, we measured the similarity between each genotype's outcome distribution and the global production distribution using the metric 1-H (Fig. 3.5a). After an evolutionary timescale of three substitutions, the outcome

distributions are more similar to the production distribution in the S map than in the experimental map (mean similarity = 0.84 in S vs. 0.78 in Exp; Fig. 3.5b), consistent with the expected stronger influence of the production of variation. Even over longer timescales, which allow genotypes to explore more of the network, the production distribution continues to shape phenotypic outcomes more strongly in the S map (Fig. 3.5c).

Second, the likely direction of phenotypic evolution depends more strongly on the starting genotype in the experimental map, consistent with the expected stronger effect of heterogeneity. We measured the similarity in the distribution of evolutionary outcomes between every pair of starting genotypes using the squared Pearson's correlation coefficient (r^2; Fig. 3.5d). After an evolutionary timescale of three substitutions, genotypes show greater differences in their distribution of evolutionary outcomes in the experimental map compared to the S map (mean $r^2 = 0.45$ in S vs. 0.32 in Exp; Fig. 3.5e). Furthermore, the expected direction of phenotypic evolution diverges more rapidly in the experimental map (Fig. 3.5f). For pairs of genotypes separated by a single substitution in the network, r^2 is only 2.5% lower in the experimental map. However, when genotypes are three steps apart, r^2 decreases by 12%, and when five steps apart, it decreases by 30%. In more epistatic maps, evolutionary outcomes diverge faster as lineages traverse the map.

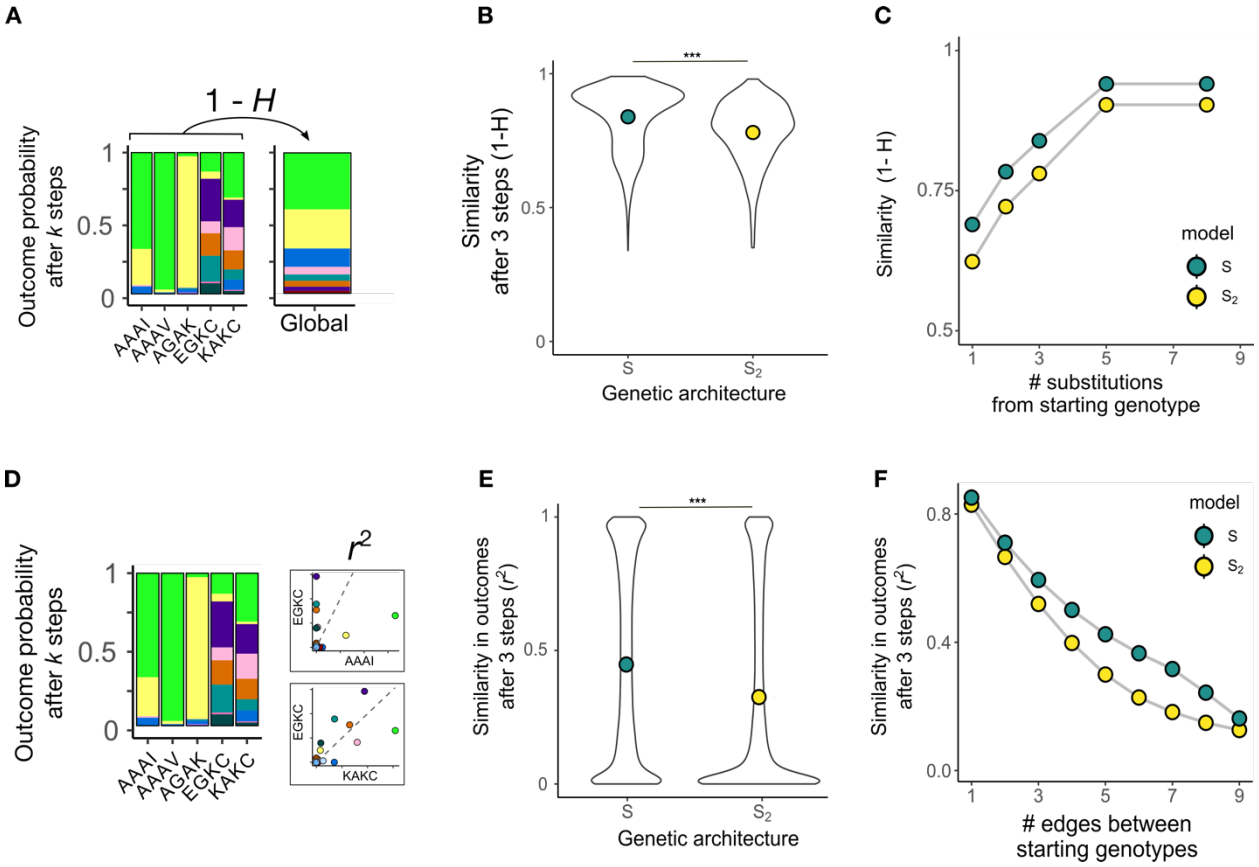


Figure 3.5. Evolutionary effects of variation in the GP map structure. (A) The distribution of evolutionary outcomes from every genotype in the GP network, after an evolutionary trajectory of *k* steps, is compared to the global production distribution using the metric 1-*H*. Barplots, distribution of evolutionary outcomes from a sample of starting genotype (left) and global production distribution (right) **(B)** Distributions of the similarity in outcomes from individual genotypes to the global production distribution after 3-step trajectories. **(C)** Change in the average similarity in outcomes to the production distribution as a function of the length of the evolutionary trajectory. **(D)** The similarity in the distribution of evolutionary outcomes after 3 steps between every pair of starting genotypes is computed as the squared Pearson's coefficient (r^2). Barplot, distribution of evolutionary outcomes from a sample of starting genotype. Scatterplots, pairwise comparison of evolutionary outcomes; dashed line, y = x. **(E)** Distribution of the similarity in evolutionary outcomes between every pair of starting genotypes in the GP network after 3 steps. **(F)** Change in the average similarity in outcomes between pairs of starting genotypes as a function of their distance in the network. *** *p*-value << 10^{-5}.

These results support our hypothesis that anisotropy plays a more dominant role in the S map, while heterogeneity shapes more strongly the outcomes in the experimental map. They also show that the limited set of possible GP map configurations leads to predictable processes of

phenotypic evolution. In maps with minimal epistasis, the global production frequency of each phenotype would contain more information to predict evolutionary outcomes, largely independent of the starting genotype. In contrast, in highly epistatic maps, this information becomes less predictive because phenotypic evolution is more context-dependent; outcomes will be more strongly shaped by the local distribution of phenotypes in the neighborhoods surrounding the starting genotypes.

3.3.7 Discussion

The modern synthesis established natural selection as the primary—if not sole—factor determining the direction of phenotypic evolution (1, 114). Implicitly, this view assumed a highly specific structure for the GP map: to allow selection to act without constraint, the GP map had to be both isotropic and homogeneous (20, 21). Such a structure would impose no bias in evolutionary outcomes, leaving selection as the dominant force. Over the past two decades, experimental and computational studies have challenged this perspective (22–24, 26, 30), demonstrating that GP maps are not random—genotypes do not produce all conceivable phenotypes equally, and some phenotypes are produced far more frequently than others. Despite this evidence, these examples do not exclude the possibility that the GP map could, in principle, possess an unbiased structure—or, more broadly, that it could adopt any configuration. If the GP map's structure were entirely malleable, selection could still exert absolute control, shaping the structural biases of the map to align with adaptive needs.

Our results establish that the GP map of SRs is constrained to certain configurations, imposing strong limits to natural selection. The pleiotropic effect of epistasis across the protein-DNA interface drives this constraint by simultaneously shaping both properties of the map, leading to structural integration—strong asymmetric and anticorrelated changes between

production and access of phenotypic variation. This phenomenon is reminiscent of the patterns of covariation that exist among traits in organisms (115), suggesting common underlying principles governing the structure of complex biological systems. Structural integration defines the space of GP map structures that can arise through changes in the genetic architecture. In doing so, it also precludes the possibility of a GP map that is simultaneously isotropic and homogeneous, and many other configurations, as well. The structure of the SR's GP map is therefore a permanent causal factor in SR's evolution—it is neither entirely malleable nor can it be evolutionarily inert.

Structural integration also shapes the influence of the GP map on phenotypic evolution. We show that variation in the map's structure influences the outcomes of evolution in predictable ways. While both properties jointly shape evolutionary outcomes (except in maximally biased maps), structural integration determines which property has a stronger influence. Maps biased along the production axis exert more uniform effects across genotypes, making phenotypic evolution more predictable. In contrast, maps biased along the accessibility axis lead to outcomes that depend more heavily on the starting genotype, rendering evolution more contingent. By providing a framework to understand these dynamics, structural integration unifies the role of epistasis in phenotypic evolution: at a broad scale, epistasis is essential to generate phenotypic diversity—it always reduces the anisotropy of the map—, while at a finer scale, it can either hinder or facilitate access to novel phenotypes, depending on the genotype's location within the map (116, 117). The genetic architecture of the map determines the particular combination of structural biases, which reflects the GP map's potential to influence evolutionary outcomes.

Features of the genetic architecture of SRs are found in many other molecular systems, so we expect epistasis to shape their GP maps in similar ways. For example, intermolecular epistasis is present in all regulatory systems involving the interaction of two or more

components—whether protein-DNA or protein-protein interactions—and it is always necessary for phenotypic changes in specificity (106, 111, 112, 118, 119). Moreover, protein-DNA regulatory systems exhibit greater variation in gene expression phenotypes when mutations epistatically interact between the two components, compared to the effects of mutations occurring in individual components (106). It is therefore likely that the effects of intermolecular epistasis on phenotype production and accessibility are common across a wide range of molecular systems.

We only considered two of the map's structural properties, specifically the distribution in the number of genotypes per phenotype and the distribution of encoded phenotypes across the network of genotypes. However, other structural properties such as the number and connectivity of functional genotypes can also vary and impact evolutionary outcomes (110, 120). GP maps are extremely complex objects, so multivariate approaches describing the extent of variation on these additional structural axes, whether they also covary, and the genetic determinants underlying these changes will be essential for a comprehensive understanding of the GP map and its role in evolution.

Finally, our findings call for a re-evaluation of the expected structure of the GP map and its influence on evolution (20). By implementing an explicit model of genetic effects (61, 121) we establish a direct link between epistasis and the macroscopic structure of the GP map. Our results offer a realistic null expectation for the structure of the SR GP map—and potentially for many other systems, as well: a fully additive genetic architecture would yield a GP map that is anisotropic and homogeneous, while a maximally epistatic architecture would result in an isotropic and heterogeneous map. This framework implies that bias in the GP map is not an

exception but the rule, that evolutionary outcomes are invariably shaped by the map's structure, and that evolution of these biases reflect meaningful biological mechanisms.

3.4 Methods

3.4.1 Model fitting of truncated genetic architectures

To characterize the genetic architecture of the AncSR2 GP map, we implemented a statistical framework based on reference-free analysis (RFA) (61, 110). RFA is an unbiased method for inferring the phenotypic effects of genetic states and their combinations for any arbitrary number of sites and states. The sequence space of the AncSR2 dataset contains all possible combinations of 20 amino acid states at 4 protein sites and 4 nucleotide states at 2 DNA sites. We implemented RFA by encoding every protein-DNA complex g as a genotype vector with 6 genetic states (4 amino acids and 2 nucleotides). RFA relates the genetic states in g to a latent phenotype s, known as the genetic score, through a linear combination of main and epistatic effects:

$$s(g) = e_0 + \sum_{i}^{n} e_i(g_i) + \sum_{i<j}^{n} e_{i,j}(g_i, g_j) + \ldots + \varepsilon \quad (1)$$

where, e_0 is the global mean genetic score across all variants, $e_i(g_i)$ is the additive effect of state g_i at site i, and $e_{i,j}(g_i,g_j)$ is the pairwise epistatic effect of states g_i and g_j at sites i and j. RFA models can thus be specified to include any arbitrary order and form of interactions. The genetic score s is related to the measured phenotype (fluorescence; $F(g)$) through the sigmoid link function

$$F(g) = L + \frac{U - L}{1 + e^{-s(g)}} + \varepsilon$$

where, L and U are the lower and upper bounds of fluorescence, respectively, and ε is experimental noise assumed to be normally distributed. Phenotype bounding between L and U arises due to limited dynamic range of the assay and can introduce spurious inferences of

epistasis; the sigmoid link function accounts for this nonlinearity, removing a source of nonspecific epistasis.

Using equation 1, we specified 7 RFA truncated models (Table 3.1). The models vary in the order of interactions and form of epistasis. We specified intramolecular epistasis as interactions between sites/states in the same molecule (i.*e.,* between amino acid states in the protein or nucleotide states in the DNA), and intermolecular epistasis as interactions between amino acid and nucleotide states. The models were fit in R using *glmnet* v4.1-6 (94). Since *glmnet* cannot perform joint estimation of the linear coefficients in *s* and the parameters of the link function, we first fitted an unregularized model 7 (Table 3.1) using nonlinear least squares regression to estimate the *L* and *U* parameters; we set these values as global parameters for the link function. We then fitted each truncated model with *glmnet* using L1 regularization to avoid overfitting and 10-fold cross validation (CV) to find the best regularization parameter (λ_{min}) for each model (Fig A2.1a); we used the value of λ_{min} to fit the full models.

3.4.2 Phenotypic variance explained by models and model terms

To assess the fraction of phenotypic variance explained by each truncated RFA model, we computed an out-of-sample R^2 from the CV models corresponding to λ_{min} (CV_{min}) as these models minimize the prediction error. We used the model coefficients estimated from each of the 10 training sets and predicted the phenotypes of the held-out sets. For each model, we report the average out-of-sample R^2 across the 10 CV_{min} models. We computed an out-of-sample R^2 including all variants and only those that are active (Fig A2.1b). Active variants are defined as those whose fluorescence is significantly higher than the average fluorescence for stop-codon variants, and thus provide quantitative variation in fluorescence.

We also computed the fraction of the total phenotypic variance of the best-fit model explained by different sets of genetic determinants (61, 110). Since the terms of an RFA model are defined relative to the global functional average, the total phenotypic variance explained by the model can be partitioned into the variances attributed to every possible combination of genetic determinants in the model. The phenotypic variance of any term is just the square of its coefficient (e_i), and the variance of any set of terms including that site or combination of sites (β) is the average of the squared coefficients:

$$Var(\beta) = \frac{1}{\left(S_A{}^{O(\beta)}\right)\left(S_N{}^{O(\beta)}\right)} \sum_i e_i^2$$

where S_A and S_N are the number of amino acid and nucleotide states, respectively, for the site (or combination of sites), and $O(\beta)$ is the order of the effect represented by β. The fraction of the total phenotypic variance explained by β is

$$F(Var(\beta)) = \frac{Var(\beta)}{\sum_i Var(\beta)_i}$$

3.4.3 Titration of epistasis into the best-fit model

To simulate quantitative variation in the magnitude of the effects of intermolecular epistatic terms, we scaled the intermolecular coefficients of the best-fit model by a constant. The composition of the model remains unchanged but the contribution of intermolecular epistasis changes, allowing us to evaluate the effects of quantitative variation in epistasis on the structure of the GP map. For each scaling procedure, we computed the new amount of intermolecular epistasis as the epistatic variance (ε_{GP}), defined as the total phenotypic variance in the map attributed to both families of specificity terms:

$$\varepsilon_{GP} = Var(\beta_{A:N}) + Var(\beta_{A:N:N}) + Var(\beta_{A:A:N})$$
$$= \left(\frac{1}{N}\sum_{i\in A}\sum_{k\in N} e(A_i{:}N_k)^2\right) + \left(\frac{1}{N}\sum_{i\in A} e(A_i{:}N_5{:}N_6)^2\right) + \left(\frac{1}{N}\sum_{i<j\in A}\sum_{k\in N} e(A_i{:}A_j{:}N_k)^2\right)$$

where, $Var(\beta_{A:N})$ is the total phenotypic variance contributed by main specificity terms, and the variance contributed by higher-order specificity terms is partitioned into $Var(\beta_{A:N:N})$—interactions between one amino acid and both nucleotide states—and $Var(\beta_{A:A:N})$—interactions between two amino acids and one nucleotide state. A_i iterates over every possible amino acid state across all 4 protein sites, and $N_k$iterates over every possible nucleotide state across the two DNA sites.

3.4.4 Classification of functional complexes

We classified protein-DNA complexes as functional if their predicted fluorescence, under each model of genetic architecture, was not significantly lower than the experimentally measured fluorescence of the AncSR2 ancestral wild-type complex GSKV:SRE. To account for model prediction error, we used the relationship between predicted fluorescence from the CV_{min} models and observed fluorescence, and computed an empirical distribution of residuals. For each complex, we sampled residuals from the distribution from within an interval of $\pm$ 0.1 fluorescence units of its predicted fluorescence. *p*-values were calculated as the proportion of bootstrap samples (n = 250) with fluorescence greater than or equal to that of the wild-type complex. We used a Benjamini-Hochberg FDR threshold of 0.25 to classify functional complexes.

3.4.5 Protein genotype networks

To describe the structure of the GP maps, we built genotype networks of the functional protein

variants predicted by each model of genetic architecture. The network includes only functional genotypes, following the model of molecular evolution in which proteins can traverse sequence space without passing through nonfunctional intermediate nodes. Two protein genotypes are connected by an edge if they differ at a single amino acid and at least one pair of codons encoding each amino acid differ by a single nucleotide. We used the R package *igraph* v1.5.1 (95) to build and analyze the genotype networks, and the software *gephi* v0.10.1 (96) for network visualization. To identify clusters of densely connected genotypes within the networks, we used the cluster_fast_greedy function from the *igraph* package, designed for very large networks. The function finds the optimal number of subgraphs (*i.e.,* genotype clusters) by directly optimizing a modularity score for the network (113).

3.4.6 Model of evolution on GP maps

To model the evolution of phenotypes on the genotype networks, we used a strong selection-weak mutation, where the mutation rate is low enough that the time to fixation of a mutation is shorter than the time between mutations (97). Thus, trajectories on a genotype landscape can be modeled as a stepwise origin-fixation process. To isolate the effect of the GP map's structure on evolution, we considered a scenario in which the fixation process is unbiased, *i.e.*, all functional genotypes have equal fitness, so the fixation probability is affected only by drift and nonfunctional genotypes are removed by purifying selection. The rate of substitution q_{ij} from protein genotype *i* to genotype *j* is equal to the mutation rate μ_{ij}. The probability that a substitution will occur from genotype *i* to *j* depends on the local structure of the genotype network around *i* (98):

$$P(i,j) = \frac{q_{ij}}{\sum_k q_{ik}} = \frac{\mu_{ij}}{\sum_k \mu_{ik}} \quad (9)$$

where k indexes all single-step neighbors of the focal node i.

We assumed that there are no biases in the nucleotide mutation process (*e.g.*, transition vs. transversion rate), so the amino acid mutation rate μ_{ij} is affected only by unequal mutational access between amino acids imposed by the genetic code. To incorporate the effect of the genetic code, each μ_{ij} is scaled by the number of possible nucleotide mutations that can change any nucleotide sequence that encodes protein genotype i to any nucleotide sequence encoding protein genotype j:

$$\mu_{ij} = \eta_{ij}^{o^*} \times \prod_{o \neq o^*} c_o \quad (10)$$

where o indexes the amino acid position, o^* is the position at which the amino acid change occurs; $\eta_{ij}^{o^*}$ is the number of single nucleotide changes between codons for the amino acid in i and the amino acid in j at site o^*, and c_o is the number of codons that encode each of the invariant amino acid states at the other sites. This allows a population fixed for a given protein genotype to neutrally explore all synonymous codons that encode that amino acid sequence.

We build a transition probability matrix (P), containing all possible $P(i,j)$'s between every pair of functional genotypes in the network, separately for the S and S_2 GP maps. We used the P matrices to specify a discrete Markov chain, where each time step represents an amino acid substitution. Genotypes that are more than one nucleotide change apart cannot access each other in a single time step, and the probability of staying in the same genotype across a single step in the Markov chain is also zero. We only considered functional genotypes in the largest connected component of the networks.

With this model, we computed the probability distribution $\pi_{(k)}$ of evolving all possible genotypes after k substitution steps given any specified set of starting genotypes:

$$\pi_{(k)} = \pi_{(0)} \times P^k$$

where P is the transition matrix with entries $P(i, j)$, $k > 0$, and $\pi_{(0)}$ is the vector of the probability distribution of genotypes at time step $k = 0$. To model evolution from every possible starting genotype i, we set the i-th entry of $\pi_{(0)}$ to 1 and all others to zero. We calculated the relative probability of evolving a given specificity phenotype after k steps by summing over all elements of $\pi_{(k)}$ that encode that specificity and normalizing by the total probability across all specific protein genotypes.

Chapter 4

Towards a multi-scale genotype-phenotype map: Bridging phenotypic variation and evolution from molecules to organisms

4.1 Summary

Phenotypic evolution in complex biological systems—macromolecules, gene regulatory networks and organisms—is studied in two largely isolated research fields: evolutionary biochemistry and developmental evolution. Yet, theoretical and experimental results from both fields have arrived at fundamentally similar conclusions—that biological systems are tightly integrated entities and that the structure of biological systems is a causal factor in their evolution, because it determines the ways in which the system can change when it is perturbed. Here we articulate the implications of integrating the conceptual and experimental approaches of evolutionary biochemistry and developmental evolution into a unified field to study phenotypic change. We show that all biological systems share fundamental ways in which they can change, offering general mechanistic principles of phenotypic modification occurring at all levels of biological organization. We also introduce the paradigm of a multi-scale genotype-phenotype map, which aims to provide a mechanistic framework to explain the production and evolution of phenotypes through the complete chain of biological causality. A conceptual and experimental integration of both fields would move us forward towards a more complete understanding of how complex biological systems have produced their astonishing diversity of phenotypes.

4.2 Introduction

Explaining how the breadth of organic forms—their shapes, functions and structures—has come

to be is a central task for evolutionary biochemists and developmental biologists alike. Evolutionary biochemistry aims to explain the historical processes and physical mechanisms by which biological molecules diversified in structure and function (122). Similarly, developmental evolution is concerned with explaining how alterations in the mechanisms of embryonic development generate organismal diversity (31, 123). Albeit at different scales, both fields seek to study the molecular or developmental origins of phenotypes, uncover the causes of phenotypic evolution, and provide explanations about the evolutionary processes that drove particular features into existence.

Both fields place the structure of the biological system—the macromolecules' 3D structure and the architecture of gene regulatory networks and, respectively—at the core of their inquiries. They provide explanations stemming from how mutations produce phenotypic change as mediated by its effects on the structure of the system, which informs the type of changes that are more or less likely, or even possible, to arise during evolution. For example, they explain the effect of a change in gene expression on development given the gene's position within a gene regulatory network, or a mutation's effect on protein function given the residue's position within the 3D structure. It is by these types of explanations that we may begin to answer questions that have puzzled evolutionary biologists, paleontologists, and naturalists for over a century: Why do some phenotypes evolve whereas other alternative feasible ones have never been observed? Why do some features of biological systems change more often than others? Why do some features remain conserved over millions of years?

This common interest in explaining the mechanisms for the evolutionary diversification of biological systems creates an avenue to expand both fields' boundaries into an encompassing and richer evolutionary framework. Unfortunately, current research in both fields occurs largely

in isolation. Developmental biologists seek to connect organismal phenotypes down to developmental changes mediated through genetics. But this last step—the modifications of gene regulatory networks—occurs essentially through biochemistry via the modification, addition or removal of gene-product interactions, which is usually treated as a black box. Conversely, evolutionary biochemists seek to connect molecular evolution and protein biochemistry to macroevolutionary processes. But the last step involves the complexity of organismal biology—development, physiology, and cell biology—which is usually left aside. As a consequence, we still lack a satisfactory mechanistic framework able to explain the evolution of phenotypes through the complete chain of biological causality: the effects of genetic changes on biophysical properties of molecules; the effects of biophysical modifications on the architecture of gene regulatory networks; and the effects of developmental changes on organismal phenotypes (Fig. 4.1).

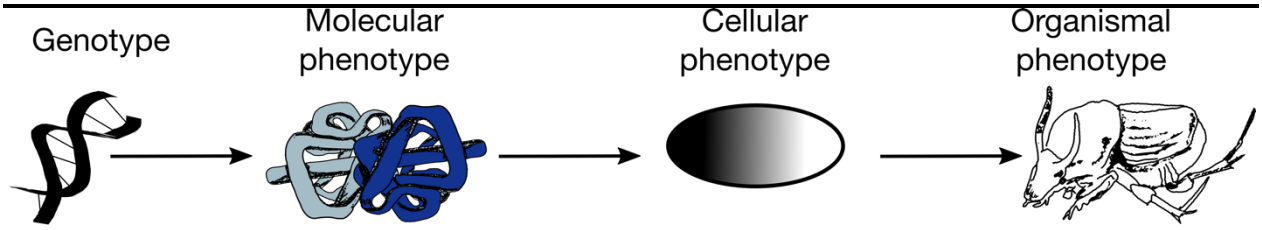


Figure 4.1. Biological chain of causality underlying phenotypic change. The genotype-phenotype map represented as a chain of connected phenotypes across multiple levels of biological organization For example, protein-protein interactions can be measured as the binding affinity of dimerization at the molecular level. Changes in protein dimerization of a transcription factor can alter the architecture of gene regulatory networks, and regulate gene expression by controlling the timing and amount of downstream gene product within cells and the overall patterns of expression in the embryo. Finally, the specific pattern of gene expression gives rise to an organismal feature.

In this perspective, we propose a unified framework integrating evolutionary biochemistry and developmental evolution to explain the historical processes and mechanisms underlying phenotypic evolution. First, we argue that the research agendas of both fields

converge on fundamentally similar questions, employing shared conceptual frameworks, albeit at different biological scales. This overlap provides fertile ground for building a conceptual bridge between the two disciplines. We synthesize theoretical and experimental findings from both fields to show that biological systems, regardless of scale, share fundamental principles governing their capacity for change. These principles arise from common architectural features that shape how systems explore phenotypic spaces through mutation, drift, and selection. Second, we introduce the paradigm of multi-scale genotype-phenotype maps to demonstrate how integrating these fields can enrich our mechanistic understanding of phenotypic evolution by identifying specific questions where one field can inform and complement the research agenda of the other. Finally, we propose exciting directions for future research programs that lie at the intersection of evolutionary biochemistry and developmental evolution.

4.3 A common research program to study phenotypic evolution

4.3.1 The origins and modification of phenotypes

Developmental evolution and evolutionary biochemistry share a key methodological foundation: a phylogenetically explicit, system-based approach to understanding phenotypic evolution. Both fields study traits in the context of their evolutionary history, seeking to explain how specific modifications to the structure or function of a system lead to the emergence of phenotypes. In evolutionary biochemistry, this involves uncovering the processes and mechanisms by which macromolecules acquire new structural and functional features—such as binding affinity, enzymatic catalysis, or multimeric states—and how these features diversify over time (122). Similarly, developmental evolution investigates how organisms modify preexisting traits or

acquire entirely new ones, such as novel cell types or organs, and explores the evolutionary mechanisms driving these changes (31, 123).

A shared goal in both disciplines is to produce a mechanistic explanation of phenotypic evolution—one that reconstructs the sequence of genetic or developmental events by which an ancestral phenotype transformed into its derived counterpart (Figure 4.2). This paradigm involves three essential steps: (1) identifying what phenotypic transformation happened, (2) formulating a hypothesis about the direction of change based on the phylogenetic distribution of the trait, and (3) experimentally testing the hypothesis to uncover the genetic and molecular basis for the observed transformation. Common questions guiding this approach include: What genetic changes are responsible for transforming an ancestral phenotype into a derived one? How many changes were required, and what were their individual and combined effects? And finally, given the genetic architecture of the trait, what evolutionary processes—such as selection or drift—likely drove these changes? We illustrate this paradigm with a well-established example from each field.

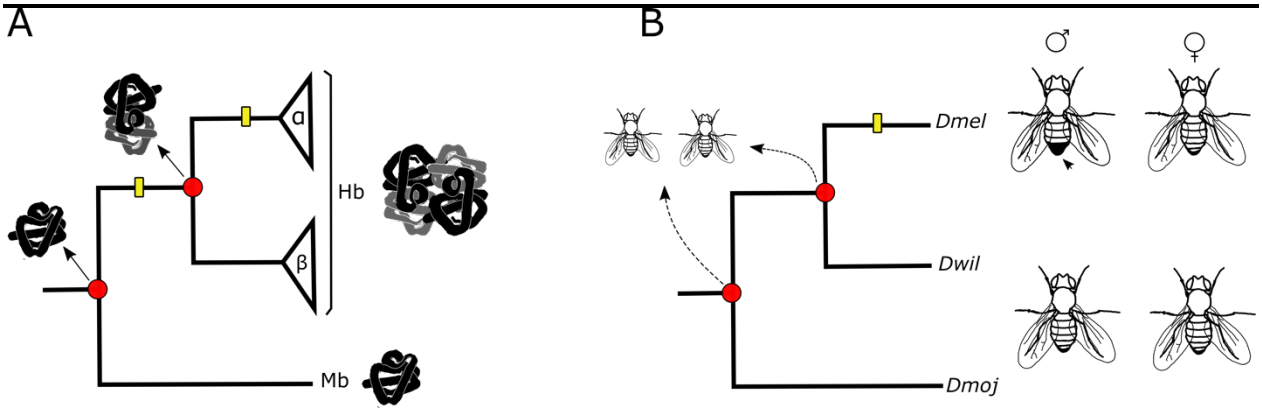


Figure 4.2. Basic research program to study phenotypic evolution. (A) Evolution of tetrameric structure in hemoglobin (Hb) from ancestral monomer. Other globin proteins, like myoglobin (Mb), have the ancestral monomeric state. (B) Evolution of dimorphic, male-specific abdominal pigmentation in *Drosophila melanogaster* (*Dmel*) from a monomorphic ancestor. Other fly species, like *D. willistoni* (*Dwil*) and *D. mojavensis* (*Dmoj*) have the ancestral monomorphic state. Red circles denote ancestral nodes; arrows from red circles point to

experimentally reconstructed ancestral states (solid arrows) or inferred ancestral states (dashed arrows); yellow bars denote genetic and developmental changes that occurred along the branch where the new phenotype evolved.

Evolution of a new protein architecture – Vertebrate globins, a family of proteins that mediate oxygen transport, exist in two main architectures: a monomeric form, where a single globin subunit binds oxygen (e.g., myoglobin), and a tetrameric form, where four subunits cooperatively bind oxygen (e.g., hemoglobin). In the tetrameric structure, oxygen binding to one subunit increases the binding affinity in the others. Pillai et al. (124) combined phylogenetics, ancestral sequence reconstruction, and biophysical experiments to investigate the evolution of this tetrameric architecture (Figure 4.1a). By reconstructing ancestral globin sequences, they showed that hemoglobin's tetrameric state evolved from an ancestral monomeric state, passing through a dimeric intermediate state. Furthermore, they show that just two substitutions were sufficient to cause the assembly of the ancestral dimers intro tetramers in the ancestor of jawed vertebrates.

Evolution of a novel color pattern – In the sexually dimorphic *Drosophila melanogaster*, males exhibit complete pigmentation in the two most posterior abdominal plates, whereas in females, pigmentation is restricted to the abdomen's posterior tip. In other *Drosophila* species, both sexes are monomorphic. Williams et al. (125) used genetic manipulations and phylogenetics to uncover when and how this dimorphic pattern evolved (Figure 4.1b). They found that dimorphic pigmentation arose approximately 30 million years ago from a monomorphically pigmented ancestor resembling the female state. The evolution of dimorphism can be recapitulated by changes in two cis-regulatory elements (CRE): One change increased the expression of a pigment repressor in females, preventing pigmentation of their abdominal plates

(125), while the other change increased the expression of a pigment-promoting enzyme in the abdomen of males (126).

By integrating genetic, molecular, and phylogenetic approaches, researchers in both fields reveal how ancestral phenotypes are modified over evolutionary time. This approach has uncovered three overarching principles that underlie the evolution of diversity across biological scales.

First, phenotypic evolution can occur via few changes by exploiting ancestral features of the system. For example, the two substitutions that enabled the transition of hemoglobin from a dimer to a tetramer did so by creating multiple favorable interactions with residues in the other subunits that were already present. Similarly, the genetic modifications in the two CREs that established the dimorphic pigmentation pattern involved the recruitment of conserved transcription factors into a novel gene regulatory network. This process of structural tinkering—modifying and co-opting ancient elements of a system to generate new features—has been observed across a wide range of systems and phenotypes (127–131). The second principle stems from the first: the genetic and developmental changes driving phenotypic transitions are contingent on the preceding state of the system, reflecting the continuous process of tinkering.

Lastly, the first and second principles establish that the internal architecture of biological systems is a causal factor in their evolution (1, 5, 122). Rather than being mere collections of independent parts shaped solely by external pressures, biological systems are tightly integrated entities. Their structure and function emerge from the interdependence of multiple components. Organisms are not simply assemblages of separate body parts (9), gene regulatory networks (GRNs) are not random aggregations of genes (132), and macromolecules are far more than

linear "strings of letters" (122). In each case, the system's architecture determines which components are most susceptible to change, which profoundly influences their evolution.

4.3.2 The distribution of phenotypic variation and its causes

A second shared goal between developmental evolution and evolutionary biochemistry is to explain the patterns of distribution of phenotypic variation in nature. Across scales, two well-established observations have emerged. First, that variation within biological systems is not distributed randomly; some parts of the system exhibit greater variability than others (11, 133–136). Second, that the collection of observed phenotypes in nature is highly sparse and nonrandom—many conceivable phenotypes never materialize or remain exceedingly rare, and those that exist are restricted to specific lineages (23, 27, 28, 43, 47, 76, 118, 137). These observations motivate two central questions: Why do biological systems vary in specific ways? And why are some phenotypes prevalent while others remain rare or absent? The first question examines how the structure of a biological system enables or restricts variation in certain components, while the second investigates how the system's architecture shapes the production of phenotypic variation according to its structural "rules." Together, these questions frame the second major shared goal of both fields.

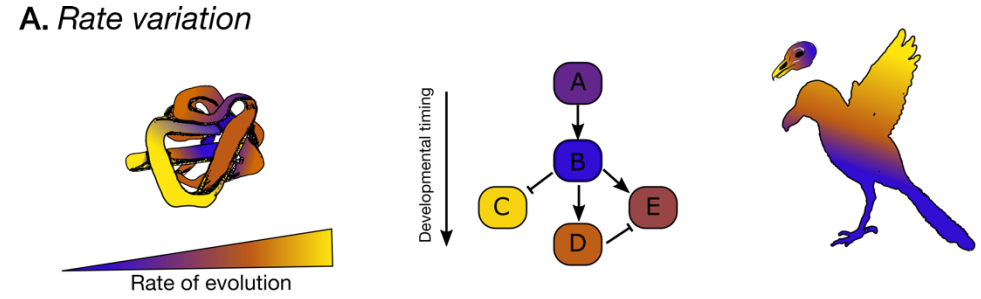
While natural selection plays a role in shaping both observations—uneven variation among parts and sparsity of phenotypes—it alone is not sufficient. The genotype-phenotype (GP) map can contribute to these patterns through at least three non-mutually exclusive ways. First, biases in mutation rates influence the likelihood of generating specific genetic variants, resulting in uneven opportunities for phenotypic innovation (14, 138). Mutational biases shapes evolutionary trajectories by determining the availability of alleles and mutational pathways

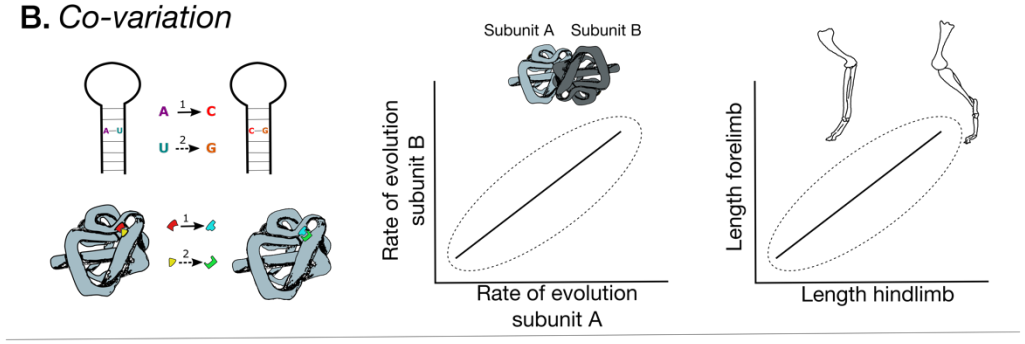
(138–140). Second, the processes that translate genetic changes into phenotypic variation, such as development and biochemistry, are not random—they tend to favor the production of certain phenotypes over others (22, 26). Additionally, mutational biases and differential production often interact, reinforcing directional trends in evolutionary change (25, 30, 139). Lastly, the organization of phenotypes across the space of genotypes creates asymmetries in the accessibility of phenotypes. Some phenotypes are more easily reachable from particular genetic starting points, limiting the evolutionary pathways available (30, 69, 141).

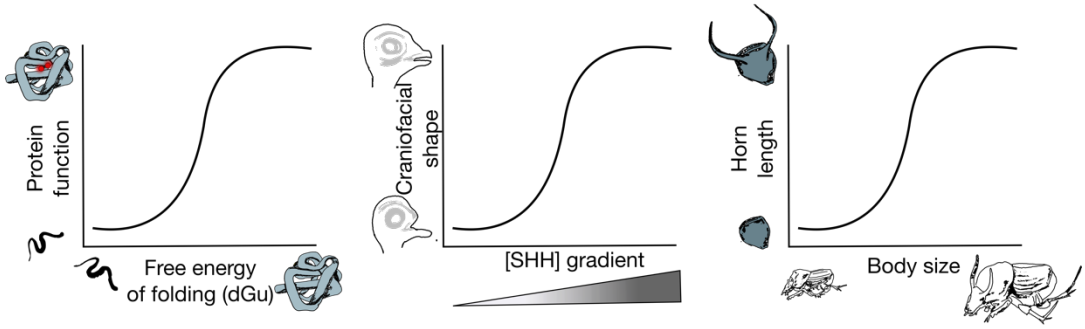
Together, these phenomena—differential introduction, production, and access to variation—shape phenotypic distributions and reflect fundamental structural features of biological systems. In the following sections, we show that, by comparing the patterns of variation found in macromolecules, GRNs and bodies, clear regularities emerge. All biological systems share five fundamental mechanisms of phenotypic change: rate-variation, co-variation, global co-variation, redundancy and entrenchment (Fig. 4.3). As a result, phenotypic evolution across biological scales also follows similar patterns.

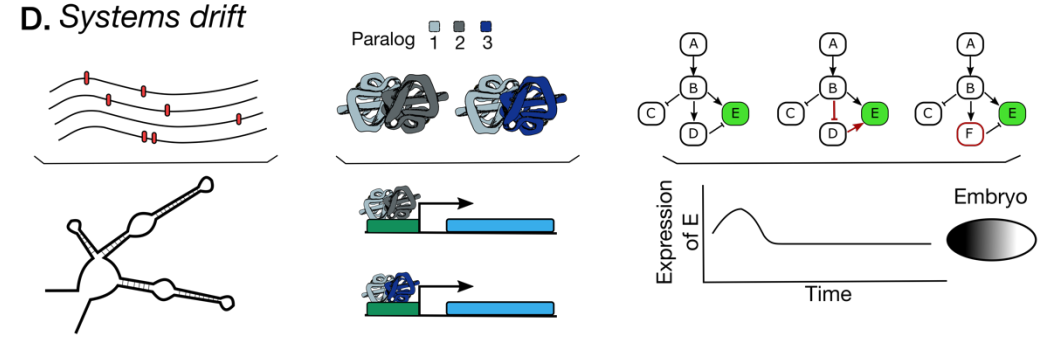
Rate-variation among parts – An emerging general pattern across biological systems is that not all parts—whether sites in a macromolecule, genes in a GRN, or body parts—evolve at the same rate. Structural and functional features of these systems can lead to differences in evolutionary rates among parts (Fig. 4.3a).

In globular proteins, sites closer to the protein core evolve more slowly than those on the surface, and sites in the core tend to be hydrophobic, while surface-facing sites are more likely to be hydrophilic (135, 142) (Figure 4.3a, left). These patterns arise from fundamental physicochemical constraints essential for proper protein folding. Hydrophobic amino acids are buried in the core to avoid water contact and prevent destabilizing thermodynamic interactions.


A. Rate variation
Rate of evolution
Developmental timing
A
B
C
D
E


B. Co-variation
A 1 C
U 2 G
Subunit A
Subunit B
Rate of evolution subunit B
Rate of evolution subunit A
Length forelimb
Length hindlimb


C. Global nonlinearities
Protein function
Free energy of folding (dGu)
Craniofacial shape
[SHH] gradient
Horn length
Body size


D. Systems drift
Paralog 1 2 3
A
B
C
D
E
F
Expression of E
Time
Embryo


E. Entrenchment
A
B
C
D
E

Figure 4.3. Shared biological mechanisms of phenotypic change across scales. (A) Evolutionary rate variation among parts of the system: Left, sites within a protein. Center, genes within a GRN. Right, body parts within a body. **(B)** Co-variation between parts of the system: Left,nucleotide sites within a stem structure in an RNA molecule and amino acid sites a protein. Center, subunits of interacting protein complexes in networks. Right, length of fore- and hindlimbs in most placental mammals. Co-variation results from epistasis, pleiotropy and phenotypic integration. **(C)** Global nonlinearities arising from the inherent coupling of two phenotypic features: Left, Free energy of folding and protein function. Center, a concentration gradient of the morphogen *sonic hedgehog* (SHH) and craniofacial shape in vertebrates. Right, allometric scaling of body size and horn length in dung beetles. **(D)** Systems drift arising from a many-to-one mapping between genotype and phenotype: Left, different RNA sequences fold into the same secondary structure. Center, different transcription factor complexes, resulting from gene duplication, can regulate the same target gene. Right, GRNs with different topologies, resulting from the modification of gene interactions, can result in the same stable pattern of gene expression. **(E)** Entrenchment of some features due to interdependence: Left, a substitution in a protein becomes entrenched (yellow) if it masks the deleterious phenotypic effect of a second substitution (purple). Right, a gene within a GRN becomes entrenched (yellow) as more interactions build upon it.

This constraint leads to stronger purifying selection at core sites, contributing to their slower evolutionary rates (142, 143). Protein function also influences evolutionary rates. In enzymes, evolutionary rates increase with the distance from the nearest catalytic residue in the active site, suggesting that long-range interactions among amino acids further constrain evolution to preserve protein function (144).

Genes within GRNs display a similar pattern. Highly connected genes, which act as hubs in the network, evolve more slowly than peripheral genes. Additionally, genes expressed earlier in development tend to evolve more slowly than those expressed later (136) (Figure 4.2a, center). These patterns reflect the hierarchical structure of GRNs. Hub genes, which integrate inputs and outputs from multiple genes, are under stronger purifying selection due to their broader pleiotropic effects, leading to slower rates of evolution (145). Similarly, early-expressed genes often have broader pleiotropic effects, influencing multiple downstream genes, while later-expressed genes typically affect fewer genes (146, 147). Consequently, there is a positive

relationship between the timing of gene expression during development and the rate of protein evolution (148, 149), although the specific roles of purifying and diversifying selection on early versus late-expressed genes remain debated.

In body parts and regions within those parts, a phenomenon called mosaicism occurs, where different traits evolve at different rates. This results in morphological features that contain both ancestral and derived characteristics (150–152) (Figure 4.3a, right). The causes of this phenomenon are still under investigation, but evidence suggests that rate variation can be influenced by the developmental origins of a trait. Traits derived from multiple cell populations or tissues tend to evolve more quickly (150, 153). In other cases, mechanical constraints, such as those required to produce functional joints in the vertebrate jaw, or ecological demands, like the shape of a bird's wing, can lead to rate variation driven by both purifying and diversifying selection (151, 152).

Rate-variation among different parts of a system has a major implication for phenotypic evolution: phenotypic variation will be overrepresented in parts that change more rapidly. Evolutionary change is likely to be more influenced by phenotypic changes arising from fast-evolving parts, rather than those evolving slowly, simply because fast-evolving parts explore more phenotypic variation.

Co-variation: Epistasis, pleiotropy and integration – Biological systems often exhibit correlated changes among their parts (Figure 4.3b). These correlations arise due to genetic or developmental interactions between traits. Structural constraints can favor correlated changes to maintain a functional architecture, while genetic and developmental processes may cause multiple traits to change together.

In macromolecules like proteins and RNA, the 3D structure is shaped by genetic interactions between sites, a phenomenon known as epistasis. Because a macromolecule's biological function depends on its structure, there is strong selection to preserve this functional architecture. Mutations that destabilize the structure often require compensatory changes at interacting sites, resulting in co-variation between them (154, 155). As a result, protein 3D structure can often be predicted from co-variation patterns between sites, as strongly co-varying sites are likely to physically interact in the folded protein (156) (Figure 4.3b, left). Furthermore, in RNA secondary structures, sites involved in Watson-Crick-Franklin (WCF) pairings—a simple form of epistasis—have lower polymorphism and higher levels of compensatory coevolution compared to sites that do not form WCF pairs (Figure 4.3b, left) (155, 157).

This pattern of co-variation extends to interacting genes within genetic networks (Figure 4.3b, center). Proteins often form complexes essential for cellular function (e.g., transcription factors, membrane channels or transporters). Selection acts on the integrity of these complexes, both by maintaining interactions between components and by ensuring proper stoichiometry. Mutations that disrupt the complex—whether at an interface or through changes in expression—are often followed by compensatory changes in interacting partners. This leads to correlated rates of evolution at both the sequence and gene expression levels (158, 159).

Co-variation also occurs at the organismal level. Phenotypic integration—the correlated changes between body parts—reflects shared underlying genetic and developmental programs (115). For example, the strong co-variation between homologous bones in the fore- and hindlimbs of quadrupedal mammals reflects the fact that both structures are serial homologs, sharing a substantial part of their developmental programs (160, 161) (Figure 4.3b, right).

Phenotypic integration therefore arises from pleiotropy, where changes in genes influencing multiple traits result in correlated changes.

Co-variation, driven by epistasis and pleiotropy, has three major implications for phenotypic variation. First, co-variation leads to nonindependence of biological parts, introducing complex nonlinearities into the mapping from genotype to phenotype (11, 103, 117, 162). Nonlinearities arising from interacting components is a pervasive feature of biological systems and is critical for understanding phenotypic variation in their structure and function from a mechanistic perspective (61, 121, 163).

Second, co-variation increases the range of phenotypic variation that a biological system can produce, compared to a scenario where each part changed independently from each other. For example, intra- and intermolecular molecular epistasis increases the number and connectivity of genotypes that produce different phenotypes, facilitating the evolution of new functions or the production of intermediate gene-expression phenotypes (106, 164). Similarly, phenotypic integration increases the magnitude of trait changes along certain directions, supporting the evolution of extreme morphologies (115).

Lastly, co-variation strongly affects the response to selection. Co-variation arising from epistasis introduces strong contingency in evolutionary change, because the effect of a change in one part depends on the state of its interacting part. Epistasis therefore shapes the possible paths that evolution can take and the rate at which a population can respond to selection (120, 165, 166). Similarly, co-variation arising from pleiotropy introduces strong biases in the direction of evolution, even under strong selection, because genetic changes cause simultaneous changes in multiple traits. Pleiotropy therefore shapes long-term responses to selection and phenotypic divergence (16, 25).

Global nonlinearities – Beyond the nonlinearities caused by interactions among parts, intrinsic nonlinear relationships between phenotypic features of a system can also give rise to nonadditivity (Figure 4.3c). In macromolecules, this phenomenon is illustrated by global epistasis. Global epistasis emerges from the nonlinear relationship between a molecule's biophysical properties (e.g., stability, solubility, or ligand affinity) and its biological properties (e.g., function or fitness) (117). A classic example is the sigmoidal relationship between protein folding stability and molecular function, a form of phenotypic integration in proteins (167). This relationship reflects a threshold-like dependence of function on stability, where small changes in stability can have disproportionately large effects on function near the threshold (Figure 4.3c, left).

Global nonlinearities are also prevalent during embryonic development, often arising from threshold-like responses of morphological features to morphogen concentration or gene expression gradients (168). A well-studied example is craniofacial morphology in vertebrates, where variation arises from the proportion of cells in mitotic zones that respond to Sonic Hedgehog (Shh) and Fgf8 signaling (169, 170) (Figure 4.3c, center).

At the organismal level, allometry—the differential scaling of morphological and physiological traits with body size—represents another prominent example of global nonlinear relationships. Allometry, a special case of phenotypic integration, reflects the co-variation generated by genetic and developmental processes regulating size and is often modeled using power or sigmoidal functions (171–173). For instance, in male dung beetles (*Onthophagus taurus*), horn length exhibits a sigmoidal, threshold-like relationship with body size: males exceeding a critical body size develop long horns, while smaller males develop no horns or only rudimentary ones (174) (Figure 4.3c, right).

Global nonlinearities have three key implications for phenotypic evolution. First, these nonlinearities can modulate phenotypic variation by introducing contingency. A given change in one phenotypic feature (e.g., folding stability, gene expression, or body size) can produce drastically different outcomes in a coupled trait (e.g., function, facial shape, or horn length) depending on the ancestral trait value along the nonlinear curve. Second, nonlinearities can act as global biases, channeling phenotypic variation along the trajectory defined by the shape of the nonlinear function (173, 175, 176). Lastly, global nonlinearities introduce an additional layer of complexity to the GP map. Global nonlinearities are characterized by a many-to-one mapping where multiple mutations, genetic factors, and developmental mechanisms may contribute to variation in one or both coupled features. Conversely, because these relationships manifest at the phenotypic level, such nonlinearities could still emerge even when variation in each feature was completely additive at the genetic level (117, 171, 177).

Systems drift: Evolution through redundancy – Another prominent feature of the GP maps in complex systems is redundancy—the fact that multiple genotypes produce the same phenotype (Figure 4.3d). This arises because the number of possible genotypes vastly exceeds the number of phenotypes. For example, many RNA sequences can fold into the same secondary structure (23, 100) (Figure 4.3d, left), and the sequence determinants of protein function can undergo dramatic turnover during evolution while maintaining the same function (57, 178). Similarly, transcriptional circuits with different regulatory network topologies can produce the same steady-state pattern of gene expression (77, 179, 180) (Figure 4.3d, center and right), and homologous body parts are often regulated by divergent developmental programs—a phenomenon known as developmental systems drift (181, 182).

Redundancy has three key implications for phenotypic evolution. First, redundancy creates an "entropic" effect: more degenerate phenotypes are be more likely to evolve simply because mutations will often produce phenotypes encoded by many genotypes. Second, redundancy implies a moderate rate of turnover in the underlying genetic and developmental mechanisms of a trait, which can lead to a decorrelation between genotype and phenotype over evolutionary time (57). Finally, redundancy shapes the exploration of the phenotypic space through mutations because genotypes with the same phenotype often form vast neutral networks. Genotypes with exclusively neutral neighbors are more robust to change, reducing the probability of phenotypic diversification, while those surrounded by non-neutral neighbors are more likely to produce variation (24, 69, 183). Furthermore, the passive exploration these networks can drive periods of prolonged evolutionary conservation followed by rapid change (79, 183, 184).

Entrenchment and ground plans – Lack of variation, or phenotypic conservation, is a pervasive evolutionary pattern across scales (5). In developmental evolution, it is exemplified by the concept of the ground plan (or Baüplan), a set of key developmental and morphological features shared by phylogenetically related organisms (185–187). For example, the anterior-posterior axis in nearly all animals is specified by the expression of Hox genes, a defining feature of the metazoan ground plan (123, 188). Similarly, in evolutionary biochemistry, proteins with conserved arrangements of secondary structures are grouped into structural families or folds (137). All bilaterian nuclear hormone receptors (NRs), for example, share a Cys2Hys2 zinc-finger fold.

While organismal and molecular ground plans reflect deeply conserved developmental or structural features, they do not imply the absence of alternative ways for building biological

systems. Neither the Hox code nor the Cys2Hys2 fold represents the only way to build complex organisms or DNA-binding proteins. For instance, flower development relies on MADS-box genes, entirely unrelated to the Hox code but still conserved across all angiosperms (189, 190). Similarly, different conserved folds are capable of DNA binding, such as the Leucine Zipper or Helix-Turn-Helix folds found in many transcription factors. The diversity of ground plans thus prompts a fundamental question: what makes a feature to become conserved during evolution? In other words, how do ground plan features emerge during evolution? Purifying selection explains why conservation is maintained, but it does not explain why a feature became essential. Increasing evidence suggests that ground plan features can be established through tinkering and subsequent entrenchment—the process by which initially neutral changes become indispensable as systems evolve around them (130, 191).

At the molecular level, entrenchment stems from epistatic interactions. Mutations that were neutral when they fixed can become increasingly deleterious to revert or change over time because they modify the phenotypic effects of other states in the protein. Epistatic dependence on preceding substitutions can therefore "lock-in" amino acid states as more interactions build around them (178, 192) (Figure 4.3e, left). For example, despite a fair amount of sequence divergence, all bilaterian NRs have four cysteine residues that are completely conserved because they coordinate a zinc atom which is essential to retain a stable domain structure, folding, and DNA-binding activity (193)—changing any of these residues would break the entire fold. Entrenchment can also operate at developmental and organismal scales (191, 194, 195). A character can become epistatically locked-in as subsequent gene interactions and developmental processes build hierarchically upon them (Figure 4.3e, right). For example, the notochord—a key signaling center for the development of axial structures in chordates—is a defining feature of the

vertebrate ground plan. The same structure, however, is variably present in ascidians, implying that it is less developmentally integrated and therefore dispensable (195). In both cases—macromolecules and organisms—entrenched features, due to the continuous accretion of parts, are therefore more likely to become ground plan characters and remain conserved over time.

Entrenchment has three major implications for phenotypic evolution. First, it limits the system's ability to explore new phenotypic variations within the constraints of the ground plan—entrenched features are for the most part invariable. Second, due to these constraints, phenotypic variation within ground plans is expected to be smaller than variation between ground plans, leading to clumped, discontinuous patterns of variation (28, 70). Third, conservation does not need to imply function—entrenched features will lead to patterns of conservation even if they were functionally inconsequential when first became established.

4.4 Multi-scale GP maps: Phenotypic evolution across biological scales

Thus far, macromolecules, GRNs, and bodies have been treated as independent entities, studied separately within each field. This approach allowed us to show that both fields share a common conceptual toolkit to study phenotypic evolution despite the different scales, that the same biological mechanisms driving phenotypic change exist across scales, and that these mechanisms result in similar patterns of phenotypic evolution. In this section we introduce an integrative, hierarchical approach to build GP maps that account for the propagation of phenotypic variation across biological levels.

The proposed GP map has two components. The first maps genotype to molecular phenotype, describing how mutations affect molecular traits via biophysical and biochemical mechanisms—such as how an amino acid change influences the free energy of binding in a

protein dimer. The second maps molecular phenotype to developmental outcome, linking biophysical changes to developmental processes through gene regulation and cell biology—for instance, how dimeric protein binding shapes tissue-specific gene expression and drives cell differentiation.

Integrating phenotypic effects into a multi-scale GP map will deepen our understanding of the mechanisms driving phenotypic evolution. It would clarify the chain of biological effects producing organismal traits, whether certain types of molecular phenotypes are more likely to drive organismal change, identify historical trajectories of sequence evolution and how their effects propagate across scales, and reveal how the relationships between scales shape the accumulation of variation at different scales.

4.4.1 Revealing the first map: the biophysical basis of developmental evolution

Molecular evolution and biochemistry are uniquely positioned to construct the first genotype-phenotype map by addressing a key question: what are the biochemical and biophysical causes of developmental changes? While developmental evolution has identified many underlying developmental events driving organismal phenotypic change, the challenge remains to uncover how these developmental changes arose. Development, mediated by gene regulatory networks, relies on molecular interactions among gene products (e.g., protein-DNA or morphogen-receptor binding). These interactions are governed by biophysical properties such as binding affinity, specificity, and folding stability. Consequently, the evolution of development—via modifications to gene regulatory networks—reflects changes in the biophysical properties of macromolecules.

In this section, we highlight examples where molecular evolution and biochemistry have elucidated the origins of specific developmental changes. Although not exhaustive, these cases

illustrate the breadth of organismal phenotypes that can be explored through this lens. We also discuss broader implications of these findings for understanding phenotypic evolution.

Modification of a phosphoswitch during the evolution of pregnancy –A pivotal evolutionary innovation in Eutherian pregnancy is the differentiation of endometrial stromal fibroblasts into decidual stromal cells (DSCs) (196). DSCs produce prolactin, a key hormone regulating numerous physiological functions, including milk production and mammary gland development. The recruitment of prolactin expression in DSCs marked a critical step in the evolution of pregnancy. Prolactin expression in DSCs is directly regulated by the mammalian transcription factor CEBPB. But how did CEBPB acquire this derived role in eutherians?

Lynch et al. (197) investigated this question using ancestral sequence reconstruction and *in vitro* assays to analyze the biochemical impact of specific amino acid substitutions in CEBPB along the lineage leading to Eutherian mammals. They found that three amino acid changes reorganized the location of critical phosphorylation sites. This reconfiguration transformed CEBPB from a repressor to an activator upon phosphorylation by a kinase expressed in DSCs, enabling it to positively regulate prolactin expression and drive this developmental innovation.

Degradation of a phosphoswitch during the evolution of the insect body plan – The evolution of novel body plans is a key driver of morphological diversity. Around 400 million years ago, six-legged insects diverged from a crustacean-like arthropod ancestor with multiple limbs along their body segments. A critical step in shaping the insect body plan was the repression of thoracic leg development, regulated by the Ultrabithorax (Ubx) and Abdominal-A Hox proteins. Unlike its counterparts in velvet worms and crustaceans, the insect Ubx protein gained the ability to repress limb primordia development (198, 199). But how did this repressive function evolve?

Ronshaugen et al. (199) combined phylogenetic analysis with *in vivo* expression assays to uncover the biochemical mechanism behind Ubx's limb-repression role in insects. They found that the C-terminal region of velvet worm and crustacean Ubx proteins contains a conserved phosphoswitch. Upon phosphorylation by a kinase, this switch inhibits a latent limb-repression domain present in arthropod Ubx proteins. In insects, however, substitutions to glutamine and alanine eroded this phosphoswitch, rendering Ubx unresponsive to phosphorylation and exposing its repressive function, which was crucial for the evolution of the six-legged insect body plan.

Protein-complex formation during the evolution of flowers – The emergence of angiosperms, around 140–250 million years ago, marked a major evolutionary innovation: the flower, a reproductive structure integrating both male (stamen) and female (carpel) organs (200). This developmental shift from the separate male and female cones of gymnosperms was a key step in angiosperm evolution. The specification of the four basic floral organs—sepals, petals, stamens, and carpels—relies on unique tetrameric protein complexes known as floral quartets, composed of MADS-box transcription factors, and each quartet is formed by a specific combination of four protein classes (A, B, C, and E). Stamen identity is governed by $B+C+E_2$ quartets, while carpels are specified by C_2+E_2 quartets (189).

In gymnosperms, however, only two simpler quartets exist: B_2+C_2 and C_4 complexes, specifying male and female cones, respectively. So how did the novel floral quartets evolve, enabling the integration of both reproductive organs within a single structure? Ruelens et al. (201) used ancestral sequence reconstruction and *in vitro* protein-protein interaction assays to identify three key changes in protein binding specificity, after a series of gene duplications, underlying the formation of stamen and carpel quartets at the origin of angiosperms. First, C-class proteins partially lost their ability to form homomers. Second, C-class proteins lost their

direct interaction with B-class proteins. Third, the interaction between B and C proteins became mediated by E-class proteins, a hallmark of angiosperm evolution. These changes made E proteins an essential scaffold for carpel and stamen quartets, enabling the combinatorial assembly of these complexes with distinct regulatory roles, ultimately allowing both reproductive organs to coexist within the same floral structure.

DNA-binding affinity and diversification of leaf shape – Plants in the Brassicaceae family exhibit remarkable variation in leaf shape. For example, the Thale cress (*Arabidopsis thaliana*) produces simple leaves, whereas its close relative, the Hairy bittercress (*Cardamine hirsuta*), develops complex leaves divided into distinct subunits called leaflets. This evolutionary divergence in leaf shape was driven by the recruitment of the Homeobox gene RCO (202), which sculpts leaflets by inhibiting growth at their flanks. But how did RCO acquire its leaf-shaping regulatory specificity? Hajheidari et al. (203) combined comparative genomics, *in vitro* protein-DNA affinity assays, and *in vivo* experiments to uncover the mechanism. They found that RCO establishes growth-inhibiting domains in developing leaflets through autoregulation, repressing its own transcription. This autoregulatory capacity evolved via changes in RCO's cis-regulatory elements, which reduced its DNA-binding affinity, fine-tuning its expression and precisely defining the regions of cellular growth inhibition.

Evolution of development and propagation of phenotypic effects – The examples discussed illustrate how biochemical and biophysical modifications shape the evolution of development, offering two key insights.

First, macromolecules influence developmental evolution through a diversity of biophysical mechanisms. This diversity stems from the vast array of genetic interactions within GRNs, providing numerous opportunities for biophysical changes to drive evolutionary shifts.

However, the evolution of pregnancy in mammals and the insect body plan—vastly different phenotypes in clades that diverged over 500 million years ago—also highlight common simple mechanisms. Developmental changes in each case were driven by modifications of phosphoswitches in regulatory proteins. These modifications provide a simple yet effective means of generating novel regulatory responses in transcriptional circuits, underscoring the potential for shared biochemical strategies across evolutionary contexts.

Second, even relatively modest biochemical and biophysical changes can have far-reaching macroevolutionary consequences, reflecting the intricate propagation of phenotypic effects. For instance, the origins of insects and flowering plants—two of the most successful evolutionary radiations—were accompanied by shifts in protein regulation and specificity. These changes profoundly reshaped phenotypic possibilities. The diversification of MADS-box genes and the assembly of floral quartets, for example, introduced entirely new axes of morphological variation, enabling flowering plants to evolve the perianth (petals) and integrate male and female reproductive organs within a single structure. This innovation expanded the morphospace available to flowering plants, granting them unique dimensions of phenotypic variation inaccessible to other clades.

In addition to creating new axes of variation, biophysical modifications strongly influence how clades explore phenotypic variation within existing dimensions. For example, regulatory changes via modification of Ubx proteins modulate the number and placement of paired appendages in insects, while regulatory changes via protein-DNA affinity in *RCO* sculpts leaf shape in plants. It is therefore likely that many other patterns of variation in organismal traits—both discrete and continuous variation— are modulated by “tunable” molecular switches.

4.4.2 Revealing the second map: the developmental effects of biophysical variation in macromolecules

Developmental evolution is uniquely positioned to construct the second GP map by addressing a fundamental question: what are the developmental effects of biophysical changes in molecules? Modifications to the biochemical and biophysical properties of developmental genes—whether in enhancers, promoters, or proteins—can drive changes in cell behavior (e.g., movement, differentiation, communication, proliferation), generate novel expression patterns, or alter adult morphology. Characterizing how a change in a macromolecule's properties maps to developmental outcomes is essential for understanding how phenotypic variation is transformed along the GP map.

In this section, we highlight two key examples where developmental and cell biology have elucidated the phenotypic effects of molecular phenotypic variation. We also discuss broader implications of these findings for understanding phenotypic evolution.

Variation in protein-DNA affinity and digit number in vertebrates – Although most vertebrates develop limbs with five digits, deviations in digit number—such as polydactyly—are common in humans, and fossil evidence reveals that stem tetrapods had seven or eight digits (204). Digit development in the limb bud is governed by the enhancer ZRS, which regulates the expression of *Shh* (Sonic hedgehog) in the posterior limb bud—a morphogen crucial for establishing digit number and identity (123, 205). Lim et al. (205) demonstrated that transcription factor binding sites within the ZRS enhancer are typically of low affinity, and mutations that increase binding affinity result in polydactyly by increasing *Shh* domain. The severity of phenotypes was modulated by changes in protein-DNA affinity: mutations that

produced a greater increase in affinity led to an increase in the number of digits and/or an increase in the number of phalanges.

Variation in tandem repeats and the diversification of facial length in placental mammals – Mammals exhibit remarkable craniofacial diversity, with variation in facial length being particularly striking. Most facial bones in mammals form through intramembranous ossification, a process where mesenchymal cells differentiate directly into osteoblasts. This differentiation is regulated by the transcription factor Runx2 (206). Upregulation of Runx2 accelerates and extends bone development, whereas downregulation truncates it. Sears et al. (207) identified that variation in the length of Runx2's polyglutamine (polyQ) and polyalanine (polyA) tandem repeats—a highly variable feature of Runx2—is associated with macroevolutionary changes in facial length. The transcriptional activity of Runx2 was influenced by the polyQ:polyA ratio: increasing polyQ repeats enhances gene expression by strengthening interactions with the cofactor CBFβ, but an excess of polyQ or polyA repeats causes protein aggregation, reducing transcriptional activity (206, 207). Variation in the polyQ:polyA ratio is positively correlated with facial length in carnivoran mammals, suggesting that these repeats act as modulators of facial morphology by altering the timing of ossification (207).

The developmental function – The examples discussed provide two key insights for understanding the evolution of development. First, developmental outputs can be mapped to variation in the biophysical properties of macromolecules. Changes in protein-DNA affinity within the ZRS enhancer modulates digit number, while changes in the strength of protein-protein interactions in Runx2 influence craniofacial length. These molecules, ZRS and Runx2, act as "tuning knobs" for organismal morphology, providing insight into a fundamental aspect of biological systems: the transformation of phenotypic effects between levels (11, 208). This

transformation—what we term the developmental function—defines the relationship between biophysical properties and developmental or morphological outcomes.

The shape of the developmental function depends on the biological processes that connect variation in one level to variation at a higher level. For example, the developmental function associated with the ZRS enhancer appears relatively monotonic: increased protein-DNA affinity generally leads to more severe polydactylous phenotypes. In contrast, the developmental function of Runx2 is nonmonotonic: increases in the polyQ:polyA ratio promote craniofacial elongation, but an excess of polyQ or polyA repeats suppresses this effect due to protein aggregation. Knowledge of the developmental function would therefore provide a direct mechanistic and quantitative model for predicting phenotypes from genotypes.

Second, knowledge of the developmental function could inform what pattern of phenotypic and genetic variation we should expect at the level of molecules. For instance, if polydactyly was deleterious in vertebrates, then purifying selection against increase-affinity mutations should maintain low levels of polymorphism in affinity. However, since many different mutations produce similar changes in affinity, polymorphism at the sequence level is likely to be higher. This is consistent with the pattern of variation found the ZRS enhancer (205). The interaction between the developmental function and selection can therefore be used to understand the accumulation of variation at different scales. Understanding the shape of these developmental functions—and providing a quantitative description of this relationship—is critical to unravel how variation propagates across biological scales and how phenotypic variation is transformed from one level to another.

4.5 Concluding remarks: Future directions and open questions

Developmental evolution and evolutionary biochemistry share a common framework for studying phenotypic evolution. Both fields reveal that molecules, gene regulatory networks (GRNs), and organisms exhibit share fundamental mechanisms of phenotypic change due to their tight structural and functional architectures. These shared mechanisms illuminate how biological systems produce phenotypic variation across all levels of biological organization. Furthermore, integrating approaches from both fields into multi-scale GP maps, our understanding of how variation propagates from molecular properties to developmental outcomes and organismal phenotypes is enriched. We conclude by identifying five promising research directions at the intersection of these fields.

Charting natural variation in molecular and developmental properties – A critical step in understanding how development evolves is to quantify the natural variation in molecular properties of key developmental regulators and map this variation to developmental outputs. While GRNs can be rewired through diverse biophysical mechanisms, it remains unclear how much polymorphism exists in these mechanisms and how it influences developmental outcomes. A systematic chart of natural variation in molecular phenotypes, coupled with their developmental effects, could reveal whether different biophysical mechanisms contain more polymorphism than others, whether certain mechanisms are more likely to drive developmental change than others or whether certain types of developmental changes are driven more by certain mechanisms than others.

Building developmental functions – A major goal in the mechanistic understanding of development is characterizing how variation from one level is transformed into variation at a higher level. This requires building quantitative models of phenotypic change integrating

structural mechanisms and developmental functions (170, 191, 208). Computational GP models based on cellular dynamics and molecular interactions have successfully explained natural morphological variation in mammalian teeth and leaf shape (22, 209) and pathogenic variation in heart cell physiology (210), by mapping morphological outcomes over a wide range of possible parameter values. However, these models still have two major limitations. First, molecular parameters are treated as genotypes instead of phenotypes. As we have argued, biochemical and biophysical features of macromolecules are phenotypes which directly arise from sequence variation. Second, due to their definition of genotype, they lack direct genetic information—genotypes do not represent nucleotide or amino acid sequences. As a result, these models cannot address how specific mutations alter molecular parameters in the model. Incorporating lower-level GP maps—from nucleotide or amino acid sequence to molecular phenotype—into these models would therefore allow for predictions of how mutations affect molecular and developmental traits, and how development could evolve via mutations.

Integrating such models with phylogenetic statistical frameworks could further enhance their utility. For example, combining knowledge from multi-scale GP maps with phylogenetic models could improve ancestral state reconstruction (ASR) and inferences of evolutionary trajectories. Recent efforts to incorporate developmental knowledge into statistical phylogenetic frameworks have been developed for modeling discrete phenotypic traits (211, 212). However, these models still rely on the statistical inference of underlying developmental or molecular parameters. Molecular ASR combined with experimental characterization of reconstructed sequences would provide direct measurement of ancient molecular phenotypes. This information could be further integrated into complete phenomenological models to predict ancient

organismal phenotypes, providing a more complete and mechanistic picture of evolutionary change.

Mapping the distribution of natural vs. possible variation – A deeper understanding of phenotypic evolution requires contrasting natural variation with the full spectrum of possible variation. For example, what is the distribution of phenotypes—biochemical, biophysical, and developmental—arising from all possible mutations, natural polymorphisms, or substitutions? Comparing these distributions can illuminate the interplay of historical contingencies, selection pressures, and intrinsic system constraints in shaping observed diversity (19, 72).

Unbiased screenings of genetic and phenotypic variation would also provide insight into the accessibility of developmental variation (69, 122). A key gap in macroevolutionary studies of phenotypic variation involving morphospaces is the lack of explicit genotype spaces (30). As a result, using morphospaces to interpret evolutionary transitions between morphologies may be misleading because there is not an explicit map to genetic changes. By mapping molecular and developmental phenotypes onto genotypic space, we would be able to identify genetic changes likely to produce specific phenotypes and infer the genetic and developmental trajectories required to transform one phenotype into another.

Achieving this goal, however, requires advances in high-throughput experimental approaches. *In vitro* assays for measuring molecular phenotypes of large genetic libraries, are currently optimized for limited properties, such as TF-DNA binding or protein-ligand interactions (51, 213). Similarly, high-throughput *in vivo* assays for assessing developmental effects rely on efficient transformation protocols, embryo screening, and imaging, which remain restricted to a few model systems (72). Expanding these tools to diverse molecular phenotypes

and non-model organisms will be essential for comprehensive studies of developmental variation.

Emergence and propagation of nonlinearities in developmental systems – Nonlinearities are pervasive in biological systems at every scale, but their hierarchical interconnections remain poorly understood. A key unresolved question is how nonlinearities at lower levels, like intra- and intermolecular interactions, propagate to higher levels, generating epistasis and phenotypic integration at organismal levels (214). For instance, how does intramolecular epistasis shape the topology of GRNs? How does the topology of GRNs affects the distribution of perturbation effects on developmental and organismal traits? Is molecular epistasis amplified or dampened as it propagates? Building multi-scale GP maps, by connecting variation across hierarchical scales, is critical to obtain quantitative models of phenotypic change based on developmental functions and to elucidate how nonlinearities emerge, propagate, and evolve across biological scales.

Origins of multicellular development – Complex development, involving the process by which a multicellular organism develops from a single cell, has evolved multiple times. However, the origins of multicellular development remain poorly understood (215). Multicellular development involves the activation of highly conserved signaling pathways (216). The assembly of these molecular circuits—rather than their modification—represents a major transition in multicellular organisms and, therefore, a key question to understand the origins of development is how do major signaling pathways assemble during evolution. Answering this question would require at least two things. First, revealing the evolutionary origins of signaling pathways requires a methodological transition from the traditional evolutionary-biochemistry approach of single-molecule phylogenetics and ASR to a multi-molecule approach. To characterize the evolution of a molecular circuit, all parts of the circuit should be studied

simultaneously. This requires addressing the phylogenetic history of each circuit's component and time-matching ancestral reconstructions across components—those representing the same phylogenetic node in the species tree. Second, understanding the evolution of a circuit requires experimentally characterizing the developmental effects of whole ancestral circuits. This would involve the implementation of developmental assays on transgenic organisms carrying reconstructed ancestral genes (217). A multi-molecule approach combined with *in vivo* assays would allow us to understanding the developmental effects of the circuit when it first became assembled, disentangle the individual developmental effects of each component, and identify the genetic and biophysical changes that caused the assembly of a new signaling pathway.

Chapter 5

Conclusion

The causal influence of the GP map on evolution has been difficult to establish because of four key challenges. First, its size is so vast that a comprehensive characterization is impossible. Second, even for partial GP maps, disentangling its influence from selection in shaping patterns of diversity has remained difficult. Third, production biases in present-day lineages may not reflect the biases that existed when the phenotypes evolved, further obscuring the link between the GP map's structure and evolutionary outcomes. Lastly, the mechanistic link between genetic architecture and the GP map's structure remains elusive, limiting our understanding of the genetic basis of production biases and their evolutionary consequences.

By experimentally characterizing an ancestral GP map of a molecular system within defined genetic and phenotypic scopes, I showed that the map is strongly anisotropic and heterogenous—there is unequal propensity to produce phenotypes and genotypes have unequal access to the encoded phenotypes. This structure strongly shaped the phenotypic outcomes that were likely to evolve from any genotype and biased the lineage-specific phenotypic outcomes that occurred during history. Furthermore, by dissecting the genetic architecture of the GP map, I uncovered a mechanistic link between epistasis and the structure of the map: epistasis reduces the anisotropy of the map but also increases its heterogeneity. Overall, these findings establish that the GP map is a causal factor in phenotypic evolution and show that its structure—and evolutionary consequences—can be understood by using appropriate models of genetic architecture.

Appendix 1

Supplementary figures and tables for Chapter 2: Bias in an ancient genotype-phenotype map caused the functional diversity of the steroid hormone receptor family

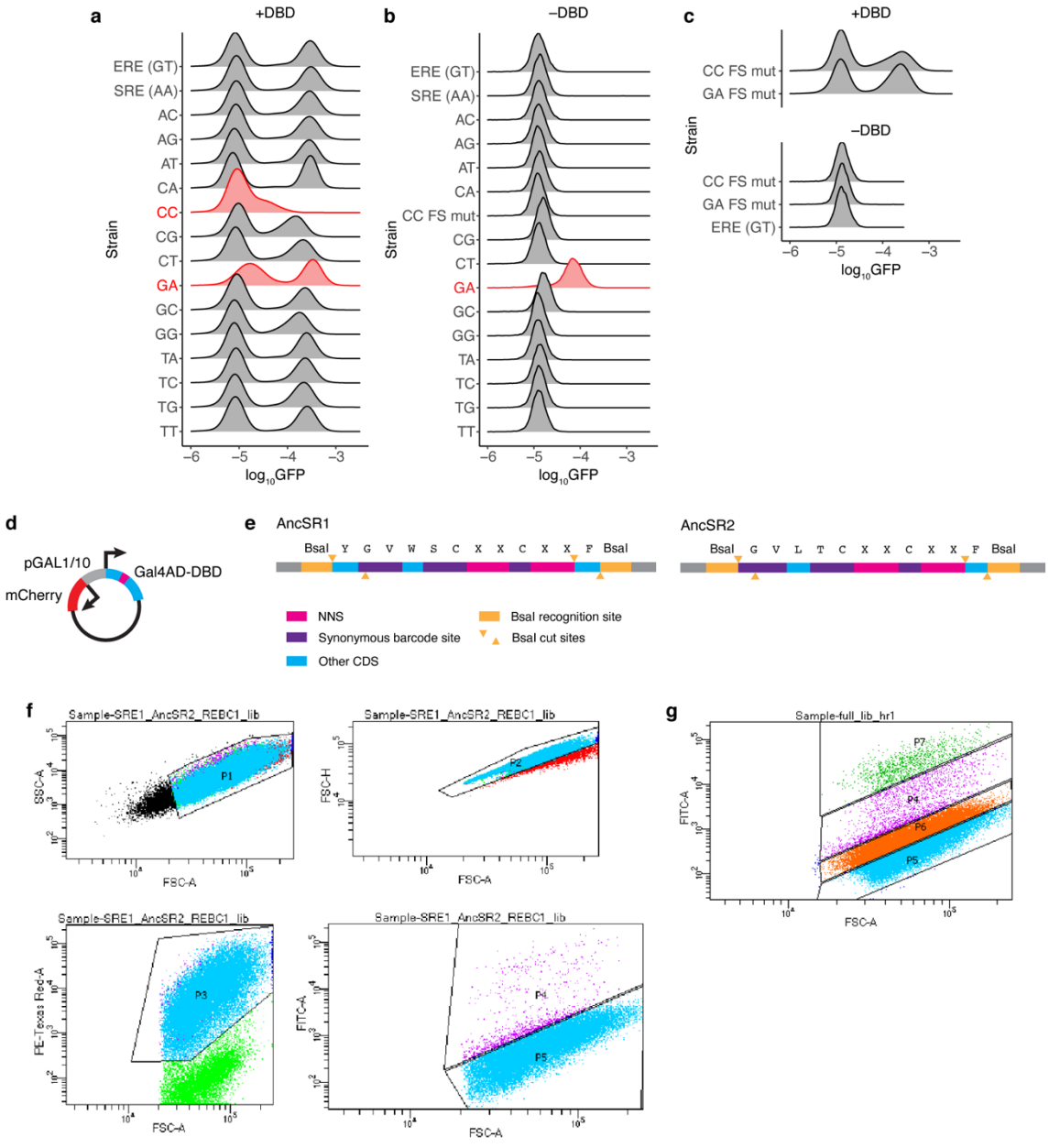


Figure A1.1. DBD library construction and sorting. a, Design of the DBD expression vector used for DMS. The SR DBD is fused to an N-terminal *S. cerevisiae* Gal4 Activation Domain. Its expression is under control of a bidirectional pGAL1/10 promoter, which simultaneously drives

mCherry expression to select cells that maintain the plasmid during sorting. **b**, Design of DBD library oligos. NNS codons (pink) were used to generate all possible combinations of amino acid mutations at the four RH scanning sites (marked as X in the amino acid sequence). For each background (AncSR1, left; AncSR2, right), we synthesized 16 libraries, each with a unique set of synonymous barcode mutations at five codons (purple, Supplementary Table 1), which allows each to be associated with one RE strain. BsaI sites (orange) were used for Golden Gate assembly into the pDBD2.1 backbone. **c–e**, Validation of the RE reporter strains. GFP fluorescence was measured by flow cytometry in each strain in the presence (+DBD) or absence (–DBD) of a universally high-affinity DBD variant (AncSR1+GGKA+11P, (53)). In each row, the left peak corresponds to autofluorescence from cells that do not express GFP, either due to lack of DBD binding or loss of the DBD expression plasmid; the right peak corresponds to cells that are expressing GFP in response to DBD-RE binding. "FS mut" denotes strains with mutations in the flank/spacer regions of the RE that correct anomalous expression patterns shown in red (see Supplementary Methods). Red strains were not used in the final DMS experiment. Experiments were conducted on the same day within each panel. **c**, Fluorescence in the presence of high-affinity DBD. **d**, Fluorescence in the absence of DBD expression plasmid. **e**, Fluorescence in the CC and GA FS mut strains, with the ERE strain included as a negative control. **f–g**, Sorting gates used for DMS. **f**, Enrichment sort gates. Homogeneous single cells were first selected by gating on FSC-A vs. SSC-A and FSC-A vs. FSC-H (top). Plasmid retention was then selected for by gating on mCherry expression (PE-Texas Red-A, bottom left). Finally, cells were sorted into GFP+ (P4) and GFP– (P5) populations (bottom right). The boundary between the GFP+ and GFP– gates was drawn to have a slope of 1.5 on a log-FSC-A vs. log-GFP (FITC) scale so that populations were sorted by GFP expression relative to cell volume. **g**, Binned sort gates. Gates P1–P3 were drawn as in **C**. Cells were then sorted into four GFP bins, which were drawn to have roughly equal heights (P5–P7). The boundaries between GFP gates were again drawn to have a log-log slope of 1.5.

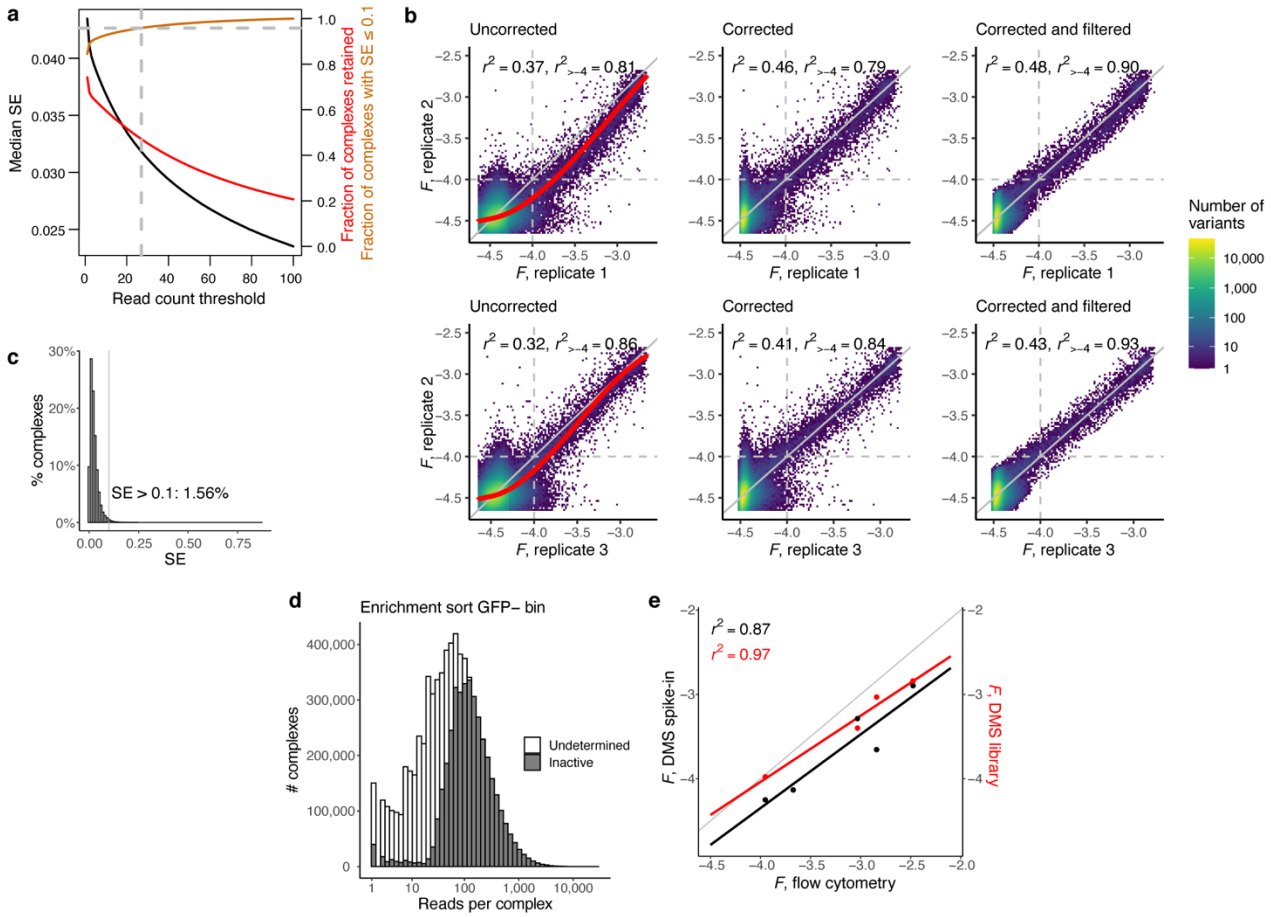


Figure A1.2. DMS data cleaning. a, Curves show characteristics of the binned sort dataset as a function of the read count threshold used to retain protein-RE complexes for further analysis (*x*-axis). Black, standard error of *F* (SE, left axis); red, complexes retained, expressed as a fraction of the number of complexes in the binned sort (right axis); gold, fraction of complexes retained that have SE ≤ 0.1 (right axis). We used a read count threshold of 27 (vertical dashed line), at which ≥95% of complexes have SE ≤ 0.1 (horizontal dashed line). **b**, Correcting and filtering estimates of *F* from the binned sort. Left, correlation in *F* between replicates before correction. Pearson's r^2 is shown for all complexes, and for the subset of complexes with $F > -4$ in both replicates, which roughly corresponds to the boundary between active and inactive complexes (gray dotted lines). Red curves, I-splines fit using complexes with SE of $F < 0.1$. Center, correlation in *F* between replicates after correcting using the I-spline transforms. Right, correlation in *F* between replicates after filtering corrected variants for SE ≤ 0.1. **c**, Distribution of SE across all complexes in the binned sort after the I-spline correction. Complexes with SE > 0.1 were discarded. **d**, Read count distribution for complexes sequenced in the enrichment sort GFP– bin. Complexes were inferred to be inactive (gray) if they were not observed in the binned sort, but had high enough inferred cell count in the enrichment sort to have been detectable in the binned sort had they been at least minimally fluorescent (see Supplementary Methods). **e**, Correlations between estimates of *F* from flow cytometry (*x*-axis) and DMS (*y*-axes). Left *y*-axis (black points) shows estimates from isogenic strains that were spiked into the DMS libraries prior to the binned sort. Right *y*-axis (red points) shows estimates from complexes that were encoded in the DMS libraries.

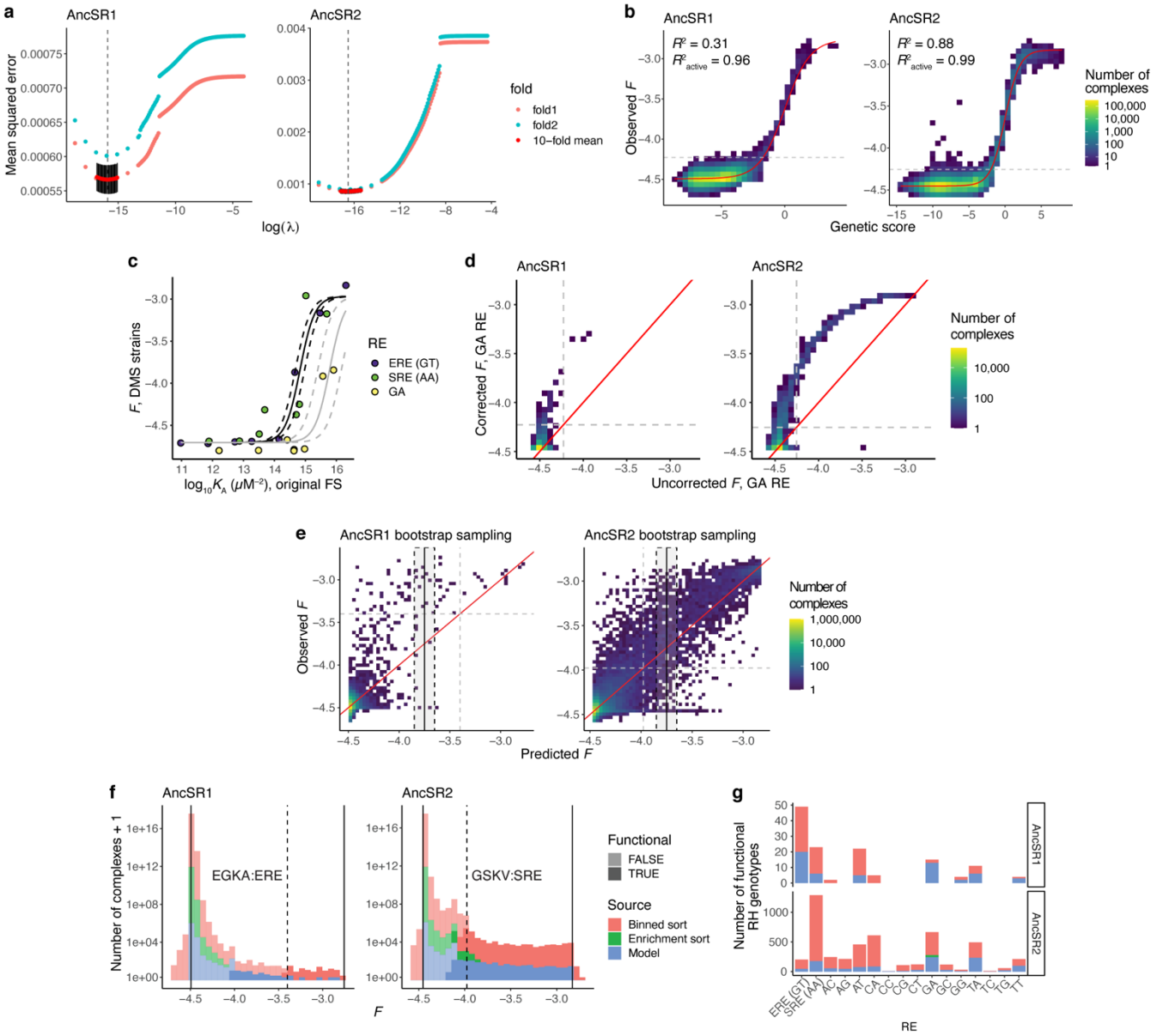


Figure A1.3. Fluorescence imputation, GA strain correction, and functional genotype classification. a, **b**, A generalized linear model that predicts the fluorescence of each protein-RE complex from its sequence was fit to the data for each background, using L2 regularization to address overfitting. **a**, Ten-fold cross-validation (CV) was used to identify the optimal L2 penalty parameter (λ). Red and black, mean and SE of the out-of-sample mean squared error (MSE) across the 10 folds. Initial range finding was performed using two folds (pink and cyan). Vertical line, λ that minimizes mean MSE. **b**, Genetic score versus observed *F* for the regularized RFA models. Red line, best-fit nonspecific epistasis function. For display, the distribution was discretized; colors show the number of variants in the interval defined by each square. Coefficient of determination (R^2) is reported for all complexes and for the subset of active complexes (above the gray line). **c**, **d**, Fluorescence correction for the GA strain. **c**, Affinity (K_A) versus *F* for a panel of DBD variants measured on ERE, SRE, and GA. Affinities, measured by fluorescence anisotropy on the three REs, all with the original flank/spacer

sequence, were previously reported(53, 58). *F* was measured by flow cytometry in the RE strains that were used for DMS, of which the ERE and SRE strains had the original flank/spacer sequence, and the GA strain had a mutated flank/spacer sequence (see Supplementary Methods, Extended Data Fig. 1c–e). Curves, best-fit sigmoidal function. The same midpoint parameter was used for ERE and SRE (black); that for GA was independently estimated (gray). Dashed lines, sigmoidal functions using 95% confidence intervals on the midpoints. **d**, GA fluorescence correction based on the affinity effect estimated in **c**. Plots show *F* before and after the correction. Dashed gray lines, mean boundary between active and null variants. Red line, $y = x$. **e**, Bootstrap sampling strategy for classifying functional complexes with model-inferred fluorescence. Plots show concatenated out-of-sample predictions versus observed *F* across all 10 CV models. Bootstrap-sampled residuals from the interval within ±0.1 units of a complex's predicted *F* were used to test whether a variant with model-inferred *F* was not significantly worse than the wild-type complex (dashed gray lines). An example for a complex with inferred $F = -3.75$ (solid black line) is shown, with the bootstrap interval shown as a shaded rectangle. Solid red line, $y = x$. **f**, Distribution of *F* across all 2,560,000 complexes in each DBD background. Solid vertical lines, upper and lower bounds of fluorescence inferred from the RFA models; dashed vertical lines, fluorescence of wild type complex (EGKA: ERE for AncSR1 and GSKV:SRE for AncSR2). Colors indicate the source from which *F* was estimated. Darker colors show functional variants, lighter colors nonfunctional. All "enrichment sort" complexes were assigned to the lower bound of fluorescence, except for GA RE variants whose fluorescence was corrected upward (**d**). Some model-predicted variants in the AncSR1 background have predicted *F* below the reference but are classified as functional, because the bootstrap test accounts for the AncSR1 RFA model's tendency to under-predict fluorescence (**e**, left). **g**, Bars show the number of functional RH variants per RE per DBD background, colored by source of *F* estimate as in **f**.

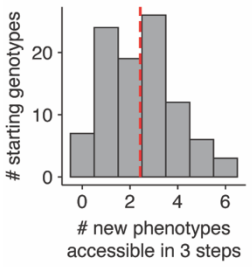


Figure A1.4. Accessible new phenotypes after 3 substitution steps in the AncSR1 network. Bars show the distribution over every starting genotype in the AncSR1 main component. Dashed line, mean.

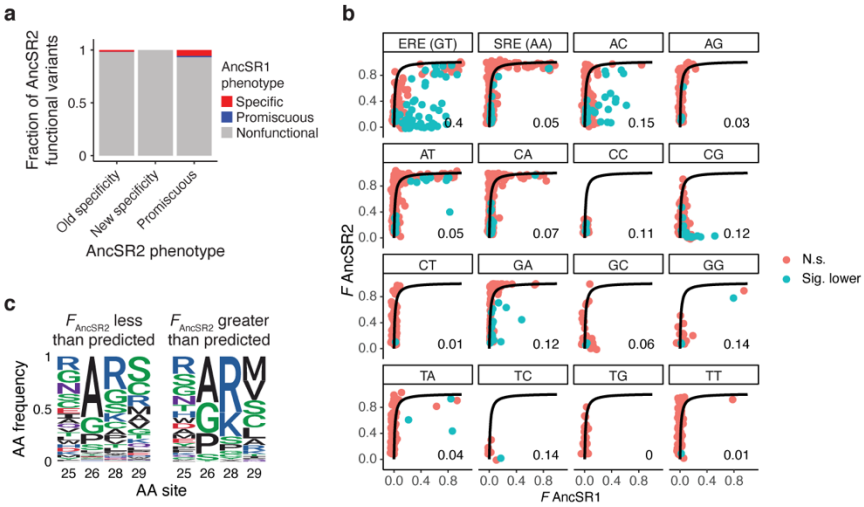


Figure A1.5. Additional analyses for effects of background substitutions on DBD-RE affinity. a, Changes in phenotype across the AncSR1-to-AncSR2 transition. Bars represent the set of protein variants in AncSR2 that have different classes of phenotypes: specificity phenotypes that were encoded in the AncSR1 map (old specificity), specificity phenotypes not encoded in the AncSR1 map (new specificity), or promiscuous in AncSR2. Colored sections show the fraction of variants in each class whose functional category in the AncSR1 background was specific, promiscuous, or nonfunctional. **b**, Plots are the same as in Fig. 6A, but split into panels by RE. Blue points, protein-DNA complexes with significantly lower fluorescence in the AncSR2 background than predicted by the model; red, all other variants. Numbers at the bottom-right of each panel show the fraction of plotted variants with significantly lower than expected AncSR2 fluorescence. **c**, Amino acid frequencies at the RH variable sites among all complexes that are significantly more (left) or less (right) fluorescent in the AncSR2 background than predicted by the ERE-specific model in Fig. 6e. To test for significance in **b** and **c**, we tested whether their Bonferroni-corrected 95% CI of fluorescence was outside of the 95% CI of the model in both the AncSR1 and AncSR2 backgrounds.

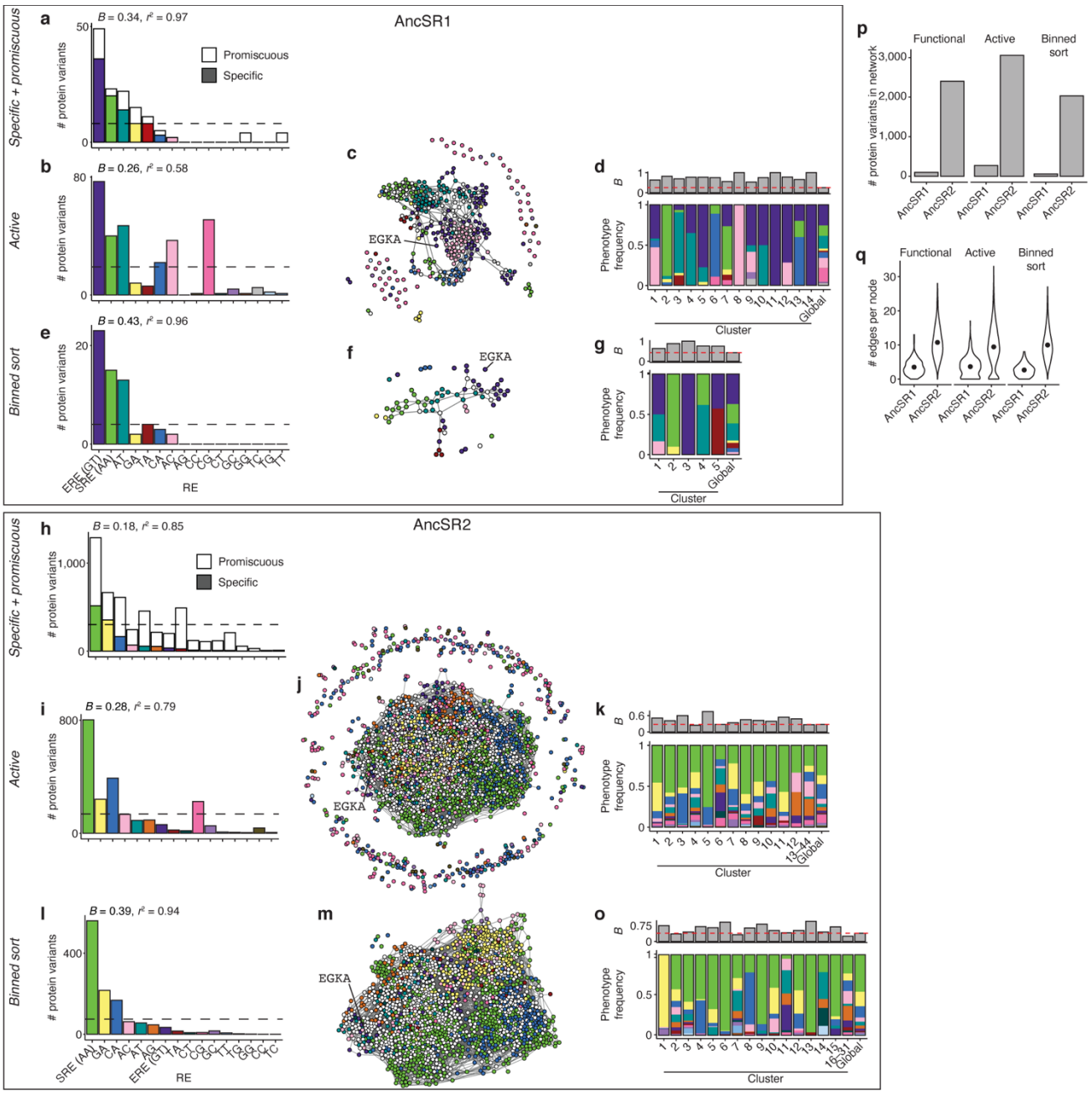


Figure A1.6. Robustness to alternative phenotype assignment methods. a, Global production distribution in the AncSR1 background, counting variants that bind specifically (colored bars) and promiscuously (white bars) to each RE. Dashed line shows the expected frequencies if the production distribution were isotropic. The bias, *B*, of the distribution and r^2 to the production distribution for specific variants (Fig. 2a) are reported. **b**, Same as in **a**, with phenotypes calculated using data from variants with fluorescence significantly higher than that of nonsense variants (active variants). **c**, Sequence space network for AncSR1 active variants. **d**, Bottom: Frequencies of specificity phenotypes within each genotype cluster in the AncSR1 active variant networks; the global production distribution is shown for comparison. Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. **e–g**, Same as in **b–d**, but with phenotypes calculated using only data from the binned sort experiment; protein-

DNA complexes without experimental fluorescence measurements were assumed to have null fluorescence. **h–o**, Same as in **a–g**, but for the AncSR2 background. Note that the active variant datasets are likely to be enriched for false positives due to misclassification of variants whose fluorescence is by chance slightly higher than the nonsense variant distribution. This may explain the high frequency of variants that do not share any mutational connections to other active variants. It may also explain the high frequency of CG-specific variants compared to the original classification scheme, since the CG yeast strain has a slightly higher null fluorescence level than most other strains (Extended Data Fig. 1c, d) and most CG-specific variants are unconnected in the active variant genotype networks. **p**, Number of protein variants in each network under different methods of phenotype assignment. "Functional" indicates the original method used in the main text; note that this yields the same number of protein variants as the "specific + promiscuous" method. **q**, Number of edges per node in each network, with the original phenotype classification method (functional) shown for comparison.

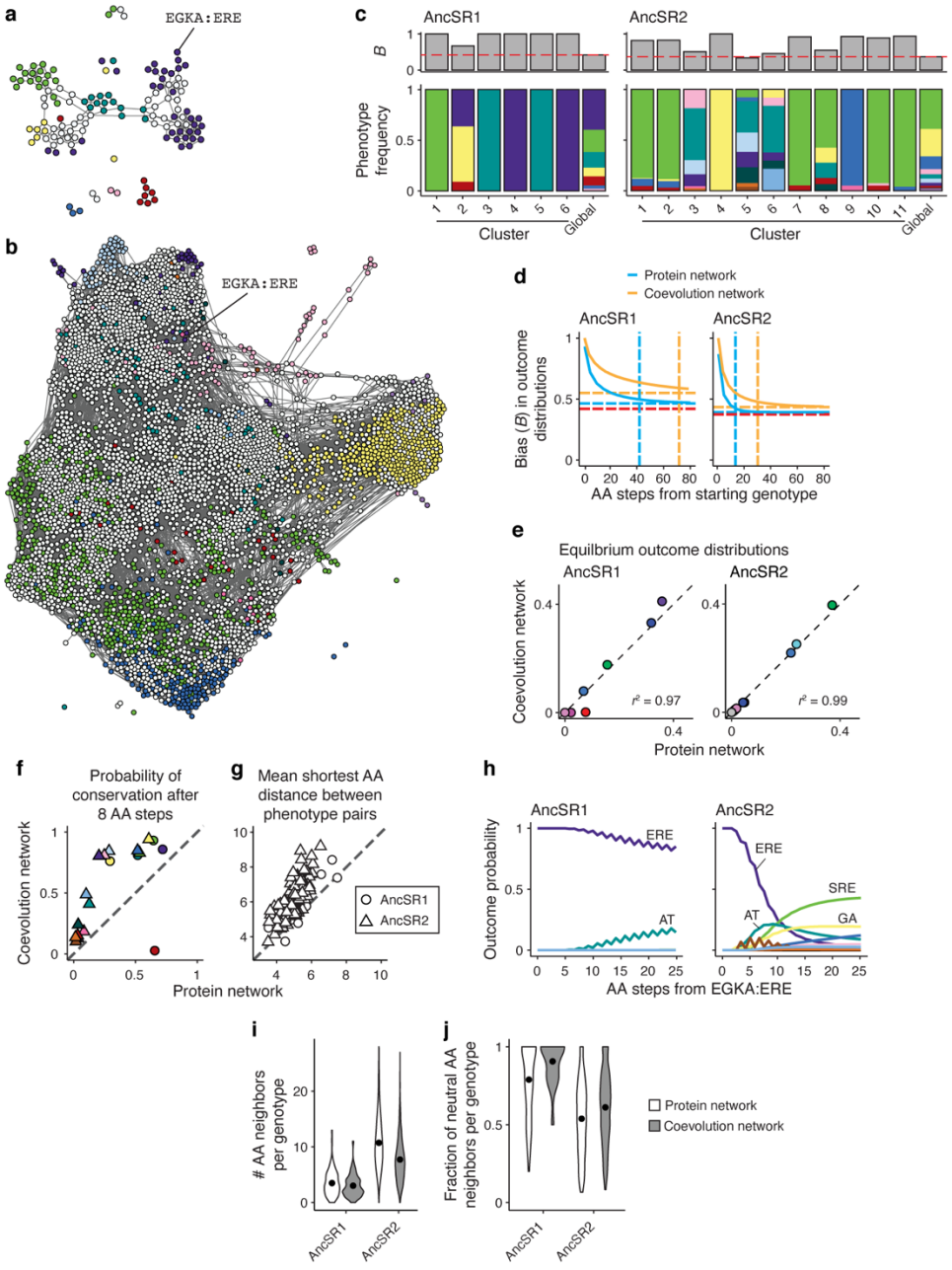


Figure A1.7. Robustness to model of evolution using joint protein-DNA networks. a, AncSR1 protein-DNA coevolution network. Nodes represent functional protein-RE complexes, colored by the RE specificity of the protein genotype; colors are as in Fig. 2b and 4b. Promiscuous protein genotypes are represented by multiple nodes, one for each RE it binds. Edges connect complexes that can be interconverted by a single nucleotide change in the RE or the coding sequence of the protein. **b**, AncSR2 protein-DNA coevolution network. **c**, Bottom: Frequencies of specificity phenotypes within each genotype cluster in the AncSR1 (left) and AncSR2 (right) coevolution networks; the global production distribution (right-most column) is shown for comparison. Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. **d**, Bias (*B*) in evolutionary outcomes as a function of the length of evolutionary trajectories. Solid curves, mean *B* across starting genotypes in the protein (cyan) or

coevolution (orange) networks. Dashed horizontal lines, *B* of the equilibrium distribution in each network; dashed horizontal red line, global bias. Vertical dashed lines show the number of substitutions required for mean *B* to reach within 0.05 units of the equilibrium value within each type of network. The equilibrium distributions are more biased in the coevolution networks, and require more amino acid substitutions to be reached, because changes in protein genotype must occur between variants that can bind to the same RE sequence. **e**, Comparison between equilibrium outcome distributions of the protein-only evolution and protein-DNA coevolution networks in each AncSR1 (left) and AncSR2 (right) backgrounds. Pearson's r^2 between the two distributions are shown. Dashed line, $y = x$. **f**, Probability of conservation of each phenotype after 8 amino acid substitution steps in the protein vs. coevolution networks. **g**, Mean shortest amino acid distance between all possible pairs of phenotypes in the coevolution vs. protein networks, calculated as in Fig. 2g. Circles, AncSR1 networks, triangles, AncSR2 networks. Dashed line, y = x. **h**, Probability of evolving each specificity phenotype as a function of the number of amino acid substitutions away from EGKA:ERE in the AncSR1 (left) and AncSR2 (right) coevolution networks. In both backgrounds, conservation is more likely at short trajectory lengths than in the corresponding protein networks (Fig. 3g, 5f), but the relative likelihood of achieving each phenotypic outcome is similar. **i**, Distribution of the number of neighbors per genotype with distinct RH sequences in each type of network. Dots, means of distributions. **j**, Distribution of the fraction of neutral neighbors per node with distinct RH genotypes in each network. Dots, means of distributions.

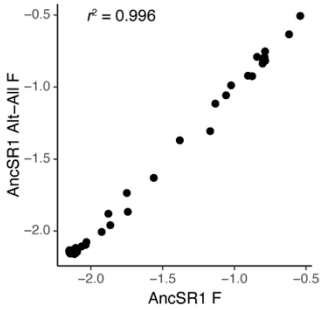


Figure A1.8. Robustness of RH mutation effects to uncertainty in ancestral reconstruction. Effects on ERE binding of all possible single amino acid mutations at the four variable RH sites in the background of the maximum *a posteriori* (MAP) wild type AncSR1 protein (*x*-axis), and in the background of the AltAll wild type AncSR1 protein, which has the second-most likely amino acid state at all sites at which the posterior probability of the MAP state is less than 0.8 (*y*-axis)(57). Pearson's r^2 is shown.

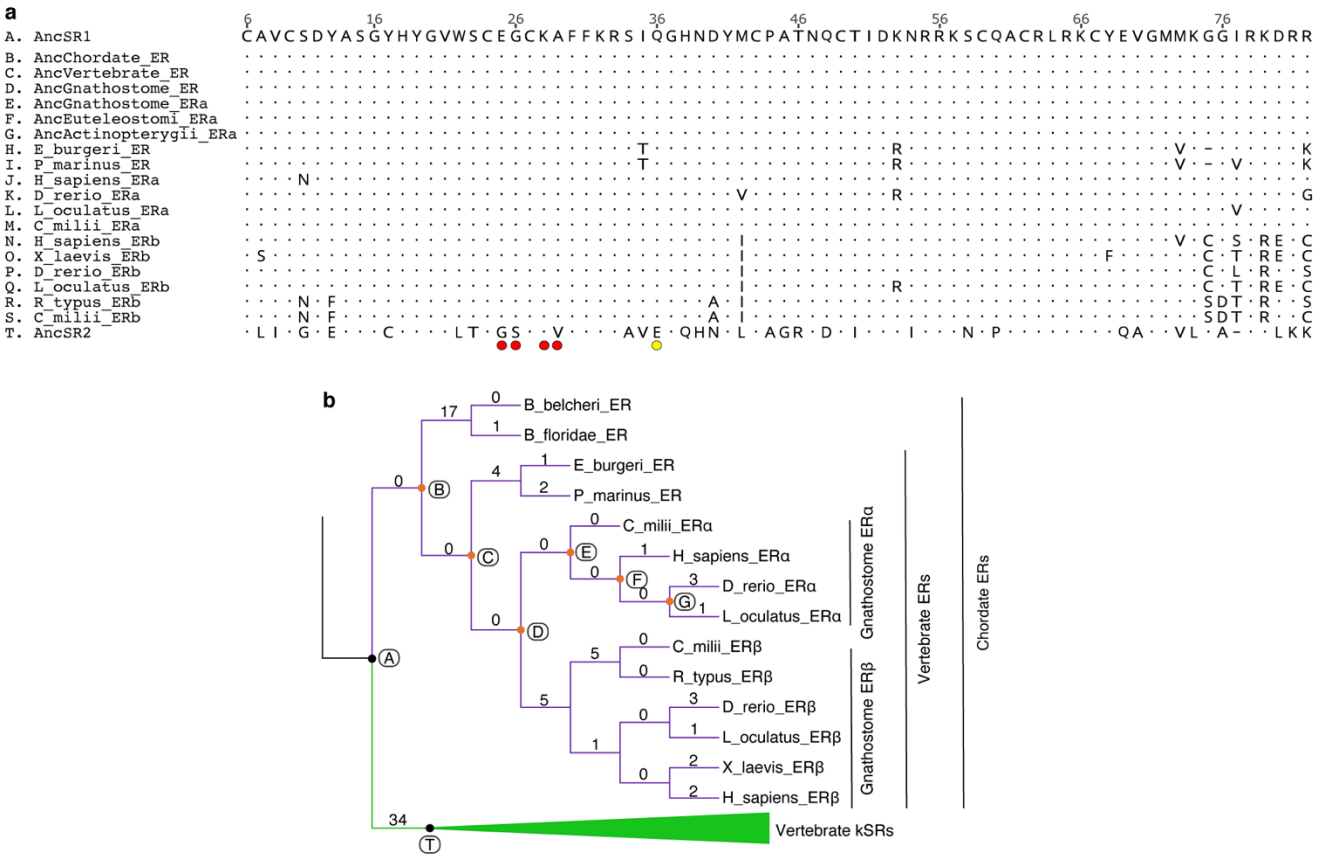


Figure A1.9. Amino acid changes along the SR phylogeny. a, Amino acid alignment of extant vertebrate ERs and the MAP protein sequences for key ancestral nodes in the SR phylogeny(57). The AncSR1 sequence is used as the reference to indicate amino acid changes; dots, same amino acid state as that in AncSR1; dashes, gaps; red circles, variable sites in DMS experiment; yellow circle, historical substitution (q36E) that likely contributed to the shift in the direction of the global bias away from ERE. **b**, Cladogram of SRs showing the number of substitutions that occurred along each branch. Letters, nodes shown in alignment in **a**; black nodes, AncSR1 and AncSR2; orange nodes, ancestral ER sequences identical to AncSR1. Branches and clades are colored according to their DNA specificity phenotype: purple, ERE-specificity; green, SRE-specificity.

Table A1.1: Synonymous RE barcodes (REBCs)

Sequences of synonymous mutations used to barcode the DBD libraries to associate them with RE strains. NNS, variable sites in the DMS experiment.

RE barcode ID	RE barcode sequence	RE
AncSR1 REBC 1	GGTGTATGGTCATGTNNSNNSTGTNNSNNS	SRE
AncSR1 REBC 2	GGGGTTTGGTCGTGTNNSNNSTGTNNSNNS	GA
AncSR1 REBC 3	GGGGTTTGGAGTTGTNNSNNSTGTNNSNNS	ERE
AncSR1 REBC 4	GGTGTATGGAGCTGTNNSNNSTGTNNSNNS	AC
AncSR1 REBC 5	GGCGTTTGGTCATGCNNSNNSTGTNNSNNS	AG
AncSR1 REBC 6	GGAGTATGGTCGTGCNNSNNSTGTNNSNNS	AT
AncSR1 REBC 7	GGTGTCTGGAGTTGCNNSNNSTGTNNSNNS	CA
AncSR1 REBC 8	GGCGTGTGGAGCTGCNNSNNSTGTNNSNNS	CC
AncSR1 REBC 9	GGCGTCTGGTCATGTNNSNNSTGCNNSNNS	CG
AncSR1 REBC 10	GGAGTGTGGTCGTGTNNSNNSTGCNNSNNS	CT
AncSR1 REBC 11	GGCGTGTGGAGTTGTNNSNNSTGCNNSNNS	GC
AncSR1 REBC 12	GGAGTCTGGAGCTGTNNSNNSTGCNNSNNS	GG
AncSR1 REBC 13	GGTGTGTGGTCATGCNNSNNSTGCNNSNNS	TA
AncSR1 REBC 14	GGGGTCTGGTCGTGCNNSNNSTGCNNSNNS	TC
AncSR1 REBC 15	GGAGTTTGGAGTTGCNNSNNSTGCNNSNNS	TG
AncSR1 REBC 16	GGGGTATGGAGCTGCNNSNNSTGCNNSNNS	TT
AncSR2 REBC 1	GTGTTGACGTGCNNSNNSTGCNNSNNS	SRE
AncSR2 REBC 2	GTTCTTACGTGCNNSNNSTGCNNSNNS	GA
AncSR2 REBC 3	GTATTAACATGCNNSNNSTGCNNSNNS	ERE
AncSR2 REBC 4	GTCCTCACATGCNNSNNSTGCNNSNNS	AC
AncSR2 REBC 5	GTACTTACATGTNNSNNSTGCNNSNNS	AG
AncSR2 REBC 6	GTGCTCACCTGTNNSNNSTGCNNSNNS	AT
AncSR2 REBC 7	GTTCTGACTTGTNNSNNSTGCNNSNNS	CA
AncSR2 REBC 8	GTGCTTACATGCNNSNNSTGTNNSNNS	CC
AncSR2 REBC 9	GTCTTAACCTGCNNSNNSTGTNNSNNS	CG
AncSR2 REBC 10	GTTCTCACCTGCNNSNNSTGTNNSNNS	CT
AncSR2 REBC 11	GTCCTGACTTGCNNSNNSTGTNNSNNS	GC
AncSR2 REBC 12	GTATTGACGTGTNNSNNSTGTNNSNNS	GG
AncSR2 REBC 13	GTTCTAACGTGTNNSNNSTGTNNSNNS	TA
AncSR2 REBC 14	GTCCTTACCTGTNNSNNSTGTNNSNNS	TC
AncSR2 REBC 15	GTGTTAACTTGTNNSNNSTGTNNSNNS	TG
AncSR2 REBC 16	GTACTCACTTGTNNSNNSTGTNNSNNS	TT

Table A1.2: Library transformation and enrichment sort statistics

Library transformation yields, glycerol stock recovery rates, and number of cells sorted for enrichment sorts.

DBD background	RE strain	RE barcode	Bacterial transfor-mation cfu (x10^7)	Yeast transfor-mation cfu (x10^7)	Top-up yeast transfor-mation cfu (x10^7)	Total yeast transfor-mation cfu (x10^7)	Enrich-ment sort batch	Glycerol stock recovery cfu (x10^7)	Enrich-ment sort GFP– cell count	Enrich-ment sort GFP+ cell count	Enrich-ment sort GFP+ proportion
AncSR1	SRE	1	5.67	3.09		3.09	1	19.4	24375871	625635	0.025
AncSR1	GA	2	2.01	2.82		2.82	1	29.6	24800681	435760	0.017
AncSR1	ERE	3	2.38	2.58		2.58	1	25.4	24681039	405514	0.016
AncSR1	AC	4	2.64	1.73		1.73	1	10.6	24592905	528709	0.021
AncSR1	AG	5	2.07	3.25		3.25	1	25.8	24662980	408313	0.016
AncSR1	AT	6	3.39	3.8		3.8	1	25.8	24806774	472083	0.019
AncSR1	CA	7	1.56	2.07		2.07	1	14.4	24352862	726379	0.029
AncSR1	CC	8	1.08	2.9		2.9	1	27.8	24658946	448824	0.018
AncSR1	CG	9	1.66	0.85	1.1	1.95	5/6	18.4	24326839	858760	0.034
AncSR1	CT	10	2.37	2.34		2.34	5/6	25.6	24736293	379479	0.015
AncSR1	GC	11	1.97	2.26		2.26	5/6	26.4	24336612	744771	0.03
AncSR1	GG	12	1.38	1.32		1.32	5/6	33.6	24841745	285197	0.011
AncSR1	TA	13	1.86	0.69	1.29	1.98	5/6	10.4	24748789	329854	0.013
AncSR1	TC	14	1.1	2.16		2.16	5/6	20.6	24558731	587774	0.023
AncSR1	TG	15	1.22	0.91	1.01	1.92	2/6	15.6	24945619	348916	0.014
AncSR1	TT	16	3.07	2.54		2.54	2/6	16.4	25226917	437954	0.017
AncSR2	SRE	1	5.24	1.79		1.79	4	17.8	24220370	847039	0.034

Table A1.2 cont: Library transformation and enrichment sort statistics

AncSR2	GA	2	3.79	1.37		1.37	4	25	24825719	391276	0.016
AncSR2	ERE	3	5.1	1.5		1.5	4	17.4	24694083	425418	0.017
AncSR2	AC	4	3.54	0.7	0.87	1.57	4	9.8	24602087	533981	0.021
AncSR2	AG	5	4.6	0.43	0.9	1.33	4	8.8	24690832	434945	0.017
AncSR2	AT	6	4.3	1.6		1.6	4	1.6	24686331	421661	0.017
AncSR2	CA	7	2.2	1.17		1.17	4	1.6	24407887	777213	0.031
AncSR2	CC	8	2.8	1.35		1.35	4	2.8	24637616	409661	0.016
AncSR2	CG	9	6.9	1.98		1.98	3	18.6	24399903	903819	0.036
AncSR2	CT	10	4.5	1.16		1.16	3	4.6	24518676	796209	0.031
AncSR2	GC	11	5.4	0.97		0.97	3	21	24214644	940935	0.037
AncSR2	GG	12	5.1	2.52		2.52	3	18	24511230	679816	0.027
AncSR2	TA	13	1.46	1.08		1.08	3	8.2	24606769	673875	0.027
AncSR2	TC	14	12.9	1.3		1.3	3	18	24040623	1105703	0.044
AncSR2	TG	15	4.6	1.5		1.5	3	10.8	24441448	617178	0.025
AncSR2	TT	16	2.7	2.03		2.03	3	20.8	24619287	854539	0.034

Table A1.3: Binned sort statistics

Glycerol stock recovery rates and number of cells sorted for binned sorts.

Binned sort replicate	Background	Libraries	Enrich-ment sort batch	Glycerol stock recovery cfu (x10^7)	Glycerol stock recovery cfu per GFP+ cells sorted	Cells sorted, bin 1	Cells sorted, bin 2	Cells sorted, bin 3	Cells sorted, bin 4	Cells sorted, total
1	AncSR1	1-8	1	2.8	6.9	68112878	86824472	5447744	3638776	164023870
1	AncSR1	15-16 / 9-14	2 / 5	2.3 / 1.2	29.2 / 3.8					
1	AncSR2	1-8	4	10.2	24					
1	AncSR2	9-16	3	4	6.1					
2	AncSR1	1-8	1	1.8	4.4	71873842	75663838	5380470	3116191	156034341
2	AncSR1	9-16	6	29.4	90.3					
2	AncSR2	1-8	4	15.6	36.8					
2	AncSR2	9-16	3	5.8	8.8					
3	AncSR1	1-8	1	1.8	4.4	66662146	87779826	7047242	4303413	165792627
3	AncSR1	9-16	6	4.8	14.7					
3	AncSR2	1-8	4	19	44.8					
3	AncSR2	9-16	3	7.2	11					
4	AncSR1	9-16	6	4.6	14.1	16233786	12616013	822095	579424	30251318

Table A1.4: Binned sort sequencing statistics

Estimated number of reads per cell across libraries, bins, and replicates. Libraries for AncSR1 REBC 9–14 and REBC 15–16 had very low and high sequencing depths, respectively, in replicate 1, so we repeated the enrichment sort for these libraries and used them for bins sort replicates 2–4.

Background	REBC	Rep 1 reads/cell	Rep 2 reads/cell	Rep 3 reads/cell	Rep 4 reads/cell
AncSR1	1	0.86	0.56	1.35	
AncSR1	2	1.16	0.8	1.83	
AncSR1	3	0.91	0.62	1.56	
AncSR1	4	1.9	1.29	3.27	
AncSR1	5	2.37	1.57	3.87	
AncSR1	6	2.39	1.68	3.81	
AncSR1	7	1.41	0.93	2.4	
AncSR1	8	2.11	1.46	3.57	
AncSR1	9	0.06	0.33	1.2	1.55
AncSR1	10	0.04	1.26	4.04	1.4
AncSR1	11	0.06	0.3	1.01	1.89
AncSR1	12	0.05	1.17	3.78	2.01
AncSR1	13	0.04	0.64	1.99	1.18
AncSR1	14	0	0.36	1.18	0.98
AncSR1	15	8.97	1.09	3.59	2
AncSR1	16	4.7	0.58	1.86	1.42
AncSR2	1	1.38	0.86	4.43	
AncSR2	2	0.84	0.57	2.74	
AncSR2	3	1.41	0.88	4.77	
AncSR2	4	1.07	0.67	3.7	
AncSR2	5	1.01	0.63	3.39	
AncSR2	6	0.99	0.64	3.08	
AncSR2	7	1.39	0.83	4.63	
AncSR2	8	1.64	1.01	5.65	
AncSR2	9	4.26	2.15	12.19	
AncSR2	10	0.77	0.44	2.16	
AncSR2	11	1.55	0.84	4.38	
AncSR2	12	0.98	0.57	2.67	
AncSR2	13	0.28	0.16	0.77	
AncSR2	14	0.58	0.32	1.58	
AncSR2	15	0.61	0.35	1.7	
AncSR2	16	0.41	0.23	1.11	

Appendix 2

Supplementary figures and tables for Chapter 3: Epistasis shapes the genotype-phenotype map via structural integration

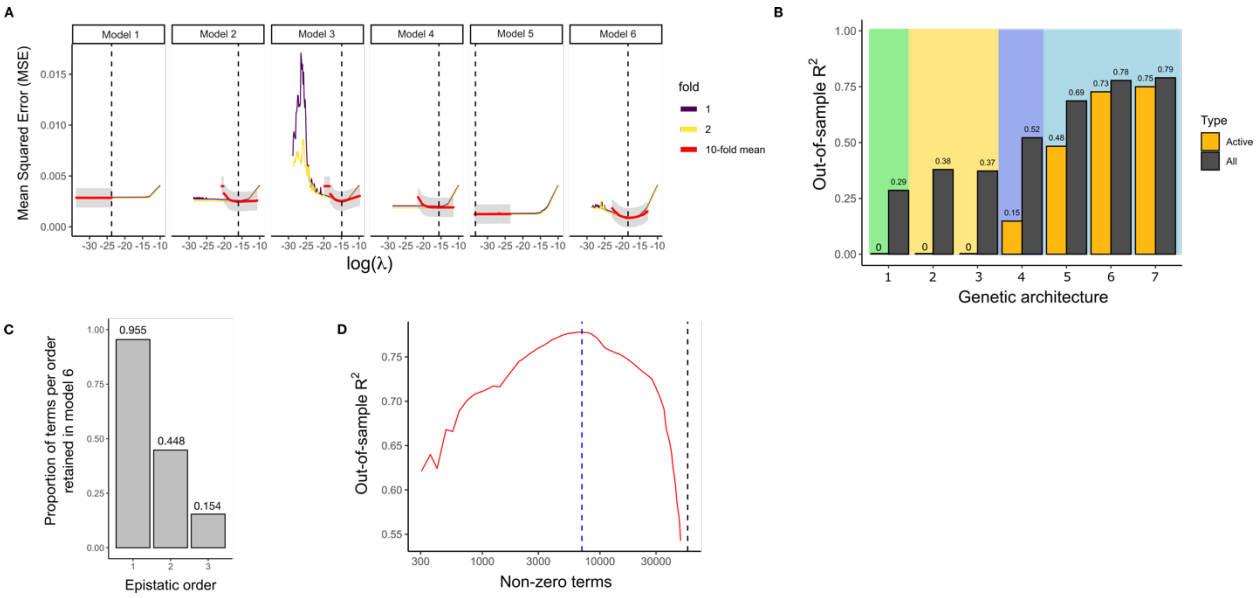


Figure A2.1. Model fitting of alternative genetic architectures. (A) L1 regularization and 10-fold cross-validation (CV) for truncated models of genetic architecture. Folds 1 and 2 show the performance of the model over a coarse range of regularization parameter (λ) values. 10 CV folds were then run over a finer scale of λ values. Red points, mean MSE across 10 CV folds; error bars, standard error of the mean MSE; vertical dashed line, value of λ that minimizes prediction error (λ_{min}). **(B)** Phenotypic variance explained by truncated models of genetic architecture evaluated by 10-fold CV. Bars, average out-of-sample R^2 across 10 CV models for active and all variants; CV models correspond to λ_{min} in **A**. Colored regions group models by the form of epistasis included; green: no epistasis, yellow: intramolecular epistasis, dark blue: intermolecular epistasis, light blue: intra + intermolecular epistasis (see Table1 for details). **(C)** Composition of the genetic architecture of the best-fit model. Bars, proportion of the total number of terms of each order that are retained (non-zeroed) after L1 regularization. **(D)** Phenotypic variance explained the best-fit model as a function of the number of regularized terms. Blue vertical line, size of the model that maximizes the explained phenotypic variance (7,627 terms). Black vertical line, full size of the model (55,624 terms).

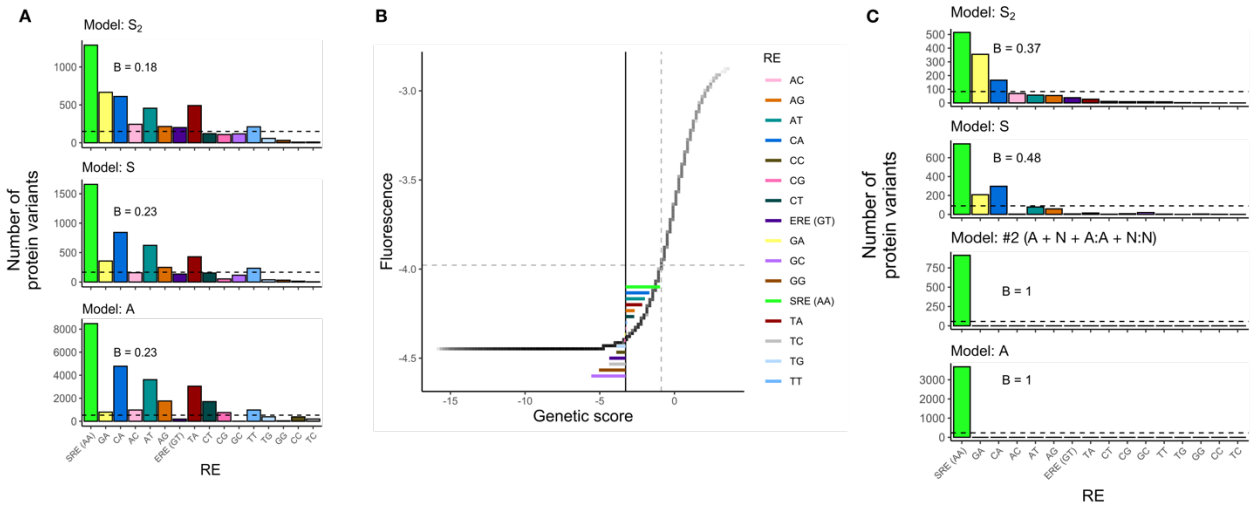


Figure A2.2. Intermolecular epistasis shapes the production of specificity phenotypes. (A) Global distribution of binding (specific and promiscuous) and global bias *B* for each model. Bars represent the number of functional protein variants that bind specifically or promiscuously to each RE. The dashed line shows the expected frequencies if the distribution were unbiased. **(B)** Relationship between the genetic score of a protein-DNA complex (the sum of the contributions of genetic states) and is fluorescence under an additive genetic architecture. Horizontal dashed line, fluorescence of the reference GSKV:SRE complex. Vertical dashed line, genetic score of the reference complex representing the threshold for functional binding. Vertical solid line, average genetic score based solely on amino acid contributions of specific protein variants. Colored lines, net contribution of an RE sequence on the average protein genetic score. Only the independent contributions of nucleotide states in SRE can bring protein genotypes close to the functional threshold. **(C)** Global production distribution and global bias *B* for each model. Intramolecular epistasis (A:A + N:N) does not influence the production of specificity phenotypes. Bars represent the number of functional protein variants that bind specifically to each RE. Global bias *B* is calculated as 1 minus the Shannon entropy (base 16) of the distribution.

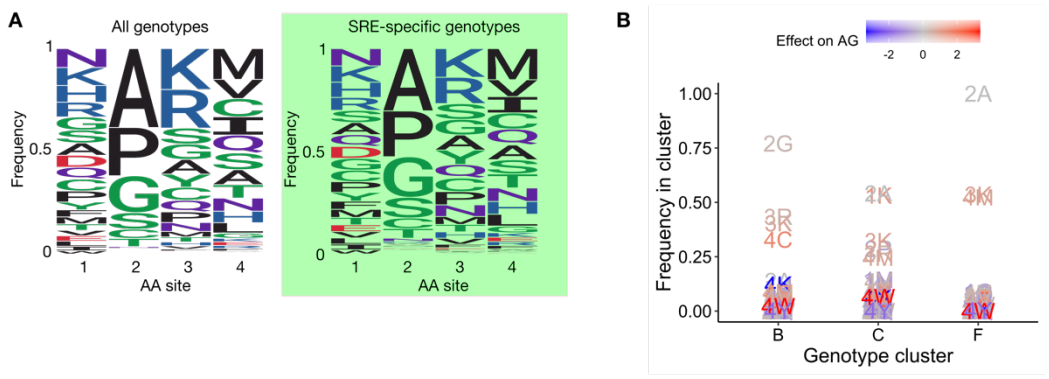


Figure A2.3. Amino acid profiles and interactions related to the heterogeneity of networks. **(A)** Amino acid profiles of all functional protein variants (left) and all SRE-specific variants (right) in the additive genetic architecture. **(B)** Frequency of amino acid contributions to AG specificity across 3 clusters in the S network. Combinations of amino acid sites and states are colored by their genetic contribution to the phenotype.

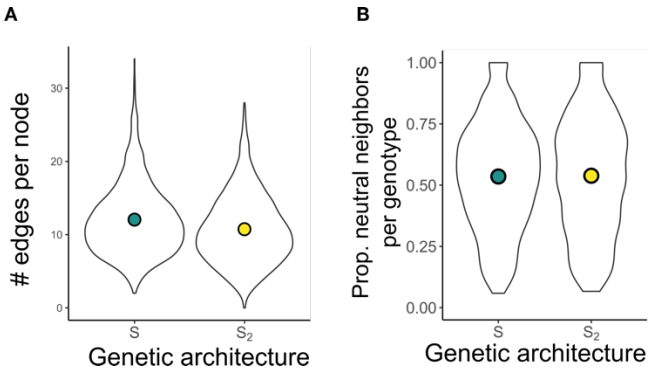


Figure A2.4. Additional structural properties of the GP maps. (A) Distributions of the number of neighbors per genotype (connectivity). **(B)** Distributions of the fraction of neighbors per genotype that have the same phenotype as the focal genotype (neutrality).

Bibliography

1. R. Amundson, *The changing role of the embryo in evolutionary thought: roots of evo-devo* (Cambridge University Press., 2005).

2. W. Arthur, *Biased Embryos and Evolution* (Cambridge University Press, 2004).

3. W. Provine, *The Origins of Theoretical Population Genetics* (The University of Chicago Press, 1971).

4. R. Riedl, A Systems-Analytical Approach to Macro-Evolutionary Phenomena. *Q. Rev. Biol.* **52**, 351–370 (1977).

5. G. P. Wagner, *Homology, genes and evolutionary innovation* (Princeton university press, 2014).

6. R. C. Lewontin, "Four complications in understanding the evolutionary process" in *SFI Bulletin*, (2003).

7. S. B. Carroll, Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* **134**, 25–36 (2008).

8. M. E. Olson, The developmental renaissance in adaptationism. *Trends Ecol. Evol.* **27**, 278–87 (2012).

9. S. J. Gould, R. C. Lewontin, The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proc. R. Soc. B Biol. Sci.* **205**, 581–598 (1979).

10. R. C. Lewontin, *The genetic basis of evolutionary change* (Columbia University Press, 1974).

11. G. P. Wagner, L. Altenberg, Perspective : Complex Adaptations and the Evolution of Evolvability. *Evolution* **50**, 967–976 (1996).

12. B. Hallgrímsson, *et al.*, "The developmental basis for evolvability" in *Evolvability: A Unifying Concept in Evolutionary Biology?*, (Vienna Series in Theoretical Biology, 2023).

13. J. Maynard-Smith, *et al.*, Developmental constraints and evolution. *Q. Rev. Biol.* **60**, 265–287 (1985).

14. A. Stoltzfus, L. Y. Yampolsky, Climbing Mount Probable: Mutation as a Cause of Nonrandomness in Evolution. *J. Hered.* **100**, 637–647 (2009).

15. A. Hodgins-Davis, F. Duveau, E. A. Walker, P. J. Wittkopp, Empirical measures of mutational effects define neutral models of regulatory evolution in Saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **116**, 21085–21093 (2019).

16. D. Schluter, Adaptive radiation along genetic lines of least resistance. *Evolution* **50**, 1766–1774 (1996).

17. D. B. Wake, A. Larson, Multidimensional analysis of an evolving lineage. *Science* **238**, 42–48 (1987).

18. R. J. Dugand, J. D. Aguirre, E. Hine, M. W. Blows, K. McGuigan, The contribution of mutation and selection to multivariate quantitative genetic variance in an outbred population of *Drosophila serrata*. *Proc. Natl. Acad. Sci.* **118**, e2026217118 (2021).

19. P. J. Wittkopp, Contributions of mutation and selection to regulatory variation: lessons from the Saccharomyces cerevisiae TDH3 gene. *Philos. Trans. R. Soc. B Biol. Sci.* **378**, 20220057 (2023).

20. I. Salazar-Ciudad, Why call it developmental bias when it is just development? *Biol. Direct* **16**, 1–13 (2021).

21. B. M. R. Stadler, P. F. Stadler, G. P. Wagner, W. Fontana, The Topology of the Possible: Formal Spaces Underlying Patterns of Evolutionary Change. *J. Theor. Biol.* **213**, 241–274 (2001).

22. I. Salazar-Ciudad, J. Jernvall, A computational model of teeth and the developmental origins of morphological variation. *Nature* **464**, 583–586 (2010).

23. P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, From sequences to shapes and back: A case study in RNA secondary structures. *Proc R Soc Lond B* **255**, 279–284 (1994).

24. T. Uller, A. P. Moczek, R. A. Watson, P. M. Brakefield, K. N. Laland, Developmental bias and evolution: A regulatory network perspective. *Genetics* **209**, 949–966 (2018).

25. P. T. Rohner, D. Berger, Developmental bias predicts 60 million years of wing shape evolution. *Proc. Natl. Acad. Sci.* **120**, e2211210120 (2023).

26. K. Dingle, F. Ghaddar, P. Šulc, A. A. Louis, Phenotype Bias Determines How Natural RNA Structures Occupy the Morphospace of All Possible Shapes. *Mol. Biol. Evol.* **39**, 1–11 (2022).

27. D. M. Raup, Geometric analysis of shell coiling: General Problems. *J. Paleontol.* **40**, 1178–1190 (1966).

28. B. Deline, *et al.*, Evolution of metazoan morphological disparity. *Proc Nat Acad Sci* E8909–E8918 (2018). https://doi.org/10.1073/pnas.1810575115.

29. J. W. Clark, *et al.*, Evolution of phenotypic disparity in the plant kingdom. *Nat. Plants* (2023). https://doi.org/10.1038/s41477-023-01513-x.

30. S. Gerber, Not all roads can be taken: Development induces anisotropic accessibility in morphospace (2014).

31. S. B. Carroll, Evolution at two levels: On genes and form. *PLoS Biol.* **3**, 1159–1166 (2005).

32. S. B. Carroll, *et al.*, Pattern Formation and Eyespot Determination in Butterfly Wings. *Science* **4917**, 1–6 (1994).

33. N. Gompel, B. Prud'homme, P. J. Wittkopp, V. a Kassner, S. B. Carroll, Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* **433**, 481–487 (2005).

34. A. I. Podgornaia, M. T. Laub, Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).

35. N. C. Wu, L. Dai, C. A. Olson, J. O. Lloyd-Smith, R. Sun, Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, 1–21 (2016).

36. H. Kemble, P. Nghe, O. Tenaillon, Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. *Evol. Appl.* **12**, 1721–1742 (2019).

37. O. Tenaillon, The Utility of Fisher's Geometric Model in Evolutionary Genetics. *Annu. Rev. Ecol. Evol. Syst.* **45**, 179–201 (2014).

38. S. Kauffman, S. Levin, Towards a General Theory of Adaptive Walks on Rugged Landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).

39. R. Lande, QUANTITATIVE GENETIC ANALYSIS OF MULTIVARIATE EVOLUTION, APPLIED TO BRAIN:BODY SIZE ALLOMETRY. *Evolution* **33**, 402–416 (1979).

40. M. H. Gromko, UNPREDICTABILITY OF CORRELATED RESPONSE TO SELECTION: PLEIOTROPY AND SAMPLING INTERACT. *Evolution* **49**, 685–693 (1995).

41. O. Cotto, T. Day, A null model for the distribution of fitness effects of mutations. *Proc. Natl. Acad. Sci.* **120**, e2218200120 (2023).

42. D. Houle, Genetic Covariance of Fitness Correlates: What Genetic Correlations are Made of and Why it Matters. *Evolution* **45**, 630–648 (1991).

43. G. J. Vermeij, Forbidden phenotypes and the limits of evolution. *Interface Focus* **5**, 20150028 (2015).

44. R. Dawkins, *Climbing mount improbable* (WW Norton & Company., 1996).

45. P. R. Grant, B. R. Grant, *40 years of evolution: Darwin's Finches on Daphne Major Island* (Princeton university press, 2014).

46. W. Arthur, The interaction between developmental bias and natural selection: From centipede segments to a general hypothesis. *Heredity* **89**, 239–246 (2002).

47. D. Jablonski, Developmental bias, macroevolution, and the fossil record. *Evol. Dev.* 103–125 (2019). https://doi.org/10.1111/ede.12313.

48. S. J. Steppan, P. C. Phillips, D. Houle, Comparative quantitative genetics: evolution of the Gmatrix. *Trends Ecol Evol* **17**, 320–327 (2002).

49. J. W. McGlothlin, *et al.*, Adaptive radiation along a deeply conserved genetic line of least resistance in *Anolis* lizards. *Evol. Lett.* 310–322 (2018). https://doi.org/10.1002/evl3.72.

50. J. C. Fay, P. J. Wittkopp, Evaluating the role of natural selection in the evolution of gene regulation. *Heredity* **100**, 191–199 (2008).

51. D. M. Fowler, J. J. Stephany, S. Fields, Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).

52. M. L. Bulyk, X. Huang, Y. Choo, G. M. Church, Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci.* **98**, 7158–7163 (2001).

53. D. W. Anderson, A. N. McKeown, J. W. Thornton, Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, e07864–e07864 (2015).

54. L. C. Wheeler, M. J. Harms, Were Ancestral Proteins Less Specific? *Mol. Biol. Evol.* **38**, 2227–2239 (2021).

55. G. K. A. Hochberg, J. W. Thornton, Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu. Rev. Biophys.* **46**, 247–269 (2017).

56. T. N. Starr, L. K. Picton, J. W. Thornton, Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).

57. Y. Park, B. P. H. Metzger, J. W. Thornton, Epistatic drift causes gradual decay of predictability in protein evolution. *Science* **376**, 823–830 (2022).

58. A. N. McKeown, *et al.*, Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* **159**, 58–68 (2014).

59. J. S. Carroll, *et al.*, Genome-wide analysis of estrogen receptor binding sites. *Nat. Genet.* **38**, 1289–1297 (2006).

60. L. C. Watson, *et al.*, The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat. Struct. Mol. Biol.* **20**, 876–883 (2013).

61. Y. Park, B. P. H. Metzger, J. W. Thornton, The simplicity of protein sequence-function relationships. *Nat. Commun.* **15**, 7953 (2024).

62. B. P. H. Metzger, Y. Park, T. N. Starr, J. W. Thornton, Epistasis facilitates functional evolution in an ancient transcription factor. *eLife* **12** (2024).

63. S. Gerber, Not all roads can be taken: Development induces anisotropic accessibility in morphospace. *Evol. Dev.* **16**, 373–381 (2014).

64. S. Psujek, R. D. Beer, Developmental bias in evolution: evolutionary accessibility of phenotypes in a model evo-devo system. *Evol. Dev.* **10**, 375–390 (2008).

65. J. Maynard-Smith, Natural Selection and the Concept of a Protein Space. *Nature* **225**, 726–734 (1970).

66. M. E. J. Newman, M. Girvan, Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113 (2004).

67. M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).

68. F. Russo, J. Williamson, Interpreting Causality in the Health Sciences. *Int. Stud. Philos. Sci.* **21**, 157–170 (2007).

69. W. Fontana, P. Schuster, Shaping space: The possible and the attainable in RNA genotype-phenotype mapping. *J Theor Biol* **194**, 491–515 (1998).

70. P. Alberch, Ontogenesis and Morphological Diversification. *Am. Zool.* **20**, 653–667 (1980).

71. A. D. Chipman, W. Arthur, M. Akam, A Double Segment Periodicity Underlies Segment Generation in Centipede Development. *Curr. Biol.* **14**, 1250–1255 (2004).

72. T. Fuqua, *et al.*, Dense and pleiotropic regulatory information in a developmental enhancer. *Nature* **587**, 235–239 (2020).

73. W. Arthur, M. Farrow, The Pattern of Variation in Centipede Segment Number as an Example of Developmental Constraint in Evolution. *J. Theor. Biol.* **200**, 183–191 (1999).

74. E. Harjunmaa, *et al.*, Replaying evolutionary transitions from the dental fossil record. *Nature* **512**, 44–48 (2014).

75. R. Galupa, *et al.*, Enhancer architecture and chromatin accessibility constrain phenotypic space during Drosophila development. *Dev. Cell* **58**, 51-62.e4 (2023).

76. E. Ferrada, A. Wagner, Evolutionary innovations and the organization of protein functions in genotype space. *PLoS ONE* **5** (2010).

77. S. Ciliberti, O. C. Martin, A. Wagner, Innovation and robustness in complex regulatory gene networks. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 13591–13596 (2007).

78. J. F. Matias Rodrigues, A. Wagner, Evolutionary Plasticity and Innovations in Complex Metabolic Reaction Networks. *PLoS Comput. Biol.* **5**, e1000613 (2009).

79. S. J. Gould, N. Eldredge, Punctuated Equilibria : The Tempo and Mode of Evolution Reconsidered. *Paleobiology* **3**, 115–151 (1977).

80. P. Alberch, E. A. Gale, A developmental analysis of an evolutionary trend: digital reduction in amphibians. *Evolution* **39**, 8–23 (1985).

81. C. Braendle, C. F. Baer, M.-A. Félix, Bias and Evolution of the Mutationally Accessible Phenotypic Space in a Developmental System. *PLoS Genet.* **6**, e1000877 (2010).

82. A. M. Phillips, *et al.*, Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife* **10**, 1–40 (2021).

83. T. N. Starr, *et al.*, ACE2 binding is an ancestral and evolvable trait of sarbecoviruses. *Nature* **603**, 913–918 (2022).

84. T. J. Ord, T. C. Summers, Repeated evolution and the impact of evolutionary history on adaptation. *BMC Evol. Biol.* **15**, 137 (2015).

85. C. J. Maclean, *et al.*, Deciphering the Genic Basis of Yeast Fitness Variation by Simultaneous Forward and Reverse Genetics. *Mol. Biol. Evol.* **34**, 2486–2502 (2017).

86. R.D. Gietz, R.A. Woods, "Yeast Transformation by the LiAc/SS Carrier DNA/PEG Method" in *Yeast Protocol, W. Xiao, Ed.*, Second edition, (Humana Press, 2006), pp. 107–120.

87. A. Yick-Lun So, C. Chaivorapol, E. C. Bolton, H. Li, K. R. Yamamoto, Determinants of Cell-and Gene-Specific Transcriptional Regulation by the Glucocorticoid Receptor. *PLoS Genet.* **3**, e94–e94 (2007).

88. W. Welboren, *et al.*, ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.* **28**, 1418–1428 (2009).

89. T. C. Scanlon, E. C. Gray, K. E. Griswold, Quantifying and resolving multiple vector transformants in S. cerevisiae plasmid libraries. *BMC Biotechnol.* **9**, 95 (2009).

90. K. Mir, K. Neuhaus, M. Bossert, S. Schober, Short Barcodes for Next Generation Sequencing. *PLOS ONE* **8**, e82933 (2013).

91. N.A. Joshi, J.N. Fass, Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (2011). Deposited 2011.

92. J. Zhang, K. Kobert, T. Flouri, A. Stamatakis., PEAR: A fast and accurate Illimuna Paired-End reAd mergeR. (2015). Deposited 2015.

93. P. J. A. Cock, *et al.*, Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

94. Jerome Friedman, *et al.*, glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.

95. Csardi, G., Nepusz, T., The igraph software package for complex network research. *InterJournal* **Complex Systems**, 1695 (2006).

96. Bastian, M., Heymann, S., Jacomy, M., Gephi: an open source software for exploring and manipulating networks. *Proc. Int. AAAI Conf. Web Soc. Media* **3**, 361–362 (2009).

97. J. Gillespie, Molecular Evolution Over the Mutational Landscape. *Evolution* **38**, 1116–1129 (1984).

98. D. M. Mccandlish, A. Stoltzfus, Modeling Evolution Using the Probability of Fixation: History and Implications. *Q. Rev. Biol.* **89**, 225–252 (2014).

99. W. Arthur, The effect of development on the direction of evolution: toward a twenty-first century consensus. *Evol. Dev.* **6**, 282–288 (2004).

100. S. E. Ahnert, Structural properties of genotype – phenotype maps. (2017).

101. D. Jiang, M. Pennell, Alternative mutational architectures producing identical M-matrices can lead to different patterns of evolutionary divergence. [Preprint] (2023). Available at: http://biorxiv.org/lookup/doi/10.1101/2023.08.11.553044 [Accessed 12 December 2024].

102. T. F. Hansen, The Evolution of Genetic Architecture. *Annu. Rev. Ecol. Evol. Syst.* **37**, 123–157 (2006).

103. P. C. Phillips, Epistasis - The essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.* **9**, 855–867 (2008).

104. E. Borenstein, D. C. Krakauer, An end to endless forms: Epistasis, phenotype distribution bias, and nonuniform evolution. *PLoS Comput. Biol.* **4**, e1000202 (2008).

105. A. G. Jones, R. Bürger, S. J. Arnold, Epistasis and natural selection shape the mutational architecture of complex traits. *Nat. Commun.* **5**, 1–10 (2014).

106. M. Lagator, S. Sarikas, H. Acar, J. P. Bollback, C. C. Guet, Regulatory network structure determines patterns of intermolecular epistasis. *eLife* **6**, 1–22 (2017).

107. A. Wagner, Genotype networks shed light on evolutionary constraints. *Trends Ecol. Evol.* **26**, 577–84 (2011).

108. J. Aguilar-Rodríguez, J. L. Payne, A. Wagner, A thousand empirical adaptive landscapes and their navigability. *Nat. Ecol. Evol.* **1**, 1–9 (2017).

109. D. M. Weinreich, R. a Watson, L. Chao, Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* **59**, 1165–1174 (2005).

110. B. P. H. Metzger, Y. Park, T. N. Starr, J. W. Thornton, Epistasis facilitates functional evolution in an ancient transcription factor. *eLife* **12**, RP88737 (2024).

111. D. W. Anderson, A. N. McKeown, J. W. Thornton, Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, 1–26 (2015).

112. A. S. B. Jalal, *et al.*, Diversification of DNA-Binding Specificity by Permissive and Specificity-Switching Mutations in the ParB/Noc Protein Family. *Cell Rep.* **32**, 107928 (2020).

113. A. Clauset, M. E. J. Newman, C. Moore, Finding community structure in very large networks. [Preprint] (2004). Available at: http://arxiv.org/abs/cond-mat/0408187 [Accessed 26 November 2024].

114. P. J. Bowler, "Variation from darwin to the modern synthesis" in *Variation: A Central Concept in Biology*, B. Hallgrímsson, B. K. Hall, Eds. (Elsevier, 2011), pp. 9–27.

115. A. Goswami, J. B. Smaers, C. Soligo, P. D. Polly, The macroevolutionary consequences of phenotypic integration: From development to deep time. *Philos. Trans. R. Soc. B Biol. Sci.* **369** (2014).

116. J. Domingo, P. Baeza-Centurion, B. Lehner, The Causes and Consequences of Genetic Interactions (Epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).

117. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci.* **25**, 1204–1218 (2016).

118. C. J. McClune, A. Alvarez-Buylla, C. A. Voigt, M. T. Laub, Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. *Nature* **574**, 702–706 (2019).

119. A. M. Bendel, *et al.*, The genetic architecture of protein interaction affinity and specificity. *Nat. Commun.* **15**, 8868 (2024).

120. D. M. Weinreich, N. F. Delaney, M. A. Depristo, D. L. Hartl, Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 2004–2007 (2006).

121. T. F. Hansen, G. P. Wagner, Modeling genetic architecture: A multilinear theory of gene interaction. *Theor. Popul. Biol.* **59**, 61–86 (2001).

122. M. J. Harms, J. W. Thornton, Evolutionary biochemistry: Revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14**, 559–571 (2013).

123. Sean B Carroll, *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom* (WW Norton and Co, 2005).

124. A. S. Pillai, *et al.*, Origin of complexity in haemoglobin evolution. *Nature* **581**, 480–485 (2020).

125. T. M. Williams, *et al.*, The Regulation and Evolution of a Genetic Switch Controlling Sexually Dimorphic Traits in Drosophila. *Cell* **134**, 610–623 (2008).

126. M. J. Roeske, E. M. Camino, S. Grover, M. Rebeiz, T. M. Williams, Cis-regulatory evolution integrated the Bric-à -brac transcription factors into a novel fruit fly gene regulatory network. *eLife* **7**, 1–28 (2018).

127. J. T. Bridgham, *et al.*, Protein evolution by molecular tinkering: Diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol.* **8** (2010).

128. J. T. Bridgham, S. M. Carroll, J. W. Thornton, Evolution of Hormone-Receptor Complexity by Molecular Exploitation. 97–101 (2006).

129. L. T. Shirai, *et al.*, Evolutionary history of the recruitment of conserved developmental genes in association to the formation and diversification of a novel trait. *BMC Evol. Biol.* **12**, 21 (2012).

130. F. Jacob, Evolution and tinkering. *Science* **196**, 1161–1166 (1977).

131. N. Shubin, C. Tabin, S. Carroll, Deep homology and the origins of evolutionary novelty. *Nature* **457**, 818–23 (2009).

132. M. Levine, E. H. Davidson, Gene regulatory networks for development. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 4936–4942 (2005).

133. R. Nielsen, Z. Yang, Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics* **148**, 929–936 (1998).

134. J. L. King, T. H. Jukes, Non-Darwinian evolution. *Science* **164**, 788–798 (1969).

135. M. Kimura, T. Ota, On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* **71**, 2848–2852 (1974).

136. I. S. Peter, E. H. Davidson, Evolution of gene regulatory networks controlling body plan development. *Cell* **144**, 970–985 (2011).

137. I.-G. Choi, S.-H. Kim, Evolution of protein structural classes and protein sequence families. *PNAS* **103**, 14056–14061 (2006).

138. L. Y. Yampolsky, A. Stoltzfus, Bias in the introduction of variation as an orienting factor in evolution. *Evol. Dev.* **3**, 73–83 (2001).

139. A. V. Cano, J. L. Payne, Mutation bias interacts with composition bias to influence adaptive evolution. *PLoS Comput. Biol.* **16**, 1–26 (2020).

140. J. F. Storz, *et al.*, The role of mutation bias in adaptive molecular evolution: Insights from convergent changes in protein function. *Philos. Trans. R. Soc. B Biol. Sci.* **374** (2019).

141. I. Salazar-Ciudad, M. Marín-Riera, Adaptive dynamics under development-based genotype-phenotype maps. *Nature* **497**, 361–364 (2013).

142. J. Echave, S. J. Spielman, C. O. Wilke, Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**, 109–121 (2016).

143. M. A. DePristo, D. M. Weinreich, D. L. Hartl, Missense meanderings in sequence space: A biophysical view of protein evolution. *Nat. Rev. Genet.* **6**, 678–687 (2005).

144. B. R. Jack, A. G. Meyer, J. Echave, C. O. Wilke, Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biol.* (2016). https://doi.org/10.1371/journal.pbio.1002452.

145. D. Vitkup, P. Kharchenko, A. Wagner, Influence of metabolic network structure and function on enzyme evolution. *Genome Biol.* **7** (2006).

146. D. H. Erwin, E. H. Davidson, The evolution of hierarchical gene regulatory networks. *Nat. Rev. Genet.* **10**, 141–148 (2009).

147. H. Hu, *et al.*, Constrained vertebrate evolution by pleiotropic genes. *Nat. Ecol. Evol.* **1**, 1722–1730 (2017).

148. J. M. Good, M. W. Nachman, Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. *Mol. Biol. Evol.* **22**, 1044–1052 (2005).

149. J. Liu, M. Robinson-Rechavi, Adaptive evolution of animal proteins over development: Support for the Darwin selection opportunity hypothesis of evo-devo. *Mol. Biol. Evol.* **35**, 2862–2872 (2018).

150. R. N. Felice, A. Goswami, Developmental origins of mosaic evolution in the avian cranium. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 555–560 (2018).

151. G. Navalón, A. Bjarnason, E. Griffiths, R. B. J. Benson, Environmental signal in the evolutionary diversification of bird skeletons. *Nature* **611**, 306–311 (2022).

152. A. Watanabe, *et al.*, Ecomorphological diversification in squamates from conserved pattern of cranial integration. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 14688–14697 (2019).

153. A. Goswami, *et al.*, Developmental origin underlies evolutionary rate variation across the placental skull. *Philos. Trans. R. Soc. B Biol. Sci.* **378**, 20220083 (2023).

154. S. Chaurasia, J. Y. Dutheil, The Structural Determinants of Intra-Protein Compensatory Substitutions. *Mol. Biol. Evol.* **39**, 1–16 (2022).

155. J. Y. Dutheil, F. Jossinet, E. Westhof, Base pairing constraints drive structural epistasis in ribosomal RNA sequences. *Mol. Biol. Evol.* **27**, 1868–1876 (2010).

156. M. A. Stiffler, *et al.*, Protein Structure from Experimental Evolution. *Cell Syst.* **10**, 15-24.e5 (2020).

157. R. Assis, Strong Epistatic Selection on the RNA Secondary Structure of HIV. *PLoS Pathog.* **10** (2014).

158. J. L. Steenwyk, *et al.*, An orthologous gene coevolution network provides insight into eukaryotic cellular and genomic structure and function. *Sci. Adv.* **8** (2022).

159. H. B. Fraser, A. E. Hirsh, D. P. Wall, M. B. Eisen, Coevolution of gene expression among interacting proteins. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9033–9038 (2004).

160. N. M. Young, G. P. Wagner, B. Hallgrímsson, Development and the evolvability of human limbs. *Proc Nat Acad Sci* **107** (2010).

161. N. M. Young, B. HallgrÍmsson, Serial Homology and the Evolution of Mammalian Limb Covariation Structure. *Evolution* **59**, 2691 (2005).

162. T. E. Hansen, The Evolution of Genetic Architecture. *Annu Rev Ecol Evol Syst* **37**, 123–157 (2006).

163. F. J. Poelwijk, V. Krishna, R. Ranganathan, The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Comput. Biol.* **12**, 1–19 (2016).

164. B. P. H. Metzger, Y. Park, T. N. Starr, J. W. Thornton, Epistasis facilitates functional evolution in an ancient transcription factor. *eLife* 2023.04.19.537271 (2023).

165. A. J. R. Carter, J. Hermisson, T. F. Hansen, The role of epistatic gene interactions in the response to selection and the evolution of evolvability. *Theor. Popul. Biol.* **68**, 179–196 (2005).

166. D. M. Tufts, *et al.*, Epistasis constrains mutational pathways of hemoglobin adaptation in high-altitude pikas. *Mol. Biol. Evol.* **32**, 287–298 (2015).

167. B. K. Shoichet, W. A. Baase, R. Kuroki, B. W. Matthews, A relationship between protein stability and protein function. *Proc Natl Acad Sci U A* **92**, 452–456 (1995).

168. J. Briscoe, S. Small, Morphogen rules: Design principles of gradient-mediated embryo patterning. *Development* **142**, 3996–4009 (2015).

169. R. M. Green, *et al.*, Developmental nonlinearity drives phenotypic robustness. *Nat. Commun.* **8** (2017).

170. N. M. Young, H. J. Chong, D. Hu, B. Hallgrímsson, R. S. Marcucio, Quantitative analyses link modulation of sonic hedgehog signaling to continuous variation in facial growth and shape. *Development* **137**, 3405–3409 (2010).

171. B. Hallgrímsson, *et al.*, Integration and the Developmental Genetics of Allometry. *Integr. Comp. Biol.* **59**, 1369–1381 (2019).

172. C. K. Mirth, W. A. Frankino, A. W. Shingleton, Allometry and size control : what can studies of body size regulation teach us about the evolution of morphological scaling relationships ? *Curr. Opin. Insect Sci.* **13**, 93–98 (2016).

173. S. J. Gould, Allometry and Size in Ontogeny and Phylogeny. *Biol Rev* **41**, 587–640 (1966).

174. D. J. Emlen, H. F. Nijhout, Hormonal control of male horn length dimorphism in the dung beetle Onthophagus taurus (Coleoptera: Scarabaeidae). *J. Insect Physiol.* **45**, 45–53 (1999).

175. Sarah M. Ardell, Alena Martsul, Milo S. Johnson, Sergey Kryazhimskiy, Environment-independent distribution of mutational effects emerges from microscopic epistasis. *Science* **386**, 87–92.

176. S. Gerber, G. J. Eble, P. Neige, Allometric space and allometric disparity: A developmental perspective in the macroevolutionary analysis of morphological disparity. *Evolution* **62**, 1450–1457 (2008).

177. G. Reddy, M. M. Desai, Global epistasis emerges from a generic model of a complex trait. *eLife* **10**, e64740 (2021).

178. T. N. Starr, J. M. Flynn, P. Mishra, D. N. A. Bolon, J. W. Thornton, Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. *Proc Nat Acad Sci* **115**, 4453–4458 (2018).

179. C. K. Dalal, A. D. Johnson, How transcription circuits explore alternative architectures while maintaining overall circuit output. *Genes Dev.* **31**, 1397–1405 (2017).

180. K. R. Wotton, *et al.*, Quantitative system drift compensates for altered maternal inputs to the gap gene network of the scuttle fly megaselia abdita. *eLife* **2015**, 1–28 (2015).

181. J. R. True, E. S. Haag, Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* **3**, 109–119 (2001).

182. K. M. Weiss, S. M. Fullerton, Phenogenetic Drift and the Evolution of Genotype Phenotype Relationships. *Theor. Popul. Biol.* **57**, 187–195 (2000).

183. A. Wagner, Neutralism and selectionism: a network-based reconciliation. *Nat. Genet.* **9**, 965–974 (2008).

184. W. Fontana, P. Schuster, Continuity in evolution: On the nature of transitions. *Science* **280**, 1451–1455 (1998).

185. E. H. Davidson, D. H. Erwin, Gene regulatory networks and the evolution of animal body plans. *Science* **311**, 796–801 (2006).

186. J. W. Valentine, *On the origin of phyla* (The University of Chicago Press, 2004).

187. K. E. Willmore, The Body Plan Concept and Its Centrality in Evo-Devo. *Evol. Educ. Outreach* **5**, 219–230 (2012).

188. J. M. W. Slack, P. W. H. Holland, C. F. Graham, The zootype and the phylotypic stage. *Nature* **361**, 490–492 (1993).

189. L. Gramzow, G. Theissen, A hitchhiker's guide to the MADS world of plants. *Genome Biol.* **11** (2010).

190. T. Honma, K. Goto, Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature* **409**, 525–529 (2001).

191. R. Riedl, A systems-analytical approach to macro-evolutionary phenomena. *Q. Rev. Biol.* **52**, 351–370 (1977).

192. P. Shah, D. M. McCandlish, J. B. Plotkin, Contingency and entrenchment in protein evolution under purifying selection. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3226–E3235 (2015).

193. D. L. Bain, A. F. Heneghan, K. D. Connaghan-Jones, M. T. Miura, Nuclear Receptor Structure: Implications for Function. *Annu. Rev. Physiol.* **69**, 201–220 (2007).

194. R. R. Schoch, Riedl's burden and the body plan: Selection, constraint, and deep time. *J Exp Zool Mol Dev Evol* **314B**, 1–10 (2009).

195. G. P. Wagner, M. D. Laubichler, Rupert riedl and the re-synthesis of evolutionary and developmental biology: Body plans and evolvability. *J. Exp. Zoolog. B Mol. Dev. Evol.* **302B**, 92–102 (2004).

196. G. P. Wagner, K. Kin, L. Muglia, M. Pavličev, Evolution of mammalian pregnancy and the origin of the decidual stromal cell. *Int. J. Dev. Biol.* **58**, 117–126 (2014).

197. V. J. Lynch, G. May, G. P. Wagner, Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. *Nature* **480**, 383–386 (2011).

198. J. K. Grenier, S. B. Carroll, Functional evolution of the ultrabithorax protein. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 704–709 (2000).

199. M. Ronshaugen, N. Mcginnis, W. Mcginnis, Hox protein mutation and macroevolution of the insect body plan. *Nature* **415**, 1–4 (2002).

200. H. Sauquet, *et al.*, The ancestral flower of angiosperms and its early diversification. *Nat. Commun.* **8** (2017).

201. P. Ruelens, *et al.*, The origin of floral organ identity quartets. *Plant Cell* **29**, 229–242 (2017).

202. D. Vlad, *et al.*, Leaf shape evolution through duplication, regulatory diversification and loss of a homeobox gene. *Science* **343**, 780–783 (2014).

203. M. Hajheidari, *et al.*, Autoregulation of RCO by Low-Affinity Binding Modulates Cytokinin Action and Shapes Leaf Diversity. *Curr. Biol.* **29**, 4183-4192.e6 (2019).

204. M. I. Coates, J. A. Clack, Polydactyly in the earliest known tetrapod limbs. *Nature* **347**, 66–69 (1990).

205. F. Lim, *et al.*, Affinity-optimizing enhancer variants disrupt development. *Nature* **626**, 151–159 (2024).

206. A. H. Newton, A. J. Pask, Evolution and expansion of the RUNX2 QA repeat corresponds with the emergence of vertebrate complexity. *Commun. Biol.* **3** (2020).

207. K. E. Sears, A. Goswami, J. J. Flynn, L. A. Niswander, The correlated evolution of Runx2 tandem repeats, transcriptional activity, and facial length in Carnivora. *Evol. Dev.* **9**, 555–565 (2007).

208. P. Alberch, From genes to phenotype: dynamical systems and evolvability. *Genetica* **84**, 5–11 (1991).

209. D. Kierzkowski, *et al.*, A Growth-Based Framework for Leaf Shape Development and Diversity. *Cell* **177**, 1405-1418.e17 (2019).

210. J. O. Vik, *et al.*, Genotype-phenotype map characteristics of an in silico heart cell. *Front. Physiol.* **2** (2011).

211. S. Tarasov, Integration of Anatomy Ontologies and Evo-Devo Using Structured Markov Models Suggests a New Framework for Modeling Discrete Phenotypic Traits. *Syst. Biol.* **68**, 698–716 (2019).

212. D. S. Porto, J. Uyeda, I. Mikó, S. Tarasov, ontophylo: Reconstructing the evolutionary dynamics of phenomes using new ontology-informed phylogenetic methods. *Methods Ecol. Evol.* **15**, 290–300 (2024).

213. A. K. Aditham, C. J. Markin, D. A. Mokhtari, N. DelRosso, P. M. Fordyce, High-Throughput Affinity Measurements of Transcription Factor and DNA Mutations Reveal Affinity and Specificity Determinants. *Cell Syst.* **12**, 112-127.e11 (2021).

214. Kryazhimskiy, Sergey, Emergence and propagation of epistasis in metabolic networks. *eLife* **10**, e60200 (2021).

215. M. Olivetta, C. Bhickta, N. Chiaruttini, J. Burns, O. Dudin, A multicellular developmental program in a close animal relative. *Nature* **635**, 382–389 (2024).

216. A. Pires-daSilva, R. J. Sommer, The evolution of signalling pathways in animal development. *Nat. Rev. Genet.* **4**, 39–49 (2003).

217. Q. Liu, *et al.*, Ancient mechanisms for the evolution of the bicoid homeodomain's function in fly development. *eLife* **7**, 1–28 (2018).