THE UNIVERSITY OF CHICAGO

THE POSSIBLE AND THE ACCESSIBLE: EPISTASIS AND CONTINGENCY IN PROTEIN SEQUENCE SPACE

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

COMMITTEE ON GENETICS, GENOMICS AND SYSTEMS BIOLOGY

BY

YEONWOO PARK

CHICAGO, ILLINOIS

AUGUST 2022

Table of Contents

List of Figures

List of Tables

Acknowledgements

The past five years have taught me emphatically that science is a social and collaborative endeavor. My first thanks go to our lab members past and present. My advisor Joe Thornton demonstrated what it is to think and write rigorously and how to move beyond one's self-perceived limit. His 45-mile bike commute everyday inspired me to take up cycling, which proved essential for mental and physical health, particularly during the pandemic. The image of scientist he embodies will be a one I look up to for the many coming years. Brian Metzger, with his 1,000-rpm mind, made key contributions to my projects, without which they would not have taken their shape. Georg Hochberg and Arvind Pillai were always in the lab; their nightly brain screen-saver talks were truly entertaining. Jaeda Patton, Santiago Herrera, and Carlos Cortez joined the lab two years after me. Without them, my laboratory life would have been desolate. Particular thanks to Jaeda for her cycling companionship. Patrick Cantwell, Ricardo Muniz-Trejo, and Max Bogan recently joined the lab. They made me realized the joy of building a relationship through science. Finally, without the zeal of Hanna Bascom, the past couple months of my Ph.D. would not have been possible. I would also like to thank my committee members—Allan Drummond, Rama Ranganathan, and Engin Özkan—whose mere existence has been an unwavering motivation for me. Finally, thanks to Sue Levison and Bonnie Brown whose administrative help was essential for my Ph.D. life.

I would also like to thank the friends from and before Chicago. Particular thanks to Stephanie Sang, Carlos Servan, and Soo Ji Kim. We began with a tiny studio without a dining table. Over the past six years we have gone on to live in a giant two-bedroom apartment, do a second Ph.D., and get a dog and a husband. I would also like to celebrate my friendship with Jong Yeon Lee spanning high school, college, and Ph.D. and two countries. Finally, Little Cat

and Meow Meow have been the dearest companion during the pandemic. These people (and cats) have kept me going in times of hardship.

All of this could not have been possible without the unwavering support of my family. I have not had a chance to see them for the past three years. Being in a boarding school and moving abroad for college and graduate school means I have been mostly absent for almost 15 years. I realized we do not have a family photo where I don't look unrecognizably youthful. My sister has moved to San Francisco a few years ago with her husband. Her mere presence in the same continent has been a great relief, although I have never expressed it. My mom wrote eleven children's fiction books over the past ten years. Her productivity has always astonished me and kept me humble. My dad visualized my work on epistatic drift in a painting, which *Science* took to introduce the paper. The colors along horizontal lines represent the coming and going of diversity in evolution; the moving curves represent the changing accessibility of evolutionary opportunities. This has become my image of evolution.

Last thanks go to the dearest. Zoe and I met in Chicago five and a half years ago. Three and a half of those years were spent in opposite sides of the globe. The pandemic meant for two entire years we could not see each other. Despite that, we persisted and married. I cannot be more excited for the shared future that lays in front of us. We embody the great principle of historical contingency, where chance caught on wing turns into necessity.

Abstract

Epistatic interactions determine the phenotypic consequences of mutations and shape the course of evolution. However, little is known about the pattern of epistatic interactions among the possible mutations within a protein and the extent and temporal dynamics with which the effects of mutations change during evolution. By using a novel method to analyze experimental mutational datasets, I show that the architecture of epistatic interactions within a protein is surprisingly simple: Knowing only the context-independent effects and pairwise interactions of amino acids is sufficient to predict the phenotype with high accuracy. I then combine ancestral protein reconstruction with deep mutational scanning to experimentally reconstruct how the effect of every possible point mutation in a protein changed during long-term evolution. The effects of most mutations changed gradually and randomly at a rate characteristic to each mutation—a pattern I call epistatic drift. Epistatic drift randomized the effects of most mutations during evolution, making the outcome of evolution highly unpredictable. The statistical regularity of epistatic drift, however, means that this unpredictability can be quantified: A probability distribution for the future effect of a mutation—therefore the timescale at which evolution becomes unpredictable—can be calculated from the rate of epistatic drift. Overall, my work reveals a simple architecture and statistical regularity of epistasis and demonstrates the pervasive historical contingency of protein evolution.

Chapter 1

Introduction

1.1 Structuralism in molecular evolution

Evolution proceeds through the origin of heritable variations in individuals due to mutations and the sorting of these variations through population genetic processes that lead to their eventual fixation or removal. While both the origin and sorting of variations contribute to the outcome of evolution, evolutionary biologists have focused on the sorting of variations as the primary cause of evolution since the development of population genetics in the early twentieth century (*1*). Organisms were regarded as black-boxes that somehow produce small heritable variations in all aspects of biology in a random manner; natural selection was considered the agent that channels these random variations into a coherent direction, culminating in adaptations that fulfill the functional needs of organisms in their environments. The neutral theory of molecular evolution, which challenged functionalists to appreciate the role of chance genetic drift in the shaping of molecular diversity, suggested that natural selection has blind spots and that mutations can have an autonomous role in sculpting molecular diversity (*2*). However, research programs for understanding what phenotypic variations can be produced by mutations and whether any bias or constraint in this process has shaped the outcome of evolution did not mature until the recent development of powerful molecular genetics tools.

Structuralism recognizes that organisms have internal structures that make certain phenotypic variations more likely to arise than others (*3*). The accessibility of a particular phenotypic variation may depend on the internal configuration of the system and therefore historically contingent. The structural properties and evolutionary histories of biological systems,

structuralists argue, are fundamental causes that shape the pattern of biological diversity. The goal of a structuralist research program is then to characterize the possible ways in which a biological system can vary through mutations and reconstruct how the accessibility of particular variations at particular historical moments shaped the course of evolution. These tasks require systematically interrogating the effects of possible mutations and reconstructing evolutionary histories, which have only recently become possible.

A paradigmatic example of structuralist reasoning is the explanation for the marginal stability of proteins (*4*). Most proteins are stable enough to be folded in their native environment but not more stable than that. Functionalists interpret marginal stability as a result of functional optimization. They argue that more stable proteins are less active and therefore disfavored by natural selection, resulting in an active protein with a minimal required stability. The validity of this argument rests on the proposed functional trade-off between stability and activity, which has not been well corroborated. An alternative explanation rests on the well documented fact that random mutations are much more likely to destabilize a protein than to stabilize it. This mutational pressure to reduce stability is counterbalanced by purifying selection for maintaining sufficient amount of folded proteins. Any extra stability is quickly eroded by mutations, which are invisible to selection as long as they do not affect the amount of folded proteins—resulting in marginal stability. The core of this argument is that phenotypic variations generated by random mutations are biased—more likely to result in unstable than stable proteins. This mutational bias, in turn, reflects the structural fact that folded proteins comprise a miniscule fraction of possible protein sequences.

Structuralist explanations based on the distribution of random mutations' effects have been put forward as alternatives to functional explanations for a number of other molecular

phenomena. The essentiality of a trait in present-day organisms is commonly cited as evidence for the adaptive origin of the trait. However, structuralist studies show that traits that arise nonadaptively can later become essential. This has been best demonstrated for protein-protein interactions (*5*, *6*). Random nucleotide changes tend to create more hydrophobic amino acids than that can be tolerated on solvent-exposed protein surfaces. Once a surface is shielded in an interface, however, it can accommodate this mutational bias and acquire hydrophobic substitutions. Over time, removing the interface becomes deleterious because it exposes the accumulated hydrophobic residues. In this way, an interface that arose nonadaptively can become essential and maintained by purifying selection. The essentiality of a present-day interface, which seems to warrant a functionalist explanation for its origin, can therefore be explained as an expected outcome of a non-adaptive mutational process.

1.2 Epistasis and historical contingency

Natural selection is the chief causal factor in the functionalist explanation of evolution. In structuralism, the phenotypic effects of mutations assume that role: In the above examples of structuralist reasoning, patterns of molecular diversity were explained by the propensity of mutations to bring about certain phenotypic changes, often through interplay with selection. A critical structuralist agenda is therefore to understand why mutations have the phenotypic effects they do.

A major determinant of the effect of a mutation in a protein is its interaction with other mutations in the same protein—called epistasis (*7*). A mutation that breaks the protein may be permitted in the background of another mutation that contributes favorable interactions or removes unfavorable interactions. Conversely, a mutation that is otherwise tolerable may

become restricted if preceded by incompatible epistatic partners. What epistatic partners are present in the genetic background at a particular point in time therefore determines the phenotypic effects of mutations at that time.

Epistatic interactions, for this reason, can cause historical contingency. Mutations may be accessible in some historical intervals but not in others depending on the epistatic partners present. The outcome of evolution can thus be contingent on the historical starting point and the intervening sequence substitutions that modulate mutational opportunities. Depending on the prevalence and magnitude of epistasis, historical contingency due to epistasis can be strong, causing the evolution of even highly adaptive function to be contingent on chance genetic drift (*8*). For example, if a mutation necessary for an adaptive function must be permitted by a particular mutation, natural selection must wait for the permissive mutation to fix by genetic drift or as a by-product of selection for another function irrespective of how beneficial the eventual function may be.

My thesis research addresses two key questions on epistasis: What is the pattern of epistatic interactions among the possible mutations within a protein? What is the extent and temporal dynamics with which the effects of mutations change during evolution due to epistatic interactions with sequence substitutions? These questions form pillars of structuralist approach to molecular evolution. However, despite numerous studies that examined epistatic interactions within proteins, they remain poorly addressed. Below I explain why this knowledge gap persists and lay out the novel approaches I developed to address these questions.

1.3 Phylogenetic deep mutational scanning

Despite the recent avalanche of works examining epistatic interactions within proteins, the role of epistasis in protein evolution remains poorly characterized. This is because no study has systematically traced how the effects of mutations change during long-term evolution. Deep mutational scanning, which combines the generation of a large number of protein variants with a high-throughput sequencing-based functional assay to phenotype thousands to millions of protein variants in a single experiment, has enabled the interrogation of epistatic interactions among the many possible mutations within a protein (*9–14*). However, studies that apply deep mutational scanning to a single protein can only indirectly suggest a role for epistasis in evolution because what is relevant for evolution is not the genetic interactions among all possible mutations but the extent to which the particular sequence substitutions that fix during evolution alter the effects of other mutations.

Comparing the effects of the same mutations among homologous proteins provided the first evidence that epistasis is relevant to protein evolution: Differences in the effects of mutations indicate that sequence substitutions that accrued during the divergence of the homologs have opened or closed mutational opportunities (*15–22*). However, because these studies only examine the effects of mutations on present-day homologs, they cannot reveal the temporal dynamics with which the effects of mutations change during evolution.

A number of studies employed ancestral sequence reconstruction to trace how the effects of a few mutations changed during evolution (*23–30*). Ancestral sequence reconstruction is a statistical method for inferring the most likely sequences of ancestral proteins by analyzing the sequences of their present-day descendants (*31*). Several case studies employed ancestral sequence reconstruction to show that mutations required for the evolution of a new function were deleterious in an ancestral background and relied on permissive substitutions, indicating strong

historical contingency on chance epistatic interactions (*8*, *23*, *25*, *27*). However, these studies examined only a small number of mutations and compared their effects between the beginning and end of a single phylogenetic interval; the overall extent of epistasis and temporal dynamics of epistasis remain unknown.

I address this knowledge gap by combining deep mutational scanning and ancestral sequence reconstruction to comprehensively trace how the effect of every possible point mutation in a protein changed along a densely reconstructed evolutionary trajectory.

1.4 Protein sequence space and genotype-phenotype map

Analyzing the architecture of epistatic interactions and its impact on evolution is facilitated by the notion of sequence space (*32*). Sequence space is a graph where each possible sequence is a node and two sequences are connected if they are related by a single mutation. Evolution by point mutation can be conceptualized as a continuous walk on this graph. Mapping each genotype to its phenotype then allows us to ask how the distribution of phenotype in sequence space affects evolutionary trajectories.

Central questions in protein biochemistry and evolution can be recast as understanding the genotype-phenotype map (*33*). For biochemistry, knowing the genotype-phenotype map means knowing the genetic architecture of a protein—how the individual amino acids and their interactions determine the phenotype. For evolutionary biology, knowing the genotype-phenotype map means knowing all possible ways in which the phenotype can vary and the accessibility of a particular phenotypic value from any starting point.

Genotype-phenotype maps are difficult to characterize because of their astronomical size. Even a 100-aa protein has 20^{100} possible genotypes, which is far greater than the number of

atoms in the universe and can never be exhaustively characterized. Analyzing the effects of mutations has the potential to describe a genotype-phenotype map using a much fewer number of parameters than there are genotypes. For example, if each of the 19 × 100 possible point mutations in a 100-aa protein act independently of each other, the phenotype of any genotype can be calculated as the sum of the contribution of each point mutation. This amounts to a dramatic reduction of information from what could be 20^{100} unique parameters to just 1,900.

Epistasis complicates the analysis of mutational effects. Epistatic interactions can be classified into different orders. Pairwise interactions account for the difference between the observed effect of a pair of mutations from the sum of their individual effects. Three-way interactions account for the difference between the observed effect of a set of three mutations from the sum of their individual effects and pairwise interactions. In general, as the order of epistatic interaction increases, the number of possible combinations of that order rises exponentially. Therefore, a key question is the extent to which high-order epistasis shapes protein genotype-phenotype maps. If epistasis is of low orders, then a genotype-phenotype map can be encoded using a relatively small number of parameters.

Recent application of deep mutational scanning has generated datasets in which a defined subset of sites within a protein has been comprehensively mutated and phenotyped. These datasets offer an opportunity to address the prevalence of epistasis and the complexity of experimental genotype-phenotype maps. Existing studies differ in their conclusion regarding epistasis because no study has comprehensively analyzed the available datasets using a common method. Furthermore, many studies have employed methods that are not designed to efficiently capture the global structures of genotype-phenotype maps, leading to overestimation of their

complexity. I develop a new method to analyze genotype-phenotype maps and apply them to available experimental datasets to address this knowledge gap.

1.5 A cautionary tale in ancestral sequence reconstruction

Ancestral sequence reconstruction is a methodological pillar of structuralism. By effectively allowing a time travel to the past, it allows us to trace the sequence changes that led ancestral proteins to evolve present-day molecular diversity and to uncover the genetic and biochemical mechanisms underlying the historical evolution of protein structures and functions. As with any powerful method of inference, ancestral reconstruction has the potential to mislead when not applied cautiously. I show that insufficient sampling of evolutionary information and failure to cross-compare with established phylogenetic relations led a group of researchers to wrongly infer the evolutionary loss of allostery as a de novo gain. I show that extensive sampling of available evolutionary information and rigorous analysis can lead to robust historical conclusions.

Chapter 2

Simplicity of experimental protein genotype-phenotype maps

2.1 Summary

Learning the mapping from genotype to phenotype is a central goal in biology. Existing studies disagree on the complexity of protein genotype-phenotype maps: Some studies report widespread, high-order epistasis but others report epistasis limited in extent and order (*12*, *34*–*38*). These studies examine different genotype-phenotype maps using different methods, often using formalisms that are inadequate for learning the global structure of a map. We developed a simple implementation of a formalism that optimally captures the global structure of a genotype-phenotype map and used it to analyze 15 empirical maps encompassing 3 to 16 sites in a diverse set of proteins. We found surprising simplicity in every dataset: Modeling only the context-independent effects and pairwise interactions of mutations is sufficient to predict phenotype with high accuracy ($R^2 > 0.92$ in cross-validation). The fraction of possible effects and interactions needed to be measured to achieve 90% prediction accuracy decreases with the size of genotype-phenotype map, reaching as low as 1 in 10,000. Consistent with this sparsity, a sparse learning method can be used to estimate the important effects and interactions from a small sample of genotypes and predict the phenotype of unobserved genotypes. Overall, our analyses reveal the extraordinary simplicity of available protein genotype-phenotype maps and suggest that sparse experimental characterization and statistical learning may be sufficient to elucidate the genotype-phenotype map of an entire protein.

2.2 Introduction

Understanding the genetic architecture of a protein—how the individual residues and their epistatic interactions determine its structure and function—requires characterizing how genetic variation maps to phenotypic variation. The genotype-phenotype map also shapes the evolutionary accessibility of phenotypic variations from an initial genotype and is therefore a central object of study in evolutionary biology.

For all but the shortest proteins, experimentally characterizing all possible genotypes is impossible because of their astronomical number. However, an exhaustive characterization may be not necessary to learn a genotype-phenotype map. For example, if all residues act independently of each other, the phenotype of any genotype can be predicted from the effect of each residue measured in a single genetic background. When the residues interact epistatically, more experiment is needed to learn their context-dependent effects. Whether it is feasible to experimentally learn a genotype-phenotype map depends chiefly on the order of epistatic interactions among residues. A complex map shaped by high-order epistasis is difficult to learn because identifying the important high-order interactions, even if they are few, requires searching a vast space of combinatorial possibilities. Therefore, a key unknown is the complexity of genotype-phenotype maps—the extent to which they are shaped by high-order epistasis.

A growing body of studies report widespread and often high-order epistasis, suggesting that genotype-phenotype maps are too complex to be learned (*9–14*, *34*, *35*, *38–46*). However, many of these studies could have overestimated the extent of epistasis because they used methods of analysis that are not designed to describe a genotype-phenotype map in the simplest possible way.

First, many studies do not account for nonspecific epistasis and therefore infer more specific epistasis than is necessary to explain the data. The nonadditivity of measured phenotype

may be due to a simple nonlinear transformation of an underlying additive phenotype (*7*, *36*, *47*). For example, mutations that act additively on the free energy of folding will act nonadditively on the proportion of folded proteins because of the Boltzmann distribution relating the two quantities. It is inefficient and misleading to explain this nonadditivity as a result of specific epistatic interactions among mutations. Unlike the Boltzmann distribution, some forms of nonspecific epistasis cannot be determined a priori. The particular setup of an experimental assay, such as the bounds of measurement or the exact way a biological quantity of interest is converted into a measurable phenotype, can give rise to nonspecific epistasis, which means that nonspecific epistasis must be learned from data. A recent analysis of three empirical genotype-phenotype maps shows that an effective model of nonspecific epistasis substantially reduces the extent of specific epistasis (*47*), necessitating a systematic re-examination of existing datasets.

Furthermore, the most commonly used method of analysis cannot efficiently capture the global structure of a genotype-phenotype map. In this method, called reference-based analysis, the genotype-phenotype map is described from the point of view of a single wild-type genotype. It proceeds by measuring the effects of point mutations and using them to predict the phenotypes of double mutants; deviations are explained by pairwise epistasis. The phenotypes of triple mutants are then predicted using the measured effects and pairwise interactions of mutations, explaining the deviations by three-way epistasis. This procedure continues to higher-order mutants until no more data are available. The result is an approximation of genotype-phenotype map that is exact in the neighborhood of wild-type genotype but may or may not extrapolate well to unobserved higher-order mutants.

Reference-based analysis interprets the local structure of a genotype-phenotype map observed in the neighborhood of wild-type genotype as a general structure of the map and

therefore can be misled by local idiosyncrasies (*48*). Suppose that two mutations interact epistatically only on the background of wild-type genotype. Reference-based analysis interprets this idiosyncratic interaction as ubiquitous, invoking it whenever the two mutations co-occur; as a result, higher-order epistasis must be invoked frequently to correct for it. This sensitivity to local idiosyncrasy can be more intuitively illustrated by the problem of approximating a curve by fitting a polynomial to a reference point: Even if the overall curve is well-approximated by a low-order polynomial, small deviations near the reference point can mislead the approach to a much higher order polynomial. What we need is a method that approximates the entire curve at once rather than prioritizing a particular region.

Two methods—Fourier analysis (*49*–*51*) and background-averaged epistasis (*48*)—were developed as alternatives to reference-based analysis, but their application has been limited. Instead of analyzing the effects of mutations with respect to a single genotype, these methods examine the average effects of residues across many genetic backgrounds, which makes them robust to local idiosyncrasies. While Fourier analysis takes a simple form when applied to a genotype-phenotype map with just two states per site (*51*), current implementations for multi-state maps rely on manipulating large Hadamard matrices (*50*) or constructing graph Fourier bases (*49*). The formal complexity and lack of an easy-to-use implementation for multi-state maps likely explain why the application of Fourier analysis has been limited to two-state maps, with only two studies reporting application to multi-state maps (*39*, *49*). A recent study shows that background-averaged epistasis describes an empirical genotype-phenotype map using much fewer parameters than reference-based analysis, revealing the simple structure of the map invisible under reference-based analysis (*37*). However, although extendable, background-

averaged epistasis has only been implemented for two-state maps, so its properties remain largely unexplored for multi-state maps.

Here, we present a simple and intuitive implementation of Fourier analysis applicable to any number of states, which we call reference-free analysis. The tractability of our implementation allows us to prove a key advantage of Fourier analysis: Among all possible ways of defining the effects and interactions of residues—including the formalisms described above and the many more that are possible—none provide a simpler description of a genotype-phenotype map than does Fourier analysis. Specifically, among all linear models that approximate the phenotype using epistatic interactions up to a given order, none are more accurate than Fourier analysis when gauged by the Pearson correlation between the predicted and observed phenotype. For example, among all possible additive models, including the first-order reference-based model under any choice of reference genotype and any scheme of averaging reference-based terms, none are more accurate than the first-order Fourier analysis. We combine reference-free formalism with a model of nonspecific epistasis to systematically characterize the complexity of empirical protein genotype-phenotype maps.

2.3 Results

2.3.1 Reference-free analysis of genotype-phenotype maps

Our goal is to learn the structure of a genotype-phenotype (GP) map by decomposing the phenotype into the contribution of individual sequence states and their interactions. We begin by outlining reference-free analysis and its key properties that make it ideal for this goal. Let $g = (g_1, \ldots, g_L)$ denote a genotype with sequence state g_i in site $i = 1$ to L. We denote the phenotype

of g as $y(g)$ and the set of all possible genotypes as G. The reference-free intercept, e_0, is defined as the average phenotype of all genotypes, written as

$$e_0 = \langle y|G\rangle.$$

The first-order reference-free term associated with state s in site i, denoted by $e_i(s)$, is defined as

$$e_i(s) = \langle y|G_i^s\rangle - e_0,$$

where G_i^s is a subset of G comprising all genotypes with state s in site i. This term quantifies how the average phenotype of all genotypes containing state s in site i differs from that of all genotypes; it can thus be interpreted as the average context-independent effect of the state.

The second-order reference-free term associated with state-pair (s_1, s_2) in site-pair (i_1, i_2), denoted by $e_{i_1,i_2}(s_1, s_2)$, is defined as

$$e_{i_1,i_2}(s_1, s_2) = \left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - e_0 - \left[e_{i_1}(s_1) + e_{i_2}(s_2)\right],$$

where $G_{i_1,i_2}^{s_1,s_2}$ is a subset of G comprising all genotypes with states s_1 and s_2 in sites i_1 and i_2, respectively. This term quantifies how the average phenotype of all genotypes containing state-pair (s_1, s_2) in site-pair (i_1, i_2) differs from that all of genotypes when accounted for the individual effects of the two states; it can be interpreted as the average epistatic effect of the state-pair.

This definition and interpretation can be extended to higher-order terms, resulting in an increasingly accurate representation of the GP map. The intercept is the crudest representation, a

single average for the entire map. Each first-order term carries information about the average phenotype of a sub-map comprising all genotypes with a particular state in a particular site. Each second-order term carries information about the average phenotype of a smaller sub-map, comprising all genotypes with a particular state-pair in a particular site-pair. Higher-order terms offer an increasingly finer representation, with the highest-order terms carrying information about individual genotypes.

The phenotype of a genotype is equal to the sum of all relevant reference-free terms:

$$y(\boldsymbol{g}) = e_0 + \sum_{i \in L} e_i(g_i) + \sum_{i_1 < i_2 \in L} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_k \in L} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}) + \cdots + e_{i_1,\ldots,i_L}(g_1, \ldots, g_L). \quad [\mathbf{1}]$$

Truncating Eq. **1** to a certain order results in an approximation of the GP map. A key question is what order of epistasis is required to accurately approximate experimental GP maps.

A critical advantage of reference-free analysis is that it offers a maximally accurate linear approximation for the GP map. Let y_k denote the phenotype predicted by a model of order k (truncation of Eq. **1** removing all higher-order terms). The accuracy of y_k can be quantified by summing the squared prediction error: $\sum_{g \in G}[y(g) - y_k(g)]^2$. We show that a reference-free model of a given order achieves the minimum total prediction error possible for any linear model of the same order (Method 2.5.6). In other words, the accuracy of a reference-free model is the maximum accuracy attainable using epistatic interactions of up to a given order; better prediction requires modeling higher-order epistasis.

Each reference-free term is a function of the phenotype of every genotype; directly calculating any term requires experimentally characterizing the entire GP map. It is possible,

however, to infer the terms from a random subset of genotypes: Eq. **1**, truncated to a desired order, can be fit to data by minimizing the prediction error across the sampled genotypes. The resulting estimates are unbiased (expected to equal the true parameter values) if genotypes are randomly sampled (Method 2.5.8).

Finally, we account for nonspecific epistasis by modeling the observed phenotype as a nonlinear function of a latent phenotype (φ) (Fig. 2.1B). We jointly infer the shape of the nonlinear function and the reference-free terms for the latent phenotype. By subjecting the latter to a lasso penalty, we identify the nonlinear function that minimizes the specific epistasis required in the latent space. Although nonspecific epistasis can in principle be complex, we find that a simple sigmoidal curve with just two parameters effectively captures the nonspecific epistasis in most experimental datasets: $y = L + \frac{(U-L)}{1+e^{-\phi}}$, where L and U represent the lower and upper bound of observed phenotype, respectively.

2.3.2 Simplicity of experimental protein genotype-phenotype maps

We assessed the complexity of 15 experimentally determined protein genotype-phenotype maps (Table 2.1). These datasets were chosen on the basis of two criteria. First, phenotype must be measured for a combinatorially complete set of genotypes, possibly missing for a random subset of genotypes. This criterion excludes datasets that are biased toward low-order mutants of a particular genotype, such as those generated by error-prone PCR or saturation point mutagenesis. Systematically assessing the contribution of each order of epistasis is not possible in such biased datasets. Second, phenotype measurement must be precise because measurement noise is indistinguishable from high-order epistasis. Datasets in which the squared Pearson correlation between measurement replicates is less than 0.9 were excluded. Each chosen

dataset is designed to be combinatorially complete over a defined subset of sites in a protein (3 to 16 sites, 42 to 100% coverage of possible genotypes). Some datasets sample all 20 amino acids, whereas others restrict the state space to two or more specific amino acids. The datasets range in size from 32 to 160,000 possible genotypes and include antibodies, enzymes, fluorescent protein, protein complex subunits, and transcription factors.

Table 2.1. Experimental GP maps analyzed in this study.

Number of genotypes	**Sampling**	**Protein**	**Phenotype**	**Reference**
2^5 (= 32)	100%	Methyl-parathion hydrolase	Catalytic activity	Anderson (*40*)
2^5 (= 32)	100%	Beta-lactamase	Antibiotics resistance (MIC)	Weinreich (*45*)
3×2^4 (=48)	100%	Dihydrofolate reductase	Antibiotics resistance (IC_{75})	Palmer (*52*)
2^{11} (= 2,048)	78.8%	Antibody CR6261	Binding affinity ($-\log K_D$) for antigen H1	Phillips (*53*)
2^{11} (= 2,048)	57.7%	Antibody CR6261	Binding affinity ($-\log K_D$) for antigen H9	Phillips (*53*)
20^3 (= 8,000)	98.5%	Antitoxin ParD3	Fitness (conferred by binding toxin ParE2)	Lite (*54*)
20^3 (= 8,000)	98.5%	Antitoxin ParD3	Fitness (conferred by binding toxin ParE3)	Lite (*54*)
2^{13} (= 8,192)	100%	Fluorescent protein	Fluorescence	Poelwijk (*37*)
$13 \times 12 \times 10 \times 6$ (= 9,360)	99.6%	Antitoxin ParD3	Fitness (conferred by binding toxin ParE2)	Aarke (*9*)
$13 \times 12 \times 10 \times 6$ (= 9,360)	98.2%	Antitoxin ParD3	Fitness (conferred by binding toxin ParE3)	Aarke (*9*)
2^{16} (= 65,536)	99.7%	Antibody CR9114	Binding affinity ($-\log K_D$) for antigen fluB	Phillips (*53*)
2^{16} (= 65,536)	90.8%	Antibody CR9114	Binding affinity ($-\log K_D$) for antigen H1	Phillips (*53*)
2^{16} (= 65,536)	96.3%	Antibody CR9114	Binding affinity ($-\log K_D$) for antigen H3	Phillips (*53*)
20^4 (= 160,000)	42.2%	Transcription factor ParB	Fitness (conferred by binding TFBS parS)	Jalal (*41*)
20^4 (= 160,000)	68.6%	Protein G B1 domain	Binding for IgG-Fc (enrichment score)	Wu (*46*)

To assess the complexity of each dataset, we sequentially fit reference-free models of increasing order, beginning with the first-order model including only the context-independent effects of amino acids that are subject to nonspecific epistasis. Each model was evaluated by cross-validation—by fitting the model after excluding a random subset of data and testing how well the excluded data can be predicted by the inferred effects (quantified by squared Pearson correlation, R^2_{CV}).

First-order reference-free models attain an R^2_{CV} greater than 0.90 for 10 of the 15 datasets (Fig. 2.1C). These 10 datasets include 16-site genotype-phenotype maps where up to 16^{th}-order epistatic interactions are theoretically possible, and a 4-site, 20-amino acid map where first-order terms make up only 0.05% of all reference-free terms.

Modeling pairwise epistasis raises R^2_{CV} to greater than 0.92 for every dataset. Modeling three-way epistasis, by contrast, offers only marginal or no improvement in fit (average increase in R^2_{CV} of 0.01). The residual of the second-order model (difference between the observed and predictive phenotype) closely follows a normal distribution, implying that they reflect measurement noise or that higher-order epistasis can be formally treated as noise. Overall, our analysis uncovers a surprising simplicity of experimental protein genotype-phenotype maps: Accurate prediction of phenotype is possible by learning just the context-independent effects and pairwise interactions of amino acids.

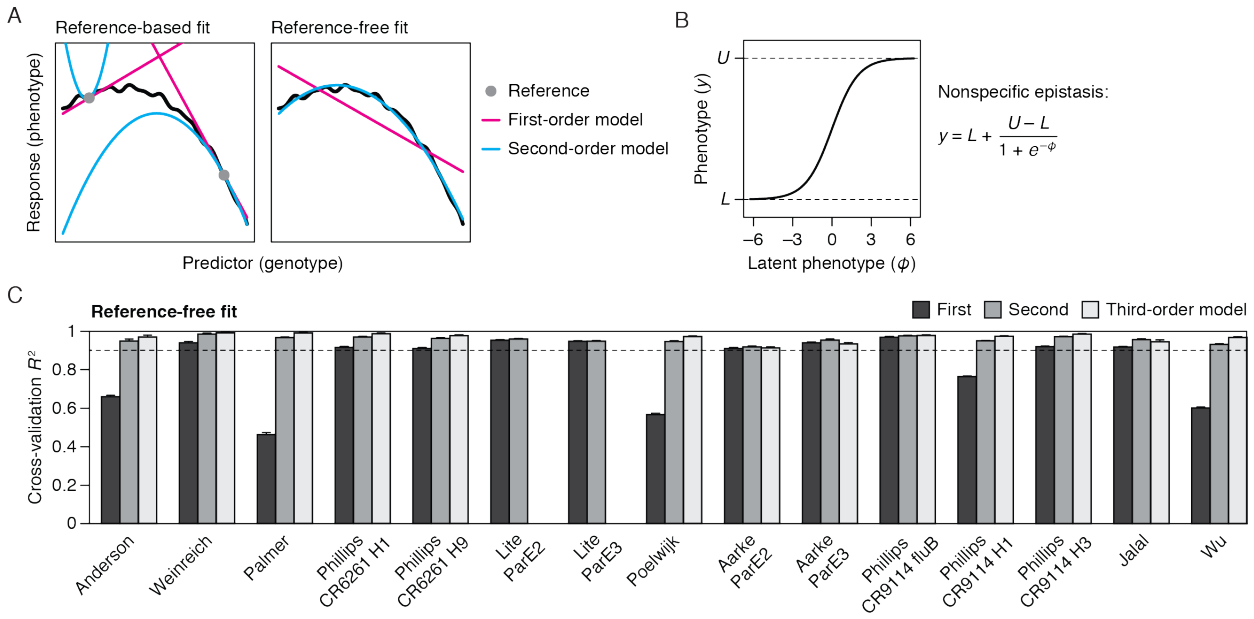

Figure 2.1. Reference-free analysis offers a simple description of experimental genotype-phenotype maps. (**A**) Schematic illustrating the different approach of reference-based and reference-free analysis for the common problem of obtaining a simple approximation for an unknown function. In reference-based analysis, an increasingly complex local fit is made around a chosen reference point. The result depends on the particular reference chosen and can be misled by local idiosyncrasies. In reference-free analysis, a fit that minimizes the global deviation is obtained. (**B**) Model of nonspecific epistasis. The observed phenotype is modeled as a sigmoidal function of a latent phenotype. The sigmoidal function has just two parameters representing the lower and upper bound of phenotype (L and U, respectively). (**C**) Analysis of 15 experimental genotype-phenotype maps. A reference-free fit of a given order was evaluated by cross-validation.

2.3.3 Widespread nonspecific epistasis

Modeling nonspecific epistasis is critical for the simple description of the genotype-phenotype maps. Without modeling nonspecific epistasis, first-order reference-free models display a median R^2_{CV} of only 0.37 across the 15 datasets, a drastic reduction from 0.92 (Fig. 2.2A). Second-order models also suffer a reduction in median R^2_{CV} from 0.96 to 0.85, leaving a fraction of phenotypic variance to be explained by higher-order epistasis.

We sought to understand how a simple sigmoidal curve with just two parameters—representing the upper and lower bound of measurement—so drastically improves the model fit.

The extent to which model fit depends on modeling nonspecific epistasis is proportional to the fraction of genotypes inside the inferred dynamic range of measurement (Fig. 2.2A). For example, in a dataset where only 0.12% of genotypes are inside the inferred dynamic range (Phillips CR9114 fluB), omission of the nonspecific epistasis model reduces the first-order R^2_{CV} from 0.98 to 0.01 (Fig. 2.2B). By contrast, in a dataset of the same size but containing 67% of genotypes inside the inferred dynamic range (Phillips CR9114 H1), first-order R^2_{CV} drops from 0.77 to just 0.61.

To explain this result, we focused on the Phillips CR9114 fluB dataset. This dataset consists of binding affinity measurement ($-\log K_D$) for variants of the antibody CR9114 against the influenza hemagglutinin subtype fluB. The vast majority of genotypes in this dataset have $-\log K_D$ close to the minimum observed value of 6, with only 0.12% having a value greater than 6.1 (Fig. 2.2C). While observed $-\log K_D$ ranges from 6.0 to 8.1 across the genotypes, the first-order prediction without nonspecific epistasis ranges from just 5.99 to 6.02 (Fig. 2.2B).

Limited dynamic range of measurement can cause systematic underestimation of mutational effects and spurious epistasis when not explicitly accounted for. For example, on the background of a genotype masked by the lower bound of measurement, all mutations with a negative phenotypic effect will appear neutral. Mutations with a positive effect will also appear neutral if the effect size is not sufficiently great. Given that 99.9% of genotypes in the fluB dataset are at the lower bound, mutations on average have no observable effect, resulting in systematic underestimation of reference-free effects and predicted phenotype. Only on the background of 0.12% of genotypes can a mutation show any observable effect—a strong background-dependence that must be explained by epistasis in the absence of a model for limited

dynamic range of measurement. Consistent with this explanation, most genotypes are predicted to have a latent phenotype that is well below the observable range (Fig. 2.2D).

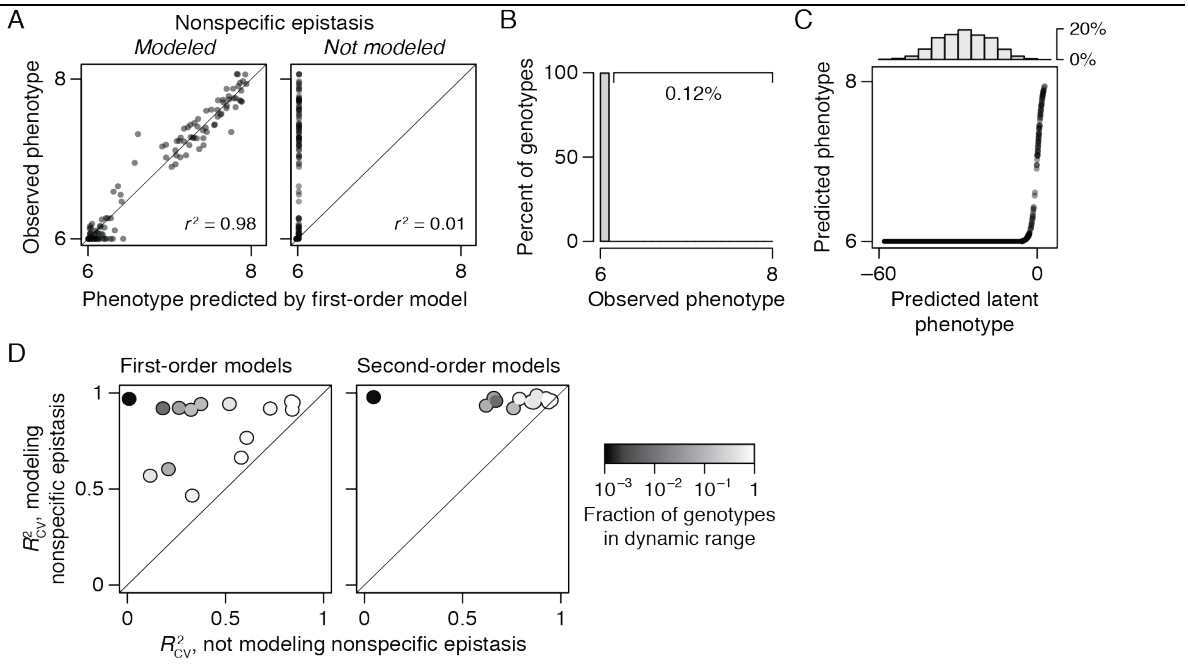

Figure 2.2. Modeling nonspecific epistasis critical for the simple description of the experimental genotype-phenotype maps. (**A**) Comparing the observed phenotype of the Phillips CR9114 fluB dataset to the prediction made by first-order reference-free model with (left) or without (right) modeling nonspecific epistasis. (**B**) Histogram of observed phenotype for the Phillips CR9114 fluB dataset. (**C**) Predicted latent phenotype and observed phenotype for the Phillips CR9114 fluB dataset based on a first-order reference-free model with nonspecific epistasis. The histogram on top shows the distribution of predicted latent phenotype across genotypes. (D) Comparing the cross-validation R^2 (R^2_{CV}) obtained using reference-free models with or without the nonspecific epistasis model. Each circle represents an experimental dataset, colored by the fraction of genotypes inside the inferred dynamic range.

2.3.4 Genetic and biophysical basis of phenotype

A key goal in analyzing genotype-phenotype maps is to explain the genetic and biophysical basis of phenotypic variation. This requires quantifying the amount of phenotypic variation caused by genetic variation in individual sites and their interactions. We show that these contributions can be directly calculated from reference-free effects (Method 2.5.7). For

example, the amount of phenotypic variance due to genetic variation in a site is equal to the mean of the square of all first-order reference-free terms for that site. Similarly, the variance due to genetic interaction between two sites is equal to the mean of the square of all second-order reference-free terms for that pair of sites.

We performed analysis of variance to identify the key sites and states that shape phenotypic variation in the experimental genotype-phenotype maps. In the Phillips CR6114 H1 dataset, which consists of binding affinity measurements for 2^{11} variants of the antibody CR6114 against the influenza hemagglutinin subtype H1, 92% of phenotypic variance can be explained by a first-order reference-free mode. Within the first-order model, effects at just 3 of the 11 mutated sites explain 94% of phenotypic variance (86% of total variance; Fig. 2.3A). These are the three mutated sites that directly contact H1 in the crystal structure of CR6114-H1 complex. All other sites, which contribute negligibly to phenotypic variation in this genotype-phenotype map, are distal to H1. Overall, context-independent effects of genetic variation at just three key sites explain this genotype-phenotype map.

In the Poelwijk dataset, which measured fluorescence of 2^{13} variants of a fluorescent protein, a second-order model is required to explain most of phenotypic variance. Within the second-order model, however, just 4 of the 13 mutated sites explain 86% of phenotypic variance (Fig. 2.3B). One site is a part of the chromophore and the others directly contact the chromophore; the rest of mutated sites are distal to the chromophore. This shows that the Poelwijk dataset can be explained by the context-independent effects and pairwise interactions of genetic variation at key sites surrounding the chromophore.

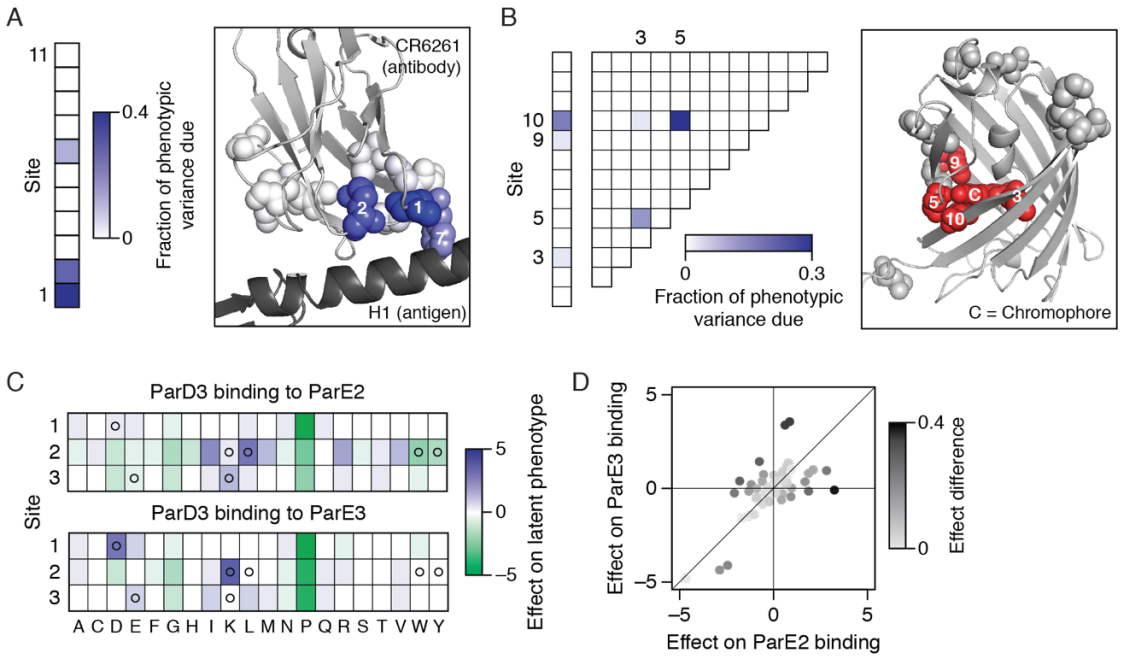

Figure 2.3. Genetic and biophysical basis of experimental genotype-phenotype maps revealed by reference-free analysis. (**A**) The fraction of phenotypic variance explained by the first-order terms in each site for the Phillips CR6261 H1 dataset is shown on the left heatmap and in the proteins structure. (**B**) The fraction of phenotypic variance explained by the first-order terms in each site and the second-order terms in each site-pair for the Poelwijk dataset is shown on the left heatmap. Only the sites with nonnegligible contribution are marked and colored red in the structure. (**C**) The inferred effects of amino acids at three sites in the ParD3 interface with ParE3 or ParE2. Amino acids with effect on ParE3 and ParE2 binding different by more than 2 latent phenotype unit are marked by black circles; these amino acids contribute to the specificity of the interface.

2.3.5 Sparsity of genotype-phenotype maps

To quantify the sparsity of each genotype-phenotype map, we first ordered the reference-free effects by their contribution to the latent phenotype. Beginning with a model that includes just one term with the largest contribution, we constructed a series of models with an increasing number of terms. We then evaluated each model by cross-validation and determined the minimal model size (fraction of possible terms included) required for R^2_{CV} of 0.9 (denoted by F_{90}; Fig. 2.4A).

We found a scaling relation between sparsity and the size of genotype-phenotype map (Fig. 2.4B). Larger maps are sparser, with F_{90} almost inversely proportional to the number of possible genotypes (N): $F_{90} \sim N^{-0.85}$. This relation holds across almost four orders of magnitude in N. This relation can partly be explained by the fact that context-independent effects and pairwise interactions are sufficient to attain R^2_{CV} of 0.9 in every dataset: The proportion of these terms out of all possible terms decreases in proportion to the size of genotype-phenotype map. However, this relation implies that a relatively constant proportion of first- and second-order terms shape the phenotype across the experimental datasets, which is surprising given their heterogeneity. An explanation for this scaling relation, to the extent that it generalizes, is yet to be found.

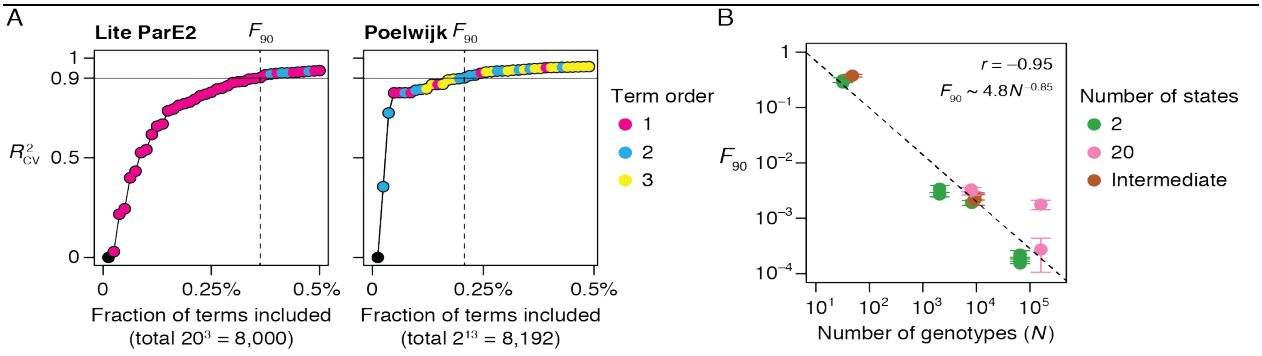

Figure 2.4. Sparsity increases with GP map size. (**A**) Quantifying sparsity with the F_{90} measure. F_{90} is defined as the minimum fraction of all possible reference-free effects that must be modeled to achieve a cross-validation R^2 greater than 0.9. To estimate F90, we ordered the reference-free effects by their contribution to the latent phenotype and tested a series of models beginning with just one effect with the largest contribution and sequentially including more effects. The fraction of terms included that first achieves R^2 of 0.9 is taken as an estimate of F_{90} (shown by the vertical dashed line). Each dot represents a model, colored by the order of the last term included. (**B**) Sparsity as a function of genotype-phenotype map size. The inferred value of F_{90} is shown as a function of the total number of genotypes in each genotype-phenotype map (N). Each dot represents a genotype-phenotype map, colored by the number of states sampled for each mutated site. Dashed line represents linear regression.

2.3.6 Learning the genotype-phenotype map by random sampling

Even if only a small fraction of reference-free terms is required to describe a genotype-phenotype map, if identifying and estimating them requires experimentally characterizing a large number of genotypes, reference-free analysis will not be useful for learning a new genotype-phenotype map. Sparse learning theory suggests that if a system with a large theoretically possible degrees of freedom is in fact determined by a relatively small number of parameters, sparse learning methods can be used to estimate them from a random sample of a size on the order of the number of important parameters.

We quantified the fraction of genotypes that must be sampled to learn each experimental genotype-phenotype map. We randomly sampled a certain fraction of genotypes, used them to infer a reference-free model, and predicted the phenotype of the rest of genotypes. By repeating this analysis for a range of sample sizes, we determined the minimal fraction of genotypes required to achieve a prediction R^2 of 0.9 (denoted by S_{90}).

S_{90} varies widely among GP maps of similar size (Fig. 2.5B). For example, the Lite ParE2 dataset (8,000 genotypes, F_{90} = 0.36%) can be accurately learned by sampling just 2.8% of all genotypes (Fig. 2.5A). However, the Aarke ParE3 dataset (9,360 genotypes, F_{90} = 0.27%) requires sampling 21% of genotypes for a comparable accuracy. This is likely caused by the different proportion of genotypes within dynamic range: Whereas 80% of genotypes are within the inferred dynamic range in the Lite ParE2 dataset, this fraction is only 14% in the Aarke ParE3 dataset. Little information about the effects and interactions of states can be obtained from genotypes masked at measurement bounds. From individual such genotypes, we can only learn that the phenotypes of certain combinations of states are too high or too low, but not their exact value. Likewise, comparison among genotypes all masked at one side of the range gives little information about the effects of mutations between them. Therefore, when the dynamic range is

limited, a larger overall fraction of genotypes must be sampled to learn the same amount of information.

Our analysis points to an efficient experimental sampling strategy as the major requirement for learning protein GP maps. Random sampling is too inefficient for learning when most genotypes are nonfunctional. Reference-based sampling, in which sampling is limited to low-order mutants of a particular genotype, cannot be used to train a reference-free model that generalizes across the GP map. Therefore, a method to enrich for functional genotypes sampled from across the GP map is required.

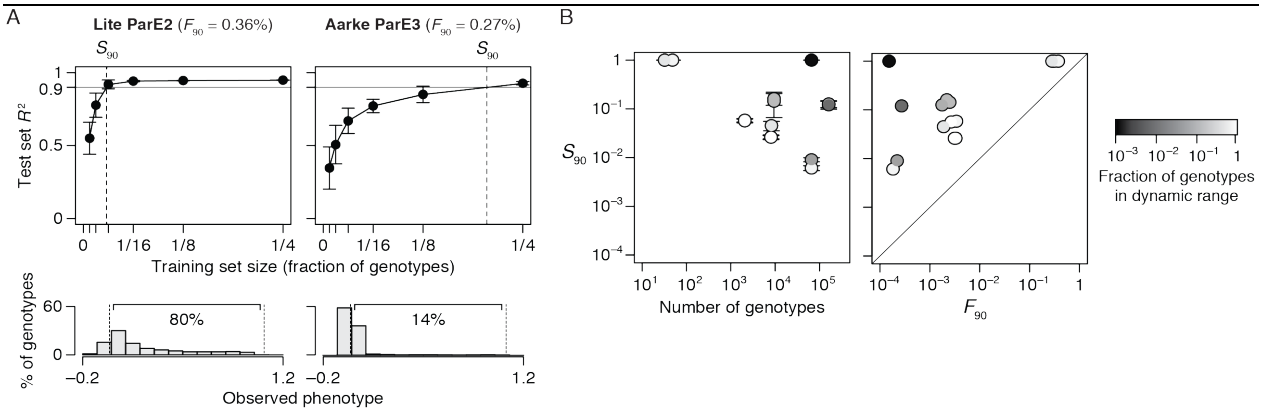

Figure 2.5. Whether a GP can be learned from a small random sample of genotypes depends critically on the dynamic range of measurement. (A) (Top) Calculating the fraction of genotypes that must be sampled to learn a reference-free model that achieves a prediction of R_2 greater than 0.9 (S_{90}, marked by the vertical dashed lines). For each training set size, 10 or more random samples of the size were obtained, used to trained a reference-free model, and R^2 of prediction calculated for the rest of genotypes. Each dot and error bars represent the mean and standard deviation of prediction R^2. (Bottom) Histogram of phenotype for each dataset, showing the percent of genotypes within the inferred dynamic range. (**B**) S_{90} as a function of GP map size (left) or F_{90} (right). Each dot represents a dataset, colored by the fraction of genotypes within the inferred dynamic range.

2.4 Discussion

Characterizing the genotype-phenotype map of proteins is a fundamental goal in molecular and evolutionary biology that is hindered by the astronomical number of possible amino-acid sequences. Decomposing the phenotype of each sequence into the individual effects and interactions of amino acids has the potential to reveal structures in a genotype-phenotype map that allows us to describe it using a much smaller number of parameters than there are genotypes. Currently, there is no analytic framework that meets the following requirements: 1) A global approximation to the genotype-phenotype map, instead of a local approximation based on an arbitrarily chosen reference genotype, must be performed; 2) Nonspecific epistasis must be explicitly modeled, which can make a simple genotype-phenotype map appear complex; 3) The formalism should be applicable to any genotype-phenotype map with any number of states (potentially variable across sites) and allow quantifying the phenotypic contribution of any amino acid or set of amino acids; 4) The formalism should be applicable to a random sample of genotypes.

Here, we present a formalism called reference-free analysis that meets these criteria. We show that this formalism is optimal—there is no other linear decomposition of a genotype-phenotype map that achieves better prediction accuracy. It enables the analysis of variance (ANOVA) framework, quantifying the fraction of phenotypic variance due to each amino acid or set of amino acids. We applied this formalism to analyze experimental mutation datasets and found surprising simplicity: For every dataset, accurate prediction of phenotype is possible by learning just the context-independent effects and pairwise interactions of amino acids. This implies that currently available deep mutational scanning methods can characterize enough genotypes to reveal the genetic architecture of proteins.

Our model of nonspecific epistasis is formally equivalent to a Boltzmann distribution describing an equilibrium between two thermodynamic states. Two-state equilibrium is widely used as a model for protein structure and function. For example, protein folding is often describable as an equilibrium between the folded and unfolded state. The proportion of folded state is given by $1/(1 + e^{\Delta G})$, where ΔG is the Gibbs free energy of folding in appropriate units. The formal similarity between the two-state equilibrium and our model of nonspecific epistasis suggests that the protein phenotypes we analyzed can be described as an equilibrium between the functional and nonfunctional state. The latent phenotype corresponds to free energy in reverse sign, with a change of 1 in latent phenotype corresponding to 2.7-fold change in the relative occupancy of the functional and nonfunctional state. Modeled this way, the genetic basis of phenotype can be described in terms of the context-independent energetic contributions of individual sites and sparse epistatic interactions that are mainly between pairs of sites.

2.5 Methods

Here we present a detailed exposition of reference-free linear decomposition, including proofs for its key properties. We begin by formally defining reference-free linear decomposition and discussing three ways of interpreting its terms. We develop the notion of generalized linear decomposition—a unified formalism for representing any linear decomposition of a GP map—and show that reference-free linear decomposition is the best linear decomposition for capturing the global structure of a GP map. We then show that the total genetic variance of a GP map can be decomposed into the contribution of each reference-free term, which enables the analysis of variance (ANOVA) framework for quantifying and comparing the importance of individual terms and measuring the complexity and sparsity of GP maps. We present a method to estimate

the terms of reference-free linear decomposition from randomly sampled genotypes and end by comparing reference-free linear decomposition to other linear methods for decomposing GP maps.

2.5.2 Notations

We consider a GP map consisting of every genotype with one of q states in each of n sites—total q^n genotypes. Although for mathematical simplicity we only consider GP maps with the same number of states in every site, reference-free linear decomposition can be applied to any discrete-state GP map. The n-tuple $\boldsymbol{g} = (g_1, \ldots, g_n)$ denotes a genotype with state g_i in site $i = 1, \ldots, n$. The phenotype of $\boldsymbol{g}$ is written as $y(\boldsymbol{g})$. The set of all genotypes is denoted by G and the set of all genotypes with states $s_1, \ldots, s_k$ in sites $i_1, \ldots, i_k$ by $G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k}$. Angled brackets indicate the mean of a quantity over a set; for example, $\langle y|G\rangle$ is the average phenotype of all genotypes. The set $N = (1, \ldots, n)$ is used to denote iteration over sites and site-combinations; for example, $\Sigma_{i\in N}$ indicates summation over all sites and $\Sigma_{i_1<i_2\in N}$ summation over all site-pairs. Likewise, $Q = (1, \ldots, q)$ is used to denote iteration over states and state-combinations.

2.5.3 Definition

We first present reference-free linear decomposition as a stepwise approximation of a GP map. Two alternative formulations are then presented.

The intercept, e_0, is defined as the average phenotype of all genotypes:

$$e_0 = \langle y|G\rangle.$$

It can be considered the best single-parameter approximation of a GP map. The first-order term for state s in site i, denoted by $e_i(s)$, is defined as

$$e_i(s) = \langle y|G_i^s\rangle - e_0.$$

It can be considered the error associated with approximating $\langle y|G_i^s\rangle$ by the lower-order term e_0. (It is convenient to consider e_i as a function of a single variable taking q values.)

The above expression extends naturally to higher-order terms. The second-order term for state-pair (s_1, s_2) in site-pair (i_1, i_2), denoted by $e_{i_1,i_2}(s_1, s_2)$, is the error associated with approximating $\left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle$ using the lower-order terms:

$$e_{i_1,i_2}(s_1, s_2) = \left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - \left[e_0 + e_{i_1}(s_1) + e_{i_2}(s_2)\right].$$

In general, the k-th-order term for state-combination (s_1, …, s_k) in site-combination (i_1, …, i_k), denoted by $e_{i_1,\ldots,i_k}(s_1, \ldots, s_k)$, is defined as

$$e_{i_1,\ldots,i_k}(s_1, \ldots, s_k) = \left\langle y\middle|G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k}\right\rangle - \left\langle y\middle|G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k}\right\rangle_{(k-1)}, \quad (1)$$

where the subscript $(k-1)$ indicates approximation by terms of order up to $(k-1)$:

$$\left\langle y \middle| G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k} \right\rangle_{(k-1)}$$

$$= e_0 + \sum_{\alpha \in K} e_{i_\alpha}(s_\alpha) + \sum_{\alpha_1 < \alpha_2 \in K} e_{i_{\alpha_1}, i_{\alpha_2}}(s_{\alpha_1}, s_{\alpha_2}) + \cdots$$

$$+ \sum_{\alpha_1 < \cdots < \alpha_{k-1} \in K} e_{i_{\alpha_1},\ldots,i_{\alpha_{k-1}}}(s_{\alpha_1}, \ldots, s_{\alpha_{k-1}}),$$

where K denotes the set ($1, \ldots, k$).

This stepwise process builds an increasingly accurate approximation of a GP map. The intercept is the crudest approximation, a single average approximating the entire map. The first-order terms can describe the average phenotype of nq sub-maps, each comprising all genotypes with a particular state in a particular site. The second-order terms can describe the average phenotype of $\binom{n}{2}q^2$ sub-maps, each comprising all genotypes with a particular state-pair in a particular site-pair. Higher-order terms offer finer descriptions, with the n-th-order terms describing the phenotype of q^n sub-maps—of every genotype.

We have so far presented the terms of reference-free linear decomposition as errors associated with lower-order approximations. They can also be interpreted as phenotypic effects. For example, the first-order term $e_i(s)$ quantifies how the average phenotype of all genotypes containing state s in site i differs from that of all genotypes; it is the average phenotypic effect of state s in site i. Similarly, the second-order term $e_{i_1,i_2}(s_1, s_2)$ quantifies how the average phenotype of all genotypes containing states s_1 and s_2 in sites i_1 and i_2 differs from that of all genotypes, after being accounted for the individual effects of the two states; it is the average epistatic effect of the state-pair (s_1, s_2) in site-pair (i_1, i_2).

Yet another formulation interprets terms of order k as measuring the context-dependence of terms of order $k - 1$. Let us re-think the definition of the first-order term

$$e_i(s) = \langle y|G_i^s\rangle - e_0.$$

$\langle y|G_i^s\rangle$ can be considered the intercept of a GP map consisting only of genotypes with state s in site i. We denote this relation by $\langle y|G_i^s\rangle = e_0|_i^s$. Thus,

$$e_i(s) = e_0|_i^s - e_0.$$

In this expression, $e_i(s)$ quantifies how different the intercept is when calculated for the complete map G versus the sub-map G_i^s—the context-dependence of the intercept. A parallel exists for second-order terms:

$$\begin{aligned} e_{i_1,i_2}(s_1,s_2) &= \left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - \left[e_0 + e_{i_1}(s_1) + e_{i_2}(s_2)\right] \\ &= \left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - \left\langle y\middle|G_{i_1}^{s_1}\right\rangle - \left\langle y\middle|G_{i_2}^{s_2}\right\rangle + \langle y|G\rangle \\ &= \left[\left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - \left\langle y\middle|G_{i_1}^{s_1}\right\rangle\right] - \left[\left\langle y\middle|G_{i_2}^{s_2}\right\rangle - \langle y|G\rangle\right] \\ &= e_{i_2}(s_2)|_{i_1}^{s_1} - e_{i_2}(s_2), \end{aligned}$$

where $e_{i_2}(s_2)|_{i_1}^{s_1}$ is the first-order term $e_{i_2}(s_2)$ of the sub-map $G_{i_1}^{s_1}$. In general,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{s_1} - e_{i_2,\ldots,i_k}(s_2,\ldots,s_k).$$

Here site i_1 is chosen as the context, but any other site can be chosen.

Finally, we express the phenotype of individual genotypes using the terms of reference-free linear decomposition. Substituting n for k in Eq. (1) and noting that $\left\langle y \middle| G_{i_1,\ldots,i_n}^{g_1,\ldots,g_n} \right\rangle$ is simply the phenotype of $\boldsymbol{g} = (g_1, \ldots, g_n)$, the phenotype of $\boldsymbol{g}$ is the sum of every reference-free term corresponding to every possible combination of states in $\boldsymbol{g}$:

$$y(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}) + \cdots + e_{i_1,\ldots,i_n}(g_1, \ldots, g_n). \quad (2)$$

2.5.4 The zero-mean property and the uniqueness of reference-free linear decomposition

Consider the number of terms in Eq. (2). There are nq first-order terms, q for each site. There are $\binom{n}{2}q^2$ second-order terms, q^2 for each site-pair. With $\binom{n}{k}q^k$ terms of order k, the total number of terms equals $(q + 1)^n$. There are more terms than genotypes! Does this mean that reference-free linear decomposition is underdetermined, that the same GP map can be decomposed in many different ways? The answer is no, there is a unique reference-free linear decomposition for each GP map. This is because the terms of reference-free linear decomposition satisfy certain constraints that keep their degrees of freedom at q^n. We refer to this property as the *zero-mean property*. In addition to ensuring the uniqueness of reference-free linear decomposition, the zero-mean property forms the basis of all the desirable attributes of reference-free linear decomposition we describe below. In fact, the zero-mean property *defines*

reference-free linear decomposition in the sense described in the section *Generalized linear decomposition*. We simply state the property here, proving it in section 2.5.12.1.

The zero-mean property of first-order terms is that the mean of all q terms for a site is 0:

$$\langle e_i|Q\rangle = \frac{1}{q}\sum_{s\in Q} e_i(s) = 0.$$

Let • denote averaging of all terms for a site. For example, $e_i(\cdot)$ is the average of the q first-order terms for site i, and $e_{i_1,i_2}(\cdot, s_2)$ is the average of the q second-order terms for site-pair (i_1, i_2) containing state s_2 in site i_2. The above equality can be restated as $e_i(\cdot) = 0$. In general, for any site-combination (i_1, …, i_k), the mean of any q terms that vary across a single site is zero:

$$e_{i_1,\ldots,i_k}(\cdot, s_2, \ldots, s_k) = e_{i_1,\ldots,i_k}(s_1,\cdot, s_3, \ldots, s_k) \ldots = e_{i_1,\ldots,i_k}(s_1, \ldots, s_{k-1},\cdot) = 0,$$

for any state-combination (s_1, …, s_k). If the q^2 second-order terms for a site-pair are arranged in a $q \times q$ matrix, rows corresponding to the states in the first site and columns to the states in the second site, the zero-mean property means that every row and column of the matrix sums to zero. Similarly, if the q^3 third-order terms for a site-triple are arranged in a $q \times q \times q$ array, every one-dimensional section (and thus every two-dimensional section and the entire array) sums to zero.

How does the zero-mean property ensure the uniqueness of reference-free linear decomposition? The q first-order terms for a site must always sum to zero, reducing their degrees of freedom to $(q - 1)$. Every row and column of the $q \times q$ matrix of q^2 second-order terms for a

site-pair must always sum to zero, reducing their degrees of freedom to $(q-1)^2$. In general, the degrees of freedom of all k-th-order terms are reduced to $\binom{n}{k}(q-1)^k$, which sum across k to q^n.

2.5.5 Generalized linear decomposition

Consider Eq. (2) on its own without the reference-free definition of the terms. It is a general formula for linearly decomposing a GP map. Since there are more terms than genotypes, a given GP map can be linearly decomposed in infinitely many ways. Constraining the terms to exhibit the zero-mean property uniquely specifies a linear decomposition: the reference-free linear decomposition. This is the sense in which the zero-mean property *defines* reference-free linear decomposition. Alternative constraints yield alternative decompositions. For example, given any genotype $\boldsymbol{r} = (r_1, \ldots, r_n)$, setting every term involving state r_i in site i to zero yields the familiar reference-based linear decomposition with $\boldsymbol{r}$ as the reference genotype.

We call Eq. (2) the generalized linear decomposition of a GP map. In section 2.5.12.2, we show that any linear decomposition of a GP map is a special case of Eq. (2) obtained by subjecting the terms to a set of constraints that reduce their degrees of freedom to q^n. The notion of generalized linear decomposition allows us to ask the following question.

2.5.6 Optimality

Among the infinitely many ways of linearly decomposing a GP map, which is the most optimal? In other words, among all sets of constraints that can be imposed on the terms of generalized linear decomposition, which set yields the most optimal decomposition?

Answering this question requires defining optimality. Given a particular linear decomposition, let $y_k(\boldsymbol{g})$ denote the phenotype of $\boldsymbol{g}$ approximated by terms of order up to k—the truncation of Eq. (2) removing all higher-order terms. The accuracy of this approximation can be quantified by summing the squared error:

$$\epsilon[y_k] = \sum_{\boldsymbol{g} \in G} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2.$$

The optimal linear decomposition is that which minimizes ε. When k equals n, ε is zero and every decomposition is exact. When k is smaller than n, different decompositions may make different approximations. We show in section 2.5.12.3 that reference-free linear decomposition minimizes ε for any k: Reference-free linear decomposition provides the most optimal linear approximation to a GP map at every order of approximation.

We note that reference-free linear decomposition is not uniquely optimal. It is possible to modify the terms of a given reference-free linear decomposition (violating the zero-mean property) without altering the predicted phenotypes. However, optimal linear decompositions thus obtained are uninterpretable. How the terms of a linear decomposition can be interpreted depends on what constraints they satisfy. For example, the first-order term $e_i(s)$ takes a different meaning depending on whether it is used in reference-free or reference-based linear decomposition. A linear decomposition may be optimal but uninterpretable if the constraints defining its terms are unknown. Reference-free linear decomposition is unique in that it is both optimal and interpretable.

2.5.7 Analysis of variance

The purpose of decomposing a GP map is to identify the major terms that define that map. The terms can be used in a simple approximation to the map and can be related to its physical basis to gain mechanistic insights. Both tasks require determining the relative importance of terms. Intuitively, larger magnitude implies greater importance. However, magnitude cannot be the sole criterion. A term of order k is involved in the phenotype of 1 in q^k genotypes; given the same magnitude, a lower-order term is more important than a higher-order term because it influences more genotypes. We need a way to compare terms across sites and orders.

We show that the analysis of variance (ANOVA) framework can be applied to reference-free linear decomposition. This makes possible such statements as "this term explains 5% of total genetic variance" or "the first-order terms together explain 80% of total genetic variance." The applicability of ANOVA is a unique feature of reference-free linear decomposition enabled by the zero-mean property.

We first define the total genetic variance,

$$V = Var(y|G) = \frac{1}{q^n}\sum_{\boldsymbol{g}\in G}[y(\boldsymbol{g}) - \langle y|G\rangle]^2,$$

which quantifies the amount of phenotypic variation caused by genetic variation. We then quantify the contribution of each reference-free term to the total genetic variance. First, we define a quantity called effect-variance. The effect-variance of a site-combination is the variance of all terms in that site-combination:

$$Var\left(e_{i_1,\ldots,i_k}|Q^k\right) = \frac{1}{q^k}\sum_{s_1,\ldots,s_k \in Q} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)^2.$$

(This equality holds because the mean of all terms in a site-combination is zero.) Intuitively, a site-combination with larger effect-variance should contribute more to the total genetic variance. This is confirmed by the following variance partition formula, which states that the total genetic variance is the sum of the effect-variance of every site-combination (section 2.5.12.4):

$$V = \sum_{i \in N} Var(e_i|Q) + \sum_{i_1 < i_2 \in N} Var\left(e_{i_1,i_2}\middle|Q^2\right) + \cdots + \sum_{i_1 < \cdots < i_k \in N} Var\left(e_{i_1,\ldots,i_k}\middle|Q^k\right) + \cdots + Var\left(e_{1,\ldots,n}\middle|Q^n\right). \quad (3)$$

The effect-variance of a site-combination is therefore its absolute contribution to the total genetic variance; its relative contribution is the effect-variance divided by the total genetic variance.

Substituting the definition of effect-variance into Eq. (3) shows that the total genetic variance is the sum of every non-intercept term, squared and normalized for its order:

$$V = \sum \frac{e^2}{q^{O(e)}},$$

where $O(e)$ is the order of e and the summation involves all terms except for the intercept. This confirms our intuition that a lower-order term is more important (explains a larger fraction of total genetic variance) than a higher-order term of the same magnitude. Note that $1/q^{O(e)}$ is the fraction of genotypes whose phenotype involves the term e.

Let V_k denote the sum of the effect-variance of every site-combination of order k. We call the sequence $\left(\frac{V_1}{V}, \frac{V_2}{V}, \ldots, \frac{V_n}{V}\right)$ the variance spectrum of a GP map. The variance spectrum quantifies the complexity of a GP map by showing what fraction of total genetic variance is due to each epistatic order. A GP map is simple if most of its genetic variance is due to low-order terms; it is complex if most of its genetic variance is due to high-order terms.

2.5.8 Estimation by random sampling

In reference-free linear decomposition, every term is a function of every genotype; exact calculation of any term requires measuring the phenotype of every genotype. However, it is possible to estimate the terms from a randomly sampled subset of genotypes. Recall that among all linear decompositions of a given order, reference-free linear decomposition minimizes the total prediction error. Given large sample size, a linear decomposition that minimizes the prediction error for sampled genotypes should be a good estimate for reference-free linear decomposition. We formalize this idea as follows. Let y_k denote the generalized linear decomposition of order k:

$$y_k(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}).$$

We find y_k that minimizes the sample prediction error:

$$\hat{y}_k = argmin \sum_{\boldsymbol{g} \in M} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2, \tag{4}$$

where M is the set of sampled genotypes. Two questions arise: Is $\hat{y}_k$ unique? Is $\hat{y}_k$ an unbiased estimate of reference-free linear decomposition? The answer to the first question is no: Due to the degeneracy of generalized linear decomposition, there are many possible linear decompositions that minimize the prediction error for a given set of genotypes. However, we show in section 2.5.12.5 that any $\hat{y}_k$ satisfying Eq. (4) can be normalized to simultaneously satisfy Eq. (4) and the zero-mean property—and there is exactly one such $\hat{y}_k$. Furthermore, $\hat{y}_k$ thus obtained is an unbiased estimate of reference-free linear decomposition: The expected value of estimated terms equals the true value.

If our goal is to use the estimated reference-free terms to accurately reconstruct the entire GP map, how many genotypes must we sample? The answer depends foremost on the simplicity of the GP map: Accurate reconstruction is possible only if the majority of genetic variance is due to terms of order low enough to be estimated from a practical sample size. A term of order k is involved in the phenotype of one in q^k genotypes. If a term is not involved in the phenotype of any sampled genotype, there is no way to estimate its value from that sample. Therefore, estimating terms of order k by random sampling requires a sample size sufficiently greater than q^k. If 50% of genetic variance is due to terms of order k or higher, reference-free terms estimated from q^k or fewer samples cannot achieve prediction accuracy greater than 50%. By contrast, if 90% of genetic variance is due to terms of order k or lower, a sample size on the order of q^k may yield prediction accuracy near 90%.

Another critical factor is the sparsity of the GP map. Consider a simple protein GP map in which most of the genetic variance is due to first- and second-order terms. For a 100-aa protein, there are still two million first- and second-order terms to estimate. If most of them have a non-negligible magnitude, accurate reconstruction of the GP map would require sampling

millions of genotypes. By contrast, if most of them are negligible and only a small fraction of them accounts for the majority of genetic variance, sparse learning methods can be used to identify and estimate the important terms from a much smaller number of genotypes.

The key requirement for sparse learning is a sparse representation. One drawback of reference-free linear decomposition is that it is not maximally sparse. Consider the GP map of a single amino acid, in which the phenotype of alanine is 1 and that of every other amino acid is –1. This is a sparse GP map in the sense that 19 out of 20 states are phenotypically equivalent. However, no term equals zero in the reference-free linear decomposition of this map: The term for alanine is 19/10 and that for every other amino acid is –1/10. In general, unless all terms for a site-combination are zero, the zero-mean property makes most of the terms non-zero; sparsity is sacrificed for interpretability.

Among the infinitely many ways of linearly decomposing a GP map, how can we find one that is both sparse and optimal? We can learn that from data. Recall the above two-step estimation procedure: We find a linear decomposition that minimizes the sample prediction error without subjecting the terms to any constraint and then normalize the terms to enforce the zero-mean property. We modify the first step by adding the lasso penalty:

$$\hat{y}_k = \text{argmin}\left\{\sum_{\boldsymbol{g}\in M}[y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2 + \lambda\sum|e|\right\},$$

where the second summation involves all non-intercept terms of y_k. The lasso penalty selects the sparsest linear decomposition among those with equal model fit. The value of λ can be chosen by cross-validation to minimize the test-sample prediction error. Because this step searches among

all possible linear decompositions, the result is a balance of optimality and sparsity that leads to maximal prediction accuracy. The terms in the resulting $\hat{y}_k$ cannot be interpreted, however, because we do not know what constraints they satisfy. The second step, by enforcing the zero-mean property, provides the interpretability.

Lastly, the dynamic range of measurement is also a critical factor determining the sampling depth required for reconstructing a GP map. If the phenotype of a sampled genotype is outside the dynamic range of measurement, we can only learn that the sum of a certain set of terms is greater or less than certain bounds. Therefore, limited dynamic range of measurement increases the sampling depth required for reconstructing a GP map.

2.5.9 Robustness to measurement noise: Unstructured GP maps

A GP map is unstructured if the phenotype of every genotype can be considered an independent sample from a random variable. An experimentally measured GP map is a superposition of a map of interest and an unstructured map generated by measurement error. Understanding how measurement error influences the estimation of reference-free linear decomposition from data requires understanding how reference-free linear decomposition treats unstructured maps.

Consider an unstructured GP map generated by sampling each phenotype independently from a random variable ω. The result of linear decomposition will vary among such maps depending on the exact instantiation of the sampling process. We are interested in the expected result across all instantiations. Let Ω denote the set of all unstructured GP maps generated from ω. Recall that V_k is the fraction of total genetic variance due to all site-combinations of order k.

We show that the expected value of V_k is proportional to the degrees of freedom of k-th-order terms (sectopm 2.5.12.6):

$$\langle V_k|\Omega\rangle \propto \binom{n}{k}(q-1)^k.$$

In other words, each degree of freedom is expected to capture the same fraction of total genetic variance; the unstructured distribution of phenotype among genotypes is reflected in the unstructured distribution of genetic variance among the degrees of freedom. Since the majority of degrees of freedom belong to higher-order terms, an unstructured GP map is highly epistatic (but not maximally epistatic, as when all genetic variance is due to n-th-order terms).

Let us now take ω as measurement error and consider the error expected in the calculation of individual reference-free terms from complete data. As shown in Appendix 6, the measurement variance of a term is given by

$$Var(\hat{e}) = \frac{(q-1)^{O(e)}}{q^n} Var(\omega).$$

This shows that the error involved in the measurement of individual terms is smaller than the error involved in the measurement of individual phenotypes (Var[ω]) and will in general be negligible for low-order terms. Reference-free linear decomposition is thus highly robust to measurement error; when reference-free terms are estimated from incomplete data, the dominant cause of error is sampling noise, not measurement error.

This is not true for every linear decomposition. For example, in reference-based linear decomposition, a term of order k is a sum or subtraction of 2^k phenotypes. The error associated with its measurement is therefore $2^k \times \mathrm{Var}(\omega)$; this is always greater than $\mathrm{Var}(\omega)$ and rises exponentially with order such that calculation from measurement is practically impossible for all but the first few orders.

2.5.10 Modeling global nonlinearity

A GP map may be a nonlinear transformation of a simple underlying GP map. It is inefficient to capture such global nonlinearity using linear decomposition, because even a simple global nonlinearity may require a large number of high-order terms; this would reduce the prediction accuracy and obscure the potentially simple structure of the GP map. Our regression-based approach can be extended to account for global nonlinearity. We assume that the measured phenotype y is a nonlinear function ψ of a latent phenotype l:

$$y(\boldsymbol{g}) = \psi[l(\boldsymbol{g})].$$

ψ and a reference-free linear decomposition for l are jointly estimated by regression:

$$\hat{\psi}, \hat{l}_k = \mathrm{argmin} \sum_{\boldsymbol{g} \in S} \{y(\boldsymbol{g}) - \psi[l_k(\boldsymbol{g})]\}^2,$$

with lasso penalty optionally added for sparse learning. This approach requires a parametric function to model ψ. In the absence of any prior knowledge about ψ, splines are the most flexible

option. However, as we showed in the main text, global nonlinearity in most experimental datasets can be effectively captured by a very simple function, the logistic curve:

$$\psi(x) = L + \frac{R}{1 + e^{-x}}.$$

This function takes only two parameters: the lower bound L and the dynamic range R. (A general logistic function also has the midpoint and steepness parameters but they are redundant with the intercept and scale of the linear decomposition.)

2.5.11 Relation to other formalisms

We end by describing the relationship between reference-free linear decomposition and four other formalisms in use: reference-based and background-averaged linear decomposition, simplex encoding, and Fourier decomposition.

In reference-based linear decomposition, each term represents the mutational effect of a state or state-combination with respect to a particular genotype. The formalism is useful for analyzing the local structure of a GP map around a particular genotype, but is not effective for learning the global structure of a GP map or predicting unobserved phenotypes. First, the formalism is not optimal; although it can be more accurate in the neighborhood of the reference genotype, its global accuracy is generally much lower than that of reference-free linear decomposition. Second, measurement of reference-based terms is highly sensitive to measurement error, as shown above. Third, reference-based terms cannot be accurately estimated by regression. For a regression estimate to be unbiased, the residual—the sum of unmodeled higher-order terms—must have an expected value of zero across sampled genotypes. This is true

for reference-free linear decomposition (Appendix 4), but is not guaranteed for reference-based linear decomposition; therefore, regression-based estimates of reference-based terms can be strongly biased by unmodeled higher-order terms.

Simplex encoding captures the global structure of a GP map in the same way as does reference-free linear decomposition and is thus optimal. Simplex encoding is also economical, using the minimum number of terms necessary (total q^n) in contrast to the extra number of terms in reference-free linear decomposition (total $[q + 1]^n$). However, simplex encoding is difficult to intuit and complicated to implement; this is likely why it has only been applied to very small GP maps with only 2 or 4 states. In addition, simplex encoding is not maximally sparse; although this is also the case for reference-free linear decomposition, our sparse learning method based on generalized linear decomposition circumvents this problem.

To illustrate the relationship between reference-free linear decomposition and simplex encoding, we consider a GP map of a single nucleotide. Reference-free linear decomposition of this map consists of five terms: the intercept e_0 and the first-order terms e(A), e(C), e(G), and e(T). These terms are related to the phenotypes as

$$y(A) = e_0 + e(A),$$
$$y(C) = e_0 + e(C),$$
$$y(G) = e_0 + e(G),$$
$$y(T) = e_0 + e(T).$$

These equations and the zero-mean property uniquely determine the five terms. The simplex encoding of the same GP map consists of four terms: the intercept e_0 and the first-order terms e(W), e(Y), and e(K). These terms are related to the phenotypes as

$$y(A) = e_0 + e(W) - e(Y) - e(K),$$
$$y(C) = e_0 - e(W) + e(Y) - e(K),$$
$$y(G) = e_0 - e(W) - e(Y) + e(K),$$
$$y(T) = e_0 + e(W) + e(Y) + e(K).$$

These four equations uniquely determine the four terms. The relationship between reference-free linear decomposition and simplex encoding can be summarized as:

$$\begin{bmatrix} e(A) \\ e(C) \\ e(G) \\ e(T) \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} e(W) \\ e(Y) \\ e(K) \end{bmatrix}.$$

Note that no matter what the values of e(W), e(Y), and e(K) are, e(A), e(C), e(G), and e(T) obtained by the above matrix multiplication satisfy the zero-mean property. Simplex encoding can be considered an efficient implementation of reference-free linear decomposition: The four linearly dependent first-order terms are encoded by three linearly independent terms in a way that guarantees the zero-mean property for the four terms. In general, the q^k reference-free terms for a site-combination of order k can be encoded by $(q-1)^k$ linearly independent terms in a way that guarantees the zero-mean property for the q^k terms. The reduction in the number of terms can be substantial. For a GP map of all combinations of 2 states in 10 sites, reference-free

linear decomposition requires $3^{10} = 59{,}049$ terms, in contrast to the minimum necessary 1,024 terms of simplex encoding.

A drawback of simplex encoding (besides its formal complexity) is that it is not maximally sparse for multi-state GP maps. A GP map is sparse when some states and state-combinations are phenotypically equivalent. In the example GP map, suppose A and C are phenotypically equivalent and that G and T are phenotypically distinct from A and C and from each other. This sparsity cannot be represented by setting the terms of simplex encoding to zero: To make $e(\mathrm{A}) = e(\mathrm{C}) \neq e(\mathrm{G}) \neq e(\mathrm{T})$, we must set $e(\mathrm{W}) = e(\mathrm{Y}) \neq 0$. Conversely, sparsity in simplex encoding does not correspond to sparsity in the GP map: Setting $e(\mathrm{W}) = 0$, for example, makes $e(\mathrm{C}) = -e(\mathrm{G})$ and $e(\mathrm{A}) = -e(\mathrm{T})$. Overall, unless all terms for a site-combination are zero, the economical encoding structure of simplex encoding makes most of the terms non-zero.

Two-state GP maps are exceptions, for which sparsity in simplex encoding coincides with sparsity in GP map. In the simplex encoding of a two-state GP map, there is a single first-order term for each site, which quantifies the phenotypic difference between the two states at that site. The phenotypic equivalence of the two states therefore corresponds to the term being zero. Similarly, the lack of epistatic interaction among a set of sites can be represented by setting the single term for that site-combination to zero. Simplex encoding offers the most optimal and sparse representation for two-state GP maps. We in fact used simplex encoding in our analysis of two-state experimental GP maps.

Until recently, the formalism for the Fourier decomposition of a GP map has been available only for two-state GP maps. In that case, Fourier decomposition is identical to simplex encoding. Formalism for multi-state GP maps based on graph Fourier decomposition has

recently been developed; whether it retains its equivalence to simplex encoding remains to be shown.

Lastly, background-averaged linear decomposition is a modification of reference-based linear decomposition: Each background-averaged term is the average of the corresponding reference-based term across all genetic backgrounds. For two-state GP maps, background-averaged linear decomposition is optimal. Because the formalism has only been developed for two-state GP maps, whether it is optimal for multi-state GP maps is currently unknown. However, model sparsity for multi-state GP maps is expected to depend on the choice of reference states. In the example single-nucleotide GP map, model is sparse when A or C is chosen as the reference state, which makes the effect of A-to-C (or C-to-A) mutation zero, but alternative choices make none of the mutational effects zero. To use background-averaged linear decomposition for multi-state GP maps, a method for identifying the best set of reference states must be developed.

2.5.12 Proofs for the key properties of reference-free linear decomposition

2.5.12.1 Zero-mean property

The zero-mean property of reference-free linear decomposition is that for any site-combination (i_1, …, i_k), the mean of any q terms that vary across a single site is zero:

$$e_{i_1,\ldots,i_k}(\cdot, s_2, \ldots, s_k) = e_{i_1,\ldots,i_k}(s_1, \cdot, s_3, \ldots, s_k) \ldots = e_{i_1,\ldots,i_k}(s_1, \ldots, s_{k-1}, \cdot) = 0,$$

for any state-combination (s_1, …, s_k). The zero-mean property is a defining feature of reference-free linear decomposition, from which its useful properties proved in the next Appendices follow. We prove it by mathematical induction.

Recall that G_i^s is the set of all genotypes with state s in site i. $G_i^1, G_i^2, \ldots, G_i^q$ are nonoverlapping sets whose union is G. It follows that the summation $\Sigma_{s\in Q}\Sigma_{\boldsymbol{g}\in G_i^s}$ is equivalent to $\Sigma_{\boldsymbol{g}\in G}$. Thus,

$$
\begin{aligned}
e_i(\cdot) &= \frac{1}{q}\sum_{s\in Q} e_i(s) \\
&= \frac{1}{q}\sum_{s\in Q}[\langle y|G_i^s\rangle - \langle y|G\rangle] \\
&= \frac{1}{q}\sum_{s\in Q}\langle y|G_i^s\rangle - \langle y|G\rangle \\
&= \frac{1}{q}\sum_{s\in Q}\frac{1}{q^{n-1}}\sum_{\boldsymbol{g}\in G_i^s} y(\boldsymbol{g}) - \langle y|G\rangle \\
&= \frac{1}{q^n}\sum_{\boldsymbol{g}\in G} y(\boldsymbol{g}) - \langle y|G\rangle \\
&= 0.
\end{aligned}
$$

We now show that if the zero-mean property holds for terms of order $k-1$, it also holds for terms of order k. Recall the definition

$$
e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)\big|_{i_1}^{s_1} - e_{i_2,\ldots,i_k}(s_2,\ldots,s_k).
$$

By the inductive hypothesis, the mean of the two $(k-1)$-th-order terms on the right-hand side is 0 across any site $i_2, \ldots, i_k$. The mean of the k-th-order term is thus 0 across any site $i_2, \ldots, i_k$. Conditioning on a site other than i_1 shows that the mean of the k-th-order term is also 0 across i_1. This completes the mathematical induction.

2.5.12.2 Generalized linear decomposition.

Let $e_{i_1,\ldots,i_k}: Q^k \longmapsto \mathbb{R}$ be a function mapping k-tuples of states into real numbers. We refer to the following expression as the k-th-order generalized linear decomposition of a GP map:

$$y_k(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}). \quad \text{(A1)}$$

We showed that both reference-free and reference-based linear decomposition can be represented as above with suitable choices of e. We now argue that any linear decomposition of a GP map can be represented as above. The truth of this statement depends on what a linear decomposition of a GP map is. Below we define a linear decomposition of a GP map in the broadest possible sense and show that it can be represented as Eq. (A1).

In the broadest sense, a linear decomposition of order 0 is any function that approximates the phenotype of every genotype by a constant. It can thus be represented as

$$y_0(\boldsymbol{g}) = e_0.$$

What is the broadest sense in which a function $y_1(\boldsymbol{g})$ is a first-order linear decomposition of a GP map? $y_1(\boldsymbol{g})$ should be able to use the information that $\boldsymbol{g}$ has the state g_i in site i and

combine that information linearly across sites to determine the phenotype of $\boldsymbol{g}$. It is not allowed, however, to use any information about what combination of states is found in what combination of sites. Any such function y_1 can be written as

$$y_1(\boldsymbol{g}) = \sum_j \lambda_j(\boldsymbol{g}),$$

where λ_j is a function defined as

$$\lambda_j(\boldsymbol{g}) = \begin{cases} \alpha_j, & \boldsymbol{g} \in G_i^s \\ \beta_j, & \boldsymbol{g} \notin G_i^s \end{cases}$$

for some state *s,* some site *i*, and some real numbers α_j and β_j. In other words, y_1 is a linear combination of an arbitrary number of functions, each of which can distinguish whether or not a genotype has a particular state in a particular site.

For any such function y_1, we can find a constant e_0 and functions $e_i: Q \longmapsto \mathbb{R}$ such that

$$y_1(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i).$$

To prove this, we first sum all functions λ that distinguish whether or not $\boldsymbol{g} \in G_i^s$ and write the sum as

$$\lambda_i^s(\boldsymbol{g}) = \begin{cases} \alpha_i^s, & \boldsymbol{g} \in G_i^s \\ \beta_i^s, & \boldsymbol{g} \notin G_i^s \end{cases}.$$

y_1 can be written as

$$\begin{aligned} y_1(g) &= \sum_j \lambda_j(\boldsymbol{g}) \\ &= \sum_{i\in N}\sum_{s\in Q} \lambda_i^s(\boldsymbol{g}) \\ &= \sum_{i\in N}\sum_{s\in Q} [\lambda_i^s(\boldsymbol{g}) - \beta_i^s + \beta_i^s] \\ &= \sum_{i\in N}\sum_{s\in Q} [\lambda_i^s(\boldsymbol{g}) - \beta_i^s] + \sum_{i\in N}\sum_{s\in Q} \beta_i^s. \end{aligned}$$

Note that

$$\lambda_i^s(\boldsymbol{g}) - \beta_i^s = \begin{cases} \alpha_i^s - \beta_i^s, & \boldsymbol{g} \in G_i^s \\ 0, & \boldsymbol{g} \notin G_i^s \end{cases}$$

and thus

$$\sum_{s\in Q} [\lambda_i^s(\boldsymbol{g}) - \beta_i^s] = \alpha_i^{g_i} - \beta_i^{g_i}.$$

The following choice therefore completes the proof:

$$e_0 = \sum_{i\in N}\sum_{s\in Q} \beta_i^s,$$

$$e_i(s) = \alpha_i^s - \beta_i^s.$$

Higher-order linear decompositions can be defined similarly. In the broadest sense, a second-order linear decomposition of a GP map can use the information that $\boldsymbol{g}$ has the state g_i in site i and state-pair (g_{i_1}, g_{i_2}) in site-pair (i_1, i_2), but cannot use any information about higher-order combinations. Any such function y_2 can be written as

$$y_2(\boldsymbol{g}) = \sum_j \lambda_j(\boldsymbol{g}) + \sum_k \mu_k(\boldsymbol{g}),$$

where λ_j is defined as above and μ_k is a function defined as

$$\mu_k(\boldsymbol{g}) = \begin{cases} \gamma_k, & \boldsymbol{g} \in G_{i_1,i_2}^{s_1,s_2} \\ \delta_k, & \boldsymbol{g} \notin G_{i_1,i_2}^{s_1,s_2} \end{cases}$$

for some sites i_1 and i_2, some states s_1 and s_2, and some real numbers γ_k and δ_k. That is, y_2 is a linear combination of an arbitrary number of functions, each of which can distinguish whether or not each genotype has a particular state or state-pair in a particular site or site-pair. For any such function y_2, we can find a constant e_0 and functions $e_i: Q \longmapsto \mathbb{R}$ and $e_{i_1,i_2}: Q^2 \longmapsto \mathbb{R}$ such that

$$y_2(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}).$$

To prove this, we again sum all functions λ that distinguish whether or not $\boldsymbol{g} \in G_i^s$ and write the sum as

$$\lambda_i^s(\boldsymbol{g}) = \begin{cases} \alpha_i^s, & \boldsymbol{g} \in G_i^s \\ \beta_i^s, & \boldsymbol{g} \notin G_i^s, \end{cases}$$

and similarly sum all functions μ that distinguish whether or not $\boldsymbol{g} \in G_{i_1,i_2}^{s_1,s_2}$ and write the sum as

$$\mu_{i_1,i_2}^{s_1,s_2}(\boldsymbol{g}) = \begin{cases} \gamma_{i_1,i_2}^{s_1,s_2}, & \boldsymbol{g} \in G_{i_1,i_2}^{s_1,s_2} \\ \delta_{i_1,i_2}^{s_1,s_2}, & \boldsymbol{g} \notin G_{i_1,i_2}^{s_1,s_2}. \end{cases}$$

Following a logic similar to above, we can choose

$$e_0 = \sum_{i \in N} \sum_{s \in Q} \beta_i^s + \sum_{i_1 < i_2 \in N} \sum_{s_1,s_2 \in Q} \beta_{i_1,i_2}^{s_1,s_2},$$

$$e_i(s) = \alpha_i^s - \beta_i^s,$$

$$e_{i_1,i_2}(s_1,s_2) = \gamma_{i_1,i_2}^{s_1,s_2} - \delta_{i_1,i_2}^{s_1,s_2}.$$

In general, a *k*-th-order linear decomposition of a GP map in the broadest sense is any function that can use information about the combination of states in any set of up to *k* sites and combine that information linearly across site-combinations to determine the phenotype. A logic similar to above can show that any such function can be written as Eq. (A1).

2.5.12.3 Optimal linear decomposition.

Recall that any *k*-th-order linear decomposition of a GP map can be written as

$$y_k(\boldsymbol{g}) = e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}).$$

Here we show that reference-free linear decomposition minimizes the sum of squared error

$$\epsilon = \sum_{\boldsymbol{g} \in G} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2,$$

for any order *k*. We prove this by showing that the partial derivative of ε with respect to each term is 0 when the terms are defined according to reference-free linear decomposition. Using *e* to denote any term,

$$\begin{aligned} \frac{\partial \epsilon}{\partial e} &= \frac{\partial}{\partial e} \sum_{\boldsymbol{g} \in G} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2 \\ &= \sum_{\boldsymbol{g} \in G} \frac{\partial}{\partial e} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2 \\ &= \sum_{\boldsymbol{g} \in G} -2[y(\boldsymbol{g}) - y_k(\boldsymbol{g})] \frac{\partial}{\partial e} y_k(\boldsymbol{g}). \end{aligned}$$

The derivative of $y_k(\boldsymbol{g})$ with respect to e is 1 if $y_k(\boldsymbol{g})$ involves e and 0 otherwise. For example, for $e = e_i(s)$, the derivative of $y_k(\boldsymbol{g})$ is 1 for all $\boldsymbol{g} \in G_i^s$ and 0 for $\boldsymbol{g} \notin G_i^s$. For $e = e_{i_1,\ldots,i_k}(s_1, \ldots, s_k)$, let G_e denote the set $G_{i_1,\ldots,s_k}^{s_1,\ldots,s_k}$. Then,

$$\frac{\partial}{\partial e} y_k(\boldsymbol{g}) = \begin{cases} 1, & \boldsymbol{g} \in G_e \\ 0, & \boldsymbol{g} \notin G_e \end{cases}.$$

The partial derivative of ε is therefore

$$\begin{aligned} \frac{\partial \epsilon}{\partial e} &= \sum_{\boldsymbol{g} \in G} -2[y(\boldsymbol{g}) - y_k(\boldsymbol{g})] \frac{\partial}{\partial e} y_k(\boldsymbol{g}) \\ &= -2 \sum_{\boldsymbol{g} \in G_e} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})] \\ &= -2q^{n-O(e)} [\langle y | G_e \rangle - \langle y_k | G_e \rangle], \end{aligned}$$

where $O(e)$ is the order of e. The partial derivative is 0 when

$$\langle y | G_e \rangle = \langle y_k | G_e \rangle.$$

In other words, an optimal linear decomposition of order k is that which accurately predicts the average phenotype of any sub-map defined by fixing up to k sites. We defined reference-free linear decomposition to achieve just that! To formally show this, we prove the following property of reference-free linear decomposition:

Lemma 1. Consider a site-combination $(j_1, \dots, j_l)$ and any combination of states therein, denoted by $(s_{j_1}, \dots, s_{j_l})$. For any site-combination $(i_1, \dots, i_k)$,

$$\frac{1}{q^{n-l}} \sum_{\boldsymbol{g} \in G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k}) = \begin{cases} e_{i_1,\dots,i_k}(s_{i_1}, \dots, s_{i_k}), & (i_1, \dots, i_k) \subseteq (j_1, \dots, j_l) \\ 0, & otherwise \end{cases}.$$

That is, the summation is nonzero if and only if (i_1, …, i_k) is a subset of (j_1, …, j_l). This follows from the zero-mean property: If any site in (i_1, …, i_k) is outside the site-combination (j_1, …, j_l), the above summation involves summing across all q states in that site.

Before proving Lemma 1, let us see how it helps us. Take any term $e = e_{j_1,\dots,j_l}(s_{j_1}, \dots, s_{j_l})$, $l \leq k$:

$$\begin{aligned} \langle y_k | G_e \rangle &= \frac{1}{q^{n-l}} \sum_{g \in G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}} y_k(\boldsymbol{g}) \\ &= \frac{1}{q^{n-l}} \sum_{g \in G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}} \left[e_0 + \sum_{i \in N} e_i(g_i) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(g_{i_1}, g_{i_2}) + \cdots \right. \\ &\quad \left. + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k}) \right]. \end{aligned}$$

Due to Lemma 1, the sum for any site-combination (i_1, …, i_m) that is not a subset of (j_1, …, j_l) is 0. Therefore, using L to denote the set (*1, 2, ..., l*),

$$\langle y_k|G_e\rangle = e_0 + \sum_{\alpha\in L} e_{j_\alpha}(s_{j_\alpha}) + \sum_{\alpha_1<\alpha_2\in L} e_{j_{\alpha_1},j_{\alpha_2}}(s_{j_{\alpha_1}}, s_{j_{\alpha_2}}) + \cdots$$

$$+ \sum_{\alpha_1<\cdots<\alpha_{l-1}\in L} e_{j_{\alpha_1},\ldots,j_{\alpha_{l-1}}}\left(s_{j_{\alpha_1}}, \ldots, s_{j_{\alpha_{l-1}}}\right) + e_{j_1,\ldots,j_l}(s_{j_1}, \ldots, s_{j_l}).$$

This, by the definition of reference-free linear decomposition, equals $\left\langle y\middle|G_{j_1,\ldots,j_l}^{s_{j_1},\ldots,s_{j_l}}\right\rangle = \langle y|G_e\rangle$:

$$\left\langle y\middle|G_{j_1,\ldots,j_l}^{s_{j_1},\ldots,s_{j_l}}\right\rangle = \left\langle y\middle|G_{j_1,\ldots,j_l}^{s_{j_1},\ldots,s_{j_l}}\right\rangle_{(l-1)} + e_{j_1,\ldots,j_l}(s_{j_1}, \ldots, s_{j_l}),$$

$$\left\langle y\middle|G_{j_1,\ldots,j_l}^{s_{j_1},\ldots,s_{j_l}}\right\rangle_{(l-1)}$$

$$= e_0 + \sum_{\alpha\in L} e_{j_\alpha}(s_{j_\alpha}) + \sum_{\alpha_1<\alpha_2\in L} e_{j_{\alpha_1},j_{\alpha_2}}(s_{j_{\alpha_1}}, s_{j_{\alpha_2}}) + \cdots$$

$$+ \sum_{\alpha_1<\cdots<\alpha_{l-1}\in L} e_{j_{\alpha_1},\ldots,j_{\alpha_{l-1}}}\left(s_{j_{\alpha_1}}, \ldots, s_{j_{\alpha_{l-1}}}\right).$$

Proof of Lemma 1.

Consider first the case when $(i_1, \ldots, i_k) \subseteq (j_1, \ldots, j_l)$. Across the set $G_{j_1,\ldots,j_l}^{s_{j_1},\ldots,s_{j_l}}$, which consists only of genotypes with states $s_{j_1}, \ldots, s_{j_l}$ in sites $j_1, \ldots, j_l$, the term $e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k})$ is a constant, $e_{i_1,\ldots,i_k}(s_{i_1}, \ldots, s_{i_k})$. This proves the first case of the lemma.

Let us now assume that m sites in $(i_1, \ldots, i_k)$ are outside $(j_1, \ldots, j_l)$. We make our notation flexible so that the order of sites can be permuted: *e.g.*, $e_{i_1,i_2}(s_1, s_2) = e_{i_2,i_1}(s_2, s_1)$. We can then write

$$(i_1, \dots, i_k) = (x_1, \dots, x_{k-m}, y_1, \dots, y_m),$$

where $x_1, \dots, x_{k-m}$ are inside $(j_1, \dots, j_l)$ and $y_1, \dots, y_m$ outside. $G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}$ can be partitioned as

$$G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}} = \bigcup_{t_1,\dots,t_m \in Q} G_{j_1,\dots,j_l,y_1,\dots,y_m}^{s_{j_1},\dots,s_{j_l},t_1,\dots,t_m}.$$

Therefore, the summation $\Sigma_{\boldsymbol{g} \in G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}}$ equals $\Sigma_{t_1,\dots,t_m \in Q} \Sigma_{\boldsymbol{g} \in G_{j_1,\dots,j_l,y_1,\dots,y_m}^{s_{j_1},\dots,s_{j_l},t_1,\dots,t_m}}$. It follows then

$$\frac{1}{q^{n-l}} \sum_{\boldsymbol{g} \in G_{j_1,\dots,j_l}^{s_{j_1},\dots,s_{j_l}}} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k}) = \frac{1}{q^{n-l}} \sum_{t_1,\dots,t_m \in Q} \sum_{\boldsymbol{g} \in G_{j_1,\dots,j_l,y_1,\dots,y_m}^{s_{j_1},\dots,s_{j_l},t_1,\dots,t_m}} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})$$

$$= \frac{1}{q^{n-l}} \sum_{t_1,\dots,t_m \in Q} \sum_{\boldsymbol{g} \in G_{j_1,\dots,j_l,y_1,\dots,y_m}^{s_{j_1},\dots,s_{j_l},t_1,\dots,t_m}} e_{x_1,\dots,x_{k-m},y_1,\dots,y_m}(s_{x_1}, \dots, s_{x_{k-m}}, t_1, \dots, t_m)$$

$$= \frac{q^l}{q^{n-l}} \sum_{t_1,\dots,t_m \in Q} e_{x_1,\dots,x_{k-m},y_1,\dots,y_m}(s_{x_1}, \dots, s_{x_{k-m}}, t_1, \dots, t_m)$$

$$= 0.$$

The last equality follows from the zero-mean property.

2.5.12.4 Variance partition

Our goal is to show that the total genetic variance,

$$V = Var(y|G) = \frac{1}{q^n}\sum_{\boldsymbol{g}\in G}[y(\boldsymbol{g}) - \langle y|G\rangle]^2,$$

can be decomposed into the effect-variance of each site-combination:

$$V = \sum_{i\in N} Var(e_i|Q) + \sum_{i_1<i_2\in N} Var\left(e_{i_1,i_2}\middle|Q^2\right) + \cdots + \sum_{i_1<\cdots<i_k\in N} Var\left(e_{i_1,\ldots,i_k}\middle|Q^k\right) + \cdots + Var\left(e_{1,\ldots,n}\middle|Q^n\right).$$

We write $\boldsymbol{g} = (g_1, \ldots, g_n)$ and consider g_i as a function that takes a value between 1 and q depending on the genotype $\boldsymbol{g}$. For any genotype $\boldsymbol{g} \in G$,

$$y(\boldsymbol{g}) = e_0 + \sum_{i\in N} e_i(g_i) + \cdots + \sum_{i_1<\cdots<i_k\in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}) + \cdots + e_{i_1,\ldots,i_n}(g_1, \ldots, g_n).$$

Substituting $\langle y|G\rangle = e_0$ and decomposing $y(\boldsymbol{g})$ as above,

$$V = \frac{1}{q^n}\sum_{\boldsymbol{g}\in G}\left[\sum_{i\in N} e_i(g_i) + \cdots + \sum_{i_1<\cdots<i_k\in N} e_{i_1,\ldots,i_k}(g_{i_1}, \ldots, g_{i_k}) + \cdots + e_{i_1,i_2,\ldots,i_n}(g_1, g_2, \ldots, g_n)\right]^2. \quad \text{(A1)}$$

To simplify this equation, we prove the following lemma.

Lemma 2. For any two distinct site-combinations $(i_1, \dots, i_k) \neq (j_1, \dots, j_l)$,

$$\frac{1}{q^n} \sum_{\boldsymbol{g} \in G} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})\, e_{j_1,\dots,j_l}(g_{j_1}, \dots, g_{j_l}) = 0.$$

Before proving Lemma 2, let us check how it simplifies Eq. (A1). Recall that $(\sum_i x_i)^2 = \sum_i x_i^2 + \sum_{i,j} x_i x_j$. Therefore, under Lemma 2, Eq. (A1) simplifies to

$$V = \frac{1}{q^n} \sum_{\boldsymbol{g} \in G} \left[\sum_{i \in N} e_i(g_i)^2 + \cdots + \sum_{i_1 < \cdots < i_k \in N} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})^2 + \cdots + e_{i_1,i_2,\dots,i_n}(g_1, g_2, \dots, g_n)^2 \right].$$

The set G can be expressed as the union $G = \bigcup_{s_1,\dots,s_k \in Q} G_{i_1,\dots,i_k}^{s_1,\dots,s_k}$ for any state-combination (s_1, …, s_k) in any site-combination (i_1, …, i_k). Therefore,

$$\begin{aligned}
\frac{1}{q^n} \sum_{\boldsymbol{g} \in G} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})^2 &= \frac{1}{q^n} \sum_{s_1,\dots,s_k \in Q} \sum_{\boldsymbol{g} \in G_{i_1,\dots,i_k}^{s_1,\dots,s_k}} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})^2 \\
&= \frac{1}{q^n} \sum_{s_1,\dots,s_k \in Q} \sum_{\boldsymbol{g} \in G_{i_1,\dots,i_k}^{s_1,\dots,s_k}} e_{i_1,\dots,i_k}(s_1, \dots, s_k)^2 \\
&= \frac{1}{q^n} \sum_{s_1,\dots,s_k \in Q} q^{n-k} e_{i_1,\dots,i_k}(s_1, \dots, s_k)^2 \\
&= \frac{1}{q^k} \sum_{s_1,\dots,s_k \in Q} e_{i_1,\dots,i_k}(s_1, \dots, s_k)^2 \\
&= Var\left(e_{i_1,\dots,i_k} | Q^k\right).
\end{aligned}$$

This proves the variance partition formula.

Proof of Lemma 2.

We make our notation flexible so that the order of sites can be permuted: *e.g.*, $e_{i_1,i_2}(s_1, s_2) = e_{i_2,i_1}(s_2, s_1)$. Assume m sites are shared between $(i_1, \dots, i_k)$ and $(j_1, \dots, j_l)$. We can write

$$(i_1, \dots, i_k) = (a_1, \dots, a_m, x_1, \dots x_{k-m}),$$
$$(j_1, \dots, j_l) = (a_1, \dots, a_m, y_1, \dots y_{l-m}),$$

where $(x_1, \dots x_{k-m}) \cap (y_1, \dots y_{l-m}) = \emptyset$. We partition G as follows:

$$G = \bigcup_{s_1,\dots,s_m \in Q} \bigcup_{u_1,\dots,u_{k-m} \in Q} \bigcup_{v_1,\dots,v_{l-m} \in Q} G_{a_1,\dots,a_m,x_1,\dots,x_{k-m},y_1,\dots,y_{l-m}}^{s_1,\dots,s_m,u_1,\dots,u_{k-m},v_1,\dots,v_{l-m}}.$$

It then follows

$$\frac{1}{q^n} \sum_{\boldsymbol{g} \in G} e_{i_1,\dots,i_k}(g_{i_1}, \dots, g_{i_k})\, e_{j_1,\dots,j_l}(g_{j_1}, \dots, g_{j_l})$$

$$= \frac{q^{n-(k+l-m)}}{q^n} \sum_{s_1,\dots,s_m \in Q} \sum_{u_1,\dots,u_{k-m} \in Q} \sum_{v_1,\dots,v_{l-m} \in Q} e_{i_1 \dots i_k}(s_1, \dots, s_m, u_1, \dots, u_{k-m}) e_{j_1 \dots j_l}(s_1, \dots, s_m, v_1, \dots, v_{l-m})$$

$$= \frac{1}{q^{k+l-m}} \sum_{s_1,\dots,s_m \in Q} \sum_{u_1,\dots,u_{k-m} \in Q} e_{i_1 \dots i_k}(s_1, \dots, s_m, u_1, \dots, u_{k-m}) \sum_{v_1,\dots,v_{l-m} \in Q} e_{j_1 \dots j_l}(s_1, \dots, s_m, v_1, \dots, v_{l-m})$$

$= 0$.

The last equality follows from the zero-mean property. A similar proof can be constructed for $m = 0$.

2.5.12.5 Estimation by random sampling

We estimate the terms of reference-free linear decomposition from a random sample of genotypes by a two-step procedure. First, we solve the following optimization problem:

$$\hat{y}_k = argmin \sum_{\boldsymbol{g} \in S} [y(\boldsymbol{g}) - y_k(\boldsymbol{g})]^2. \quad \text{(A3)}$$

For any solution $\hat{y}_k$, we normalize the terms to simultaneously satisfy Eq. (A3) and the zero-mean property. We show in Lemma 3 that this normalization is always possible. Given Lemma 3, we show that normalized $\hat{y}_k$ is an unbiased estimate of reference-free linear decomposition. To do so, we reformulate Eq. (A3) as a standard linear regression problem. Let y be a vector of sampled phenotypes and β a vector of linear decomposition terms, both arranged in any order. We write

$$y = X\beta + \epsilon, \quad \text{(A4)}$$

where X is the design matrix specifying how the terms encode the phenotypes. The error ε is the sum of all unmodeled higher-order terms and measurement error. The solution to this linear regression is not unique because X is a singular matrix (due to the degeneracy of generalized

linear decomposition). We make X non-singular by building in the zero-mean property. For example, we eliminate the column of X corresponding to $e_i(q)$ by coding it as

$$e_i(q) = -\sum_{t \in Q \setminus q} e_i(t).$$

Similarly, we eliminate the column of X corresponding to $e_{i_1,i_2}(s_1, q)$ by coding it as

$$e_{i_1,i_2}(s_1, q) = -\sum_{t \in Q \setminus q} e_{i_1,i_2}(s_1, t).$$

In general, every term involving state q in any site can be eliminated by coding it as a linear combination of terms involving states 1 to $q - 1$ according to the zero-mean property. The design matrix thus obtained is non-singular and can be used to infer for all terms containing states 1 to $q - 1$. Estimates for terms containing state q can be calculated post-hoc using the zero-mean property. The terms thus obtained are identical to the terms that would result from the two-step procedure (due to Lemma 3). Given that least-squares estimates based on a non-singular design matrix is unique, this proves the uniqueness of normalized $\hat{y}_k$.

Finally, we show below that the expected value of the error ε in Eq. (A4) is zero across randomly sampled genotypes. Since the design matrix is non-singular and the errors are unbiased, by the Gauss-Markov theorem the regression estimates for terms containing states 1 to $q - 1$ are unbiased. The post-hoc estimates for terms containing state q are also unbiased because they are linear combinations of terms containing states 1 to $q - 1$.

For a model of order k, the error for a genotype $\boldsymbol{g}$ is given by

$$\epsilon(\boldsymbol{g}) = \sum_{i_1 < \cdots < i_{k+1} \in N} e_{i_1,\ldots,i_{k+1}}(g_{i_1}, \ldots, g_{i_{k+1}}) + \cdots + e_{1,\ldots,n}(g_1, \ldots, g_n).$$

Since genotypes are randomly sampled, it suffices to show that the expected value of ϵ is zero across all genotypes: $\langle \epsilon | G \rangle = 0$. We prove a stronger result:

$$\begin{aligned}
\left\langle e_{i_1,\ldots,i_k} \middle| G \right\rangle &= \frac{1}{q^n} \sum_{\boldsymbol{g} \in G} e_{i_1,\ldots,i_k}(g_1, \ldots, g_k) \\
&= \frac{1}{q^n} \sum_{s_1,\ldots,s_k \in Q} \sum_{\boldsymbol{g} \in G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k}} e_{i_1,\ldots,i_k}(g_1, \ldots, g_k) \\
&= \frac{1}{q^n} \sum_{s_1,\ldots,s_k \in Q} \sum_{\boldsymbol{g} \in G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k}} e_{i_1,\ldots,i_k}(s_1, \ldots, s_k) \\
&= \frac{1}{q^k} \sum_{s_1,\ldots,s_k \in Q} e_{i_1,\ldots,i_k}(s_1, \ldots, s_k) \\
&= 0.
\end{aligned}$$

It follows that $\langle \epsilon | G \rangle = 0$.

Lemma 3. Post-hoc enforcement of zero-mean property.

Consider enforcing the zero-mean property on a first-order linear decomposition without altering the predicted phenotypes (hereafter called "normalizing"). To normalize the terms for site *i*, we first subtract from each term the mean of all terms at site *i*:

$$\delta_i(s) = e_i(s) - e_i(\cdot) \Longrightarrow \delta_i(\cdot) = 0.$$

Using $\delta_i(s)$ in place of $e_i(s)$ alters the predicted phenotype of every genotype by $-e_i(\cdot)$. This can be corrected by adding $e_i(\cdot)$ to the intercept. Overall, the following modifications normalize any first-order linear decomposition:

$$\delta_i(s) = e_i(s) - e_i(\cdot),$$
$$\delta_0 = e_0 + \sum_{i \in N} e_i(\cdot).$$

Similarly, the following modifications normalize any second-order linear decomposition:

$$\delta_{i_1,i_2}(s_1, s_2) = e_{i_1,i_2}(s_1, s_2) - e_{i_1,i_2}(\cdot, s_2) - e_{i_1,i_2}(s_1, \cdot) + e_{i_1,i_2}(\cdot, \cdot),$$
$$\delta_i(s) = [e_i(s) - e_i(\cdot)] + \sum_{j \in N \setminus i} \left[e_{i,j}(s, \cdot) - e_{i,j}(\cdot, \cdot)\right],$$
$$\delta_0 = e_0 + \sum_{i \in N} e_i(\cdot) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(\cdot, \cdot).$$

For any third-order linear decomposition:

$$\begin{aligned}\delta_{i_1,i_2,i_3}(s_1, s_2, s_3) \\ &= e_{i_1,i_2,i_3}(s_1, s_2, s_3) - \left[e_{i_1,i_2,i_3}(\cdot, s_2, s_3) + e_{i_1,i_2,i_3}(s_1, \cdot, s_3) + e_{i_1,i_2,i_3}(s_1, s_2, \cdot)\right] \\ &+ \left[e_{i_1,i_2,i_3}(\cdot, \cdot, s_3) + e_{i_1,i_2,i_3}(\cdot, s_2, \cdot) + e_{i_1,i_2,i_3}(s_1, \cdot, \cdot)\right] - e_{i_1,i_2,i_3}(\cdot, \cdot, \cdot),\end{aligned}$$

$$\delta_{i_1.i_2}(s_1, s_2) = e_{i_1,i_2}(s_1, s_2) - \left[e_{i_1,i_2}(\cdot, s_2) + e_{i_1,i_2}(s_1, \cdot)\right] + e_{i_1,i_2}(\cdot,\cdot) + \sum_{i_3 \in N \setminus i_1,i_2} \left[e_{i_1,i_2,i_3}(s_1, s_2, \cdot) - e_{i_1,i_2,i_3}(\cdot, s_2, \cdot) - e_{i_1,i_2,i_3}(s_1, \cdot, \cdot) + e_{i_1,i_2,i_3}(\cdot, \cdot, \cdot)\right],$$

$$\delta_i(s) = [e_i(s) - e_i(\cdot)] + \sum_{j \in N \setminus i} \left[e_{i,j}(s, \cdot) - \epsilon_{ij}(\cdot, \cdot)\right] + \sum_{j<k \in N \setminus i} \left[e_{i,j,k}(s, \cdot, \cdot) - \epsilon_{i,j,k}(\cdot, \cdot, \cdot)\right],$$

$$\delta_0 = \epsilon_0 + \sum_{i \in N} e_i(\cdot) + \sum_{i_1 < i_2 \in N} e_{i_1,i_2}(\cdot, \cdot) + \sum_{i_1 < i_2 < i_3 \in N} e_{i_1,i_2,i_3}(\cdot, \cdot, \cdot).$$

Normalization formulae for a higher-order linear decomposition can be found by a similar logic.

Directly applying normalization formulae for higher-order linear decompositions can be cumbersome. We describe a simple alternative. We first normalize the highest-order terms (order k) without correcting for the altered phenotypes. This can be done by Eq. (A5) below. Then, let y_k denote the phenotype predicted by the original linear decomposition and z_k the phenotypic contribution of the normalized k-th-order terms. We must modify the lower-order terms so that their total phenotypic contribution is $y_k - z_k$. This can be done by using regression to find a linear model of order $k - 1$ whose predicted phenotype is $y_k - z_k$. Such a linear model exists because post-hoc enforcement of zero-mean constraint is always possible. Terms of order $k - 1$ can then be normalized, using regression to find a linear model of order $k - 2$ that corrects for the altered phenotypes.

To show how terms of order k can be normalized, we introduce a notation: $e_{i_1,\ldots,i_k}(s_1, \ldots, s_k)_{j_1,\ldots j_l}$ denotes the mean of $e_{i_1,\ldots,i_k}(s_1, \ldots, s_k)$ across sites $j_1, \ldots, j_l$. For example,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)_{i_1} = e_{i_1,\ldots,i_k}(\cdot,s_2,\ldots,s_k),$$

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)_{i_1,i_2} = e_{i_1,\ldots,i_k}(\cdot,\cdot,s_3,\ldots,s_k).$$

Normalization of second-order terms can be restated as

$$\delta_{i_1,i_2}(s_1,s_2) = e_{i_1,i_2}(s_1,s_2) - \sum_{\alpha\in(1,2)} e_{i_1,i_2}(s_1,s_2)_{i_\alpha} + \sum_{\alpha_1<\alpha_2\in(1,2)} e_{i_1,i_2}(s_1,s_2)_{i_{\alpha_1},i_{\alpha_2}}.$$

Denoting the set (*1, ..., k*) by *K*, terms of order *k* can be normalized by

$$\delta_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) + \sum_{l\in K}(-1)^l \sum_{\alpha_1<\cdots<\alpha_l\in K} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)_{i_{\alpha_1},\ldots,i_{\alpha_l}}. \quad \text{(A5)}$$

2.5.12.6 Unstructured GP maps

Recall that the contribution of a reference-free term *e* to the total genotypic variance is $e^2/q^{O(e)}$. Our goal is to calculate the expected value of e^2 when the GP map is generated by sampling each phenotype independently from a random variable ω. We write the expected value as $\langle e^2|\Omega\rangle$, where Ω is the set of all GP maps generated from ω. Below we show that

$$\langle e^2|\Omega\rangle = \frac{(q-1)^{O(e)}}{q^n} Var(\omega). \quad \text{(A6)}$$

Since the genotypic variance due to all terms of order *k* is

$$V_k = \sum_{i_1<\cdots<i_k \in N} Var(e_{i_1,\ldots,i_k}|Q^k) = \sum_{i_1<\cdots<i_k \in N} \frac{1}{q^k} \sum_{s_1,\ldots,s_k \in Q} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)^2,$$

it follows from Eq. (A6) that

$$\langle V_k|\Omega\rangle = \sum_{i_1<\cdots<i_k \in N} \frac{1}{q^k} \sum_{s_1,\ldots,s_k \in Q} \frac{(q-1)^k}{q^n} Var(\omega) = \binom{n}{k} \frac{(q-1)^k}{q^n} Var(\omega) \propto \binom{n}{k} (q-1)^k.$$

We prove Eq. (A6) by mathematical induction. Without loss of generality, we assume the expected value of ω to be 0. The expected value of any average phenotype is then 0 and thus that of any reference-free term: $\langle e|\Omega\rangle = 0$. We now calculate $Var(e|\Omega) = \langle e^2|\Omega\rangle - \langle e|\Omega\rangle^2 = \langle e^2|\Omega\rangle$.

We first consider the intercept:

$$\begin{aligned} Var(e_0|\Omega) &= Var\left[\frac{1}{q^n}\sum_{\boldsymbol{g}\in G} y(\boldsymbol{g}) \,\middle|\, \Omega\right] \\ &= \frac{1}{q^{2n}} \sum_{\boldsymbol{g}\in G} Var[y(\boldsymbol{g})|\Omega] \\ &= \frac{Var(\omega)}{q^n}. \end{aligned}$$

This is Eq. (A6) for the case $O(e) = 0$. Let us now assume that Eq. (A6) holds for terms of order $k - 1$. First, recall the expression

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{s_1} - e_{i_2,\ldots,i_k}(s_2,\ldots,s_k),$$

which writes a term of order k as a function of terms of order $k-1$. We prove that for any $j \neq i_1,\ldots,i_k$,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = \frac{1}{q}\sum_{t\in Q} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)|_j^t. \tag{A7}$$

That is, a term of order k for the complete map G is the average of the same term calculated for the sub-maps defined by conditioning on any site outside the k focal sites. Before proving Eq. (A7), let us see how it helps us:

$$\begin{aligned} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) &= e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{s_1} - e_{i_2,\ldots,i_k}(s_2,\ldots,s_k) \\ &= e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{s_1} - \frac{1}{q}\sum_{t\in Q} e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t} \\ &= \frac{q-1}{q} e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{s_1} - \frac{1}{q}\sum_{t\in Q\setminus s_1} e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t}. \end{aligned}$$

Note that the q terms $e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t}$, $t = 1, \ldots, q$, are probabilistically independent of each other because they involve genotypes from disjoint sub-maps. Furthermore,

$$\begin{aligned} Var\left[e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)\middle|\Omega\right] &= Var\left[\frac{1}{q}\textstyle\sum_{t\in Q} e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t}\middle|G\right] \\ &= \frac{1}{q} Var\left[e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t}\middle|\Omega\right], \end{aligned}$$

where t in the last term can be any one of the q states. This implies that for any such t,

$$Var\left[e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)|_{i_1}^{t}\middle|\Omega\right] = q \times Var\left[e_{i_2,\ldots,i_k}(s_2,\ldots,s_k)\middle|\Omega\right] = \frac{q(q-1)^{k-1}}{q^n} Var(\omega).$$

From the aforementioned probabilistic independence, it follows that

$$Var\left[e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)\middle|\Omega\right] = \left[\left(\frac{q-1}{q}\right)^2 + \frac{q-1}{q^2}\right] \times \frac{q(q-1)^{k-1}}{q^n} Var(\omega)$$

$$= \frac{(q-1)^k}{q^n} Var(\omega).$$

We now turn to proving Eq. (A7). We can rewrite $e_i(s)$ purely in terms of averaged phenotypes:

$$e_i(s) = \langle y|G_i^s\rangle - \langle y|G\rangle.$$

Similarly,

$$e_{i_1,i_2}(s_1,s_2) = \left\langle y\middle|G_{i_1,i_2}^{s_1,s_2}\right\rangle - \left\langle y\middle|G_{i_1}^{s_1}\right\rangle - \left\langle y\middle|G_{i_2}^{s_2}\right\rangle + \langle y|G\rangle.$$

In general,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)$$

$$= \left\langle y \middle| G_{i_1,\ldots,i_k}^{s_1,\ldots,s_k} \right\rangle - \sum_{\alpha_1<\cdots<\alpha_{k-1}\in K} \left\langle y \middle| G_{i_{\alpha_1},\ldots,i_{\alpha_{k-l}}}^{s_{\alpha_1},\ldots,s_{\alpha_{k-l}}} \right\rangle + \sum_{\alpha_1<\cdots<\alpha_{k-2}\in K} \left\langle y \middle| G_{i_{\alpha_1},\ldots,i_{\alpha_{k-2}}}^{s_{\alpha_1},\ldots,s_{\alpha_{k-2}}} \right\rangle$$

$$- \cdots.$$

More compactly,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = \sum_{0\le l\le k} (-1)^{k-l} \sum_{\alpha_1<\cdots<\alpha_l\in K} \left\langle y \middle| G_{i_{\alpha_1},\ldots,i_{\alpha_l}}^{s_{\alpha_1},\ldots,s_{\alpha_l}} \right\rangle.$$

What is important for us is that $e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)$ is a linear combination of the average phenotype of every possible sub-map of *G* defined by fixing the states at one or more of the *k* sites $i_1, \ldots, i_k$. Any such average phenotype can be decomposed by fixing the state at another site $j \neq i_1,\ldots,i_k$. For example,

$$\left\langle y \middle| G_{i_1}^{s_1} \right\rangle = \frac{1}{q} \sum_{t\in Q} \left\langle y \middle| G_{i_1,j}^{s_1,t} \right\rangle.$$

Therefore,

$$e_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = \frac{1}{q} \sum_{t\in Q} \sum_{0\le l\le k} (-1)^{k-l} \sum_{\alpha_1<\cdots<\alpha_l\in K} \left\langle y \middle| G_{i_{\alpha_1},\ldots,i_{\alpha_l},j}^{s_{\alpha_1},\ldots,s_{\alpha_l},t} \right\rangle$$

$$= \frac{1}{q} \sum_{t\in Q} e_{i_1,\ldots,i_k}(s_1,\ldots,s_k)|_{G_j^t}.$$

Chapter 3

Epistatic drift causes gradual decay of predictability in protein evolution

This work was published as "Yeonwoo Park, Brian P. H. Metzger, and Joseph W. Thornton, Epistatic drift causes gradual decay of predictability in protein evolution. Science *376, 823-830 (2022)."*

3.1 Summary

Epistatic interactions can make the outcomes of evolution unpredictable, but no comprehensive data are available on the extent, direction, rate, and consequences of changes in the effects of mutations as protein sequences evolve. Here we characterize the temporal dynamics of epistatic change by using deep mutational scanning to measure the functional effect of every possible amino-acid mutation in a phylogenetic series of reconstructed ancestral and extant proteins, using the steroid receptor DNA-binding domain as a model. Across a 700-million-year historical trajectory, the effects of most mutations became completely or partially decorrelated from their initial effects. Epistatic interactions caused windows of evolutionary accessibility for most mutations to open and close transiently, shaping the historical fate not only of the mutations that fixed during history but also the far greater number that never did. Most mutations' effects evolved under Brownian motion: gradual change without directional bias, at a rate that was largely constant across time but varied dramatically among mutations, indicating a neutral process caused by many weak interactions. Protein sequences therefore drift inexorably into contingency and unpredictability, but that process itself is statistically predictable, given sufficient phylogenetic and experimental data.

3.2 Introduction

A mutation's evolutionary fate depends on its phenotypic effects. If the effects are stable over time, knowledge of them in the present can help predict the future course of evolution and explain the causes of evolutionary change in the past. Epistatic interactions, however, may cause a mutation's effects to change over time and its evolutionary accessibility to become contingent on the particular sequence changes that preceded it during history (*7*, *55*).

Despite a recent tide of information about epistatic interactions within proteins, we lack a comprehensive understanding of changes in the effects of mutations (the set of all potential amino-acid changes) caused by interactions with substitutions (the subset of mutations that fix during evolution). What fraction of mutations change in their effects over evolutionary time, and how drastically? Do they change gradually or episodically, and at what rate? What are the consequences for evolutionary outcomes? Deep mutational scanning (DMS) experiments have detected epistasis among mutations within present-day proteins (*11–13*, *42–44*, *46*), but these studies do not address interactions with historical substitutions or reveal changes in mutations' effects over evolutionary time. Some mutations have different effects when introduced into various present-day proteins, implying epistatic interactions with the substitutions that occurred as these proteins diverged from each other (*16–18*, *20*, *21*), but without polarizing and calibrating these differences with respect to time, it is not possible to illuminate the rate, direction, or regularity of the process by which mutations' effects changed during evolution. Ancestral protein reconstruction studies have shown that the effects of particular mutations changed during particular phylogenetic intervals (*23*, *25–27*, *29*, *30*, *56*, *57*), but these works have examined only the beginning and end of an interval and therefore cannot reveal the temporal dynamics of epistasis.

Here we address this knowledge gap by using DMS to comprehensively assess the effect of introducing every possible amino-acid mutation into a series of reconstructed ancestral and extant proteins along a densely sampled phylogenetic trajectory. We used as a model the DNA-binding domain (DBD) of steroid hormone receptors, a family of essential transcription factors in bilaterian animals that mediate the actions of sex and adrenal steroids by binding to specific DNA sequences and regulating the expression of target genes (*58–60*). This approach allowed us to measure changes in the functional effect of every possible amino-acid mutation during a series of defined intervals across 700 million years of DBD evolution. To analyze these data, we developed a quantitative framework that treats each mutation's effect as a trait that evolves probabilistically on a phylogeny, which we used to characterize the temporal dynamics, evolutionary consequences, and underlying genetic architecture of epistatic interactions.

3.3 Results

3.3.1 Phylogenetic deep mutational scanning

We first inferred the phylogeny of steroid and related receptors (Fig. 3.1A and A1.1) and reconstructed the maximum a posteriori protein sequences of 7 ancestral DBDs: the ancient progenitor protein whose duplication and divergence gave rise to the first steroid receptor (AncNR3), the ancestor of all extant steroid receptors (AncSR, which existed in the ancestor of all bilaterians), and 5 descendants of AncSR along two lineages – one leading to human glucocorticoid receptor (GR) and the other to the steroid receptor of the annelid *Capitella teleta*, which are among the most diverged of all functionally characterized extant DBDs (Fig. 3.1B and A1.2). These 9 DBDs are separated by 8 phylogenetic intervals, each involving 3 to 42%

sequence divergence. We constructed a yeast strain carrying a GFP reporter driven by a DNA response element for these DBDs and confirmed that all reconstructed ancestral DBDs bind to it, as expected based on prior studies (*60*). GFP fluorescence in this strain correlates well with binding affinity previously measured using fluorescence anisotropy (Fig. A1.4D).

For each of the 7 ancestral and 2 extant DBDs, we generated a library of variants that contains all 19 possible amino-acid mutations at all 76 sites (Fig. A1.3). We used a bulk assay of fluorescence-activated cell sorting (FACS) coupled with deep sequencing to quantify the GFP fluorescence of each variant with very high repeatability (r^2 = 0.99 across 3 replicates; Fig. 3.1C, A1.4, and A1.5). We calculated the effect of a mutation as the difference in the mean $\log_{10}$-GFP fluorescence (ΔF) between variants that differ by a single amino acid; we applied this approach to all mutations from the wild-type amino acid in any of the 9 DBDs to all other 19 amino acids. Differences in the effect of a mutation between successive nodes on the phylogeny ($\Delta\Delta F$) indicate that the mutation interacts with historical substitutions that occurred during that interval (Fig. 3.1D). We normalized mutations' effects to remove global background-dependence caused by different wild-type activity levels (Fig. A1.6); after this correction, differences in a mutation's effect among the 9 DBDs are attributable to specific epistatic interactions with intervening substitutions on the phylogeny.

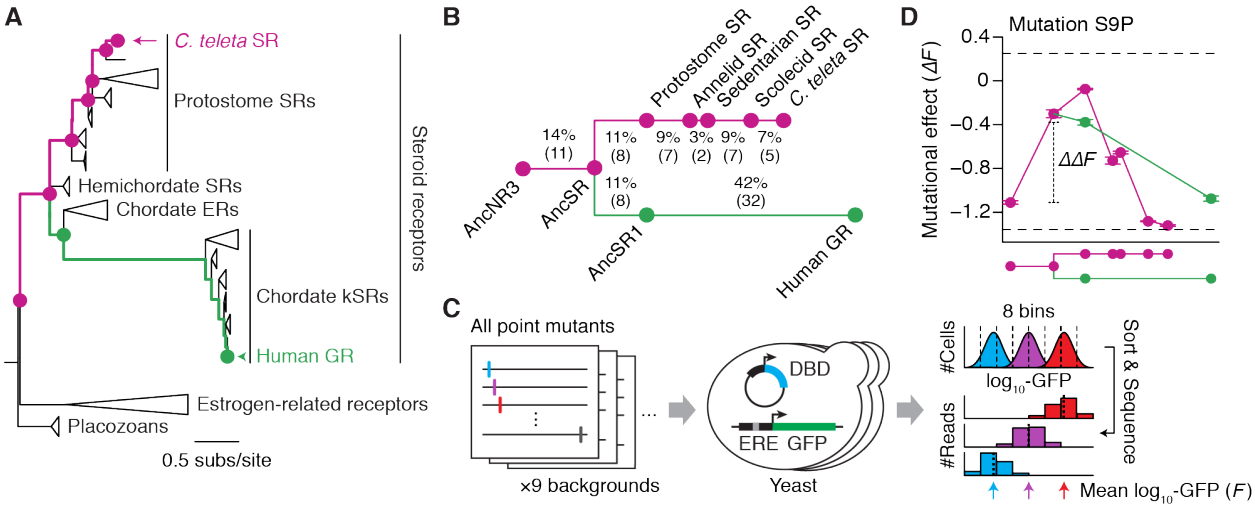

Figure 3.1. Phylogenetic deep mutational scanning. (**A**) Phylogeny of the DNA-binding domain (DBD) of steroid and related receptors. Circles, DBDs characterized here by deep mutational scanning. SRs, steroid receptors; ERs, estrogen receptors; kSRs, ketosteroid receptors—including glucocorticoid receptor (GR). Complete phylogeny in fig. S1. (**B**) Phylogenetic relations among the 9 characterized DBDs. Colors distinguish trajectories to *C. teleta* SR and human GR. Sequence divergence (percent) and number of sequence differences (parentheses) in each interval are shown. (**C**) Sort-seq assay for DBD activity. For each DBD, a library containing all possible single-amino acid mutations was generated using microarray-based synthesis and cassette assembly (fig. S3) and cloned into yeast carrying a GFP reporter; ERE, estrogen response element. Activity of each mutant was measured by sorting the library of cells into fluorescence bins, inferring the distribution of each mutant among bins by sequencing, and calculating the mean $\log_{10}$-GFP fluorescence (F). Hypothetical distributions for 3 variants with high, medium, and low F are shown. (**D**) Tracing epistatic change in mutational effect across the phylogeny using example mutation S9P. The effect on each DBD's activity (points) was quantified as the change in mean $\log_{10}$-GFP fluorescence (ΔF). Horizontal axis, each DBD in order on the phylogeny, positioned by sequence divergence and colored by trajectory. $\Delta\Delta F$, change in the mutations effect between a pair of DBDs, caused by epistatic interactions with intervening substitutions. Error bars, SEM ($n = 3$). Dashed lines, upper and lower measurement bounds.

3.3.2 Pervasive random changes in the effects of mutations

To analyze the evolutionary dynamics of epistasis over time, we adapted a classic quantitative framework for modeling trait evolution on phylogenies (*61*, *62*), including the extent, direction, and rate of evolutionary change of the trait, the underlying genetic architecture, and the relative roles of selection and genetic drift. Our approach treats the phenotypic effect of

each mutation as a trait that changes probabilistically across phylogenetic intervals, allowing us to ask these questions about epistatic change during historical DBD evolution.

Sixty percent of all mutations display significantly different effects among the 9 backgrounds, and 22% differ in the direction of their effects (FDR ≤ 0.1; Fig. 3.2A). Most of the mutations that show no evidence of epistasis destroy protein function regardless of genetic background (ΔF always at the lower bound of measurement, –1.3). Only 5% of mutations have a nondestructive effect that did not vary significantly across the phylogeny.

Epistatic changes occurred during all 8 phylogenetic intervals (Fig. 3.2B and A1.7A). Even in the shortest interval – during which there were only two sequence substitutions – the effects of more than 200 mutations changed significantly. During the other intervals, even more mutations changed in effect. On average, each substitution is associated with significant changes in the effects of about 60 mutations (Fig. A1.7B).

These epistatic changes were unbiased over time. Changes in the effects of mutations ($\Delta\Delta F$) are distributed almost symmetrically around 0 (mean = –0.01; Fig. 3.2C). The fraction of mutations that reduce activity was nearly constant among the 9 intervals, as was the fraction of mutations that destroy activity (Fig. 3.2D). No individual mutations had effects that changed with a significant bias in either direction over time (Fig. A1.7C). These data indicate that directional selection did not drive long-term epistatic changes in the effects of mutations, and mutational robustness did not change systematically over time. Further, the variance of the distribution of $\Delta\Delta F$ in each interval increased linearly with sequence divergence, rather than plateauing (Fig. 3.2, E and F), suggesting no role for stabilizing selection to maintain the effects of mutations within defined limits.

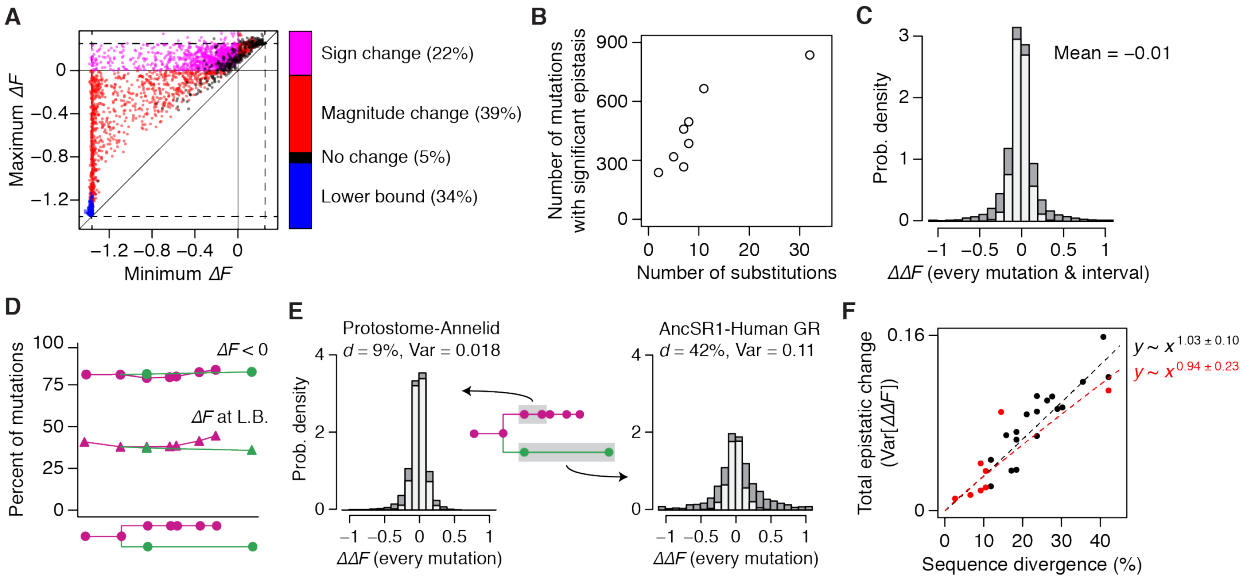

Figure 3.2. Pervasive random changes in the effects of mutations. (**A**) Maximum and minimum effect of each mutation (points) across the 9 DBDs, colored according to the stacked column at right, which shows the proportion of mutations in four categories: pink, significant effect of DBD background on ΔF and the sign of ΔF different between the maximum and minimum; red, significant effect of background but no sign difference; black, no significant effect of background and ΔF within measurement limits; blue, ΔF at the lower bound of measurement in all 9 DBDs. Significance was evaluated by Welch's ANOVA, Benjamini-Hochberg FDR ≤ 0.1. (**B**) Number of mutations in each phylogenetic interval that changed significantly in ΔF (t-test between parent and child node, FDR ≤ 0.1), plotted versus the number of amino acids that diverged in the interval. (**C**) Distribution of epistatic change in the effect of every mutation during every phylogenetic interval ($\Delta\Delta F$). Dark grey, $\Delta\Delta F$ significantly different from 0. Mutations always at the lower bound of measurement were excluded. (**D**) Fraction of mutations in each DBD with $\Delta F < 0$ (circles) or ΔF at the lower bound of measurement (triangles). (**E**) Distribution of $\Delta\Delta F$ of all mutations for the protostome-annelid interval or the AncSR1-human GR interval. The variance of the distribution (Var) quantifies the total epistatic change in the effects of all mutations during an interval. d, sequence divergence. (**F**) Total epistatic change as a function of sequence divergence across the phylogeny. Red dots, each of the 8 independent phylogenetic intervals between characterized DBDs; black, all composite intervals. Dashed lines, best-fit power function for all (black) or the 8 independent intervals (red).

3.3.3 The effects of most mutations drifted gradually

To test whether epistatic change was gradual or episodic, we fit probabilistic models of trait evolution to the trajectory of changes in the effect of each mutation. Brownian motion

represents a simple model of gradual evolution at a constant rate without directional bias: changes in the trait value among phylogenetic intervals are normally distributed when normalized for the length of the interval, with a mean change of zero and constant variance per unit sequence divergence (which represents the rate of evolution). In the alternative model of episodic evolution, the normalized variance is a free parameter for each interval, which allows the rate to differ among intervals (Fig. 3.3A). We fit both models to the 8 $\Delta\Delta F$ values of each mutation and used a likelihood-ratio test to compare the fit of the two models.

We found that the Brownian motion model was the best-fit model for 92% of mutations that changed epistatically (Fig. 3.3, B and C), irrespective of whether mutations' effects changed rapidly or slowly (Fig. 3.3D). For the 8% of mutations best fit by the episodic model, effects were nearly constant in most intervals with dramatic changes during one or a few intervals. The functional effects of most mutations therefore evolved as a random variable that changes gradually along the phylogeny at a characteristic rate and without bias. We call this process epistatic drift.

Phylogenetic cross-validation confirmed that the effects of most mutations evolved at a steady rate across the phylogeny (Fig. 3.3E). For each mutation, we predicted the epistatic change expected in each of the 8 intervals given the rate of epistatic change estimated from the 7 other intervals and then compared these predictions to experimental observations, pooling mutations with similar estimated rates (Fig. 3.3F). Predicted and observed epistatic changes were strongly correlated (Spearman's $\rho \geq 0.94$ for every interval; Fig. 3.3G), indicating that mutations' relative rates of epistatic change did not strongly vary along the phylogeny. The absolute rate of epistatic change, however, was systematically faster than predicted in some intervals and slower in others: the mean rate of epistatic change for all mutations in each interval ranges from 0.7 to

1.4 of the average across the phylogeny (Fig. 3.3H). Epistatic change in the effect of each mutation therefore varies stochastically across intervals (consistent with Brownian motion), but this variation is correlated among mutations; as a result, the total amount of epistatic change across all mutations is systematically greater in some intervals than others. This pattern is likely to arise because the total epistatic change depends on the particular substitutions that fixed during an interval, and some substitutions are more epistatic than others, interacting more strongly or with a larger number of mutations. The mean rate of epistatic change was not systematically different during intervals following gene duplications.

These observations have two major implications for evolution and the genetic architecture of epistatic interactions (Fig. 3.3I). First, epistatic interactions within the DBD are dense: most mutations' effects changed gradually because of weak interactions with many substitutions, and each substitution typically modified the effects of many mutations (Fig. 3.2B). If most epistatic changes were triggered by rare, large-effect modifiers, the distribution of $\Delta\Delta F$ would be enriched near zero and at extreme values, a pattern that we observed for only a small fraction of mutations. Most historical contingency is therefore the cumulative result of many small-effect epistatic modifications. Against this background of gradual epistatic drift, a few mutations occasionally undergo dramatic changes in their effects.

Second, some mutations are more epistatically sensitive than others, with effects that diverged more rapidly as substitutions accumulated. Conversely, some substitutions are more epistatic than others, changing the effects of more target mutations or causing changes of greater magnitude. As a result, there are systematic differences among intervals in the average rate of epistatic change across all mutations.

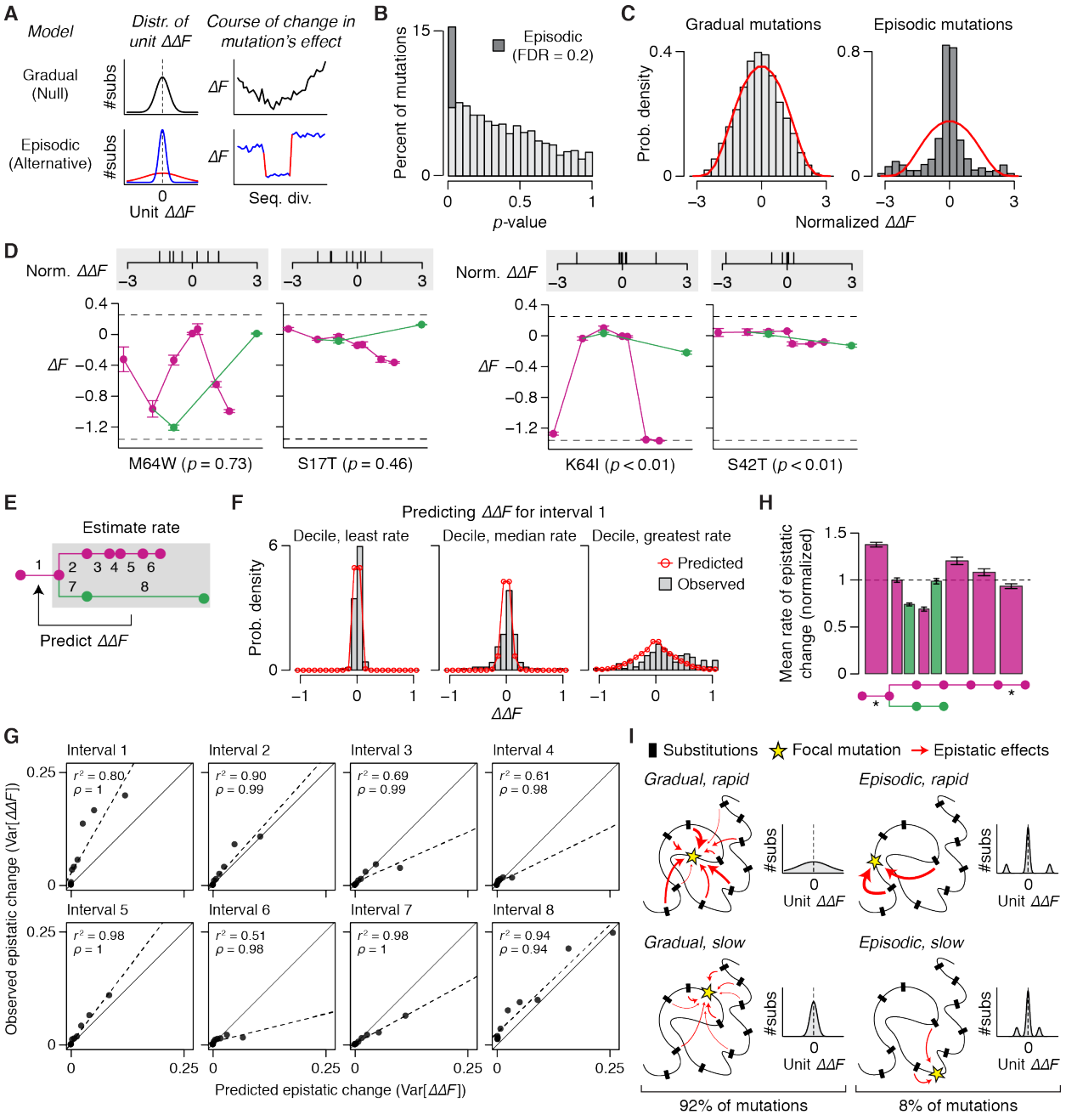

Figure 3.3. Effects of most mutations changed gradually at characteristic rates. (**A**) Models of the tempo of epistatic change. Null model, the amount of change in a mutation's effect per substitution in an interval (unit $\Delta\Delta F$) is randomly drawn from a normal distribution centered at 0; the variance is the same among intervals, so the mutation's effect changes gradually at a constant expected rate as substitutions accrue. Alternative model, the variance may differ among phylogenetic intervals (blue vs. red), leading to episodic changes in a mutation's effect. (**B**) Distribution of the *p*-value of the likelihood-ratio test (LRT) comparing gradual and episodic models for each mutation. Darker grey, mutations for which the gradual model is rejected (FDR ≤ 0.2). Mutations always at the lower bound of measurement were excluded from this analysis. (**C**) Distribution of the normalized amount of epistatic change in each interval, for all mutations

better fit by the gradual model (*left*) or the episodic model (*right*). Normalized $\Delta\Delta F$, $\Delta\Delta F$ of a mutation in an interval divided by $\sigma d^{1/2}$, where σ is that mutation's average rate of epistatic change and d is the length of the interval. Gray columns, observed data; red line, distribution expected under the null model. (**D)** Trajectory of changes in the effect of two example mutations that are better fit by the gradual model *(left*) or episodic model (*right*); in each category, one evolves rapidly and the other slowly. Each mutation's p-value in the LRT is shown; gray box, normalized changes in the mutation's effect across each of the 8 intervals. (**E**) Phylogenetic cross-validation. In the example shown, $\Delta\Delta F$ in interval 1 is predicted from the average rate of epistatic change measured across intervals 2-8 (grey box). (**F**) Distribution of observed $\Delta\Delta F$ during interval 1 (gray columns) and predicted by cross validation (red line). Mutations were grouped into deciles by their rate of epistatic change across intervals 2-8; predictions are shown for deciles with the slowest, median, or fastest rates. (**G**) Mutations' relative rates of epistatic change are consistent across phylogenetic intervals. Points, deciles of mutations grouped by the predicted rate of epistatic change; observed epistatic change in an plotted against that predicted by cross-validation. r, Pearson's correlation coefficient; ρ, Spearman's rank correlation; dashed line, linear regression. (**H**) Among-interval differences in average rate of epistatic change. Each column shows the mean rate of epistatic change of all mutations in one phylogenetic interval, normalized so that the mean across all intervals equals 1. Error bars, estimated standard deviation obtained by bootstrap-resampling of mutations. Asterisks, intervals immediately following gene duplication. **(I)** Inferring the architecture of epistatic interactions between substitutions (black boxes) and a focal mutation (star) from phylogenetic DMS. *Left,* gradual changes in the mutation's effect during evolution arise if many substitutions act as epistatic modifiers (arrows, with thickness showing the strength of interaction), yielding a normal distribution of $\Delta\Delta F$ per substitution. *Right,* episodic changes arise from interactions with only a few substitutions, yielding a distribution heavy at zero and the tails. In either case, strong vs. weak interactions cause rapid *(top)* vs. slow (*bottom)* epistatic change. The fraction of all mutations in each category in our experiments is shown.

3.3.4 Mutations vary in memory length and the timescale of contingency

Because the effect of each mutation drifts at random at a steady rate, there should be a characteristic time period after which the mutation's effect can no longer be reliably predicted from its known effect at some other time. We call this period the mutation's memory length, the measure of which is the memory half-life – the amount of sequence divergence over which the correlation of a mutation's effect is reduced by half. To estimate the memory half-life, we partitioned mutations into deciles by their rate of epistatic change and calculated for each decile how correlated the effects of mutations are between each pair of DBDs (Fig. 3.4, A to D). We modeled the correlation coefficient as an exponentially decaying function of sequence

divergence. We then used this relationship to estimate the memory half-life of each mutation from its rate of epistatic change.

Reflecting the wide variation in the rate of epistatic change among mutations, memory half-lives range from just 3% sequence divergence to virtually infinite (Fig. 3.4E). Mutations with the shortest half-lives therefore forget the effects they had in the past after just a few sequence substitutions at other sites: at any moment, their effect and likely fate depend primarily on the substitutions that occurred most recently during their history.

Relative to the timescale of DBD evolution, about one fourth of all mutations have short memories (half-life < 50% sequence divergence); in this group, the effects in present-day human GR are almost completely independent of their initial effects in AncSR ($r^2 = 0.10$, Fig. 3.4F). 20% of mutations have medium memory (half-life 50 to 200%), with present-day effects that can be partially predicted from their initial effects ($r^2 = 0.68$). The remaining 54% of mutations have long memories (>200% divergence) and interacted negligibly with historical substitutions, retaining their initial effects throughout DBD evolution ($r^2 = 0.98$).

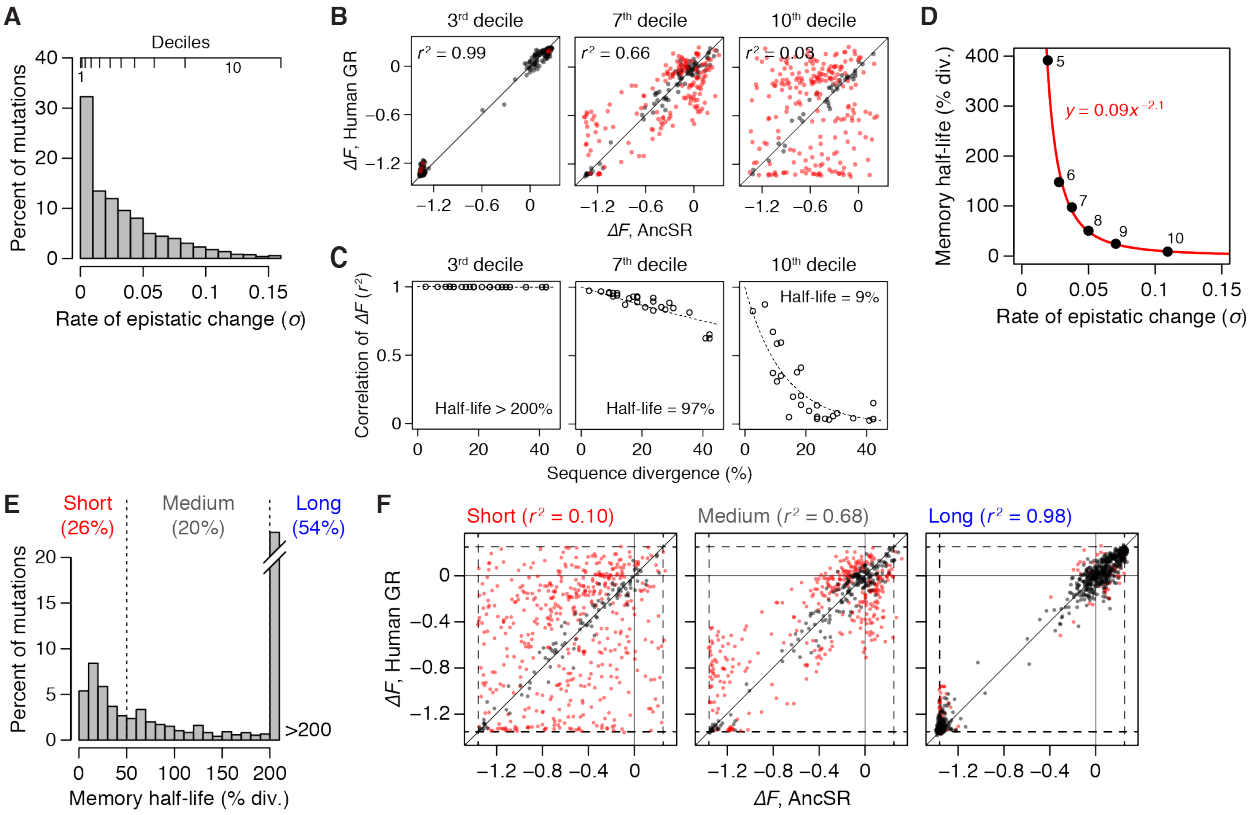

Figure 3.4. Memory length of mutations and the timescale of historical contingency. (**A-D**) Measuring the memory length of mutations. (**A**) Mutations were grouped into deciles by their rate of epistatic change (σ, expected standard deviation of $\Delta\Delta F$ per 1% sequence divergence). (**B**) The effects of mutations in each decile were compared between every pair of DBDs; shown are comparisons between AncSR and human GR (42% divergence). (**C**) The squared Pearson correlation coefficient (r^2) for each DBD pair was plotted against the sequence divergence of that pair. Dotted line, best-fit exponential decay curve; memory half-life, sequence divergence at which $r^2 = 0.5$. (**D**) Relationship between the rate of epistatic change and memory half-life inferred by fitting a power function (red) to the mean rate of epistatic change and memory half-life of the deciles. This relationship was used to calculate the memory half-life of each mutation from its rate of epistatic change. **E)** Distribution of memory half-life among mutations. Mutations were classified into short-, medium-, and long-memory categories using cutoffs of 50% and 200% divergence. (**F**) Comparing the effects of mutations between AncSR and human GR (42% divergence) for each memory category. Red dots, mutations with significant difference in ΔF (t-test, FDR ≤ 0.1); black, no significant difference.

3.3.5 Contingency of historical sequence evolution

We next focused on the subset of mutations that occurred during historical DBD evolution. We first assessed the functional effects of the 79 substitutions that occurred during the

phylogenetic intervals that we experimentally characterized (Fig. 3.5, A and B). When measured in the ancestral background in which they historically occurred, substitutions that reduce activity by $\Delta F < -0.2$ were nearly absent; of the few exceptions, most fixed during intervals immediately after gene duplication (Fig. A1.8A). This represents a 29-fold depletion compared to the set of all mutations, the majority of which have $\Delta F < -0.2$. These results imply that the DBD evolved primarily under purifying selection against mutations that strongly reduce activity, and they establish $\Delta F = -0.2$ as a boundary that roughly defines the evolutionary accessibility of mutations under purifying selection.

Epistasis shaped the fate of most historical substitutions, which occurred during limited windows when they were transiently accessible. Of all substitutions that fixed between AncSR and any extant steroid receptor on our phylogeny, 43% have short or medium memories (Fig. 3.5C). Among the short-memory substitutions, the majority were inaccessible in AncSR ($\Delta F_{\text{AncSR}} < -0.2$), implying that they became accessible in one or more descendant proteins because of permissive epistatic substitutions, which render otherwise deleterious mutations neutral or advantageous. The remaining short-memory substitutions were accessible in AncSR ($\Delta F_{\text{AncSR}} \geq -0.2$), but almost all of these became subsequently inaccessible because of restrictive substitutions, which render previously neutral or advantageous substitutions deleterious (fig. S8B). By contrast, 95% of long-memory substitutions were accessible in AncSR and remained so across the entire phylogeny (Fig. A1.8B). Medium-memory substitutions displayed an intermediate pattern. The evolutionary fate of long-memory substitutions could therefore have been reliably predicted from their initial effects, but the accessibility of substitutions with short or medium memory depended on other substitutions that occurred during history.

Epistasis also shaped the fate of the many mutations that did not become substitutions. Of all short-memory mutations that were accessible in AncSR, 90% became inaccessible in one or more descendant proteins, indicating that evolutionary paths to them were closed by restrictive substitutions (Fig. 3.5D). Conversely, 55% of the short-memory mutations that were inaccessible in AncSR subsequently became accessible because of permissive substitutions. Overall, two-thirds of short-memory mutations and one-third of medium-memory mutations changed in accessibility among the 9 DBDs we tested, with each category of mutations being accessible in 2.4 and 4.9 of the 9 DBDs on average (Fig. 3.5E).

These data indicate that epistatic interactions with the particular set of substitutions that occurred along the phylogeny contingently determined the evolutionary fate not only of the mutations that fixed historically because of permissive substitutions, but also of those that did not have the opportunity to fix because of restrictive substitutions. Studying only the sequence changes that occurred during evolution therefore underestimates the role of historical contingency: doing so cannot detect the many evolutionary roads that were closed off contingently, but which could have been taken if the trajectory of sequence changes at interacting sites had unfolded differently.

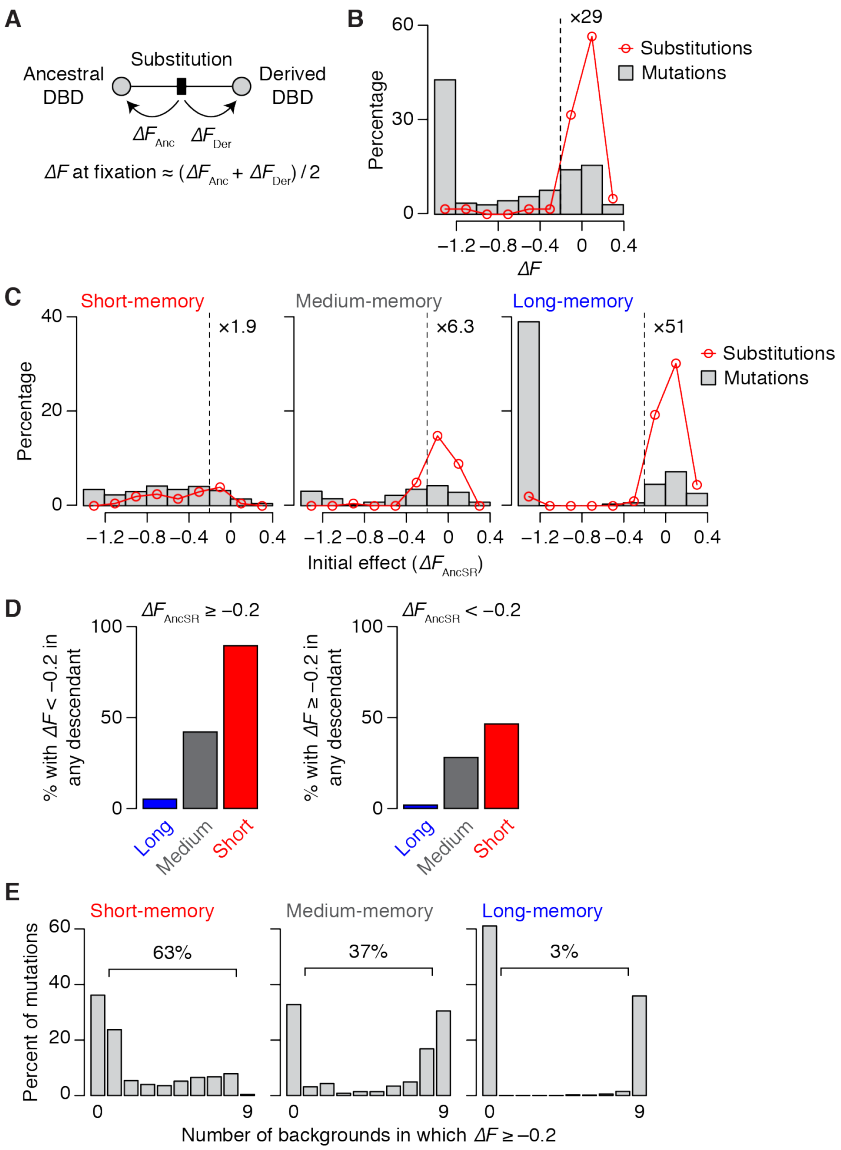

Figure 3.5. Impact on sequence evolution of memory length and initial functional effect. (**A**) The effect of a substitution at the time it fixed during history was calculated as the mean of Δ*F*s measured by DMS in the nearest ancestral and descendant nodes. (**B**) Comparing the effects of the 79 substitutions that occurred along the phylogenetic trajectories we characterized to the effects of all possible mutations. Substitutions are 29-fold enriched for $\Delta F \geq -0.2$ compared to mutations, providing an estimate of the threshold of accessibility during DBD evolution. (**C**) Distribution of the initial effect (Δ*F* on AncSR) of 275 substitutions that fixed between AncSR and any extant DBD in our phylogeny. Distributions are shown by memory half-life category. Enrichment of substitutions with $\Delta F \geq -0.2$ relative to mutations is shown. (**D**) *Left,* proportion of initially accessible mutations ($\Delta F_{\text{AncSR}} \geq -0.2$) that become inaccessible in at least one descendant DBD. *Right*, proportion of initially inaccessible mutations that become accessible in at least one descendant DBD. (**E**) Distribution of the number of characterized DBDs in which each mutation is accessible ($\Delta F \geq -0.2$), classified by memory-length category. The percentage of mutations that were accessible in some but not all DBDs is shown.

3.3.6 Causes of variation in memory length

Finally, we sought to identify the factors that determine a mutation's memory length. Some variation in memory length is attributable to the sequence site at which a mutation occurs: the median memory half-life of mutations to any of the 19 mutant states at the same site varies among sites from 11% to >200% divergence (Fig. 3.6A). But this variation is not associated with any obvious structural or functional properties: the median memory half-life of a site is poorly correlated with relative solvent accessibility, rate of substitution, rate of substitution at physically adjacent sites, distance to the DNA-binding residues, and distance to the dimerization interface ($r^2 < 0.1$ for every factor; Fig. A1.9). Further, memory length varies extensively within each site, with 59 of 76 sites in the DBD containing both short- and long-memory mutations. As a consequence, predicting a mutation's memory half-life by the median of all mutations at that site achieves r^2 of only 0.25 (Fig. 3.6B).

Another possibility is that certain types of mutations (to and from the same pair of states) might be consistently associated with memory length, irrespective of the sites at which they occur. But predicting the memory length of individual mutations from the median memory length of all mutations of the same type at any site achieved r^2 of only 0.13 (Fig. 3.6C). Explaining memory length variation therefore requires analysis of each particular mutation at each site in the protein.

Estimating a mutation's memory length requires experiments in multiple genetic backgrounds across a phylogenetic trajectory. But how many backgrounds are necessary? When the rate of epistatic change of mutations is estimated from 2 backgrounds randomly chosen from the 9 we assayed, the correlation with the rate measured using all 9 backgrounds is on average r^2

= 0.40, and the rate of epistatic change is systematically underestimated (Fig. 3.6, D and E). The correlation improves as more backgrounds are sampled and reaches 0.8 when estimates are based on 5 backgrounds. A moderate number of experiments is therefore sufficient to provide a rough estimate of a mutation's rate of epistatic change and hence its memory length.

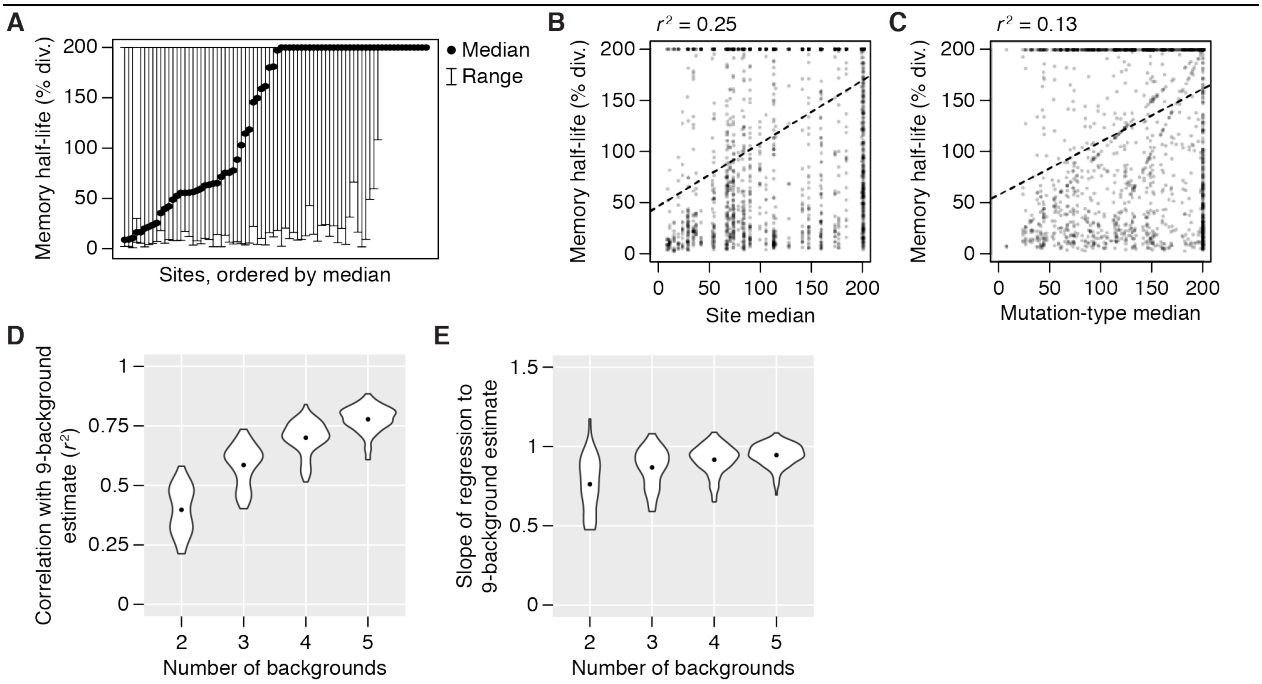

Figure 3.6. Variation of memory half-life of mutations among and within sites. (**A**) Distribution of memory half-life among sites. Each line shows the range of memory half-life of all mutations at one site in the DBD sequence. (**B**-**C**) Predicting the memory half-life of a mutation (points) by the median memory half-life of all possible mutations at the same site (**B**) or by the median of mutations of the same type (between the same wild type and mutant amino acid) at all sites (**C**). Dashed line, linear regression. (**D**-**E**) Effect of number of DBDs characterized by DMS on estimates of rate of epistatic change. The rate of epistatic change of every mutation was estimated using a subset of the 9 DMS experiments; the relationship between the estimated rate from each subset to that estimated from all 9 experiments was analyzed by linear regression. The graphs show the distribution of correlation coefficient **(D)** and best-fit regression slope (**E**) across every possible subset of a given size.

3.3.7 Robustness to uncertainty in ancestral sequence reconstruction

Our conclusions are robust to uncertainties in the reconstruction of ancestral sequences. The ancestral DBDs were generally inferred with high confidence and contained zero to 10 sites at which more than one amino acid state is plausible. For each ancestral DBD, we generated an "Alt-All" reconstruction, which contains the alternative plausible amino acid at all ambiguously reconstructed sites (*63*); this sequence represents the least likely of all plausible reconstructions and allows a conservative estimate of robustness to sequence uncertainty. We then constructed a complete DMS library of each Alt-All protein and repeated all of our experiments and analyses. Although the effects of some mutations differ between the Alt-All and maximum-posterior-probability reconstructions, all conclusions concerning the temporal dynamics of epistasis were unchanged (Fig. A1.10).

3.4 Discussion

Prior experimental studies have identified cases in which the functional effects of a few mutations changed dramatically during particular intervals of evolutionary history (*8*, *21*, *23–27*). Our observations show that such rare, large-effect epistatic modifications occur against a background of pervasive gradual drift in the effects of the majority of mutations (*20*, *29*, *64–69*). Most epistatic changes across DBD history were of small magnitude when they occurred, but across an evolutionary trajectory of moderate length (<50% sequence divergence), they were sufficient to completely or partially decorrelate the effects of the majority of mutations from their initial effects and dramatically alter the set of available opportunities for future sequence change. Because the fold and function of all proteins depend on interactions among many residues, we expect that epistatic drift will be a widespread feature of protein evolution, but the temporal

dynamics and distribution of memory lengths may depend on each protein's structural architecture, function, and the selective regime under which it evolved.

Our findings establish strong limits on the ability to predict future evolution and interpret evolutionary history, but they also provide a quantitative framework for understanding those limits. Classical evolutionary theories assume that the constraints imposed by purifying selection do not change as sequences diverge, so the effects and evolutionary fate of mutations can be predicted or retroactively inferred based on their effects measured in the present. Our results show that this assumption of constancy and independence is wrong for about half of DBD mutations, which have short or medium memories. Because epistatic modification occurs at a mostly constant rate for each mutation, however, an estimate of memory length from experimental data across phylogenetic time can quantify the extent to which any mutation's effect can be predicted at any point in time, either future or past. Further, although point projections of the effects of short- and medium-memory mutations across long timescales are unreliable, a probability distribution of those effects can be generated if we know any mutation's memory length and its effect at some other time. Ancestral protein reconstruction can replace predictions with experimental knowledge, but only for proteins in the past.

A probabilistic description of contingency and uncertainty using memory length does not require detailed knowledge of the particular genetic interactions that cause epistatic change. If we had microscopic knowledge of all the interactions that modify each mutation's effect and a dense phylogenetic reconstruction of past trajectories of sequence change, we could reliably predict the effect of every possible mutation in any genetic background. But even complete knowledge like this would not be sufficient to predict future evolutionary trajectories: the accessibility of each future mutation depends on the chain of epistatic substitutions that occur

before it, many of which will occur by chance. We can use experimental and phylogenetic data to tame evolutionary uncertainty by recognizing and quantifying it, but the future will always surprise.

3.5 Methods

3.5.1 Phylogenetics and ancestral sequence reconstruction

We focused on reconstructing the trajectories leading from AncNR3 to *C. teleta* SR and human GR because they are among the most diverged of functionally characterized extant SRs and can therefore let us trace epistatic changes over a large span of sequence divergence. We expanded a previously inferred phylogeny of steroid and related receptors (*5*, *70*) by sampling additional sequences from protostomes to break the long branch leading to *C. teleta* SR. To sample from species without high-quality genome assemblies, we *de novo* assembled transcripts from transcriptome data available in the NCBI SRA database. To selectively assemble SR transcripts, we used TBLASTN to extract RNA-seq reads with similarity to *C. teleta* SR; reads were then assembled using Velvet/Oases (*71*, *72*), and protein-coding regions were identified using TransDecoder (*73*).

DBD amino-acid sequences were aligned using MUSCLE (*74*). DBD sequences are short and therefore contain limited signal of phylogenetic relationship, so we used a constrained maximum likelihood (ML) strategy. We imposed as a constraint the topology previously inferred from full-length receptor sequences (*5*, *70*) and inferred the placement of the newly sampled sequences on this topology by ML using RAxML (v 8.2.11) (*75*). All branch lengths and substitution parameters were optimized using the best-fit model of sequence evolution under the Akaike information criterion, which was the LG substitution matrix with ML amino acid

frequencies and a four-category gamma-distributed among-site rate variation (LG+X+G). Two modifications were made to the resulting constrained ML phylogeny to reflect established taxonomic relationships: 1) Annelid and mollusk SRs were grouped together to the exclusion of priapulid SRs; 2) *Magelona berkeleyi* and *M. pitelkai* SRs were grouped together as the basal-branching annelid SRs, in accordance with phylogenomic analyses (*76*). Branch lengths and model parameters were then reoptimized, and the resulting phylogeny (shown in Fig. A1.1) was used for ancestral sequence reconstruction.

The maximum a posteriori (MAP) ancestral sequences and the distribution of posterior probability over states at each site were inferred using PAML (v 4.8a) (*77*). The 7 ancestral nodes were well-reconstructed, with the posterior probability of the MAP state averaged across all sites ranging from 0.90 to 1.00 among the 7 nodes (Fig. A1.2). A site was considered ambiguously reconstructed if an alternative state has posterior probability (PP) greater than one-fourth the PP of the MAP state. Of the 7 ancestral DBDs, one was ambiguously reconstructed at zero sites and one contained a single ambiguous site; the others contained 2 to 10 ambiguous sites. To test whether our findings are robust to statistical uncertainty in the reconstruction of ancestral proteins, we generated the "Alt-All" reconstruction for each of the 5 nodes with multiple ambiguous sites. The Alt-All sequence contains the MAP state at all unambiguously reconstructed sites and the alternative plausible amino acid at every ambiguous site; among the ensemble of plausible reconstructions, Alt-All is the most distant from the MAP sequence and therefore allows a conservative test of robustness of functional inference to statistical uncertainty about the ancestral sequence (*63*).

We generated libraries containing every possible amino-acid mutation at every site in the DBD, using each of the reconstructed MAP and Alt-All proteins, *C. teleta* SR, and human GR as

genetic background. All of these proteins bind to an inverted palindrome of AGGTCA (estrogen response element, ERE) except for human GR, which binds to AGAACA (*78*). To investigate how historical substitutions during long-term DBD evolution changed the effect of every point mutation on a single function, we used a version of human GR that contains 3 mutations in its recognition helix that revert these sites to their ancestral states, yielding a protein (GR-3rh) that binds to ERE (*60*).

3.5.2 Design of comprehensive point mutant libraries

We developed a specific and cost-efficient method for generating all point mutants for multiple genetic backgrounds (Fig. A1.3). DBD is 76 amino acids in length, so the library for each background contains 76 × 19 = 1,444 mutants. To produce libraries containing all desired point mutants with low error rate, we divided the DBD coding sequence into four cassettes of approximately equal length. We then used array synthesis to produce an oligonucleotide pool containing all point mutants for each cassette for each protein. Each cassette pool was assembled together with a sequence-verified wild-type version of the three other cassettes for that protein. Compared to array-synthesizing full-length coding sequences, the cassette strategy reduces the error rate by a factor of four and increases the expected fraction of error-free sequences from approximately 10% to 60%.

To reduce synthesis cost, we used degenerate oligonucleotides. At each site, all 19 amino-acid mutants were encoded using the following 8 codons: NAG, NAT, NCG, NCT, NGG, NGT, NTG, and NTT. This set of codons is equivalent to NNK encoding. Compared to uniquely encoding each amino acid with a separate oligonucleotide, this strategy reduces the number of

oligonucleotides for synthesis by more than half and allows us to synthesize oligonucleotides required for all backgrounds in one standard 12,472-feature microarray (CustomArray).

To assemble each mutant oligonucleotide with the appropriate wild-type sequence, each cassette in each DBD was synthesized with a unique pair of primer-binding sites. This allowed us to specifically amplify the variants of any desired cassette in any DBD background out of the total pool of synthesized oligonucleotides. Cassettes and wild-type sequences were then assembled using type IIS restriction-ligation: each fragment was synthesized to contain recognition sequences for particular type IIS restriction enzymes, which cut DNA outside the recognition site and thereby allow a seamless coding sequence to be recovered upon ligation.

To reduce the probability that variants are misassigned because of post-assay sequencing errors, we added synonymous differences to differentiate similar variants from each other, using a strategy we developed to do so as parsimoniously as possible. This "color barcoding" strategy increases the Hamming distance h between all mutants to ≥ 2, which makes it impossible for a single sequencing error to cause misassignment. Our procedure first generates a graph in which each mutant is a node and two mutants are connected if their Hamming distance $h = 1$. Colors are then assigned to the nodes in the graph, using as few colors as possible so that no connected nodes share the same color; mutants of the same color are separated by $h > 1$ and need not be further differentiated. Synonymous differences are then introduced into all nodes that share the same color; this differentiates nodes of different colors from each other without adding distance between nodes that already have $h > 1$. To further prevent misassignment, we also redundantly synthesized every mutant by identifying two residues that contain a four-fold degenerate amino acid (A, G, L, P, R, S, T, or V) and encoding that mutant using all 16 synonymous codons at these sites.

After assembly, the libraries of mutant DBDs were cloned into the expression vectors described below and transformed into NEB 5-alpha Electrocompetent *E. coli*. We obtained $>10^6$ unique transformants per DBD, equivalent to >600-fold coverage of the library.

We used deep sequencing to evaluate the quality of the assembled libraries (Fig. A1.3, C to E). We found that every desired mutant was represented for every genetic background. The read counts of mutants were distributed tightly around the median (95% of mutants with a read count within 7-fold of the median). 53% of reads were free of error. As expected, errors occurred predominantly at the microarray-derived sites rather than within the wild-type cassettes.

3.5.3 High-throughput functional characterization using sort-seq

3.5.3.1 Improved fluorescence reporter assay for steroid receptor DBD activity

We improved a previously developed fluorescence reporter assay (*14*), in which a galactose-inducible DBD expression vector is transformed into yeast cells carrying an ERE-driven genomically integrated GFP (Fig. A1.4A). The first improvement was to increase the assay's precision by boosting signal and reducing noise caused by variation in the DBD expression level. Constitutive expression of DBD is cytotoxic, necessitating conditional induction, but the original version of the assay used a yeast strain with a defective galactose induction pathway, resulting in slow, weak, and noisy expression of DBD. We engineered a new yeast strain from a recently isolated, minimally modified *S. cerevisiae* strain (YPS1000 MATa ho::KMX) (*79*), which displays a robust galactose response. We replaced the *KMX* gene in the *HO* locus with a cassette containing a yeast-enhanced GFP under control of a minimal *CYC1* promoter containing four repeats of ERE (AGGTCAcagTGACCT) and the hygromycin

resistance gene (*HYG*) for selection. Correct genomic integration was verified by Sanger sequencing.

The second improvement was to account for the fact that yeast cells frequently lose the DBD expression vector. Although derived from the pRS413 plasmid containing a centromeric sequence, we found that even under antibiotic selection 30-50% of cells in a culture lack the expression vector. We therefore engineered an expression vector that allows us to isolate plasmid-carrying cells from those that lost it. Using pRS413 as backbone, we used the galactose-inducible bidirectional promoter pGAL1/GAL10 to simultaneously express DBD and mCherry. The DBD and mCherry expression levels are tightly correlated (Fig. A1.4B), allowing the isolation of plasmid-carrying cells based on mCherry fluorescence. As in the original version of the assay, DBD is expressed as a C-terminal fusion to the SV40 nuclear localization peptide and Gal4 transcriptional activation domain. The expression vector is maintained under G418 selection. The expression vector, named pDBD2, was assembled using Gibson assembly and verified by Sanger sequencing.

To validate this new assay, we cloned 10 previously characterized DBD variants into pDBD2 and transformed each construct into the new reporter strain using the lithium acetate/single-stranded carrier DNA/polyethylene glycol method. Single transformant colonies were inoculated in 3 mL YPD supplemented with 200 μg/mL G418 (YPD+G418). After overnight growth at 30°C, 225 rpm, cells were back-diluted to 0.25 OD_{600} in 3 mL YPGal supplemented with 200 μg/mL G418 (YPGal+G418). After 6 hr of incubation at 30°C, 225 rpm, 0.5 mL culture was pelleted, washed with PBS, and resuspended in 0.5 mL PBS. Flow cytometry was performed using BD LSR Fortessa 4-15. Singlet cells within a conservative size range were isolated by gating on FSC-A/SSC-A and FSC-H/FSC-A plots. mCherry-positive cells were

isolated under 488 nm excitation and 640 nm emission. GFP fluorescence of these cells was measured under 488 nm excitation and 561 nm emission, taking the mean $\log_{10}$-fluorescence of 10,000 cells as readout. Segmented linear regression of mean fluorescence to ERE affinity previously measured by fluorescence anisotropy using purified DBD (*39*) shows a high correlation (Fig. A1.4D).

3.5.3.2 Large-scale yeast transformation

High-throughput functional characterization was performed for 20 DBD libraries, of which 14 are included in this study. The 20 libraries were pooled in equimolar ratio and transformed into the yeast reporter strain following a standard yeast electroporation protocol. A series of 7 transformations yielded 5.8×10^6 unique transformants, corresponding to >100-fold coverage of the libraries. 200 OD_{600}·mL of transformed cells (~10^9 cells) were harvested in 1.5 mL of 25% glycerol, flash-frozen in liquid N_2, and stored at –80°C.

Sort-seq requires each cell to express exactly one DBD variant. It was previously shown that the approximately dozen cell divisions between transformation and sorting are sufficient to resolve multiple transformants for a pRS413-based expression vector (*14*).

3.5.3.3 Fluorescence-activated cell sorting

The 200 OD_{600}·mL glycerol stock of transformed cells was thawed on ice, inoculated in 200 mL YPD supplemented with 25 μg/mL chloramphenicol (YPD+chlor), and recovered for 2 hr at 30°C, 225 rpm. Serial dilution and plating of an aliquot showed that ~4×10^7 cells were alive after recovery, well-above the number of unique transformants (5.8×10^6). 200 μg/mL G418 was added after recovery and cells were further grown for 17 hr at 30°C, 225 rpm; 50

OD_{600}·mL of cells were washed in PBS and transferred to 200 mL YPGal+G418+chlor and grown for 6 hr. All cells were then pelleted, washed with PBS, and resuspended in 20 mL PBS for sorting.

We performed 3 replicate sort-seq experiments. For each experiment, two 200 OD_{600}·mL glycerol stocks were prepared as above, 3 hr apart. Cells from the first stock were sorted for 3 hr and then cells from the second stock were sorted for 3 hr. Sorting was performed using BD FACSAria II. Singlet, mCherry-positive cells were isolated by gating as described above and sorted into equal-width bins spanning the range of $\log_{10}$-GFP fluorescence (488 nm excitation and 561 nm emission, divided by FSC-A to normalize for cell size). We used 12 bins for the first replicate but found that using 6 bins would have yielded essentially identical results; 8 bins were used for the second and third replicates. Table S2 lists the number of cells sorted in each sort-seq experiment. After 3 hr of sorting, cells were seeded in YPD+G418+chlor at 10^6 cells per 10 mL and grown overnight at room temperature, 225 rpm, until OD_{600} ~3. Keeping the cells at this density (logarithmic growth phase) ensures the relative proportion of mutants in the overnight culture to mirror those immediately after sorting, which is necessary for accurately inferring the mean fluorescence by deep sequencing. 5 OD_{600}·mL of overnight culture per 10^6 seeded cells was pelleted; plasmids were extracted using the protocol of ref. (*80*), using one column for every 10 OD_{600}·mL of culture. Quantitative PCR showed that plasmid extraction was almost perfectly efficient, yielding ~10^8 copies per 10 OD_{600}·mL culture.

3.5.3.4 Deep sequencing

We used Illumina NextSeq 2 × 150 nt paired-end sequencing for single coverage of the DBD coding sequence. A critical determinant of Illumina sequencing quality is nucleotide

diversity. When reads have very similar sequences, difficulties in cluster identification lead to suboptimal data-quality and low yield. To increase nucleotide diversity, we generated 8 types of amplicons. The sequenced portion of each amplicon begins with 3, 4, 5, or 6 degenerate nucleotides (N) and ends with 7, 6, 5, or 4 Ns (total 10 Ns per amplicon). This ensures that both the forward and reverse reads begin with maximal nucleotide diversity and the internal sites, due to frameshift, to also exhibit increased diversity. Furthermore, amplicons were oriented so that sequencing proceeds from the N- to C-terminus of the DBD for 4 amplicons types and the reverse direction for the other four. We also followed Illumina's recommendation to spike in PhiX library at 25% final concentration. Overall, these strategies resulted in excellent sequencing quality: Over 95% of clusters passed the quality filter, yielding >450 million reads.

We performed one NextSeq sequencing reaction for each sort-seq experiment. Sort bins were labeled using 6-nt barcodes. Amplicons were generated through two-step PCR, first appending the staggered Ns and sort-bin barcodes and next appending the adaptors required for Illumina sequencing. First-round PCR was conducted in 8 reactions per sort bin, corresponding to the 8 amplicon types described above. Each reaction was scaled to a volume of 10 μL per 10^6 sorted cells; 2 μL of the extracted plasmids were used per 10 μL total volume. NEB Q5 High-Fidelity DNA polymerase was used for 15 cycles of amplification at standard conditions (except for 60°C annealing temperature). The 8 PCR products for each bin were pooled and purified using Zymo Research DNA Clean & Concentrator. Second-round PCR was scaled to a volume of 25 μL per 10^6 sorted cells, using half the first-round PCR product as template; 15 cycles of amplification were performed at 68°C annealing temperature. After purification, amplicons for each bin were quantified using the Qubit Fluorometer (Invitrogen), pooled at concentrations proportional to the number of cells in each bin, and gel-purified using Qiagen MinElute Gel

Extraction kit. We followed Illumina's standard protocol for preparing 1.8 pM library for the NextSeq 550 system, spiking-in PhiX Control v3 at 25% final concentration.

3.5.3.5 Mapping sequence reads to mutants

Sequence reads derive from known mutant sequences, possibly with errors. We developed an algorithm to accurately map sequence reads to mutants. First, reads of correct length were compared to every mutant in the library and mapped to the mutant with the least Hamming distance. Reads mapping equally to more than one mutant were discarded. Sequence differences between a read and the mapped mutant indicate errors.

Reads with sequencing errors can be retained, but those with synthesis errors should be discarded or remapped. Sequencing errors occur independently for each read, whereas synthesis errors recur among all reads deriving from the same erroneously synthesized mutant. Because the number of reads we obtained ($\sim 10^8$) exceeds the number of unique transformants in our library ($\sim 10^7$), any synthesis error is expected to recur in multiple reads. We therefore considered errors that occur more than once among all reads mapped to the same mutant as synthesis errors. Reads containing nonsynonymous synthesis errors were discarded, as well as reads with more than three errors of any kind. 57% of amplicon reads were retained after this procedure, which is close to the expected fraction of error-free sequences given the per-nt error rate of the synthesis procedure.

3.5.3.6 Estimating mean fluorescence from sequence reads

To infer the mean fluorescence of cells containing each variant DBD, we used a simple and accurate nonparametric approach (fig. S4F). This strategy reflects the fact that the shape of the

distribution of cells across fluorescence bins varies widely among different variants (fig. S4C), making it difficult to find a family of parametric distributions that can be fit to all mutants. We first calculated $c_{m,b}$, the number of cells expressing mutant m sorted into bin b. $c_{m,b}$ can be estimated from $r_{m,b}$, the read count for mutant m in bin b:

$$(1)\ c_{m,b} = \frac{r_{m,b}}{\sum_m r_{m,b}} \times c_b,$$

where c_b is the total number of cells sorted into bin b (recorded during sorting) and Σ_m denotes summation over all mutants. The mean $\log_{10}$-fluorescence of the mutant (F_m) is given by

$$(2)\ F_m = \frac{\sum_b c_{m,b} \times \varphi_{m,b}}{\sum_b c_{m,b}},$$

where Σ_b denotes summation over all bins, and $\varphi_{m,b}$ is the mean $\log_{10}$-fluorescence of all mutant m cells sorted into bin b. $\varphi_{m,b}$ cannot be directly measured by sort-seq, so we approximate $\varphi_{m,b}$ by φ_b, the mean $\log_{10}$-fluorescence of all cells of any genotype sorted into bin b:

$$(3)\ F_m \approx \frac{\sum_b c_{m,b} \times \varphi_b}{\sum_b c_{m,b}}.$$

Because our sort bins are sufficiently narrow, $\varphi_{m,b}$ is close to φ_b for all mutants (Fig. A1.4G), making this approximation highly accurate.

3.5.3.7 Data cleaning

Since our library is generated by an equivalent of NNK encoding, each amino acid mutant is represented by up to three synonymous codon mutants (Fig. A1.5A). In addition, each codon mutant is represented by 16 synonymous variants at two four-fold degenerate sites (called barcode variants). We averaged the mean $\log_{10}$-fluorescence (F) of barcode variants into one value for each codon mutant, and averaged the mean fluorescence of synonymous codon mutants into one value for each amino acid mutant. Before this averaging, we removed outliers among barcode variants. For each barcode variant, we calculated the difference between its F and the average F of the 15 other barcode variants for the same codon mutant (δ). Because the combined read count of 15 variants is generally much greater than that of a single variant, δ approximates the deviation of measured F from true F due to sampling noise. Our goal is to identify and remove outliers with unusually large sampling noise. We grouped δ by the read count of the corresponding barcode variant, obtaining a distribution of δ for each read count (fig. S5, B and C). Outliers were detected as extreme values in this distribution, defined using the interquartile range: Given Q_1 and Q_3 (first and third quartiles) and a constant k, values outside the interval $[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$ are considered outliers. We chose the value of k that minimizes the standard deviation of mean fluorescence among synonymous codon mutants (Fig. A1.5D).

Mutants generally had higher fluorescence in the second and third sort-seq replicates than in the first (fig. S5E). This is due to variations in experimental conditions among replicates, such as subtle differences in cell growth and induction conditions. Because our goal is to compare the effects of mutations across different genetic backgrounds under the same experimental conditions, we removed these systematic differences before combining the replicates. We inferred a monotonic spline that relates the mean fluorescence between replicates, implemented as a nonnegative linear combination of I-splines using the R package splines2 (*81*). A 5-

parameter cubic spline with external boundaries at the data minimum and maximum and a single internal boundary at the midpoint was sufficient to capture the systematic differences, with more complex splines resulting in essentially the same fit. The mean fluorescence of the second and third replicates was transformed by applying the inferred spline.

3.5.4 Removing nonspecific epistasis

When the effect of a mutation is compared across genetic backgrounds with different basal levels of GFP fluorescence, the mutation's effect on GFP fluorescence can vary among the backgrounds even if its biochemical effects—such as effects on DNA affinity and protein stability—do not. This is because GFP fluorescence is a nonlinear function of the underlying biochemical parameters. This nonspecific form of epistasis arises whenever the genetic backgrounds vary in GFP fluorescence, regardless of the mutations involved (*7*, *47*). Furthermore, it is conditioned by the particular experimental setups, such as the DBD expression level and the dynamic range of measurement imposed by the particular yeast strain, reporter gene, and instrument used. We modeled and removed this form of epistasis to isolate the specific epistatic interactions between mutations and historical substitutions as follows.

In our yeast reporter assay, steroid receptor DBD drives GFP expression by binding to upstream EREs. Cellular GFP level increases with the fraction of upstream EREs bound by DBD (f_B). Under a simple binding equilibrium, f_B is related to the DBD-ERE binding constant (K) and cellular DBD concentration ([P]) as:

$$(4)\ f_B = K[P] / (1 + K[P]).$$

Our assay measures the mean $\log_{10}$-GFP fluorescence (F), which is related to f_B by complex cellular and measurement processes. Eq. (4) implies that F is a function of the product $K[P]$, not of K or $[P]$ individually. We thus write $F = g(K[P])$, where g is an unknown, monotonically increasing function. The biochemical effect of a mutation can be quantified as fold-change in $K[P]$. The effect of a mutation on GFP fluorescence (ΔF) can therefore be written:

$$(5)\ \Delta F = g(\alpha K[P]) - g(K[P]),$$

where α quantifies the mutation's biochemical effect. Eq. (5) shows that even if α is independent of genetic background, ΔF can vary among DBDs with different basal activities ($K[P]$). Thus, not accounting for activity differences among wild-type DBDs can cause nonspecific epistasis that affects all mutations.

We removed nonspecific epistasis in two ways. First, we minimized differences in reporter activity among the 14 wild-type DBDs by modulating their expression level using variant promoters, terminators, and ribosome-binding sites (Fig. A1.6, B and C; sequences of expression vectors provided in a supplementary Excel file). In this way, we normalized all wild-type DBDs to activate approximately the same level of GFP ($F_1 = \ldots = F_n$, or $K_1[P_1] = \ldots = K_n[P_n]$).

Second, we developed an analytic method for removing nonspecific epistasis, because our experimental adjustments reduced but did not eliminate all activity differences among the wild-type DBDs; furthermore, some mutations occur at sites that were substituted between DBDs, and we need a method to account for the difference in starting sequence state between backgrounds.

We developed a method to measure and then account for differences in the effect of a mutation caused by differences in the activity of wild-type proteins (Fig. A1.6, D to F). Let W_1 and W_2 denote the GFP fluorescence of two wild-type proteins and F_1 and F_2 the GFP fluorescence of two mutants created by introducing the same mutation into the two backgrounds. Given W_1, F_1, and W_2, what is F_2? As shown above, F can be written as $g(K[P])$, where g is an unknown, monotonically increasing function. We assumed the following formula for g:

$$(6)\ F = F_L + (F_U - F_L) \times K^n[P]^n / (1 + K^n[P]^n),$$

where F_L and F_U are the lower and upper bounds of measurement, respectively, and n is an unknown constant to be estimated from data. We can use Eq. (6) to derive a formula for F_2 in terms of W_1, F_1, and W_2. To simplify the formula, we define normalized fluorescence, $F^* = (F - F_L) / (F_U - F_L)$, which ranges from 0 (when $F = F_L$) to 1 (when $F = F_U$). We then obtain:

$$(7)\ F_2^* = \beta F_1^{*n} / (1 + [\beta - 1]F_1^{*n}), \text{ where } \beta = (W_2^* / W_1^{*n}) \times ([1 - W_1^{*n}] / [1 - W_2^*]).$$

To validate Eq. (7), we generated two deep mutational scan datasets for a single DBD, one by expressing it using the weakest of the DBD expression vectors used for the 14 genetic backgrounds and the other by using the strongest. We fit Eq. (7) to the resulting dataset—1,444 pairs of F_1 and F_2—using the observed values of W_1 and W_2 and the best-fit values of F_L, F_U, and n determined by orthogonal regression. With just 3 free parameters, Eq. (7) accurately describes the relationship between F_1 and F_2 across all mutants (Fig. A1.6E). We set $W_2 = -0.79$ (the average fluorescence of the 14 genetic backgrounds analyzed in this study) and obtained F_2 for

each mutation using the measured fluorescence as W_1 and F_1. In effect, every mutation was characterized at an expression level where the $\log_{10}$-GFP fluorescence of the wild-type DBD is –0.79.

3.5.5 Defining mutations and their effects

There are 149 wild-type states (at the 76 DBD sites) in the 9 experimentally characterized DBDs. We estimated the effect of all possible amino acid mutations that could have occurred along the phylogenetic trajectory: this set of mutations consists of 149 × 19 = 2,831 mutations from each of the 149 wild-type states to each of the 19 other amino acids. We quantified the effect of a mutation as $\log_{10}$-fold change in GFP fluorescence (ΔF). We did not analyze mutations for which the standard error of observed ΔF among the 3 sort-seq replicates is greater than 0.1 (Fig. A1.5F).

3.5.6 Probabilistic models for the temporal dynamics of epistasis

3.5.6.1 Likelihood-ratio test for the tempo of epistatic change

We used probabilistic models to test whether epistatic change is gradual or episodic (Fig. 3.3, A to D). Let $\Delta\Delta F_{ik}$ denote the amount of change in the effect of mutation i across phylogenetic interval k. The null model posits that $\Delta\Delta F_{ik}$ is a random sample from a normal distribution with a mean of 0 and a variance that is proportional to the amount of sequence divergence across the interval (d_k):

(8) $\Delta\Delta F_{ik} \sim N(0, \sigma_i^2 d_k) + \varepsilon_{ik}$,

where σ_i^2 is a constant representing the mutation's rate of epistatic change and ε_{ik} stands for measurement noise. Under this model, the expected epistatic change (Var[$\Delta\Delta F_{ik}$]) increases linearly with sequence divergence at a rate of σ_i^2 that is constant across the phylogeny but can vary among mutations. This model corresponds to the Brownian motion model of trait evolution, which describes a trait evolving gradually at a constant rate and randomly without any bias or constraint.

The alternative model we used allows the rate of epistatic change to vary not only among mutations but also among phylogenetic intervals:

(9) $\Delta\Delta F_{ik} \sim N(0, \sigma_{ik}^2 d_k) + \varepsilon_{ik}$,

where σ_{ik}^2 represents the mutation's rate of epistatic change in the particular phylogenetic interval. This model corresponds to a variable-rate model of trait evolution, describing a trait that evolves according to Brownian motion but with a rate that can change over time. Note that the null model is nested within the alternative model – it is a special case of the alternative model that constrains σ_{ik} to be the same for all k.

We performed a likelihood-ratio test to determine whether the null model (rate constancy) can be rejected in favor of the alternative model (rate variation) for a given mutation. We fit each model to the 8 $\Delta\Delta F$ values of a mutation. The maximum likelihood estimate for the rate of epistatic change under the null model is

(10) $\hat{\sigma}_i^2 = \frac{\sum_k (\Delta\Delta F_{ik})^2}{n}$,

where Σ_k denotes summation across all phylogenetic intervals and n is the number of intervals. The log-likelihood of the null model is

(11) $l_i^{Null} = -\frac{n}{2} - \frac{n}{2}\log(\hat{\sigma}_i^2)$.

The maximum likelihood estimate for the rate of epistatic change under the alternative model is

(12) $\hat{\sigma}_{ik}^2 = (\Delta\Delta F_{ik})^2$,

and the log-likelihood of the alternative model is

(13) $l_i^{Alt} = -\frac{n}{2} - \frac{1}{2}\sum_k \log(\hat{\sigma}_{ik}^2)$.

We calculated the likelihood-ratio statistic $S_i = 2(l_i^{Alt} - l_i^{Null})$. Statistical significance was determined using parametric bootstrap: The expected distribution of S_i if the null hypothesis is true was determined by simulating 500 sets of n $\Delta\Delta F$ values under the best-fit null model and calculating S_i for each bootstrap replicate (denoted S_i*); the p-value of the observed S_i is the fraction of boostrap replicates for which $S_i^* \geq S_i$.

$\Delta\Delta F$ values that equal 0 because they are masked by the lower bound of measurement were excluded. Mutations for which 3 or more of the 8 $\Delta\Delta F$ values are masked by the lower bound of measurement were not analyzed. Eqs. (10-13) are simplified formula valid only in the absence of measurement noise ($\varepsilon_{ik} = 0$). We assumed that ε_{ik} is normally distributed with a mean of 0 and the standard deviation that equals the standard error of $\Delta\Delta F_{ik}$ based on the 3 sort-seq replicates.

Since the complete likelihood equation involving measurement noise cannot be algebraically solved, we used a numerical optimizer (the *optim* function in R) to obtain the maximum likelihood estimate of the rate of epistatic change and the log-likelihood of each model.

3.5.6.2 Analysis of normalized $\Delta\Delta F$

The variance of $\Delta\Delta F_{ik}$ depends on the mutation and phylogenetic interval, making it difficult to compare the shape of $\Delta\Delta F$ distribution across mutations and intervals. To address this problem, we defined normalized $\Delta\Delta F$:

$$(14)\quad \Delta\Delta F_{ik}^{Norm} = \frac{\Delta\Delta F_{ik}}{\hat{\sigma}_i\sqrt{d_k}},$$

where $\hat{\sigma}_i$ is the mutation's rate of epistatic change inferred under the Brownian motion model (Eq. [10]). The variance of the normalized $\Delta\Delta F$ values of a mutation is 1. In particular, under the Brownian motion model, the expected distribution of normalized $\Delta\Delta F$ is close to the standard normal distribution. (It is exactly the standard normal distribution if $\hat{\sigma}_i$, the estimated rate of epistatic change, in the denominator of Eq. [14] is replaced by σ_i, the unknown true rate.) We obtained the distribution of normalized $\Delta\Delta F$ expected under the Brownian motion model (the red curve in Fig. 3.3C) by simulating data under the Brownian motion model parametrized by the observed rates of epistatic change.

3.5.6.3 Phylogenetic cross-validation of the Brownian motion model

Phylogenetic cross-validation of the Brownian motion model (Fig. 3.3, E to G) was performed as follows. Let $\sigma_i^{(-k)}$ denote a mutation's rate of epistatic change inferred after excluding interval k. We used $\sigma_i^{(-k)}$ to predict the distribution of $\Delta\Delta F_{ik}$:

(15) $\Delta\Delta F_{ik} \sim N(0, [\sigma_i^{(-k)}]^2 d_k) + \varepsilon_{ik}$.

Mutations were ordered into 10 groups based on the value of $\sigma_i^{(-k)}$. The predicted epistatic change (average of $[\sigma_i^{(-k)}]^2 d_k + \varepsilon_{ik}$ across all mutations in the group) was compared to the observed epistatic change.

3.5.6.4 Systematic among-interval variation in the rate of epistatic change

In Fig. 3.3H, we asked whether the effects of mutations changed systematically more quickly or slowly in certain phylogenetic intervals. To detect such systematic rate variation, we modified the Brownian motion model:

(16) $\Delta\Delta F_{ik} \sim N(0, [\lambda_k \sigma_i]^2 d_k) + \varepsilon_{ik}$,

where λ_k, a constant specific to each phylogenetic interval, represents mutation-wide acceleration or deceleration in the rate of epistatic change. We jointly inferred the maximum-likelihood estimates of σ_i and λ_k, constraining the mean of λ_k across the 8 intervals to 1.

3.5.7 Measuring memory half-life

If we could measure the effect of a mutation at the beginning and end of many independent phylogenetic intervals of length *d,* we could calculate the correlation between the initial and final effects—$r^2(d)$. Repeating this analysis for many values of *d* would reveal the rate at which the mutation's effect is randomized by epistatic interactions with sequence substitutions; the correlation can be modeled as an exponentially decaying function of sequence divergence,

$$(17)\quad r^2(d) = 2^{-d/m},$$

where *m* is the mutation's memory half-life.

In fact, for each mutation, we have measurements of initial and final effects across only 8 intervals, each with a different *d*. We reasoned that mutations with the same rate of epistatic change should have the same memory half-life; instead of analyzing $r^2(d)$ for each mutation across many intervals of same *d*, we therefore calculated it for many mutations (of similar rate of epistatic change) across one interval, and then modeled the decay of this correlation with *d* as described above.

We ordered mutations into deciles based on the rate of epistatic change measured using the Brownian motion model (σ_i; Eqs. 8 and 10). Beginning with the 10% of mutations with the greatest rate, we calculated how correlated the effects of mutations are between every pair of ancestral and derived DBDs (squared Pearson correlation coefficient, corrected for attenuation due to measurement noise; Fig. 3.4, A to D). We then fit the exponential decay function in Eq. (17) to estimate the memory half-life of each group. We identified the quantitative relationship

between memory half-life and rate of epistatic change by fitting a power function, and used this relationship to convert the rate of epistatic change of a mutation into memory half-life.

3.5.8 Analysis of historical sequence substitutions

Because each of the 8 phylogenetic intervals we examined contains more than one substitution, we cannot ascertain the exact genetic background in which any of the 79 substitutions in the trajectory occurred. We estimated the effect that a substitution had at the time of fixation by averaging two measured effects—one measured in the nearest ancestral DBD and the other in the nearest descendant DBD (Fig. 3.5A). We excluded the 19 substitutions for which the two ΔF values differ by more than 0.2.

To analyze historical contingency (Fig. 3.5C), we reconstructed the sequences of all ancestral nodes between AncSR and the present-day DBDs in our phylogeny and identified 275 sequence substitutions, the vast majority of which occurred outside the 8 experimentally characterized intervals. The memory half-life of these substitutions was determined by analyzing their effects along the experimentally characterized trajectories.

Variation of memory half-life among and within sites

Structural analyses were based on the crystal structure of human estrogen receptor 1 (Protein Data Bank ID: 1HCQ) and of human GR (1GLU). Relative solvent accessibility of each site was calculated based on the two structures using the DSSP algorithm (*82*) and averaged. Relative rate of substitution at each site was inferred using PhyML (v 3.3) (*83*) as the posterior mean rate under the gamma-distributed among-site rate variation model. To determine the rate of substitution at physically adjacent sites (R_{adj}), the distance between every pair of DBD

residues—defined as the distance between the geometric mean of the side chain atoms—was calculated based on the two structures and averaged. Sites within 5.9 Å of each other were considered physically adjacent, and R_{adj} of a site was calculated by summing the relative rate of substitution of all adjacent sites. The 5.9 Å cutoff was chosen because it maximized the correlation between R_{adj} and the median memory half-life of a site. The distance of a site to the recognition helix ($_{19}$EGCKAFFKRSIQ$_{30}$ in human ER1 and $_{19}$GSCKVFFKRAVE$_{30}$ in human GR) was defined as the minimum distance to any residue in the recognition helix. Dimerization interface was defined as sites in a DBD monomer within 5.9 Å of any residue in the other monomer; the distance of a site to the interface was defined as the minimum distance to any residue in the interface.

Chapter 4

Comment on "Ancient origins of allosteric activation in a Ser-Thr kinase"

This work was published as "Yeonwoo Park, Jaeda E. J. Patton, Georg K. A. Hochberg, and Joseph W. Thornton, Comment on 'Ancient origins of allosteric activation in a Ser-Thr kinase', Science *370:6519, eabc8301 (2020)."*

4.1 Introduction

Hadzipasic et al. used ancestral sequence reconstruction to identify historical sequence substitutions that putatively caused Aurora kinases to evolve allosteric regulation. We show that their results arise from using an implausible phylogeny and sparse sequence sampling. Addressing either problem reverses their inferences: allostery and the amino acids that confer it were not gained during the diversification of eukaryotes but were lost in a subgroup of Fungi.

4.2 Results and discussion

How allosteric regulation of proteins arose during evolution is a critical question in evolutionary biochemistry. Using ancestral sequence reconstruction (ASR) and biochemical experiments, Hadzipasic et al. (*84*) claim to have identified historical sequence substitutions that caused the acquisition of allostery from a non-allosteric ancestor during the evolution of Aurora kinase (AURK), a eukaryotic cell cycle regulator that is allosterically activated in animals by the TPX2 protein. Inferred ancestral sequences are conditional upon the set of extant sequences used and the phylogeny that describes their relationships, but Hadzipasic et al.'s sequence sampling was extremely sparse, and the phylogeny they used is implausible. We therefore investigated these shortcomings and their effects on the reconstruction of AURK evolution.

The phylogeny inferred by Hadzipasic et al. (Fig. 4.1A) is highly incongruent with the established phylogeny of eukaryotes (Fig. 4.1B). The Hadzipasic et al. phylogeny groups animals and plants together to the exclusion of fungi, but the monophyly of Opisthokonta (fungi, animals, and their unicellular relatives) has been extensively corroborated (*85*, *86*). Hadzipasic et al. also place microsporidians as the most basally branching eukaryotic lineage, despite strong evidence for their inclusion within Fungi (*87*). These incongruences are crucial to the claims of Hadzipasic et al., because the nodes on their phylogeny between which allostery is claimed to have evolved represent ancestral species that in fact never existed: Aur_{ANC2} , the non-allosteric precursor, would be AURK in the last common ancestor of all eukaryotes except microsporidians, and Aur_{ANC3} , the first allosteric protein, would be AURK in the last common ancestor of animals and plants to the exclusion of fungi.

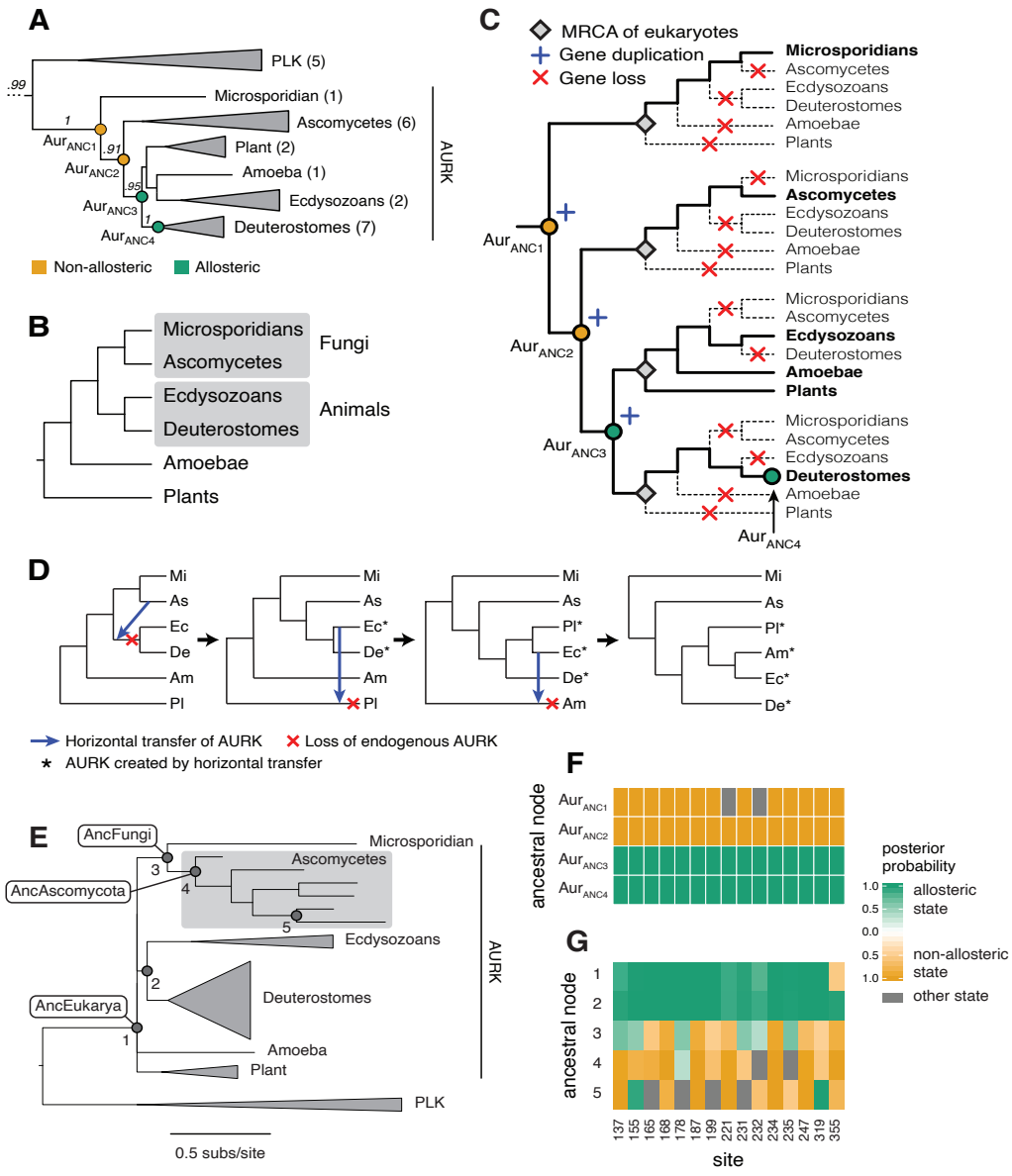

Figure 4.1. A plausible phylogeny reverses Hadzipasic *et al.*'s ancestral reconstructions. (**A**) The phylogeny of AURKs and their nearest outgroup (PLKs) used by Hadzipasic *et al.* Parentheses indicate the number of sequences in each clade. Circles mark the experimentally characterized ancestors, colored by presence/absence of the allosteric response to TPX2. Labels show inferred posterior probability of each clade. (**B**) The established phylogeny of the taxa in panel A. (**C**) Minimum number of gene duplications and losses required to reconcile panel A phylogeny with panel B. **D)** Minimum number of gene transfer and replacement events required to reconcile panel A phylogeny with panel B. Other scenarios with an equal or greater number of events are also possible. (**E**) AURK phylogeny when sequences in Hadzipasic *et al.* were reanalyzed given the constraint in B. Ancestral sequences reconstructed in panel G are labeled. **F**) Ancestral reconstruction on the phylogeny of Hadzipasic *et al.* (panel A). Inferred ancestral states are displayed for a group of 15 sites that experimentally confer allostery when the states from Aur_{ANC3} (green) replace those in Aur_{ANC2} (orange). Gray, other amino acid state. Row

labels correspond to nodes in panel A. Site numbers based on human AURKA. **G)** Maximum *a posteriori* ancestral states reconstructed on the constrained AURK phylogeny in panel E. Sites, states, and colors are as in panel F. Shading shows the posterior probability of each state.

Hadzipasic et al. suggest that their AURK gene tree might be incongruent with the accepted species phylogeny because of gene duplication and loss, but this scenario is implausible: it requires an elaborate history of three gene duplications before the most recent common ancestor (MRCA) of eukaryotes, followed by 14 gene losses distributed so precisely that only a single resulting paralog has been retained in every eukaryote that has been sequenced (Fig. 4.1C). Hadzipasic et al. also suggest horizontal gene transfer (HGT) as a possible cause, but this would require a complex scenario in which every single AURK sequence on the phylogeny except for one descends from an HGT event, with every transfer replacing the recipient's original copy and leaving no trace of the event in any extant genome (Fig. 4.1D); this scenario is especially implausible because HGT between multicellular eukaryotes is rare (*88*).

A more likely cause of the incongruence of Hadzipasic et al.'s tree with the species phylogeny is long branch attraction (LBA) (*89*). The branches leading to microsporidians and ascomycetes may have been moved from their established position near animals towards the root, where they attach to an extremely long branch leading to the nearest outgroup (PLKs). Microsporidians have previously been found to be subject to systematic LBA that moves them to an artifactual position as basal eukaryotes, especially when sampling in the Fungi is sparse (*90*). Strong support for misplaced branches is consistent with systematic bias caused by LBA (*90*).

We therefore repeated ASR using Hadzipasic et al.'s sequence set of AURKs and PLKs, but we constrained the phylogeny to follow established species relationships (Fig. 4.1B, E). We focused on the 15 sequence states from Aur_{ANC3} that experimentally confer allostery when introduced into the non-allosteric Aur_{ANC2} (Fig. 4.1F). We found that the direction of these

substitutions is almost completely reversed compared to the trajectory proposed by Hadzipasic et al. (Fig. 4.1E-G). The deepest AURK ancestor (AncEukarya) now contains 14 of the 15 states associated with allostery and only one of the non-allosteric states; the other 14 non-allosteric states were all gained within the Fungi. Repairing the major topological errors in Hadzipasic et al.'s phylogeny is therefore sufficient to remove the evidence for their paper's central claims.

We next studied the effect of improved sequence sampling. AURKs are present across eukaryotes, but the sequence set analyzed by Hadzipasic et al. included only 19 AURKs; all but three of these were from animals and fungi, which account for only a small fraction of eukaryotic diversity. Within fungi, only ascomycetes and a single microsporidian were represented, and only a single species each of plant and amoeba were included. We therefore acquired and aligned 324 AURK and 315 PLK protein sequences, broadly sampled from five major eukaryotic taxa (Fig. 4.2A): Fungi, Holozoa (animals and unicellular relatives), Archaeplastida (plants and green and red algae), Amoebozoa (amoebae), and SAR (stramenopiles, alveolates, and rhizarians). Within Fungi, we included 137 AURKs from numerous taxonomic groups to better resolve the phylogenetic position of Fungi and the amino acid states within it.

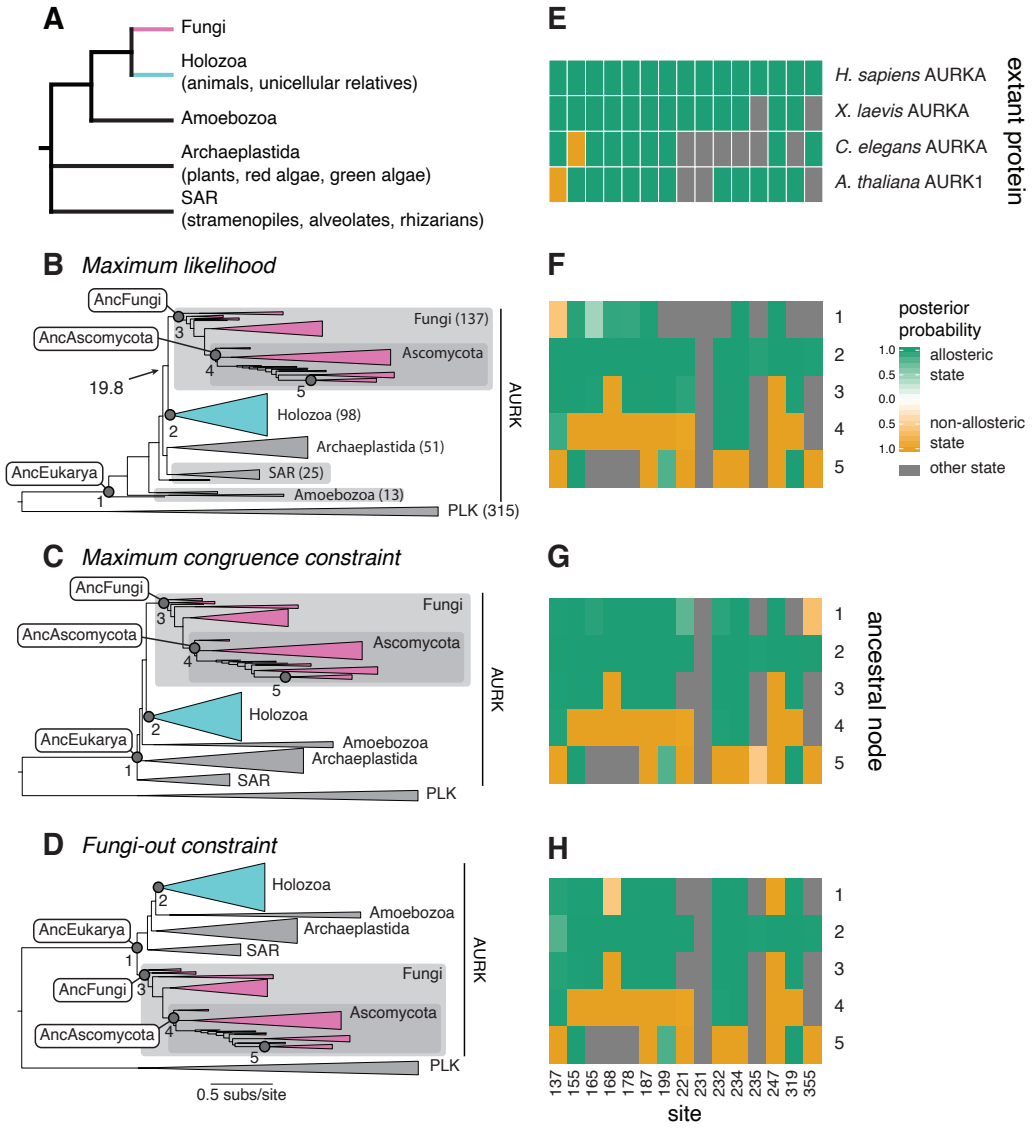

Figure 4.2. Improved sequence sampling reverses Hadzipasic *et al.*'s ancestral reconstructions. (**A**) The established phylogeny of the major eukaryotic groups. Polytomy, branching order not established. (**B**) Maximum likelihood phylogeny when AURK and PLK are densely sampled. Numbers in parentheses indicate the number of sequences in each group. Node labels, reconstructed ancestral sequences. Fungi and Holozoa are pink and cyan, respectively. Branch label with arrow, approximate likelihood ratio statistic for Fungi+Holozoa ($p < 0.01$). (**C**) ML phylogeny given the constraint in panel A. (**D**) ML phylogeny given the constraint in A, except Fungi are constrained to split first. **(E)** Most states that confer allostery in Hadzipasic et al. are not conserved in extant AURKs that are allosterically regulated by TPX2. Green and orange, allosteric and non-allosteric states from Hadzipasic et al.; gray, other states. Site numbers are in panel H. **(F, G, H)** Reconstructed sequences on the phylogenies in panels B, C, and D, respectively. The maximum *a posteriori* states at the 15 sites that experimentally confer allostery/nonallostery are shown, colored as in panel E and shaded by their posterior probability. Row labels correspond to ancestral nodes in panels B-D.

We used this alignment to reconstruct ancestral sequences on three phylogenies: 1) the unconstrained maximum likelihood (ML) phylogeny, which recovers almost all the established species relationships – including the sister relationship of Fungi and Holozoa – except that Amoebozoa and some SAR sequences are pulled towards the root (Fig. 4.2B); 2) the "maximum congruence" (MC) phylogeny, which is constrained to reflect established species relationships among the major groups (Fig. 4.2A, C); and 3) a "Fungi-out" phylogeny, which has the same constraints, but with Fungi as the first-branching eukaryotic lineage (Fig. 4.2D). The likelihood difference between the ML and MC tree is not significant ($p = 0.36$, Shimodaira-Hasegawa test (*91*)), and the latter requires no auxiliary events like gene duplications/losses or horizontal transfers, so we consider the MC phylogeny to be the best supported. The Fungi-out phylogeny is implausible, but it allows us to isolate the effect of improving sampling on ASR by imposing the critical features of the Hadzipasic et al. phylogeny.

On all three trees, AncEukarya again has predominantly allosteric states and only one or two of the 15 non-allosteric states that Hadzipasic et al. inferred as ancestral (Fig. 4.2F-H). All other non-allosteric residues are again derived within Fungi. This result arises because the allosteric states are found not only in animals and plants but also in non-ascomycete fungi and other eukaryotic groups, which Hadzipasic et al. did not include. Improved sequence sampling alone, even on the Fungi-out phylogeny, is therefore sufficient to reverse the direction of evolution of the experimentally important substitutions compared with that inferred by Hadzipasic et al.

On the most plausible MC phylogeny, AncEukarya contains the allosteric state at 11 of 15 sites. The four missing states are not universally required for allostery, because they are

absent in one or more extant allosteric AURKs (Fig. 2E) (*92–94*). The best-supported hypothesis is therefore that AURK of AncEukarya was allosteric, and this feature was lost along the lineage leading to ascomycetes; experiments will be necessary for a direct test. This scenario is consistent with the taxonomic distribution of AURK's allosteric effector TPX2. Hadzipasic et al. claim that TPX2 evolved after the origin of the AURK protein and before the emergence of allostery, but a reciprocal BLAST search identifies TPX2 orthologs in all major eukaryotic taxa, including all fungal groups except ascomycetes (Supplemental Data Table). The history of TPX2 therefore tracks exactly with the best supported history of AURK allostery: presence in the eukaryotic ancestor, loss in ascomycetes.

Finally, our analysis indicates that the basal placement of fungi in the phylogeny of Hadzipasic et al. is likely attributable to LBA. The first line of defense against LBA is improved sampling to break up long branches (*95*). When we analyzed more sequences with greater taxonomic diversity – including numerous fungal groups that branch off the established phylogeny between Microsporidiae and ascomycetes, as well as basally branching groups within the other high-level eukaryotic taxa – support for Hadzipasic et al.'s topology was eliminated, and the canonical position of Fungi was restored with strong support (Fig. 4.2B). One reason for the sparse sampling in Hadzipasic et al. may have been the use of software to co-estimate phylogeny and alignment, which is computationally demanding and therefore limited to very small datasets; although co-estimation is appealing in theory, the AURK sequences align with little ambiguity, and the compromised sampling necessitated by this approach led to severe phylogenetic error.

This case illustrates the importance of sound phylogenetic practice when employing ASR. Comprehensive sequence sampling is essential, especially from taxa that attach to the

phylogeny near the nodes of interest and that can break up long branches. Single-protein datasets may not have sufficient signal to resolve difficult phylogenetic problems or overcome LBA, so congruence with well-established relationships should be assessed, and the effect of imposing those relationships on the reconstruction should be explored. Confidence in the functional properties of reconstructed ancestral proteins should always be assessed by examining the distribution of functions among extant sequences across the phylogeny; if a very non-parsimonious history is implied, extra scrutiny is warranted. In the current case, characterization of other extant AURKs, particularly in non-ascomycete fungi, Amoebozoa, and SAR is essential. These kinds of practices can provide multiple safety checks against erroneous inference by ASR.

4.3 Methods

4.3.1 Ancestral sequence reconstruction using Hadzipasic et al. sequences under congruence constraint (Fig. 4.1E, G)

We acquired the AURK and PLK sequences analyzed by Hadzipasic *et al*. We aligned them using MUSCLE (v3.8.425) (*74*), removed sequence-specific insertions and ambiguously aligned sites, and trimmed the N- and C-termini, matching the sequence boundaries set by Hadzipasic *et al*. We used RAxML (v8.2.12) (*75*) to infer the constrained ML phylogeny, imposing the constraint shown in Fig. 1B, and used PAML (v4.8) (*77*) to perform ASR. For both phylogenetics and ASR, we used the same model of sequence evolution as used by Hadzipasic *et al*. (LG + G + X, four gamma rate categories).

4.3.2 Phylogenetics and ASR with improved sequence sampling (Fig. 4.2)

To obtain a broad sample of eukaryotic AURK and PLK sequences, we used a reciprocal best-hit protein BLAST strategy using the NCBI protein database (*96*). Human AURKA and PLK4 were used as query sequences. Taxonomically restricted BLAST searches were conducted that together encompassed all species within the five major eukaryotic kingdoms/subkingdoms (Fungi, Holozoa, Amoebozoa, Archaeplastida, and SAR). BLAST hits of anomalous length (<250 or >600 amino acids for AURK, <250 or >1000 amino acids for PLK) were discarded. Redundant sequences were eliminated at similarity cutoff 0.85 using CD-HIT (v4.8.1) (*97*). Each remaining BLAST hit was then used as query in a reciprocal BLAST search against human proteins, and all sequences for which the best hit in humans was an AURK or PLK were retained.

Sequence alignment of these hits was performed hierarchically using MUSCLE software. We first aligned sequences from within defined profile groups of species (each usually a superphylum or phylum). We trimmed the N- and C-termini, leaving sites corresponding to human AURKA sites 133 to 383, and then removed sites representing species-specific insertions and ambiguously aligned sites. We discarded sequences that were missing 10 or more consecutive amino acids present in the majority of other sequences. We then inferred the phylogeny of the profile group using FastTree (v2.1.11) (*98*). To minimize long branch attraction, we removed all sequences or groups of sequences subtended by branches of length >0.5. We also removed sequences/small groups that were assigned to entirely different phyla (e.g., annelid sequences placed inside the molluscs, or green algae sequences placed inside land plants), as well as taxon-specific paralogs with long branches that were pulled outside of the entire profile group being aligned. We then used profile-profile alignment in MUSCLE to

progressively align the group-specific alignments to each other, yielding a global AURK/PLK alignment.

We used RAxML to infer the ML AURK-PLK phylogeny from this global alignment, using the best-fit model of evolution (LG + G + X). For all RAxML analyses, we iterated topology search 50 times using different random number seeds, and chose the iteration with the highest likelihood. On the ML phylogeny, AURKs from a few lower-level groups of Ecdysozoa and Platyhelminthes subtended by long branches were placed in kingdoms other than the animals; drastic long-branch misplacements also moved a few small groups of AURKs from Fungi and Alveolates into other kingdoms/superphyla and affected some PLK sequences. These sequences were removed to yield the final alignment, and the analysis was repeated to infer the final ML phylogeny. Approximate likelihood ratio test was performed using PhyML (v3.3) (*83*). For the maximum congruence constraint analysis, we imposed the topological constraint shown in Fig. 2A and used RAxML to perform phylogenetic analysis to find the ML tree, branch lengths, and other parameters given this constraint. We used a similar approach to find the ML tree consistent with the Fungi-out constraint (the same constraint in Fig. 2A, except that Fungi are the most basally branching group). Ancestral sequences were inferred using the marginal reconstruction algorithm in PAML using LG + G and the amino acid frequencies inferred on the ML tree by RAxML.

The Shimodaira-Hasegawa test was used to evaluate relative support for the ML vs. MC trees. We used the R package phangorn to execute the SH test (*99*), comparing the ML tree found in the unconstrained ML search to the ML tree from the MC-constrained search (lnL -121973.5 and -121992.0, respectively); this returned a nonsignificant result (p-value=0.36). The heuristic searches may not have identified the globally optimal tree in each case, so we also

compared the tree from the search iteration with the highest likelihood in the unconstrained ML analysis to the tree recovered from the iteration with the lowest likelihood in the MC-constrained analysis (lnL=-122032.7, *p*-value=0.21).

Chapter 5

Conclusion

Epistatic interactions among mutations were thought to be pervasive and complex, rendering any attempt to explain the genetic basis of protein function unsuccessful. The effects of mutations were thought to change idiosyncratically during evolution, without any regularity to enable prediction. These conclusions were based on methods of analysis inadequate for uncovering the global regularities of genotype-phenotype maps. Moreover, historical studies have been limited to measuring the effects of only a few mutations across single historical intervals, which can only generate highly idiosyncratic patterns of epistasis.

By analyzing experimental mutational datasets using an optimal method, I showed that high-order interactions contribute negligibly to phenotype in these datasets and that context-independent effects and pairwise interactions that contribute significantly to phenotype are sparse. Furthermore, by measuring the effect of every possible mutation across many historical intervals and analyzing this data under a statistical framework, I uncovered simple statistical regularities that underlie the apparent idiosyncratic epistatic changes. Overall, these findings show that protein genotype-phenotype maps may be simple enough to be learned and explained and the role of epistasis in historical contingency can be quantitatively modeled and studied.

Appendix 1

Supplementary figures for chapter 3

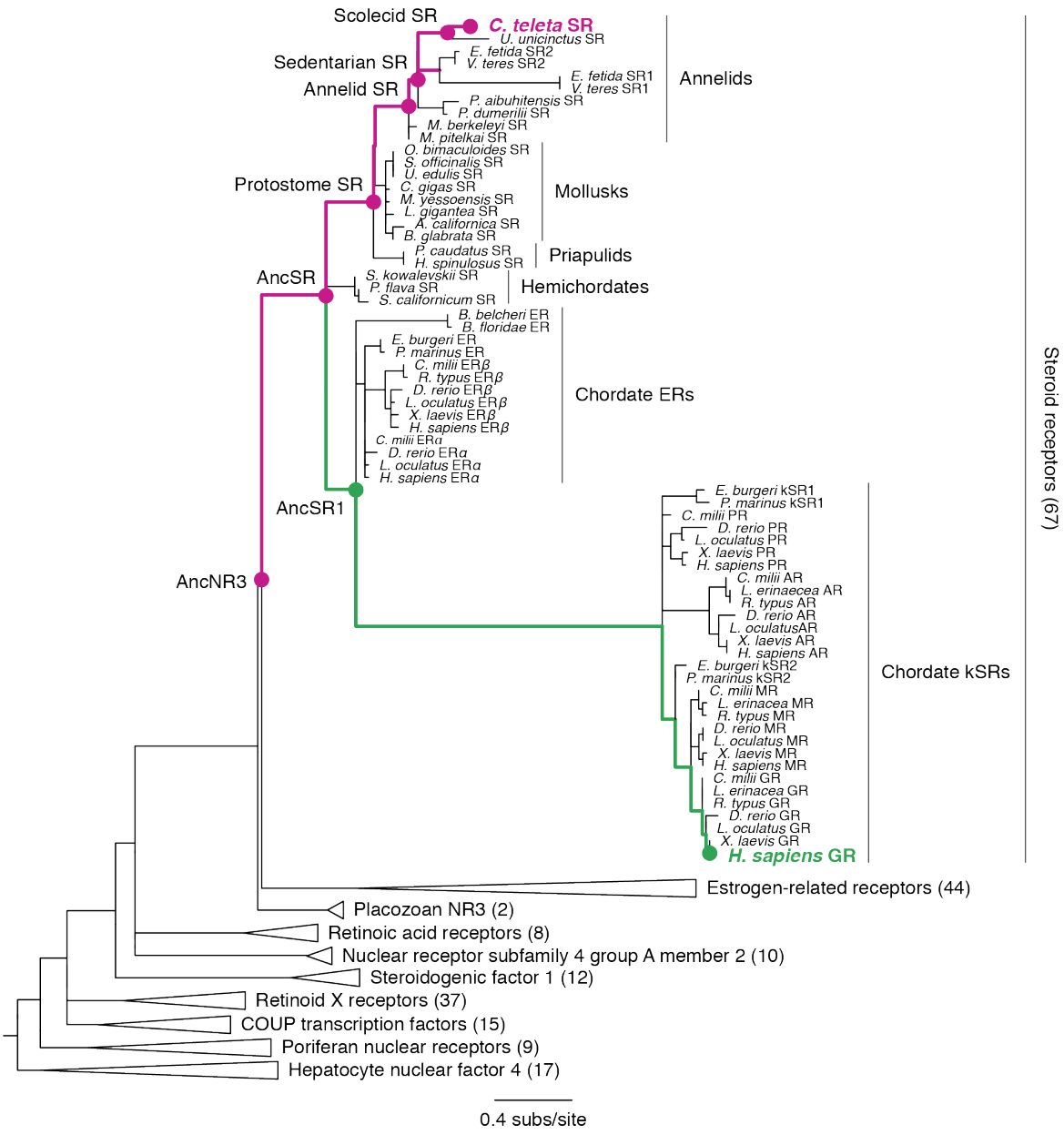

Fig. A1. Phylogeny of the DNA-binding domain of steroid and related receptors. Magenta and green, trajectories of reconstructed ancestral and extant sequences studied here. Parentheses, number of sequences in each clade. The phylogeny is rooted on hepatocyte nuclear receptor 4 proteins, the earliest-diverging group of nuclear receptors.

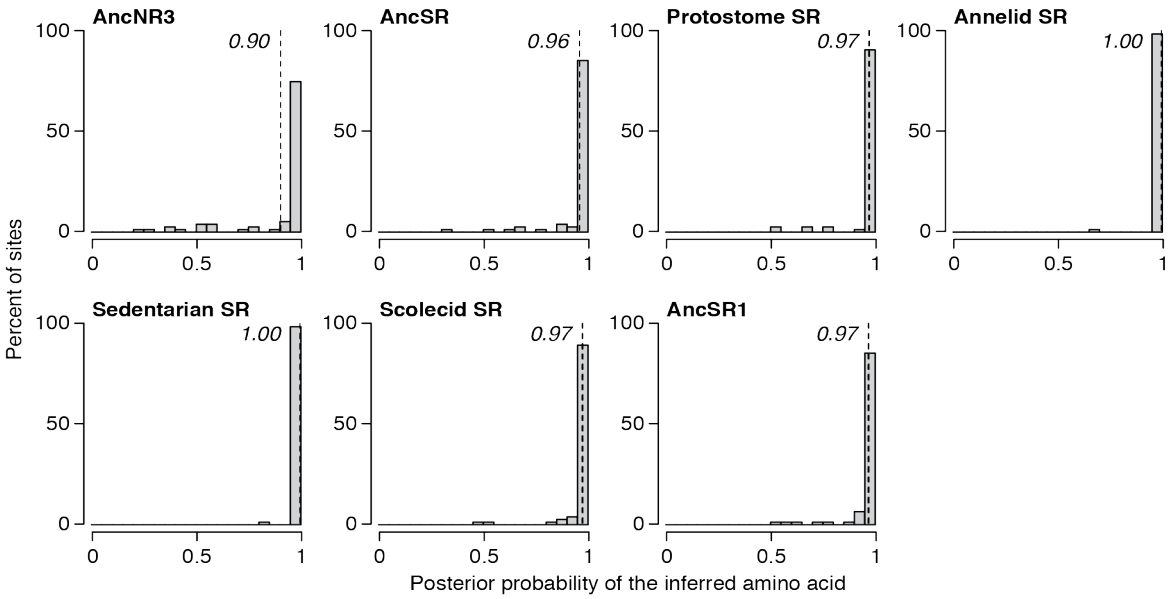

Fig. A2. Statistical support for reconstructed ancestral sequences. Each histogram shows the distribution of the posterior probability of the inferred amino acid at the 76 sites of an ancestral DBD characterized in this study. Dotted line and italic, average posterior probability of the 76 sites of the corresponding DBD.

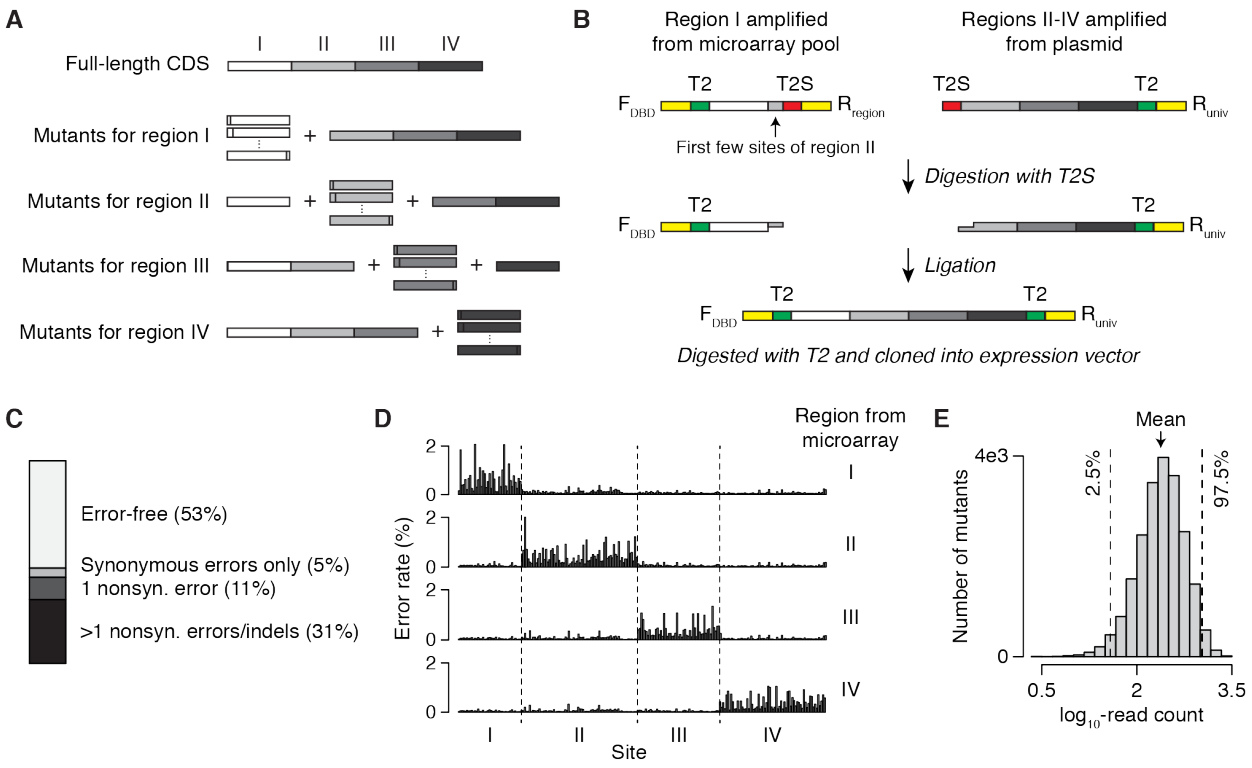

Fig. A3. Construction and validation of comprehensive point mutant libraries. (**A**) The 228-nt DBD coding sequence was divided into 4 regions (I to IV); variant oligonucleotides for each region were synthesized using a microarray and ligated to wild-type oligonucleotides for the 3 other regions, which were generated by PCR from a sequence-verified template. Because only ~25% of each full-length coding sequence derives from the microarray, both the synthesis cost and effective error rate are reduced by ~4-fold. (**B**) Selective amplification and seamless ligation of oligonucleotides. Oligonucleotides for each region in each DBD were flanked by a specific pair of primer-binding sites (PBS) for selective amplification from the microarray-generated pool of all variant oligonucleotides. F_{DBD}, forward PBS specifying the DBD; R_{region}, reverse PBS specifying the region; R_{univ}, shared reverse PBS. Type IIS restriction enzymes (T2S) cut DNA outside their recognition site, allowing digested oligonucleotides to be assembled into uninterrupted full-length coding sequences. T2, conventional palindromic restriction enzyme. (**C**-**E**) Assembled libraries were sequenced using 2 × 300 nt Illumina MiSeq paired-end sequencing to determine the frequency of synthesis errors and the relative abundance of mutants. (**C**) Synthesis error profile. Percentages indicate the fraction of reads in each error category. (**D**) Rate of mismatch error at each site, shown separately for four types of coding sequences defined by the region deriving from the microarray. Dotted lines, region boundaries. (**E**) Distribution of the read count of every mutant. No mutant was missing, and 95% of mutants had a read count within 7-fold of the median (range = [3, 3042]; median = 252; 95% range = [38, 1072]).

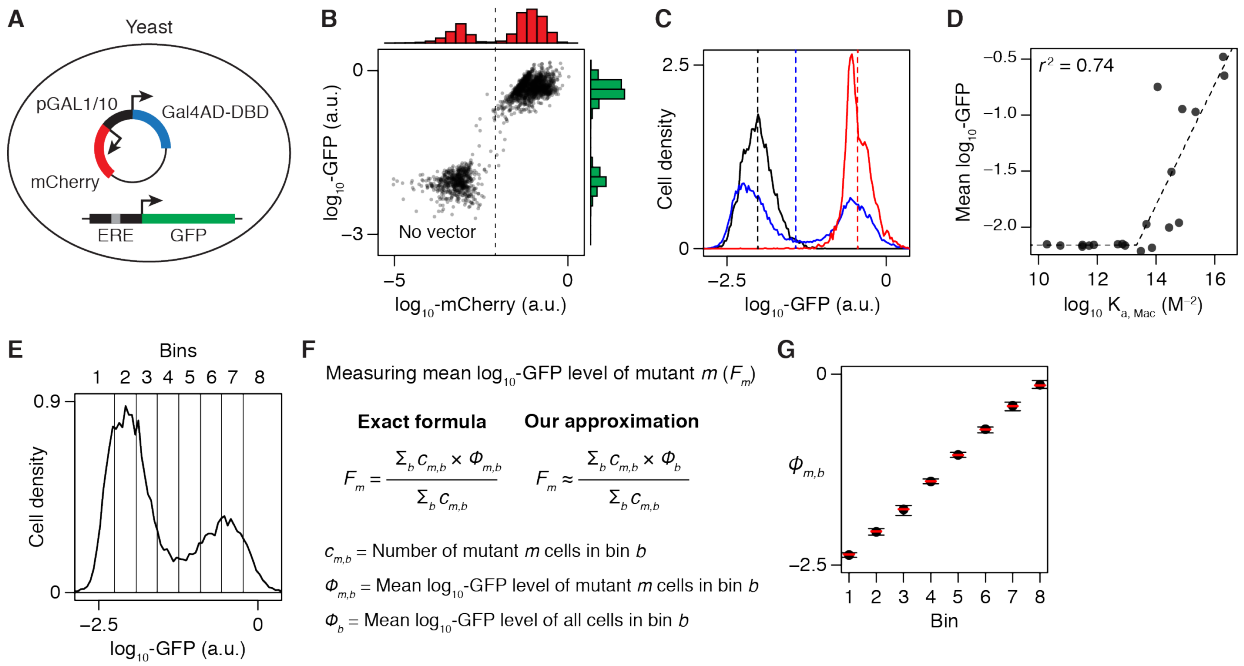

Fig. A4. Functional characterization of DBD libraries using sort-seq. (**A**) Fluorescence reporter assay for DBD activity. The yeast reporter strain contains a genomically integrated GFP reporter driven by the estrogen response element (ERE, an inverted palindrome of AGGTCA). The DBD expression vector contains the DBD fused to the Gal4 transcriptional activation domain and the fluorescent protein mCherry, both driven by a galactose-inducible bidirectional promoter pGAL1/10. (**B**) By only analyzing mCherry-positive cells, cells that lose the expression vector during growth are excluded from measurement. Each dot shows the GFP and mCherry fluorescence of a single cell (among a population of cells transformed with the same expression vector) measured by flow cytometry. Bars, histogram of GFP (green) or mCherry (red) fluorescence. Dashed line distinguishes mCherry-positive cells (right) from mCherry-negative cells (left). (**C**) Fluorescence distribution of 3 isogenic cell populations, each expressing a DBD variant: black, no activation; blue, weak activation; red, strong activation. Dashed lines, the mean $\log_{10}$-GFP fluorescence of each population (F), the final readout of the assay. (**D**) Comparing the activities of 20 DBD variants to their DNA affinities previously measured *in vitro* (*39*). $K_{a,\,Mac}$, macroscopic binding constant. Dashed line, best-fit segmented linear regression. This result shows that the dynamic range of our reporter assay spans ~10^3-fold change in $K_{a,\,Mac}$. (**E**) Fluorescence distribution of the sorted cell population expressing the DBD libraries. Vertical lines, sort bin boundaries. (**F**) Nonparametric method for estimating the mean $\log_{10}$-GFP level of a mutant from the sort-seq measured distribution of cell number across bins. (**G**) $\varphi_{m,b}$ (mean $\log_{10}$-GFP level of mutant m cells in bin b) of 64 randomly chosen DBD variants determined by flow cytometry of individual isogenic populations. Error bars, 95% range. Red line, φ_b (mean $\log_{10}$-GFP level of all cells in bin b). This result shows that $\varphi_{m,b}$ is close to φ_b regardless of mutants, indicating that the approximation in (**F**) is highly accurate.

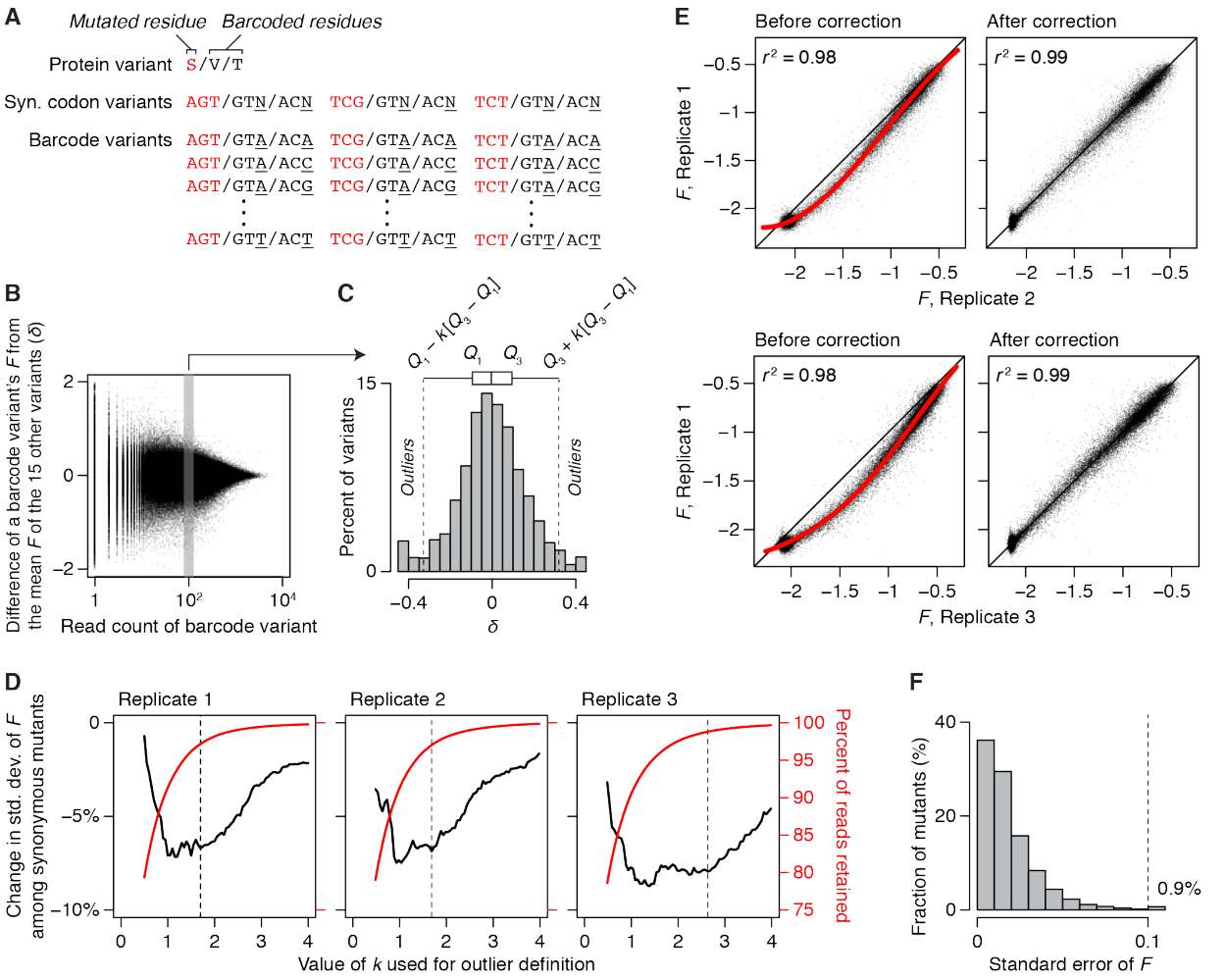

Fig. A5. Sort-seq data cleaning. (**A**) Degenerate encoding of variants in libraries. Each protein variant has one nonsynonymously mutated residue with a specified amino acid change (red) and two fourfold-degenerate residues into which synonymous mutations are introduced for internal barcoding (black). The nonsynonymous residue is encoded by up to 3 codon mutants (columns), each of which is encoded by up to 16 synonymous barcode variants (rows). (**B-D**) Removing outlier barcode variants. (**B**) For each barcode variant, we calculated the difference between its F and the mean F of the 15 other barcode variants of the same mutant (δ). Shown is every barcode variant's δ plotted against its read count. (**C**) To define outliers, we examined the distribution of δ among barcode variants with similar read counts; variants at the tails of this distribution were considered outliers. Specifically, given a constant k and the first and third quartiles of the distribution (Q_1 and Q_3), values outside the interval $[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$ were defined as outliers and removed. Retained barcode variants were pooled, resulting in a single value of F for each codon mutant. (**D**) Determining the optimal value of k. We chose a value of k that 1) minimizes the standard deviation of F (σ) among synonymous codon mutants and 2) removes the least number of reads. Dashed lines, optimal value of k for each sort-seq replicate. (**E**) Correcting for systematic variation in fluorescence among sort-seq replicates. Each plot compares the fluorescence of every mutant between a pair of sort-seq replicates. Fluorescence is systematically higher in replicates 2 and 3 compared to 1, reflecting differences in experimental conditions. We fit a spline function with 5 parameters to capture this nonlinearity (red curve) and used it to normalized the fluorescence of replicates 2 and 3. (**F**) Distribution of the standard error of F of every mutant. The 0.9% of mutants with the standard error >0.1 were excluded from analyses.

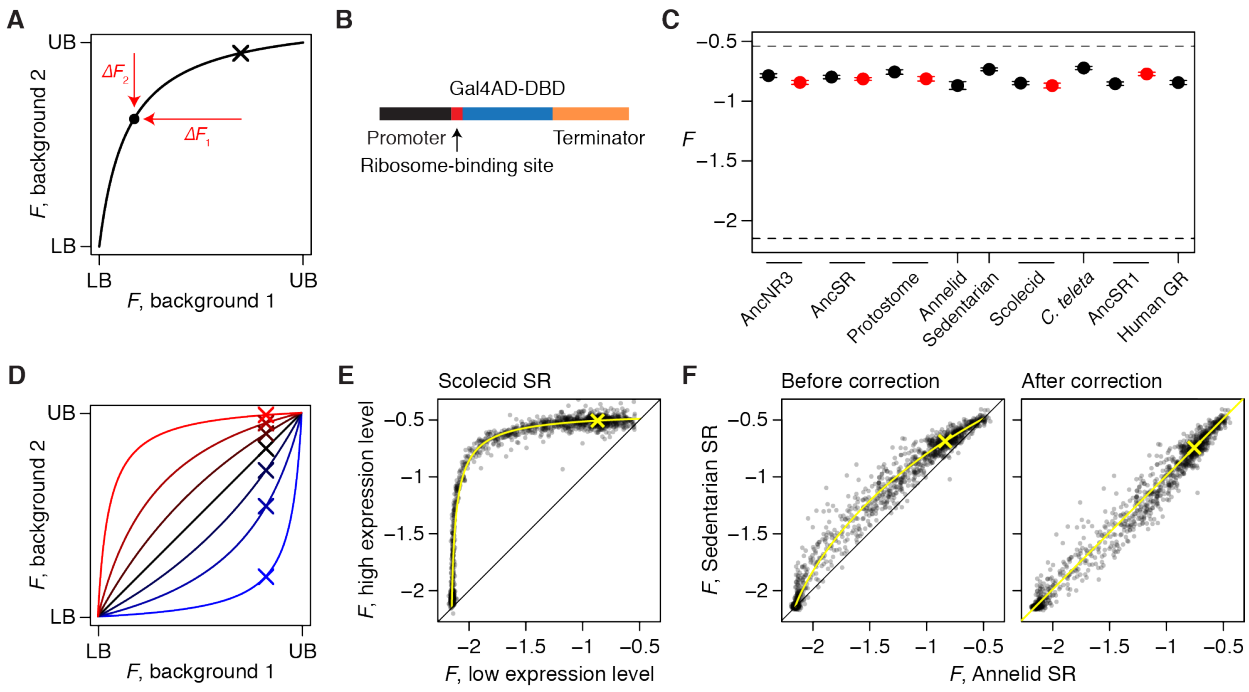

Fig. A6. Removing nonspecific epistasis. (**A**) Illustration of nonspecific epistasis. Curve shows a hypothetical relationship between the phenotype of the same set of mutants when measured in two different genetic backgrounds. Cross shows the activity of the wild-type protein, and dot shows that of a particular mutant; red horizontal and vertical arrows show the effect of the mutation in backgrounds 1 and 2, respectively. LB and UB, lower and upper bound of measurement. (**B**) Different promoters, ribosome-binding sites, and terminators were used to adjust the expression levels of the 14 wild-type DBDs so that they have similar levels of GFP activation in our assay. (**C**) Phenotype of the 14 wild-type DBDs after normalizing their expression level. Black, MAP reconstructions; red, Alt-All reconstructions (only for ambiguously reconstructed nodes). Error bars, 95% confidence interval. (**D**) Graphical depiction of model to remove nonspecific epistasis. Each curve shows the global shift in mutant phenotype predicted when measured in background 1 vs. 7 possible cases of background 2; red and blue, highest and lowest wild-type activity, respectively. Each curve is defined by a single free parameter: the difference in the phenotype of the two wild-type proteins (cross). The curve can then be used to correct for the expected difference in activity between the two backgrounds for any mutation. (**E**) Experimental validation of the nonspecific epistasis model. Two complete single-replacement DMS datasets were generated for a single DBD (scolecid SR) by expressing it in low- vs. high-expression vectors. Each dot shows the phenotype of a mutant measured in the two conditions. Yellow curve, nonspecific epistasis model determined by the observed position of the wild-type activity (cross). (**F**) Removing nonspecific epistasis from dataset. Each plot compares the phenotype of every point mutant measured in annelid SR vs. sedentarian SR. Left, measured phenotypes before removing nonspecific epistasis; yellow curve, nonspecific epistasis model determined by the wild-type activity (cross). Right, corrected phenotypes; identity line is shown in yellow.

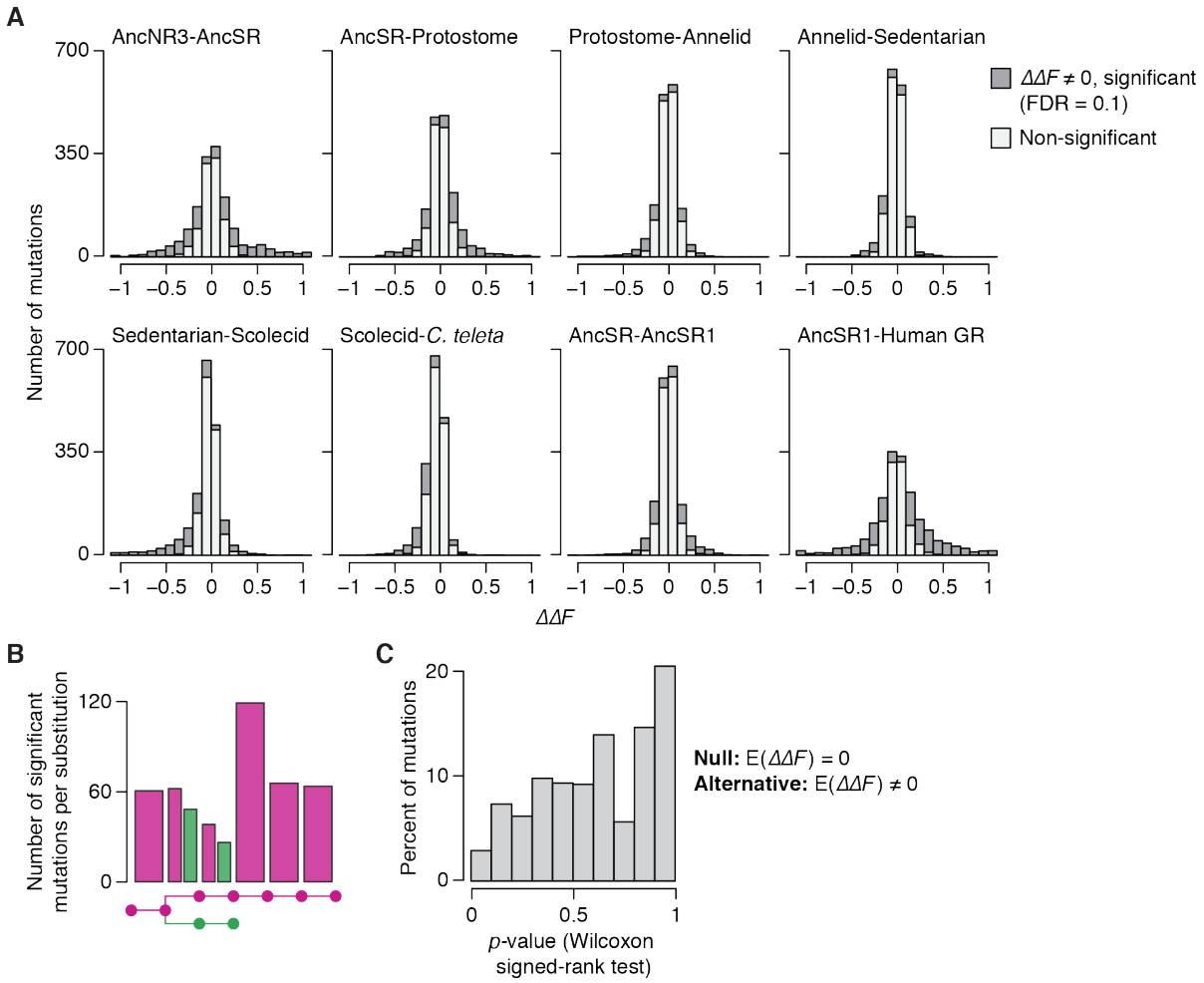

Fig. A7. Distribution of epistatic change for individual phylogenetic intervals. (**A**) Distribution of epistatic changes in the effect of every mutation ($\Delta\Delta F$) during each of the 8 experimentally characterized phylogenetic intervals. Dark grey, $\Delta\Delta F$ significantly different from 0 (*t*-test, FDR ≤ 0.1). Mutations always at the lower bound of measurement were excluded. (**B**) Number of mutations with significant epistatic change in an interval divided by the number of substitutions in that interval (based on Fig. 2B). Each column represents one phylogenetic interval. (**C**) Wilcoxon signed-rank test was used to test whether the mean of the 8 $\Delta\Delta F$ values of each mutation is significantly different from 0. Significant directional bias was not observed for any mutation even at the false discovery rate of 0.5.

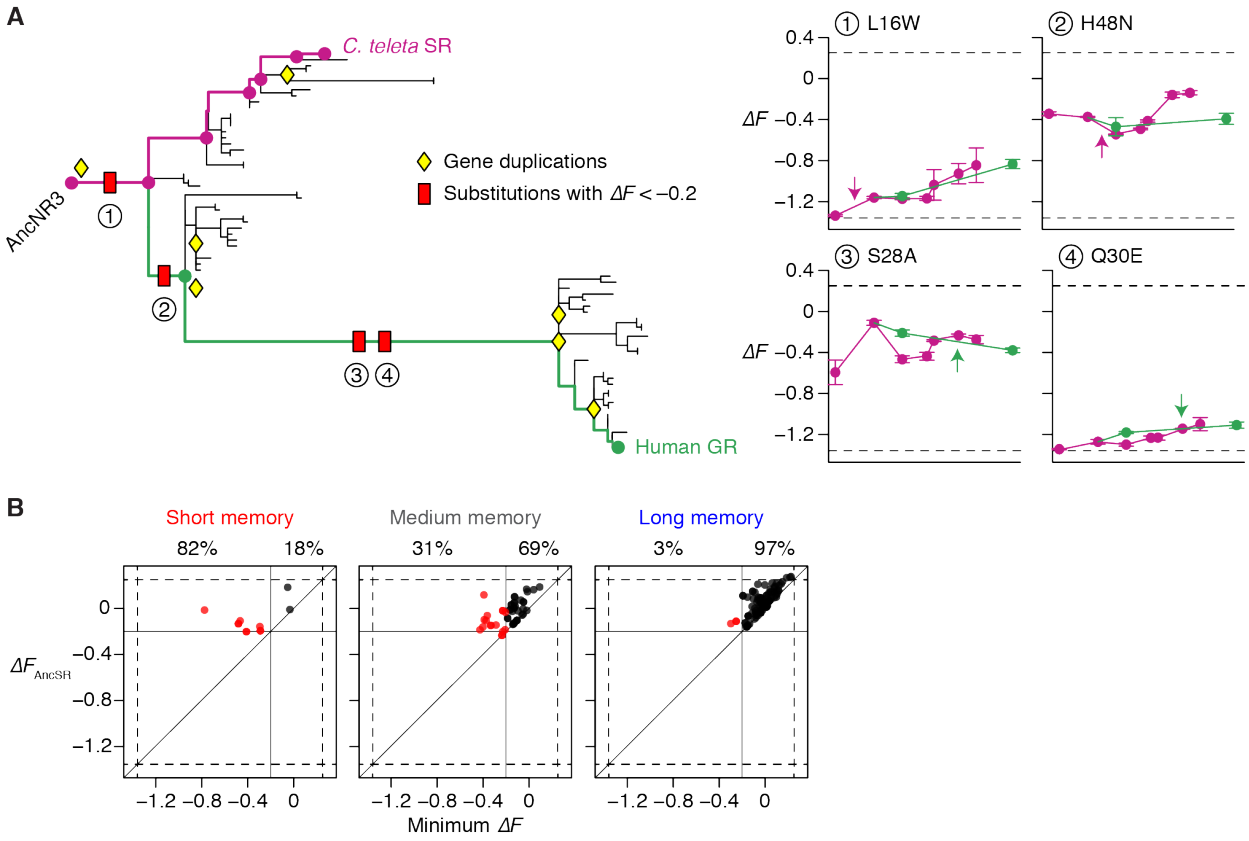

Fig. A8. Analysis of historical sequence substitutions. (**A**) Phylogenetic intervals during which historical substitutions with $\Delta F < -0.2$ occurred. Of the 79 substitutions along the trajectories from AncNR3 to *C. teleta* SR and human GR, four have estimated $\Delta F < -0.2$ at the time of fixation (Fig. 5, A and B). The interval in which each of the substitutions occurred is marked by red squares in the left phylogeny. Three of them (labeled 1, 3, and 4) occurred in intervals immediately following gene duplication, marked by yellow diamonds. The effect of each substitution on the 9 experimentally characterized DBDs is shown on the right, arranged along the horizontal axis and colored as in Fig. 1D. Arrow, interval in which the substitution occurred. (**B**) Among the 275 substitutions that occurred between AncSR and any extant DBD in our phylogeny, many were initially accessible ($\Delta F_{\text{AncSR}} \geq -0.2$). For each such substitution, we calculated the minimum effect it has on the experimentally characterized descendant DBDs (minimum ΔF) and plotted it against its initial effect, grouping the substitutions by memory category. Red, minimum $\Delta F < -0.2$ (t-test, FDR ≤ 0.1). Percentages show the fraction of substitutions in each memory category with minimum $\Delta F < -0.2$ (left) or ≥ -0.2 (right).

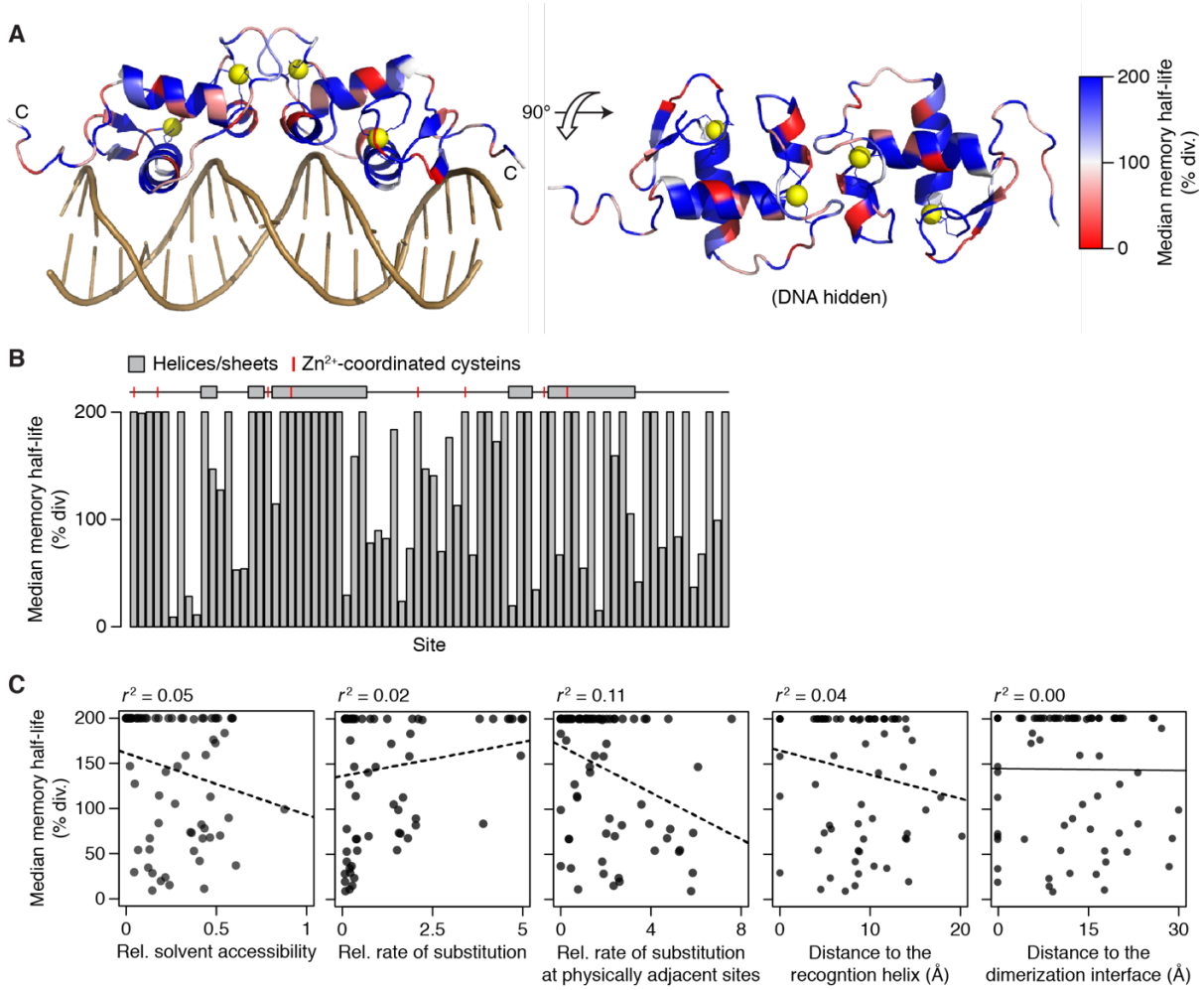

Fig. A9. Explaining variation of memory half-life among sites. (**A**) Median memory half-life of all mutations within each site shown on the crystal structure of human GR DBD bound to DNA (Protein Data Bank ID: 1GLU). Residues are colored according to the color gradient at right; yellow spheres, cysteine-coordinated Zn^{2+} ions. (**B**) Median memory half-life of each site. Secondary structure is shown above. (**C**) Median memory half-life of each site plotted against 5 structural/functional indices. Squared Pearson correlation coefficient and best-fit linear regression are shown.

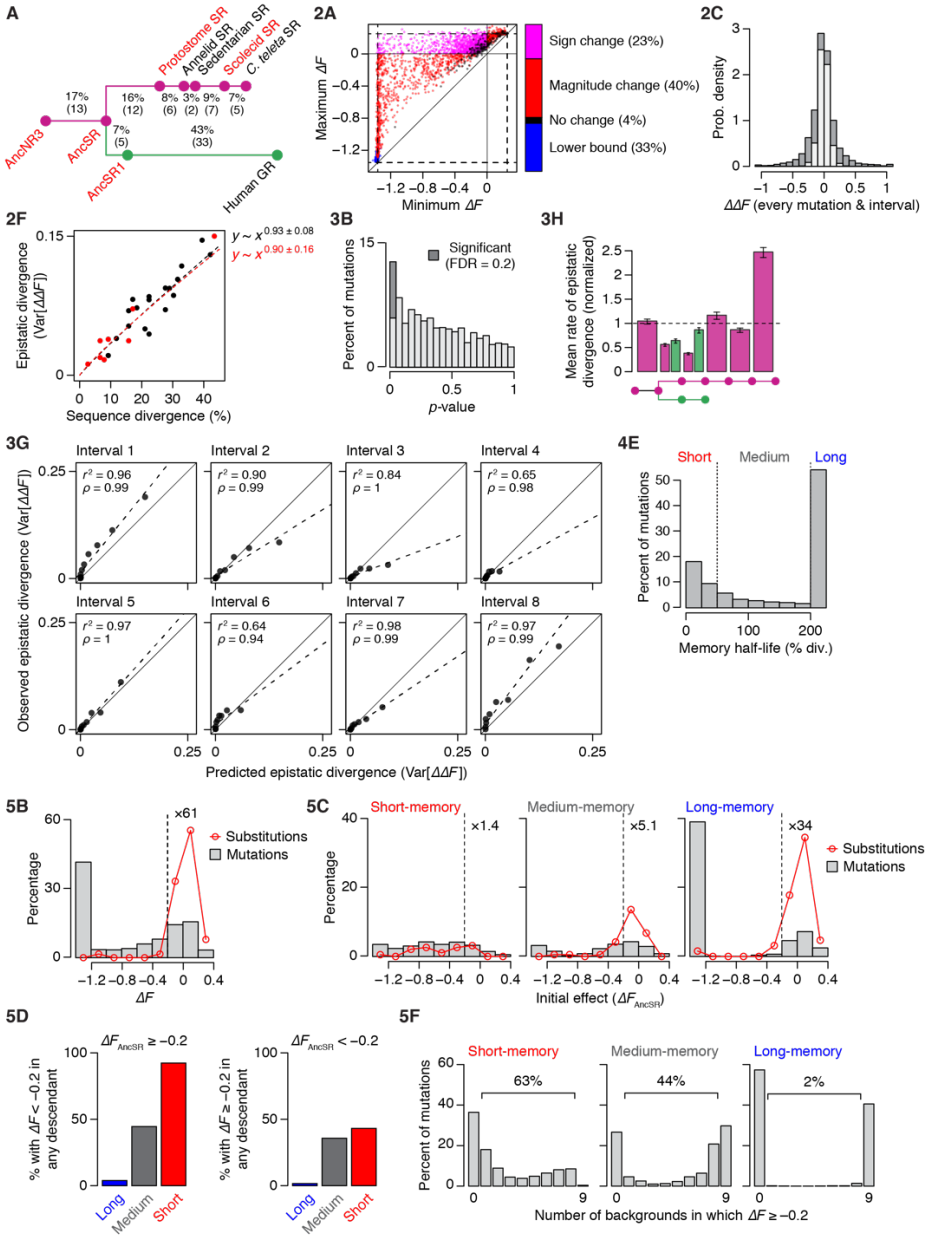

Fig. A10. Robustness of functional inference to uncertainty in ancestral sequence reconstruction. (**A**) Phylogenetic relations among the 9 characterized DBDs, shown using the Alt-All sequences for the 5 ambiguously reconstructed DBDs marked in red. (**2A-5F**) Analyses performed using MAP ancestral reconstructions shown in the paper's main figures (labels) were repeated using the Alt-All ancestors.

Bibliography

1. R. Amundson, *The changing role of the embryo in evolutionary thought: roots of evo-devo* (Cambridge University Press, 2005).
2. M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1983).
3. G. P. Wagner, in *Homology, genes, and evolutionary innovation*, Ed. (princeton university press, 2014),
4. D. M. Taverna, R. A. Goldstein, Why are proteins marginally stable. *Proteins: Structure, Function, and Bioinformatics* **46**, 105-109 (2002).
5. G. K. A. Hochberg *et al.*, A hydrophobic ratchet entrenches molecular complexes. *Nature* **588**, 503-508 (2020).
6. L. Schulz, F. L. Sendker, G. K. A. Hochberg, Non-adaptive complexity and biochemical function. *Current Opinion in Structural Biology* **73**, 102339 (2022).
7. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein science* **25**, 1204-1218 (2016).
8. M. J. Harms, J. W. Thornton, Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512**, 203-207 (2014).
9. C. D. Aakre *et al.*, Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594-606 (2015).
10. C. Bank, R. T. Hietpas, J. D. Jensen, D. N. Bolon, A systematic survey of an intragenic epistatic landscape. *Mol Biol Evol* **32**, 229-238 (2015).
11. G. Diss, B. Lehner, The genetic landscape of a physical interaction. *Elife* **7**, e32472 (2018).
12. C. A. Olson, N. C. Wu, R. Sun, A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol* **24**, 2643-2651 (2014).
13. V. O. Pokusaeva *et al.*, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet* **15**, e1008079 (2019).
14. 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).
15. O. Ashenberg, L. I. Gong, J. D. Bloom, Mutational effects on stability are largely conserved during protein evolution. *Proc Natl Acad Sci U S A* **110**, 21071-21076 (2013).
16. M. B. Doud, O. Ashenberg, J. D. Bloom, Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Mol Biol Evol* **32**, 2944-2960 (2015).
17. H. K. Haddox, A. S. Dingens, S. K. Hilton, J. Overbaugh, J. D. Bloom, Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife* **7**, e34420 (2018).
18. A. S. Kondrashov, S. Sunyaev, F. A. Kondrashov, Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* **99**, 14878-14883 (2002).

19. R. J. Kulathinal, B. R. Bettencourt, D. L. Hartl, Compensated deleterious mutations in insect genomes. *Science* **306**, 1553-1554 (2004).

20. M. Lunzer, G. B. Golding, A. M. Dean, Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* **6**, e1001162 (2010).

21. C. Natarajan *et al.*, Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324-1327 (2013).

22. V. A. Risso *et al.*, Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol Biol Evol* **32**, 440-455 (2015).

23. J. D. Bloom, L. I. Gong, D. Baltimore, Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328**, 1272-1275 (2010).

24. Z. D. Blount, C. Z. Borland, R. E. Lenski, Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proceedings of the National Academy of Sciences* **105**, 7899-7906 (2008).

25. J. T. Bridgham, E. A. Ortlund, J. W. Thornton, An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* **461**, 515-519 (2009).

26. L. I. Gong, M. A. Suchard, J. D. Bloom, Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2**, e00631 (2013).

27. E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, J. W. Thornton, Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544-1548 (2007).

28. M. L. Salverda *et al.*, Initial mutations direct alternative pathways of protein evolution. *PLoS Genet* **7**, e1001321 (2011).

29. 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. *Proceedings of the National Academy of Sciences* **115**, 4453-4458 (2018).

30. V. C. Xie, J. Pu, B. P. H. Metzger, J. W. Thornton, B. C. Dickinson, Contingency and chance erase necessity in the experimental evolution of ancestral proteins. *Elife* **10**, e67336 (2021).

31. M. J. Harms, J. W. Thornton, Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* **20**, 360-366 (2010).

32. J. M. Smith, Natural selection and the concept of a protein space. *Nature* **225**, 563-564 (1970).

33. S. Manrubia *et al.*, From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Phys Life Rev* **38**, 55-106 (2021).

34. D. M. Weinreich, Y. Lan, C. S. Wylie, R. B. Heckendorn, Should evolutionary geneticists worry about higher-order epistasis. *Curr Opin Genet Dev* **23**, 700-707 (2013).

35. Z. R. Sailer, M. J. Harms, High-order epistasis shapes evolutionary trajectories. *PLoS Comput Biol* **13**, e1005541 (2017).

36. Z. R. Sailer, M. J. Harms, Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps. *Genetics* **205**, 1079-1088 (2017).

37. F. J. Poelwijk, M. Socolich, R. Ranganathan, Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications* **10**, 1-11 (2019).

38. J. Domingo, G. Diss, B. Lehner, Pairwise and higher-order genetic interactions during the evolution of a tRNA. *Nature* **558**, 117-121 (2018).

39. 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 (2015).

40. D. W. Anderson, F. Baier, G. Yang, N. Tokuriki, The adaptive landscape of a metallo-enzyme is shaped by environment-dependent epistasis. *Nat Commun* **12**, 3867 (2021).

41. 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).

42. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537-1551 (2013).

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

44. K. S. Sarkisyan *et al.*, Local fitness landscape of the green fluorescent protein. *Nature* **533**, 397-401 (2016).

45. 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**, 111-114 (2006).

46. 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**, (2016).

47. J. Otwinowski, D. M. McCandlish, J. B. Plotkin, Inferring the shape of global epistasis. *Proceedings of the National Academy of Sciences* **115**, E7550-E7558 (2018).

48. F. J. Poelwijk, V. Krishna, R. Ranganathan, The Context-Dependence of Mutations: A Linkage of Formalisms. *PLoS Comput Biol* **12**, e1004771 (2016).

49. D. H. Brookes, A. Aghazadeh, J. Listgarten, On the sparsity of fitness functions and implications for learning. *Proc Natl Acad Sci U S A* **119**, (2022).

50. G. D. Stormo, Maximally efficient modeling of DNA sequence motifs at all levels of complexity. *Genetics* **187**, 1219-1224 (2011).

51. E. D. Weinberger, Fourier and Taylor series on fitness landscapes. *Biological cybernetics* **65**, 321-330 (1991).

52. A. C. Palmer *et al.*, Delayed commitment to evolutionary fate in antibiotic resistance fitness landscapes. *Nat Commun* **6**, 7385 (2015).

53. A. M. Phillips *et al.*, Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *Elife* **10**, (2021).

54. T. V. Lite *et al.*, Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *Elife* **9**, (2020).

55. Z. D. Blount, R. E. Lenski, J. B. Losos, Contingency and determinism in evolution: Replaying life's tape. *Science* **362**, eaam5979 (2018).

56. C. Natarajan *et al.*, Predictable convergence in hemoglobin function has unpredictable molecular underpinnings. *Science* **354**, 336-339 (2016).

57. M. Karageorgi *et al.*, Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature* **574**, 409-412 (2019).

58. G. K. Whitfield, P. W. Jurutka, C. A. Haussler, M. R. Haussler, Steroid hormone receptors: evolution, ligands, and molecular basis of biologic function. *Journal of cellular biochemistry* **75**, 110-122 (1999).

59. C. Helsen *et al.*, Structural basis for nuclear hormone receptor DNA binding. *Molecular and cellular endocrinology* **348**, 411-417 (2012).

60. 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).

61. P. H. Harvey, M. D. Pagel, *The comparative method in evolutionary biology* (Oxford university press Oxford, 1991).

62. L. Harmon, Phylogenetic comparative methods: learning from trees. (2018).

63. G. N. Eick, J. T. Bridgham, D. P. Anderson, M. J. Harms, J. W. Thornton, Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Molecular biology and evolution* **34**, 247-261 (2017).

64. I. S. Povolotskaya, F. A. Kondrashov, Sequence space and the ongoing expansion of the protein universe. *Nature* **465**, 922-926 (2010).

65. M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, F. A. Kondrashov, Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538 (2012).

66. D. D. Pollock, G. Thiltgen, R. A. Goldstein, Amino acid coevolution induces an evolutionary Stokes shift. *Proceedings of the National Academy of Sciences* **109**, E1352-E1359 (2012).

67. P. Shah, D. M. McCandlish, J. B. Plotkin, Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences* **112**, E3226-E3235 (2015).

68. Z. R. Sailer, M. J. Harms, Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences* **114**, 11938-11943 (2017).

69. A. J. Morrison, D. R. Wonderlick, M. J. Harms, Ensemble epistasis: thermodynamic origins of nonadditivity between mutations. *Genetics* **219**, iyab105 (2021).

70. J. T. Bridgham *et al.*, Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS biology* **8**, e1000497 (2010).

71. D. R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**, 821-829 (2008).

72. M. H. Schulz, D. R. Zerbino, M. Vingron, E. Birney, Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086-1092 (2012).

73. B. J. Haas *et al.*, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494-1512 (2013).

74. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

75. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

76. A. Weigert *et al.*, Illuminating the base of the annelid tree using transcriptomics. *Molecular biology and evolution* **31**, 1391-1401 (2014).

77. Z. Yang, PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-1591 (2007).

78. J. Zilliacus, A. P. Wright, U. Norinder, J.-A. Gustafsson, J. Carlstedt-Duke, Determinants for DNA-binding site recognition by the glucocorticoid receptor. *Journal of Biological Chemistry* **267**, 24941-24947 (1992).

79. C. J. Maclean *et al.*, Deciphering the genic basis of yeast fitness variation by simultaneous forward and reverse genetics. *Molecular biology and evolution* **34**, 2486-2502 (2017).

80. D. M. Fowler, J. J. Stephany, S. Fields, Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nature protocols* **9**, 2267-2284 (2014).

81. W. Wang, J. Yan, Shape-Restricted Regression Splines with R Package splines2. *Journal of Data Science* **19**, (2021).

82. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules* **22**, 2577-2637 (1983).

83. S. Guindon *et al.*, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321 (2010).

84. A. Hadzipasic *et al.*, Ancient origins of allosteric activation in a Ser-Thr kinase. *Science* **367**, 912-917 (2020).

85. S. L. Baldauf, J. D. Palmer, Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* **90**, 11558-11562 (1993).

86. J. del Campo, I. Ruiz-Trillo, Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. *Mol Biol Evol* **30**, 802-805 (2013).

87. P. J. Keeling, N. M. Fast, Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annu Rev Microbiol* **56**, 93-116 (2002).

88. S. M. Soucy, J. Huang, J. P. Gogarten, Horizontal gene transfer: building the web of life. *Nat Rev Genet* **16**, 472-482 (2015).

89. H. Philippe *et al.*, Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol* **9**, e1000602 (2011).

90. H. Brinkmann, M. van der Giezen, Y. Zhou, G. Poncelin de Raucourt, H. Philippe, An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst Biol* **54**, 743-757 (2005).

91. H. Shimodaira, M. Hasegawa, Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Molecular biology and evolution* **16**, 1114-1114 (1999).

92. P. A. Eyers, E. Erikson, L. G. Chen, J. L. Maller, A novel mechanism for activation of the protein kinase Aurora A. *Current Biology* **13**, 691-697 (2003).

93. E. Tomaštíková *et al.*, TPX2 protein of Arabidopsis activates Aurora kinase 1, but not Aurora kinase 3 in vitro. *Plant molecular biology reporter* **33**, 1988-1995 (2015).

94. N. Özlü *et al.*, An essential function of the C. elegans ortholog of TPX2 is to localize activated aurora A kinase to mitotic spindles. *Developmental cell* **9**, 237-248 (2005).

95. S. M. Hedtke, T. M. Townsend, D. M. Hillis, Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol* **55**, 522-529 (2006).

96. E. W. Sayers *et al.*, Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **48**, D9-D16 (2020).

97. W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

98. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

99. K. P. Schliep, phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592-593 (2011).