THE UNIVERSITY OF CHICAGO

THE GENETIC ARCHITECTURE AND EVOLUTIONARY CONSEQUENCES OF A TRANSCRIPTION FACTOR-DNA BINDING MAP

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

JAEDA E. J. PATTON

CHICAGO, ILLINOIS

JUNE 2025

Abstract

The genotype-phenotype (GP) map describes the association between genetic and phenotypic variation in a biological system. It delimits the phenotypic variation that a system can produce and its accessibility from any given genotype, thus determining the capacity of a system to evolve. The GP map is itself determined by the genetic architecture of the system—the set of rules by which genotype is transformed into phenotype. Deep mutational scanning (DMS) now allows for construction of large empirical GP maps for protein biochemical phenotypes, but studies to date have only studied variation in functions that exist in modern proteins; we still lack a comprehensive understanding of the ability of systems to produce functions that could theoretically exist but are not observed in nature. In this dissertation, I use DMS to interrogate the capacity of the steroid hormone receptor protein to produce specificity for all possible DNA substrates that vary at two functionally important nucleotide sites, most of which are not bound by modern steroid receptors. I perform this experiment in the background of two reconstructed ancestral steroid receptors, allowing me to compare the distribution of outcomes that could have been produced with the phenotypes that evolved in the proteins' descendants. I then analyze the genetic architecture of this system using a generalized linear modeling framework. I find that the ancestral GP maps were strongly biased in the specificities they were able to produce, and that these biases are congruent with the distribution of specificities seen in their descendants. I also find that the genetic architecture of binding and specificity is largely determined by interactions between individual amino acids and nucleotides. Overall, my work shows how the genetic mechanisms that produce phenotypic variation shaped the evolutionary history of a protein family.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I am lucky to have found such a wonderful community of colleagues, mentors, and friends in my time in the Thornton lab. Joe, thanks for taking me into your lab and supporting me in so many ways throughout my journey. You taught me how to be a better writer and communicator, a more thoughtful scientist, and most importantly, to listen to my intuition both inside and outside the lab. Gracias a mi pez, Santi, for agreeing to embark on this crazy journey with me. You kept it so real and professional throughout, and have been a great friend for the past seven years. Carlos, it's been a pleasure going through it with you, and I'm so proud of the scientists we've both become. Yeonwoo, you were such a kind and thoughtful mentor and friend, and some of my best memories in Chicago are of exploring the city with you as pandemic bike buddies. To the OGs, Mo and Georg, thank you for welcoming me into the lab with open arms as a young, wide-eyed first year student. And to the more recent Thorntonites—especially Ricardo, Max, and Emily—thanks for helping me through my geriatric PhD years, when I often just needed someone to complain and have fun with.

I had so many other great friends and colleagues in the UChicago genetics and Darwin communities who were a constant source of fun and support over the years—thanks especially to Shreya, Jennifer, Astra, Viv, Abhimanyu, Will, Paula, and Darren. Sue, thanks for being so on top of everything and for helping me out when I tore my ACL. To the members of the HG/GGSB DEI Committee, thanks for being a place to work on making our community better. To the Ducks softball team, thanks for the fun, beer, and burgers. To the University Chorus and Mollie Stone, thank you for creating a space for making music while learning about so many cultures and world traditions.

My friends outside of UChicago helped keep me tethered to the real world. Big shoutout

to the Crammies—I can't wait for all of the reunions, weddings, and babies to come. To my high school friends Ashaen and Jeena, thanks for still being rad after all these years. To Renni, you're the best friend I could ask for—thanks for listening to all of my relationship troubles, for putting up with my bullshit, and for being someone who always makes me laugh. And to Joe Parker, thanks for being such a kind and inspiring scientist and for giving me something to look forward to in my postdoc.

Kyra and Kian, thanks for not completely hating me even when I'm mean to you. You guys are cool. Mom and Dad, thanks for being there for me, and for raising me to be a proud, hardworking, honest, and independent scientist. I love you all.

Arvind, thank you for being a constant source of conversation, comfort, silliness, and inspiration these past few months. You are my favorite person, and I couldn't ask for someone better to start this next chapter with.

Chapter 1: Introduction

1.1 Genotype-phenotype maps

Nearly a century ago, Sewall Wright introduced the adaptive landscape as a conceptual tool for studying how the genetic composition of a population changes in response to selection pressures[1]. An adaptive landscape is a description of the association between genotype (or allele frequencies) and fitness in a biological system. Adaptive landscapes are made up of fitness "peaks" and "valleys," connected to one another via single-step mutations between genotypes. Adaptation can be thought of as climbing up an adaptive peak, similar to how a climber scales a mountain. The concept of adaptive landscapes has since been generalized to genotype-phenotype (GP) maps, in which fitness is replaced with a phenotype (or phenotypes) of interest[2]. GP maps have been used extensively in theoretical and empirical studies to understand how the genetic architecture of biological systems shapes their phenotypic evolution[2]. Although originally introduced in an adaptive framework, much of the power of Wright's metaphor has thus come from its utility as a tool of structuralist research programs, which emphasize the role of biological structure rather than adaptive value in interrogating the causes of evolutionary outcomes[3,4].

On the theoretical side, much of the research in GP maps and adaptive landscapes has concerned the effects of epistasis on the accessibility of fitness peaks. Epistasis here is defined as a deviation from genetic additivity—if the effects of two individual alleles are a and b, respectively, and the combination of the two alleles is different from $a + b$, then the alleles are said to interact epistatically. If two alleles that are both individually beneficial are deleterious in combination (termed reciprocal sign epistasis), then the presence of each allele restricts access to

the other[5]. In this way, epistasis can alter the accessibility of adaptive peaks. Classical theoretical work has conceived of epistasis as introducing randomness into the effects of mutations[6–9], which has led to an inference of epistasis as creating "rugged" fitness landscapes where there are many local peaks and traversal between them is difficult[7–11]. However, these models are agnostic to actual biological mechanism, making their validity in real adaptive landscapes uncertain.

One shortcoming of the adaptive landscape conception of evolution is the lack of consideration for neutral mutations. Maynard Smith formalized a more neutralist theory of evolution on GP maps in his "sequence space" model of protein evolution[12]. In this model, protein evolution proceeds through a network of genotypes separated by single-step mutations, all of which maintain protein function; deleterious mutations are purged by purifying selection. This theory has been elaborated into a "neutral network" theory of protein evolution by Wagner[13], in which sequence space is made up of networks of genotypes all connected by neutral mutations; neutral networks encoding different phenotypes are connected at certain access points, through which phenotypic transitions can occur (adaptively or themselves neutrally).

Within the past 20 years, GP maps have begun to be amenable to experimental characterization at increasingly large scales, enabling empirical examination of the role of GP map structure on evolutionary trajectories. Weinreich and colleagues introduced the strategy of characterizing combinatorially complete GP maps, in which all combinations of alleles for a subset of sites and states in a protein sequence are assayed for function[14]. This strategy allows for analysis of all possible mutational trajectories between a starting and ending genotype. The first combinatorially complete GP maps were small in scale, comprising usually five to six sites with biallelic sequence variation[14–18]. These revealed that GP maps tended to be highly epistatic, causing only a subset of mutational trajectories to be available under positive or purifying

selection.

More recently, the development of deep mutational scanning (DMS)[19] has enabled the characterization of combinatorially complete GP maps at much larger scale. DMS is an experimental strategy in which large libraries of genetic variants are phenotyped using deep sequencing-based assays. Combinatorial DMS studies have shown that while adaptive landscapes do indeed involve extensive epistasis, most genotypes are still accessible to one another via either adaptive or neutral mutational trajectories[20–28]. In particular, studies that include not just biallelic sequence variation but combinations of all 20 amino acids (or all 4 nucleotide bases in the case of nucleic acids) have shown that the high dimensionality of sequence space is key to high accessibility, since mutational paths that are inaccessible due to epistasis are often accessible via mutations to other possible sequence states[25,26,29].

Most combinatorial DMS studies map genetic variation to variation in only a single phenotype, such as binding to a particular substrate, but a few have examined accessibility between qualitatively different phenotypes by mapping variation in binding to multiple substrates (if the substrate bound is taken as the phenotype of a protein variant)[20,21,26]. These studies have found that, in line with neutral network theory, evolution must often proceed through a series of neutral steps before new phenotypes can be reached, and new phenotypes can often only be accessed via a few genetic intermediates. However, a shortcoming of these studies is that they only consider access to new phenotypes that have evolved in functionally or evolutionarily related proteins. This strategy cannot address the extent to which phenotypes that are not seen in existing proteins are evolutionary accessible, or whether they never could have been produced in the first place. Answering this question would address a deeper question in evolutionary biology about the role of phenotype production bias in shaping the phenotypes that

we observe in the world today.

1.2 Production bias

One of the most obvious features of biological diversity is that forms and functions are unevenly distributed across the tree of life. Photosynthesis, for example, is restricted to plants and some unicellular organisms, and powered flight to certain groups of animals. What explains the biased distribution of phenotypic diversity among taxonomic lineages? In the Neo-Darwinian tradition, selection is seen as the predominant force that imposes lineage-specific directionality in evolution[30,31]. However, more recent theoretical work has challenged this assumption, pointing to the mutation process as a critical orienting factor in evolution[32]. The argument goes that under conditions of weak mutation and strong selection[33]—satisfied in most natural populations except in special cases such as viruses—new alleles are introduced more-or-less sequentially into populations, so that most loci are fixed for one allele at any given time. If the mutation process is biased such that some alleles are produced more frequently than others, then this will in turn bias the alleles that actually go to fixation in the population, even in the face of natural selection—a maximally fit allele cannot fix if it is never produced in the first place.

Abundant studies in paleontology and evo-devo have provided evidence that biases in the production of variation may shape the distribution of phenotypes we observe in nature. Observations of extant and fossil taxa have found that morphological diversity often does not span the total range of theoretically possible forms[34]. One classic example is that of coiling morphology in animals with shells[35]—species belonging to different taxa tend to occupy difference regions of the "morphospace" of possible forms, suggesting that there may be taxon-specific aspects of development that favor the production of certain shell shapes, and some

regions of morphospace are not occupied at all, suggesting that these shapes may not be able to be produced at all. However, these patterns could also be due to selective constrains that make some shapes incompatible with survival, perhaps in lineage-specific ways. More direct evidence for the role of developmental biases comes from studies assessing the impact of direct experimental perturbations on phenotype. For example, Alberch and Gale showed that the same application of a mitotic inhibitor to the developing limb bud produced different patterns of digit loss in frogs and salamanders, which mirror evolutionary trends in digit loss in the two clades of amphibians[36]. The difference in response to perturbation is underlain by divergence of the developmental pathways specifying digit formation. More recently, mammalian tooth development has been used as a model system to study developmental bias, since the developmental pathways specifying tooth morphology are well understood. Computational studies have shown that perturbations of biochemical parameters of the system reproduce the distribution of tooth morphologies in extant species and in the fossil record[37,38]—the developmental architecture of the system shapes the distribution of phenotypes that can be produced by disturbing it.

A disadvantage of using morphology to study the phenotype production process is that the genetic basis of morphological phenotypes is often poorly understood and highly complex. It is therefore difficult to study how production bias manifests at the level of sequence space. Developmental pathways are composed of interactions between individual proteins, and the mapping of sequence to protein function is often highly degenerate—many sequences can often encode the same level of protein function, and by extension there may be many mutations that confer the same change in function. Furthermore, not all variation in phenotype is likely to be immediately accessible from any given genetic starting point, either due to the genetic code or to

epistasis. To study how the genetic architecture of biological systems influences their evolutionary potential, it is therefore essential to characterize phenotype production at the level of sequence space.

Studying molecular phenotypes—such as protein-ligand binding or enzyme function—can overcome this hurdle because the genotype-phenotype relationships are easier to characterize, and the space of relevant genetic variation is much easier to define. A series of pioneering studies into the distribution in sequence space of RNA secondary structures—which can be computationally predicted with high confidence—was the first to demonstrate the utility of this approach. These studies showed that the distribution of secondary structures that can be produced by randomly sampling RNA sequence space is highly biased—some structures are produced much more often than others[39]—and the production distribution correlates well with the distribution of structures observed in nature[40]. Additionally, sequences encoding the same secondary structure form vast neutral networks that are connected to other neutral networks encoding different structures[39]; mutations tend to convert between similar structures more often than highly distinct ones[41]. This work has thus provided evidence that the biased production and connectivity of phenotypes in sequence space has likely shaped the evolution of RNA secondary structures.

The work on RNA secondary structure has aided our understanding how molecular functions are distributed in sequence space, but these studies lack phylogenetic context, meaning that they cannot link the patterns they see in the GP map to patterns of lineage-specific phenotypic distributions observed in nature. To accomplish this goal, we can make use of ancestral sequence reconstruction (ASR), a phylogenetic technique that allows for statistical inference of the protein sequences that were the ancestors of modern-day proteins[42]. Ancestral

proteins can be synthesized and studied experimentally to characterize their biochemical properties, enabling empirical study of the historical evolution of protein phenotypes. By applying DMS to ancestral proteins, we can study the distribution of biochemical phenotypes that could have been produced by mutation during evolutionary history[26,43]; this distribution can then be compared to the actual phenotypes that evolved in the ancestral proteins' descendant lineages. If the possible and actual evolutionary outcomes are highly congruent, then it is likely that the production process played a large role in the extant phylogenetic distribution of protein phenotypes.

In Chapter 2 of this dissertation, I use this approach to ask how the GP map of an ancestral transcription factor-DNA binding system shaped the phenotypes that evolved in its descendants. I expand traditional combinatorial DMS methods to study how not only systematic variation in the protein sequence, but also in its DNA response element sequence affected the ability of the two molecules to bind each other to regulate gene expression. This is the first DMS study to ask how genetic variation in one molecule (the protein) affects binding to all possible genetic variants of a given class of substrate molecules (the DNA). By performing this experiment in two different ancestral protein backgrounds that precede and follow a historical change in DNA specificity, I show how unequal production and accessibility of DNA specificity phenotypes is caused by the structure these two GP maps, how these properties evolve between the two ancestral maps, and how they are sufficient to explain the phylogenetic distribution of DNA specificity phenotypes in their descendant lineages.

1.3 The genetic architecture of GP maps

What is the genetic basis of phenotypic variation in GP maps? In other words, what are the rules

that link genotype to phenotype in a system? With the wealth of DMS studies now available, considerable work has gone into developing mathematical models to learn these relationships from empirical data. Of particular interest is the question: how much and what forms of epistasis are needed to accurately describe genotype-phenotype relationships? This question has implications for both evolution and protein engineering—if genotype-phenotype relationships can mostly be described by additive (non-epistatic) effects summed across individual sequence states, then GP maps should be less rugged and therefore more navigable, and new protein sequences with desired functions should be easy to design. If, by contrast, epistasis is widespread, high-order (involving three or more sequence states), and/or of large magnitude relative to additive effects, then GP maps should be rugged and less navigable, and new proteins should be difficult to design.

Accumulating evidence suggests that molecular GP maps are surprisingly simple in their epistatic architectures. One source of this growing consensus is the development of more robust methods for modeling specific interactions between sequence states, termed specific epistasis[44]. All methods for detecting specific epistasis share the same general framework of linearly decomposing the effects of individual sequence states and their epistatic interactions on phenotype[11,45–47]. Specific epistasis can thus be partitioned into different "orders" of interaction between different numbers of sites (pairwise interactions, third-order, etc.)[11]. Different methods differ with respect to their assignment of a reference genotype and method of inference (direct calculation vs. regression)[46,47]. This methodological inconsistency has led to varying conclusions about the amount of specific epistasis present in molecular GP maps, with some studies finding evidence of extensive high-order (>2nd order) epistasis[20,21,23,24,48–50] and others arguing for simpler genetic architectures[22,47,51,52]. Recently, Park et al. demonstrated that one method, called

reference-free analysis (RFA), produces the most accurate estimates of epistasis in the face of noisy data and model truncation[47]. This robustness arises from the fact that effects at each interaction order are defined with respect to the mean predicted phenotype across all possible genotypes from lower-order effects; other methods define effects with respect to a given state or genotype at each interaction order, and are thus more sensitive to missing data and measurement noise for those genotypes or states. RFA inferred the simplest genetic architectures in a meta-analysis of 20 empirical combinatorial GP maps, with a median of 91% of phenotypic variance explained by main effects of amino acids and only 5% explained by pairwise interactions, compared to <20% combined for main and pairwise effects using the least accurate method. The finding of predominantly low-order epistatic landscapes is consistent with evidence from other studies that main and pairwise effects are sufficient to predict key protein properties, including structural contacts[53,54] and enzymatic function[55]. The extent and complexity of specific epistasis is thus likely to be less pervasive than previously thought.

Additional evidence for the simplicity of epistatic relationships comes from the wider adoption and development of methods for modeling nonlinear genotype-phenotype relationships, a phenomenon known as global or nonspecific epistasis[44]. Nonspecific epistasis arises when mutations act additively on an unobserved trait that maps nonlinearly to a higher-level observed trait, causing apparent nonadditivity. It often has a biological basis in nonlinear relationships between bulk molecular properties—such as stability[56–65], ligand affinity[51,62,66,67], cooperativity[67], or translation rate[68,69]—and molecular function or organismal fitness, but it can also arise from experimental sources, such as measurement bounds. While there is no consensus on best practices for modeling nonspecific epistasis, numerous studies have found that doing so can drastically reduce the complexity of specific epistasis inferred in GP maps[45,47]. In the same study

described above, Park et al. found that a simple sigmoid function—which accounts for measurement bounds and relationships between free energy and occupancy in two-state thermodynamic systems—captured on average >80% of observed epistasis across datasets[47]. The sigmoid may thus be a good general-use function for modeling nonspecific epistasis.

These findings suggest that the genetic architecture of molecular GP maps is dominated by additive effects of sequence states and nonspecific epistasis, an encouraging prospect for our ability to infer and interpret the mechanisms underlying genotype-phenotype relationships. They do not, however, mean that specific epistasis is unimportant for function or evolution. For example, a recent study that used RFA to infer the genetic architecture of protein specificity for two different DNA motifs found that while the majority of genetic variance was explained by additive effects, specific epistasis was necessary for almost all mutations that altered specificity[52]. A view of protein structure-function relationships that synthesizes the tension between epistatic simplicity and complexity is therefore needed.

The development of methods for inferring the genetic architecture of genotype-phenotype relationships also provides an opportunity to study the genetic basis of production bias in GP maps. Studies of developmental and molecular systems have shown how variation in the molecular architecture that specifies phenotypic output can shape the phenotypes produced by mutation—for example, genetic circuits with different regulatory architectures have been found to produce different patterns of gene expression upon mutation[70]. However, these studies only link phenotypic output to variation in biochemical parameters of the system, not to variation in genetic sequence. RFA can overcome this gap by allowing us to parse how individual sequence states and their interactions contribute to variation along multiple phenotypic axes—such as the ability of a protein to bind to different substrates. Further, the phenotypic “axis” can be encoded

as a variable in the RFA formalism, such that the additive effect of a substrate, as well as interactions between the substrate and sequence states in the protein, can be directly estimated[51,52]. This approach allows for partitioning of the effects of sequence variation in the protein into effects on nonspecific binding—the average effect of a sequence state across all substrates—and on specificity to individual substrates, which are given by the substrate-protein interaction effects. How nonspecific and substrate-specific effects relate to production bias in GP maps has not yet been examined.

In Chapter 3, I apply the RFA formalism to the combinatorial intermolecular DMS dataset generated in Chapter 2 to study the genetic basis of binding and specificity, and production bias, in a protein-DNA binding system. I decompose the additive and epistatic effects of variable sequence states in the protein and DNA to their effects on binding, examining the contributions of different orders of genetic effects, including intramolecular and intermolecular epistasis, to variation in binding and specificity. This work represents a considerable advance over previous studies examining the genetic architecture of specificity in DMS datasets, which only systematically varied the sequence of one binding partner. By using a dataset in which the sequences of both partners are varied systematically, I am able to pinpoint the epistatic interactions between individual amino acid residues and DNA bases that mediate specificity, and ask how the genetic architecture of the system shapes the space of specificity phenotypes that can be generated by variation at the mutated sites.

1.4 Ancestral sequence reconstruction and phylogenetic error

The experiments in Chapter 2 are performed in the background of reconstructed ancestral proteins, allowing characterization of how historical production biases influenced the distribution

of extant protein phenotypes. ASR is a statistical technique which takes as input a multiple sequence alignment (MSA), a phylogeny that relates the sequences in the MSA, and a model of sequence evolution, and produces a maximum likelihood estimate of the sequences that existed at ancestral nodes in the phylogeny. The accuracy of the inferred ancestral sequences relies on the accuracy of the three inputs. In particular, the phylogeny should represent our best hypothesis as to the history of gene duplications, losses, and speciation events that produced the proteins in the MSA. Tree topologies are themselves inferred using the MSA and an evolutionary model and can be subject to phylogenetic error. One common cause of phylogenetic error is poor sequence sampling; this can lead to a phenomenon termed long-branch attraction (LBA)[71] where sequences with very little detectable homology at phylogenetically informative sites are inferred as closely related to each other, often with high statistical support[72]. LBA can be mitigated by improving sequence sampling to break up long branches in the phylogeny[73], or by constraining the tree topology to reflect accepted taxonomic relationships.

In Chapter 4, I examine a case in which LBA led to the inference of non-historical ancestral protein sequences and erroneous inference of the gain of allostery in a eukaryotic protein family. I show that by either constraining the tree topology or improving sequence sampling to break up long branches, the inferred history of allosteric evolution is reversed—the states that confer allostery are present in the eukaryotic common ancestor, and the states that do not confer allostery are restricted to a subgroup of Fungi; allostery was most likely present ancestrally and lost in this clade. This study highlights the importance of broad sequence sampling and rigorous evaluation of tree topology when performing ASR.

1.5 Author contributions

The experiments, analyses, and writing for Chapters 2 and 4 of this dissertation were conducted in collaboration with other graduate student members of the Thornton lab. For Chapter 2, I jointly conceived of the project, conducted the experiments, performed the analyses, and wrote the manuscript together with Santiago Herrera-Álvarez. For Chapter 3, I performed the analyses under the mentorship of Yeonwoo Park, and we wrote the manuscript together. In the published versions of these chapters, I am listed as co-first author, with author order determined by either last name or by coin flip. The analysis and writing for Chapter 3 were performed on my own.

Chapter 2: Ancient biases in phenotype production drove the functional evolution of a protein family

2.1 Summary

Biological systems are biased in the phenotypes that they can produce by mutation, but the nature of these biases and their causal role in the evolution of extant phenotypic diversity remains unclear. There are two major challenges: 1) It is difficult to isolate the effect of the genotype-phenotype (GP) map from that of natural selection in causing natural patterns of diversity, and 2) most extant phenotypes evolved long ago in species whose GP maps cannot be recovered. Using reconstructed ancestral transcription factors as a model to address this problem, we created libraries containing all possible amino acid combinations at historically variable sites in the proteins' DNA binding interface (the genotypes) and measured their capacity to bind to response elements containing all possible combinations of nucleotides at historically variable sites in the DNA (the phenotypes). The ancestral proteins we used existed during an ancient phylogenetic interval when a new phenotype—specificity for a new response element—evolved. We found that the two ancestral GP maps were strongly anisotropic—encoding phenotypes with unequal frequencies—and heterogeneous in the phenotypes accessible around each genotype. In each map, the particular forms of these properties steered evolution toward the lineage-specific phenotypes that evolved during history. Our findings establish that ancient properties of the GP relationship were causal factors in the evolutionary process that produced the present-day patterns of functional conservation and diversity in this protein family.

2.2 Introduction

Countless conceivable lifeforms have evolved rarely or never, and those that exist are mostly restricted to specific lineages[35,74–76]. No flying vertebrates have two pairs of wings, for example, and no turtles or frogs fly. What explains the patchy distribution of phenotypes in nature? Classical explanations focus on the influence of selection[77,78], but the propensities of biological systems to produce phenotypic variation could also shape evolutionary outcomes. A phenotype can become fixed in an evolving population only if it is first generated by mutation. If biological systems are more likely to produce some phenotypes than others[79–85], and if these propensities change over time as lineages diverge[34,86], then some phenotypes will be more likely to evolve in some taxa than in others, even in the absence of selection.

Most patterns of phenotypic variation observed in nature could arise from production biases, natural selection, or both, and disentangling their past influences is challenging[84,87–90]. Ideally, we would isolate the phenotype production process by directly characterizing the complete genotype-phenotype (GP) map, which maps all possible combinations of mutations to the phenotypes they encode. This would allow us to precisely quantify the ability of a genetic system to produce phenotypic variation, both on a global scale and by mutation from each particular genotype. But the number of possible genotypes and phenotypes is far too vast, so characterizing GP maps requires reducing its dimensionality while maintaining combinatorial completeness. We reasoned that this goal should be achievable for functionally important sites within a protein and for particular biochemical phenotypes that mediate the protein's biological functions. The scope of relevant genetic variation can be defined as all combinations of all 20 amino acid states at sequence sites that determine the protein's phenotype of interest[27,25,26,22,91]. This kind of combinatorial library of protein variants can be engineered and experimentally

characterized using deep mutational scanning (DMS)[19]. Combinatorial DMS studies to date have assayed only one or a few phenotypes that exist in extant proteins[91], so they cannot directly address the role of production bias in evolution because they leave out alternative possible phenotypes that might have evolved but which we do not observe; to understand why evolution turned out as it did, we must characterize the propensity of mutations to produce all possible phenotypes that a system could conceivably produce. The scope of biochemical phenotypes can be delimited as the ability of a protein to bind all possible substrates of a defined class—for a transcription factor this would entail all response elements produced by all combinations of nucleotides at defined sites[92,93,17,94,95]. A complete combinatorial GP map within defined scopes could thus be achieved by expanding DMS experiments to profile all relevant substrates.

The phenotypes of extant lineages evolved long ago, so understanding the causal role of phenotype production in historical evolution requires GP maps to be characterized as they existed in the deep past. Ancestral protein reconstruction[42] can address this problem by estimating the sequences of ancient proteins using statistical phylogenetic methods; the reconstructed protein can then be used as the background for a DMS experiment. Characterizing GP maps across a phylogenetic time series of reconstructed ancestral proteins[26,43] would reveal how biases in phenotype production may have changed over time and whether these biases are congruent with the trajectories of phenotypic evolution that actually unfolded during history.

Here we apply this approach to assess how the structure of the GP map shaped functional diversification in the steroid hormone receptor protein family. We experimentally characterize GP maps of the binding interface of two reconstructed ancestral steroid hormone receptor DNA binding domains (SR DBDs) and their ability to encode specific recognition of all possible DNA response element sequences, with the scope limited to sites in the protein-DNA interface that

varied during historical evolution. We then analyze these maps to understand 1) how they could shape potential phenotypic outcomes of evolution on short and long timescales, 2) characterize the mechanisms that changed key features of the maps across evolutionary time, and 3) assess the impact of the maps on the historical evolutionary processes that yielded the lineage-specific patterns of DNA specificity in extant steroid hormone receptors.

2.3 Results

2.3.1 Two complete ancestral GP maps

SRs are a family of transcription factors that regulate physiological and reproductive biology in bilaterian animals. Most bilaterian taxa have a single SR, which specifically binds to inverted palindromes of the motif AGGTCA, called the estrogen response element (ERE; Fig. 2.1A). In chordates, a gene duplication of the ancestral SR (AncSR1) produced two major SR classes, which have different DNA specificity phenotypes: chordate estrogen receptors (ERs) retain the ancestral ERE specificity, but a novel specificity for a palindrome of AGAACA, called the steroid response element (SRE), evolved in the lineage leading to AncSR2, the common ancestor of the chordate ketosteroid receptors (kSRs; Fig. 2.1A, B)[67]. Specificity for DNA is determined primarily by the amino acid sequence of a recognition helix (RH) that binds in the DNA major groove[96,97]. AncSR1 and AncSR2 DBDs differ by 34 amino acid replacements, but experiments on the reconstructed proteins established that three amino acid changes in the RH were the primary cause of the evolution of SRE specificity[67].

To understand how phenotype production may have shaped the evolution of SR-DBD specificity, we characterized combinatorially complete GP maps of the DBD-response element (RE) interface at the key ancestral timepoints AncSR1 and AncSR2. The scope of genotypes is

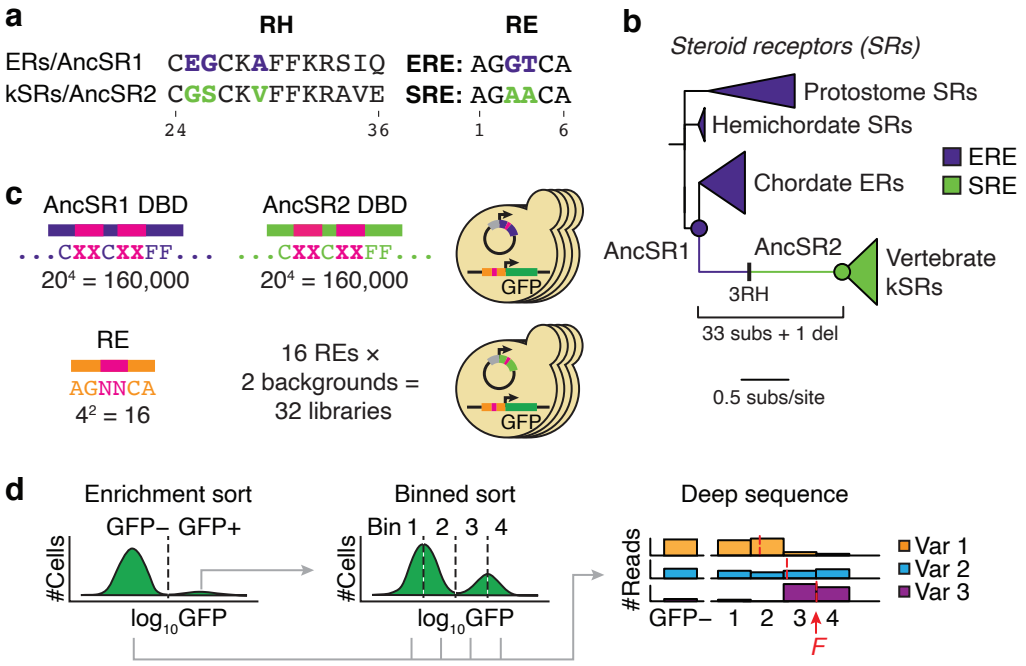

Figure 2.1. Characterizing ancestral GP maps using multi-phenotype DMS. (**A**) Amino acid sequence of the recognition helix (RH) in extant and ancestral steroid receptor (SR) proteins and the sequence of the RE they bind to. Colored residues are responsible for differences in protein-RE specificity. (**B**) Phylogeny of SRs. Each clade of proteins is colored by the RE sequence it recognizes. In chordates, a historical transition from ERE to SRE specificity occurred along the branch between AncSR1 (the common ancestor of all chordate SRs) and AncSR2 (the common ancestor of vertebrate kSRs). The number of historical sequence changes along the AncSR1-AncSR2 branch is shown; three of these in the recognition helix (RH) caused the specificity switch[67]. (**C**, **D**) DMS experiment to assay effects of RH genotype on binding to variable REs. (**C**) We built combinatorial libraries of all combinations of 20 amino acid states at four variable sites in the RH (pink Xs), using the rest of the AncSR1 and AncSR2 DBDs as backgrounds (top left). These were transformed into 16 S. cerevisiae strains, each containing one of the 16 possible RE motifs (pink Ns, bottom left) genomically integrated upstream of a GFP reporter gene (right). (**D**) We assayed binding of DBD-RE complexes using FACS coupled with deep sequencing. For each library, we performed an initial enrichment sort to select for GFP+ cells. We then grew up the selected cells, pooled them across the 32 libraries, and resorted them into four fluorescence bins in triplicate (binned sort). Sorted cells were deep sequenced to estimate the mean $\log_{10}$GFP (F) of each combination of protein and RE genotypes.

all possible $20^4 = 160{,}000$ amino acid variants at four variable sites in the recognition helix—the three that changed between AncSR1 and AncSR2, plus one other that varies in the broader nuclear receptor family (Fig. 2.1C). The scope of specificity phenotypes consists of all $4^2 = 16$ possible RE sequences that can be produced by all combinations of nucleotides at the two base

positions that vary between ERE and SRE; although extant SRs are not known to bind to many of these REs, there is no a priori biophysical reason that any of them should be impossible to bind. These two maps of the recognition helix-RE interface can be thought of as submaps within the much larger GP map of the entire DBD, which are connected by the 31 other "background" substitutions that occurred between the AncSR1 and AncSR2 proteins (Fig. 2.1B).

We engineered two protein libraries, each containing all 160,000 variants of the recognition helix in the background of either the AncSR1 or AncSR2 DBD, along with 16 yeast strains, each containing a GFP reporter driven by one of the REs (Fig. 2.1C, Fig. A.1A–E). We transformed each RE strain separately with the two protein libraries, with barcodes to mark the strain and the ancestral background, for a total of 5.12 million protein-DNA complexes. We used an initial round of fluorescence-activated cell sorting to enrich the yeast libraries for GFP-positive cells, pooled the enriched libraries, sorted cells in three replicates by their fluorescence, and sequenced the sorted bins (Fig. 2.1D, Fig. A.1F, G). Using this strategy, we obtained empirical fluorescence estimates for the majority of complexes with good replicability ($r^2 = 0.92$ across replicates, excluding complexes at the lower bound of fluorescence; Fig. A.2). Fluorescence of the remaining complexes was predicted using a generalized linear model trained on the experimental data (Methods, Fig. A.3A–D)[47,52].

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

2.3.2 Anisotropy in the AncSR1 GP map

The probability that a phenotype will evolve equals the probability that it will be produced by mutation times the probability that, once produced, it will be fixed. If some phenotypes are more likely to be produced than others, then the GP map will bias evolutionary outcomes relative to the null expectation that all phenotypes are equally likely to evolve[80]. The GP map would not bias evolutionary outcomes if and only if it had two properties: isotropy—all phenotypes are encoded with equal probability—and homogeneity—all starting genotypes in the map produce the same distribution of phenotypes by mutation[98–101]. Anisotropy will make those phenotypes that are more likely to be produced more likely to evolve; heterogeneity will cause the probability that each phenotype will be produced—and hence evolve—to change as lineages diverge from each other across the map.

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

Anisotropy in the AncSR1 map imposes hard limits on phenotypic evolution. The majority of the 16 possible phenotypes could never evolve in this map, even if they conferred

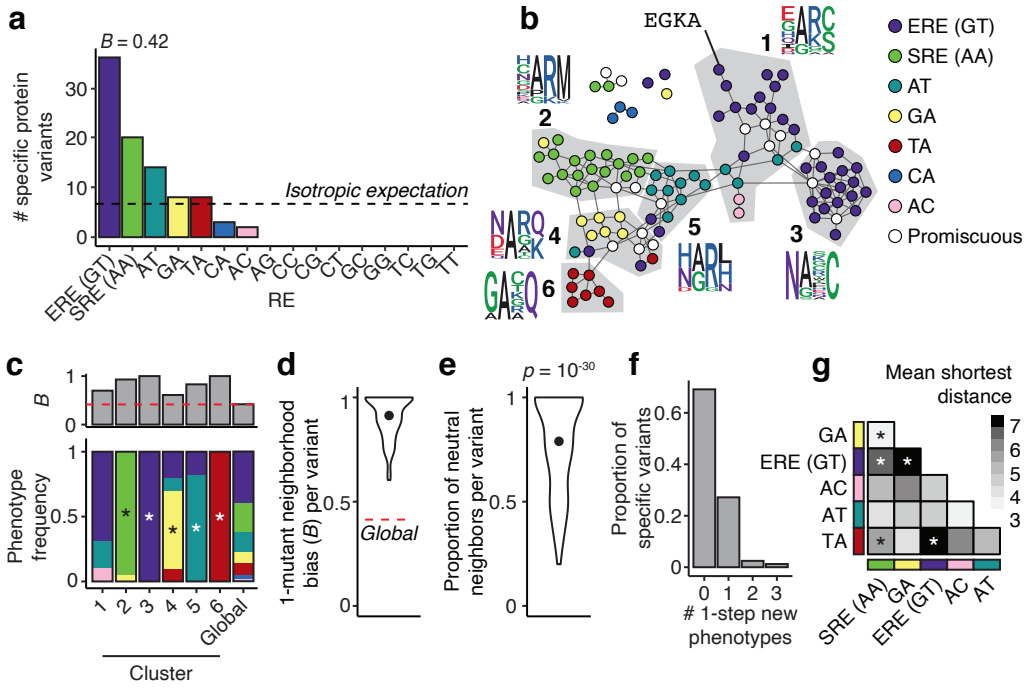

Figure 2.2. Global and local bias in the AncSR1 GP map. (**A**) Global production distribution in the AncSR1 GP map. Bars represent the number of protein variants that bind specifically to each RE. The dashed line shows the expected frequencies if the distribution were unbiased. *B,* phenotype bias,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

strong fitness advantages. The direction of anisotropy is also congruent with evolutionary

history: the phenotypes that evolved historically in the two lineages descending from AncSR1 are also the most frequently encoded.

2.3.3 Heterogeneity in the AncSR1 GP map

We next assessed the homogeneity of the AncSR1 GP map using Maynard-Smith's classic network model of sequence space[12]. Each functional protein variant is a node with its experimentally defined phenotype. Nodes are connected by edges if their amino acid sequences can be interconverted by a single nucleotide change given the standard genetic code. Nonfunctional variants are excluded from the network, based on the assumption that they will be removed quickly from evolving populations by purifying selection.

We found that the distribution of phenotypes in AncSR1 sequence space is strongly heterogeneous. Although the majority of functional genotypes (91%) and phenotypes (6 of 7) are mutually connected in a single main network component, each phenotype tends to be sequestered in a local region (Fig. 2.2B). Using a community structure detection algorithm[102], we found that the main network component can be partitioned into six clusters of genotypes that have dense connectivity within clusters and weak connectivity between (Fig. 2.2B). The phenotype bias B within every cluster is higher than the bias of the global production distribution, and 5 of 6 clusters are significantly enriched for a single specificity phenotype, which differ among all 5 clusters (Fig. 2.2C). The clumpy distribution of phenotypes in sequence space arises from the simple fact that similar genotypes, which are connected to each other in sequence space, are likely to encode similar phenotypes (Fig. 2.2B, logos).

Because of this heterogeneity, the propensity to produce phenotypes depends strongly on the particular genotype occupied at the time. The one-mutant neighborhood around every

genotype has extremely high bias (mean $B = 0.91$; Fig. 2.2D), indicating that individual genotypes can access much less phenotypic variation than is encoded across genotype space as a whole. Most mutations are phenotypically neutral (79% of edges; Fig. 2.2E), and most genotypes can directly access at most one new phenotype (Fig. 2.2F). The historical starting genotype (EGKA), for example, has access to only one functional neighbor, which also has ERE specificity. Another consequence of heterogeneity is that phenotypes, aggregated over the genotypes that encode them, are differentially accessible to each other, with substantial variation in the number of mutations required to transform each phenotype into the others (Fig. 2.2G). For example, SRE-specific protein genotypes are directly accessible from nodes encoding specificity for AT and GA, but they are multiple substitutions away from all ERE-specific genotypes, which can reach SRE-specificity only via intermediates with these phenotypes (Fig. 2.2B).

The GP map of AncSR1 is therefore strongly anisotropic and heterogeneous, and these properties impose strong biases on the production of phenotypes. Anisotropy across the GP map as a whole favors production of the historical phenotypes ERE and SRE and entirely prevents the production of most conceivable phenotypes. Heterogeneity further restricts the number of accessible phenotypes from each particular genotype, favoring conservation over the evolution of new phenotypes, including conservation of ERE specificity from the historical genotype EGKA.

2.3.4 The GP map shapes phenotypic outcomes of evolution

To characterize how the structural properties of the AncSR1 GP map could influence the outcomes of evolution, we modeled evolution on the network of functional amino acid genotypes as a discrete-time Markov chain from every possible starting genotype given a variable trajectory length. Each time-step in a trajectory is an amino acid substitution, the probability of which is

weighted by the number of single-nucleotide mutations that can mediate it; the relative probability of evolving a given phenotype at the end of the trajectory is the sum of the probabilities of evolving all genotypes that encode it. This model, in which all functional phenotypes have equal fitness, corresponds to neutral molecular evolution with purifying selection[12,103]: it represents a null scenario in which the fixation process imposes no biases on evolutionary outcomes except to prevent the loss of function. Using this model allows us to isolate the influence of phenotype production propensities imposed by the GP map's structure.

We first computed the equilibrium distribution of phenotypic outcomes after an infinite number of substitutions. This represents the limiting case at which the distribution of outcomes is insensitive to the starting genotype and does not change with additional substitutions. The equilibrium outcome distribution is well correlated with the global frequency distribution of phenotypes (Fig. 2.3A, $r^2 = 0.82$), reflecting the map's anisotropy. However, there are differences: the equilibrium distribution is more biased than the global production distribution ($B = 0.46$), and whereas ERE and SRE specificity are the two most frequently encoded phenotypes, ERE and AT specificity are the most likely equilibrium outcomes (Fig. 2.3A). This difference arises because most AT-specific genotypes are located centrally within the network, while SRE-specific genotypes are in a more peripheral cluster (Fig. 2.2B) and are therefore less likely to be occupied. The heterogeneous connectivity of functional genotypes and anisotropy in phenotype production therefore create bias in evolutionary outcomes, even over infinitely long timescales.

On finite timescales, heterogeneity strongly affects evolutionary outcomes. After 3 substitutions, for example—the shortest path between the historical ancestral and derived genotypes—the outcome distributions are very strongly biased (mean $B = 0.8$ across starting genotypes, Fig. 2.3B), because most genotypes can reach only a few new specificity phenotypes

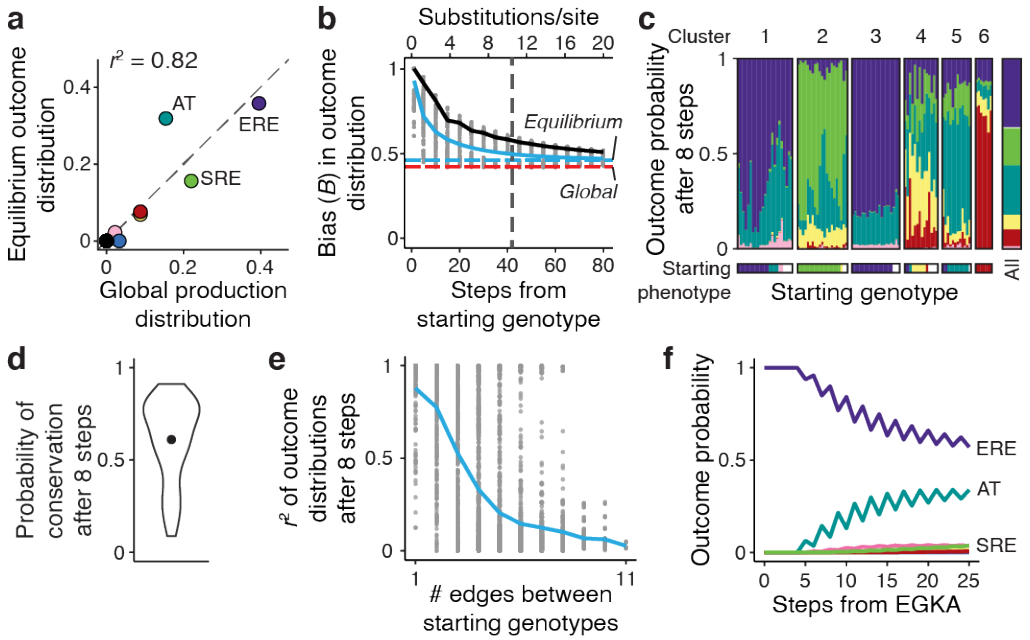

Figure 2.3. The AncSR1 GP map biases evolutionary outcomes towards phenotype conservation. (**A**)Comparison between the global production distribution and the long-term equilibrium distribution of phenotypic outcomes in the AncSR1 main network. Each dot shows the frequency of one specificity phenotype in the two distributions. Black dot at the origin represents nine phenotypes not encoded in the map. Dashed gray line, $y = x$. Squared Pearson's correlation coefficient is shown. (**B**) Strength of bias (B) in evolutionary outcomes as a function of the length of evolutionary trajectories. Each gray dot shows the B of the outcome distribution for trajectories of a given number of substitutions starting from one node on the main network component. Solid blue and black lines show the mean across all starting genotypes and from EGKA, respectively. Dashed horizontal red and cyan lines show B of the global production distribution and the equilibrium distribution, respectively. Vertical dashed line shows the number of substitutions required for mean B to reach within 0.05 units of the equilibrium value. The secondary x-axis (above) shows the trajectory length as substitutions per site. (**C**) Distribution of evolutionary outcomes after 8 substitution steps from every starting genotype in the AncSR1 main network component, organized by the cluster of the starting genotype (top). Bottom bar shows the phenotype of each starting genotype. Bars at right show the average outcome distribution for all starting genotypes. (**D**) Distribution of the probability of phenotype conservation after 8 substitution steps across all specific starting genotypes in the AncSR1 main network component. Dot shows the mean. (**E**) Evolutionary outcomes become less similar as starting genotypes diverge from each other. Each dot shows the similarity of the distributions of phenotypic outcomes (Pearson's r^2) of 8-step trajectories starting from a pair of genotypes, versus the number of network edges between the pair. Blue line, mean similarity across all pairs of starting genotypes. (**F**) Probability of evolving each specificity phenotype starting from EGKA as a function of the number of substitutions.

by a path of this length (Fig. A.4). The bias in outcomes gradually decays as trajectories get

longer, but it takes 42 substitutions (10.5 per site) for the mean bias to decrease to within 0.05

units of the equilibrium (Fig. 2.3B, vertical dashed line). By comparison, the maximum root-to-tip branch length in the steroid receptor DBD phylogeny (Fig. 2.1B), which spans over 500 million years of evolution, is just 2.2 substitutions per site. The phenotypes likely to evolve on phylogenetically relevant timescales are therefore strongly biased by heterogeneity in the GP map.

Another consequence of heterogeneity is that outcomes are strongly contingent on the genetic starting point. Consider a trajectory length of 8 substitutions—long enough for new phenotypes to become accessible from most starting points, but not so long that the influence of local heterogeneity is lost. At this timescale, genotypes differ dramatically in the distribution of phenotypes that evolve from them (Fig. 2.3C). Much of this variation is explained by the genotype cluster to which the starting node belongs (Fig. 2.3C), because evolutionary trajectories rarely jump between weakly connected clusters and clusters are strongly enriched for individual phenotypes. Even at this timescale, the structure of the GP map favors conservation of the starting phenotype on average (Fig. 2.3D), but when new phenotypes evolve, these too differ strongly among starting genotype (Fig. 2.3C).

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

as they traverse the GP map.

2.3.5 The AncSR1 GP map favored historical conservation of ERE specificity

Anisotropy and heterogeneity have a very strong and long-lasting impact on the outcomes of evolutionary trajectories that begin from the historical genotype of the recognition helix in AncSR1 (EGKA). It takes 80 substitutions for the bias in phenotypic outcomes from this starting point to decay to within 0.05 units of equilibrium, almost double the average across genotypes (Fig. 2.3B, blue vs. black solid lines). It takes at least 5 substitutions for any new specificity phenotype to be accessed, and even after 8 substitutions the probability of conserving ERE specificity is still 0.90 (Fig. 2.3F). The AncSR1 GP map heavily favors phenotypic conservation from the historical starting genotype across phylogenetically relevant timescales. Biases in phenotype production imposed by the GP map are therefore congruent with the long-term historical conservation of ERE specificity in the lineages that descend from AncSR1 and lead to modern-day estrogen receptors.

The historical outcome that evolved in AncSR1's other descendant linage—acquisition of SRE specificity in the kSR clade—was very unlikely on phylogenetic timescales. SRE-specific genotypes are distant from EGKA (Fig. 2.2B), so the probability of evolving SRE specificity after eight substitutions is only 0.0008 (Fig. 2.3F), despite the fact that this is the second-most frequently encoded specificity phenotype in the network overall. The only specificity with moderately high probability of evolving at this timescale is AT (Fig. 2.3F). Heterogeneity in the genetic neighborhood around EGKA therefore overrides the global propensity for SRE specificity, making the historical outcome that occurred in the kSR clade extremely unlikely.

2.3.6 Evolution of a different GP map in AncSR2

Given that SRE specificity was unlikely to be produced by mutation from the ancestral genotype in the AncSR1 map, how could this phenotype have historically evolved in the kSRs? We reasoned that the GP map must have changed along the branch leading to AncSR2 on which SRE specificity was acquired. Previous experiments showed that the background substitutions that occurred outside the recognition helix during this interval had a nonspecific permissive effect on both ERE and SRE activation, which allowed the protein to tolerate the historical substitutions and other mutations in the RH (Fig. 2.1B)[26,67]. We predicted that the background substitutions had a similarly permissive effect across all REs, increasing the number of functional genotypes in the map and the number of phenotypes they encode, including SRE specificity and others.

To assess this hypothesis, we characterized the GP map of the RH sites in AncSR2 and compared it to the map in the AncSR1 background. As predicted, the number of functional genotypes and phenotypes both massively increased (Fig. 2.4A, B). There are 2,407 functional protein genotypes in the AncSR2 map, an increase of >20-fold over the AncSR1 background. Fourteen of the 16 possible specificity phenotypes are now encoded, twice as many as in the AncSR1 background (Fig. 2.2A, 2.4A). The background substitutions therefore dramatically expanded the functional genetic and phenotypic variation that can be produced within the recognition helix.

Connectivity between genotypes in the map increased, reducing the impact of heterogeneity and facilitating access to new phenotypes. In the AncSR2 network, all but five of the 2,407 functional nodes are connected in a single main component (Fig. 2.4B), and the mean number of edges per node is 10.7, a three-fold increase compared to the AncSR1 network (Fig. 2.4C). Genotype clusters are still present, but bias within clusters is weaker than in the AncSR1

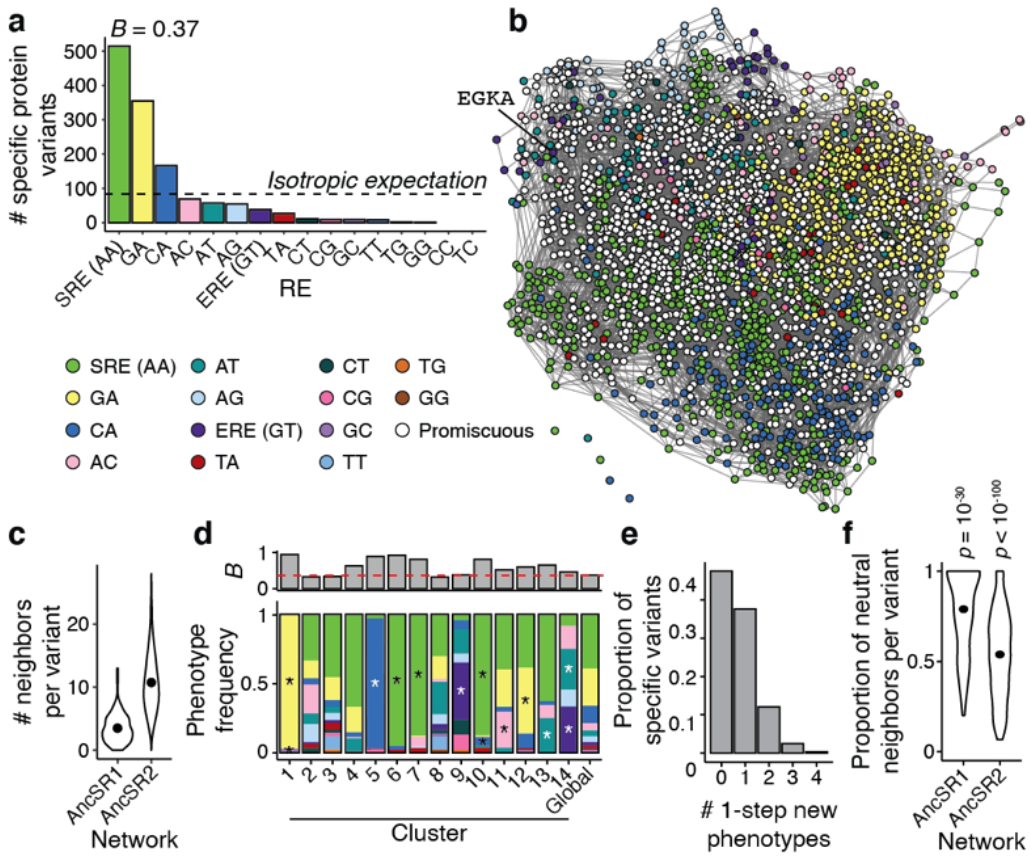

Figure 2.4. Global and local bias and connectivity changed in the AncSR2 GP map. (**A**) Global production distribution and global *B* of the AncSR2 GP map. (**B**) Sequence space network of the AncSR2 GP map. (**C**) Number of one-step neighbors per protein variant in each network. Dots show the mean of each distribution. (**D**) Bottom: Frequencies of specificity phenotypes within each genotype cluster (1–14, ordered by decreasing size); the global production distribution is shown for comparison. Only the 14 largest clusters, which contain >90% of genotypes, are shown. Asterisks, phenotypes significantly enriched within a cluster relative to the global production distribution (one-sided Fisher's exact test, $p < 0.05$ after Bonferroni correction). Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. (**E**) Distribution of the number of new phenotypes accessible within one mutation, across all RE-specific protein variants in the AncSR2 main component. (**F**) Proportion of neutral neighbors per RE-specific variant in the main network component of the AncSR1 and AncSR2 maps. Dots show the mean. *p-value,* probability that the mean would be at least as great as observed if phenotypes were randomly reassigned in the main component of each map (AncSR1 $n = 91$, AncSR2 $n = 2{,}402$).

map (Fig. 2.2C, 2.4D). As a consequence, genotypes have more access to new phenotypes: >50% of genotypes in the AncSR2 map can access between 1 and 4 new phenotypes within a single mutation (Fig. 2.4E, compare to Fig. 2.2E), because genotypes are typically connected to far

more non-neutral neighbors (Fig. 2.4F).

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

2.3.7 The AncSR2 GP map favored evolution of SRE specificity

These changes in the AncSR2 GP map dramatically altered the likely phenotypic outcomes of evolution. At long-term equilibrium using our Markov model and the AncGR2 map, the most likely evolutionary outcome is now SRE specificity, with a probability close to 40% (Fig. 2.5A, compared to <20% in the AncSR1 map). At moderate timescales as well, SRE specificity is the most likely outcome across the majority of starting genotypes (Fig. 2.5B). The probability of evolving new phenotypes overall is considerably higher in the AncSR2 network compared to AncSR1 (mean probability of conservation after 8 steps 0.47 in AncSR2 but 0.61 in AncSR1, Fig. 2.3D, 2.5C).

These changes in evolutionary outcomes are attributable to the increased connectivity of the AncSR2 network and the shift in the global production distribution. From any starting point, the increase in functional nodes and connectivity allows access to far more genotypes and phenotypes (Fig. 2.5D). As a result, the influence of local heterogeneity is lost faster, and trajectories more rapidly converge on the equilibrium distribution (Fig. 2.5E), which more closely resembles the production distribution than in the AncSR1 background (Fig. 2.5A).

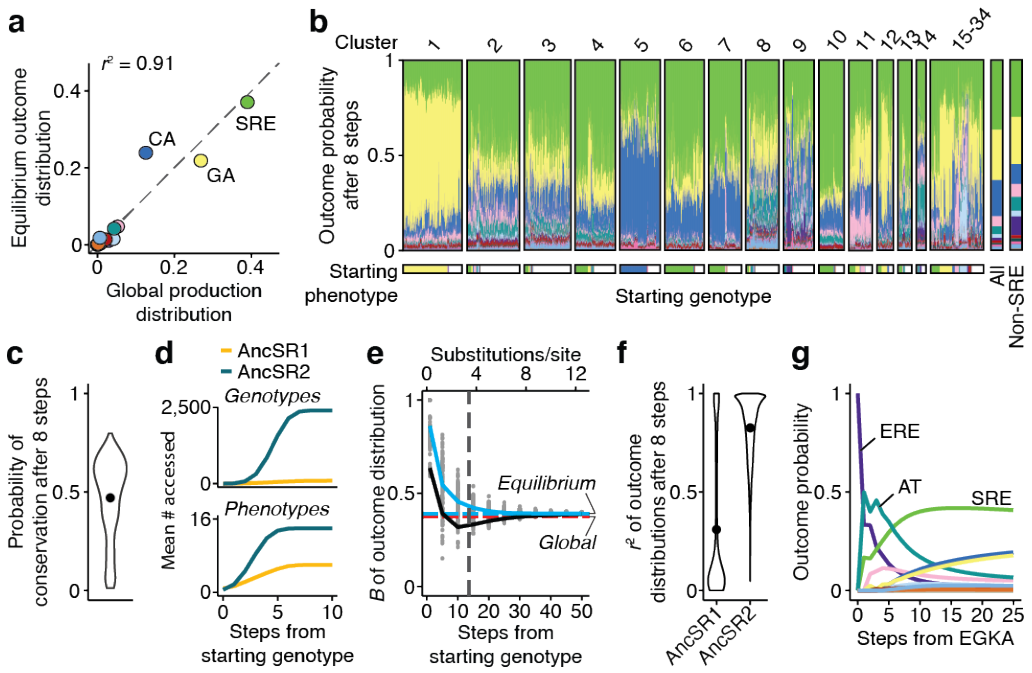

Figure 2.5. The AncSR2 GP map biases evolutionary outcomes towards SRE specificity. (**A**) Comparison between the global production distribution and the long-term equilibrium distribution of phenotypic outcomes in the AncSR2 main network. Dashed gray line, $y = x$. (**B**) Distribution of evolutionary outcomes after 8 substitution steps from every starting genotype in the AncSR2 main network component, organized by the cluster of the starting genotype (top). Bottom bar shows the phenotype of each starting genotype. Bars at right show the average outcome distribution for all starting genotypes and all non-SRE-specific starting genotypes, respectively. (**C**) Distribution of the probability of phenotype conservation after 8 substitution steps across all specific starting genotypes in the AncSR2 main network component. Dot shows the mean. (**D**) Number of genotypes (top) and phenotypes (bottom) accessible as a function of the length of evolutionary trajectories. Lines show the mean across all starting genotypes in each network. Gold, AncSR1 network; teal, AncSR2 network. (**E**) Strength of bias (B) in evolutionary outcomes as a function of the length of evolutionary trajectories. Lines and colors are the same as in Fig. 2.3B. (**F**) Distribution of the similarity in outcome distributions (Pearson's r^2) for 8-step trajectories starting from all pairs of genotypes in the AncSR1 and AncSR2 main networks. Dots show means. (**G**) Probability of evolving each specificity phenotype starting from EGKA as a function of the number of substitutions.

Evolutionary outcomes are also more similar across pairs of starting points than they were in the AncSR1 map (Fig. 2.5F). Combined with the shift in the global production distribution, this causes SRE specificity—which was already the second-most likely outcome in the AncSR1 map—to become the most likely outcome from a majority of starting points in the AncSR2 background.

From the historical RH genotype EGKA (the AncSR2 protein with the RH states reverted to their ancestral states), the likely outcomes of phenotypic evolution are dramatically different than in the AncSR1 map. EGKA is much less mutationally isolated in the AncSR2 network, so the probability of conserving ERE specificity after 8 substitutions drops from 0.9 in the AncSR1 map (Fig. 2.3F) to 0.07 in the AncSR2 map (Fig. 2.5G). The probability of evolving new specificity phenotypes on moderate timescales increases accordingly: after just three steps, two new phenotypes—including SRE specificity—are more likely than conservation of ERE. By six steps, SRE specificity becomes the most likely of all phenotypic outcomes. Heterogeneity around the ancestral genotype could also influence the evolutionary transition between ERE (GT) and SRE (AA). Of the two possible intermediate REs—GA and AT—the transition likely involved AT as an intermediate phenotype, despite GA being favored by the anisotropy of the GP map.

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

2.3.8 Simple biophysical mechanisms changed the GP map

Finally, we sought insight into the biophysical mechanisms that changed the structure of GP map of the recognition helix between AncSR1 and AncSR2. Although our experiments provide a functional rather than biophysical readout, different biophysical mechanisms predict different patterns of functional change between the AncSR1 and AncSR2 maps. We therefore analyzed the change in fluorescence of each protein-DNA complex variant between the two backgrounds

to identify potential biophysical mechanisms and considered them in light of existing crystal structures. We found evidence for two major mechanisms.

First, the background substitutions between AncSR1 and AncSR2 appear to have caused a universal increase in affinity across all protein-DNA complexes. Previous experiments and crystal structures showed that 11 of the background substitutions improve affinity on both ERE and SRE by generating favorable interactions not involving the two variable RE bases[26,67].. We therefore hypothesized that affinity increased universally for all amino acid variants across all 16 REs. To test this hypothesis, we fit a simple model in which fluorescence in each ancestral background is a function of a complex's affinity, and affinity is scaled by a constant factor in AncSR2 relative to AncSR1 (Methods). The model fits the data very well ($r^2 = 0.95$; Fig. 2.6A), with an estimated 70-fold universal improvement in affinity in the AncSR2 background. This apparent increase in affinity explains the vast increase in functional genotypes and specificity phenotypes between AncSR1 and AncSR2, because many protein-DNA complexes that had weak affinity in the AncSR1 background—and were therefore nonfunctional—bind strongly enough in AncSR2 to produce functional levels of fluorescence. The number of promiscuous protein variants also increases, because many variants cross the threshold for functionality on multiple REs (Fig. A.5A). A universal improvement in affinity explains not only the increased size but also the greater connectivity of the AncSR2 network: the background substitutions do not qualitatively change the amino acid determinants of binding but instead make them less stringent (Fig. 2.6B), so many of the newly functional nodes in AncSR2 are close neighbors of those that were already functional in AncSR1, with an average gain of 11 new neighbors per node (Fig. 2.6C).

The second apparent mechanism is that the background substitutions negatively affect

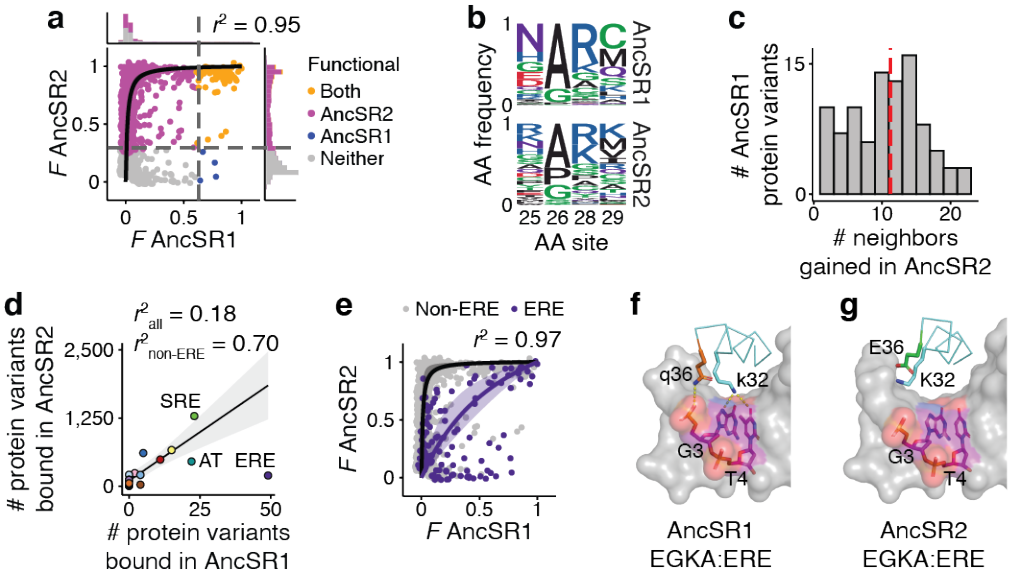

Figure 2.6. Nonspecific effects of background substitutions on DBD-RE affinity. (**A**) Fluorescence of each complex in the AncSR1 vs. AncSR2 background, scaled between the upper and lower bounds for each background. Curve shows best-fit model assuming that the affinity of every complex in the AncSR2 background is related to its affinity in the AncSR1 background by the same scaling factor. Shaded region around the curve (barely visible) shows bootstrapped 95% confidence interval (CI). The Pearson's r^2 between the data and model predictions is shown (n = 2,627). Histograms show distribution of F in each background. Dashed lines show the fluorescence of the wild type complex in each background (AncSR1-EGKA:ERE or AncSR2-GSKV:SRE). Colors indicate the backgrounds in which each genotype is functional. (**B**) Amino acid frequencies at the variable RH sites across all functional protein variants in the AncSR1 and AncSR2 maps. (**C**) Distribution of the number of neighbors gained in the AncSR2 background across all functional protein variants in the AncSR1 background that remain functional in the AncSR2 background. Dashed line, mean. (**D**) Correlation between the number of protein variants bound per RE in each background. Black line, linear fit to all REs except ERE; shaded region, 95% CI. (**E**) Same as **A**, but fitting a model in which the background substitutions affect affinity of all variants for ERE by one scaling factor and for all other REs by a different scaling factor. Purple, observed fluorescence and best-fit model predictions for ERE complexes; gray, for non-ERE complexes. (**F**) Crystal structure of the AncSR1-EGKA protein in complex with ERE (PDB 4OLN). The RH backbone is shown as a ribbon, with key side chains shown as sticks. The gray surface shows ERE, with variable bases and backbone as sticks. In this complex, glutamine (q) at site 36 forms a hydrogen bond (yellow dashed line) with the DNA backbone, and lysine (k) at site 32 forms two hydrogen bonds to the ERE-specific bases G and T. (**G**) Same as **F**, but with the AncSR2-EGKA crystal structure (PDB 4OND). Substitution to glutamic acid (E) at site 36 abolishes the ancestral hydrogen bond to the DNA backbone and results instead in electrostatic repulsion from the backbone. This deforms the recognition helix, abolishing the hydrogen bonds between K32 and the G and T bases. In **F** and **G**, lowercase letters represent ancestral amino acid states, and uppercase derived.

specific binding to ERE, shifting the anisotropy in the global production distribution away from

ERE and leaving SRE as the most-encoded phenotype in the AncSR2 background. A universal affinity increase predicts that the number of variants with every specificity phenotype should increase proportionally across the AncSR1-AncSR2 interval; this pattern holds, but ERE is an outlier, with far fewer variants than would be expected given the pattern for other phenotypes (Fig. 2.6D). Moreover, ERE complexes exhibit notably lower fluorescence in the AncSR2 background than predicted by a universal increase in affinity (Fig. A.5B). We estimated the effect of the background substitutions on ERE affinity by incorporating a background-by-ERE interaction term into our affinity-fluorescence model; adding this parameter improves the fit to the data ($r^2 = 0.97$), with the background substitutions improving ERE affinity by an estimated 2.3-fold, compared to 99-fold for all other REs (Fig. 2.6E). The extent of the relative reduction in fluorescence differs among protein variants, however, suggesting additional specific interactions between background substitutions and amino acids in the recognition helix (Fig. A.5C). Crystal structures of the EGKA:ERE complex[67] suggest one possible structural mechanism for the global reduction in ERE affinity: one of the background substitutions (q36E) deforms the protein backbone of the recognition helix, abolishing two hydrogen bonds that are formed between a conserved residue and bases in the ERE (Fig. 2.6F, G). Corroborating this mechanism, the background substitutions also shift the bias of the global production distribution away from AT specificity (Fig. 2.6D), and this is the only other RE that can form these hydrogen bonds.

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

evolving the new specificity for SRE.

2.3.9 Robustness to assumptions

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

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

Finally, we addressed uncertainty about the ancestral sequences. AncSR1 and AncSR2 DBD reconstructions have very high confidence, containing just five and zero ambiguously reconstructed sites, respectively[43]. Experimental data from a prior single-mutant DMS study show that the effects of mutations in the RH are virtually identical when they are introduced into the AncSR1 background or into an alternative reconstruction of AncSR1 that incorporates all plausible alternative amino acids at the ambiguously reconstructed sites ($r^2 > 0.99$; Fig. A.8)[43]. The very limited uncertainty about the AncSR1 ancestral sequence is therefore likely to have

little or no effect on our conclusions.

2.4 Discussion

2.4.1 The GP map was a cause of historical phenotypic evolution

Our data establish that anisotropy and heterogeneity in the two ancestral GP maps studied were causal factors in the historical lineage-specific evolution of DNA specificity. Establishing causality in a multifactorial framework requires 1) evidence that a putative cause increases the probability of the outcome(s) of interest, and 2) evidence for a specific mechanism by which the cause affects the outcome's probability[104]. Concerning the first requirement, we showed that the particular forms of anisotropy and heterogeneity in the AncSR1 map increased the probability that ERE specificity would be evolutionarily conserved relative to a map that did not have those properties, and those in the AncSR2 map increased the probability that SRE specificity would be acquired. The second requirement is satisfied by a simple axiom of population genetics: the probability that a phenotype will evolve is the product of its probability of production and its probability of fixation under the influence of selection and drift. If the structural properties of the GP map increase the probability that a phenotype is produced, then the probability that that phenotype will evolve must also increase.

A cause must precede its effect. The biases in phenotype production that favored the conservation of ERE specificity in AncSR1 are ancestral to the ER lineage in which that outcome occurred (Fig. A.1B). The structure of the map persisted unchanged for hundreds of millions of years of phenotype conservation, because zero amino acid changes anywhere in the DBD occurred along the descendant branches leading from AncSR1 to ERα in the ancestor of all bony vertebrates. Even most present-day ERa DBDs contain zero or at most a single substitution

relative to AncSR1 (Fig. A.9). As for the acquisition of SRE specificity in the AncSR2 lineage, SRE specificity was already favored in the AncSR1 map as the second-most encoded phenotype. Further, the massive increase in connectivity of the AncSR2 map, which dramatically increased the propensity for new phenotypes to evolve, must have been acquired before SRE specificity actually evolved, because the recognition helix substitutions that conferred SRE specificity during history cannot be tolerated unless the background substitutions that nonspecifically increased DNA affinity occurred first[67]. Our experiments do not resolve whether the third major property of the AncSR2 map—a shift in the direction of anisotropy away from ERE specificity that further enhanced the propensity to encode SRE specificity—occurred before or after this phenotype was historically acquired.

We do not argue that selection played no historical role in the evolution of specificity. It seems likely that purifying selection would have favored conservation of ERE specificity in the chordate ERs, and positive selection could have contributed to fixation of SRE specificity in the AncSR2 lineage. If so, however, selection would have further increased the probability of outcomes that were already favored by the phenotype production propensities caused by the GP map's structure. Our argument is agnostic as to the evolutionary causes of the nonspecific increase in DNA affinity that caused the expansion of the AncSR2: the substitutions outside the RH could have been driven by selection for affinity (not specificity), but they also could have been fixed through a process of systems drift under purifying selection to maintain occupancy on DNA, such as might occur if receptor expression and affinity drifted in opposite directions.

Our data show that the forms of anisotropy and heterogeneity in the AncSR1 and AncSR2 GP maps can cause strong biases in the outcomes of phenotypic evolution. For example, some phenotype production propensities that we observed are absolute. There are 9 specificity

phenotypes that cannot be encoded at all in the AncSR1 GP map, and two cannot be encoded in AncSR2; these could phenotypes never evolve, no matter how large a fitness benefit they might in principle confer. The shorter-term biases imposed by heterogeneity are also absolute in many cases: from every starting point, the vast majority of phenotypes are impossible to produce directly by mutation, and most require many substitutions before they become accessible. Selection would therefore be powerless to fix these phenotypes over short or medium timescales. The structure of the GP map limited evolution to a small subset of possible phenotypes; history, further influenced by selection and chance, played out within this set.

2.4.2 Generality

Our experiments addressed a small number of sites in one protein-DNA interface, but there is evidence that features similar to those we observed in the steroid receptor GP map affect many biological systems and their evolution across levels of organization. Anisotropy is apparent in other molecular[39,105] and developmental systems[106,107,37,108], and the resulting biases are often congruent with natural patterns of diversity[38,40,109,110]. Heterogeneity also appears to be widespread, because most random mutations are phenotypically neutral if they are tolerated[108,111–114], and long-term phenotype conservation is widespread in the fossil record[115]. When new phenotypes are acquired, identical perturbations often yield different phenotypes in different lineages[36,116–118], and convergent evolution becomes less likely among distantly related lineages[119]. As lineages evolve across their GP maps, their biology inevitably changes, imposing new biases on the production and future evolution of genotypes and phenotypes. It therefore seems likely that anisotropy and heterogeneity are near-universal characteristics of GP maps[80,99,101], and that the biases these properties create have shaped large-scale patterns of

phenotype conservation and lineage-specific evolutionary change across the tree of life.

Our study differs in kind from previous combinatorial DMS studies, which have addressed the distribution in sequence space of just one or a few phenotypes that are encoded by extant proteins, rather than the space of all possible phenotypes [22,26,27,117,120–122]. These studies have shown that sequence landscapes are rugged, so the probability of reaching particular genotypes encoding those phenotypes may depend on the starting point and intermediate mutational steps. Because those studies take the phenotypic "destination" for granted, they cannot address why those phenotypes, rather than other conceivable outcomes, exist at all.

Our work shows that as a protein or other biological system moves through sequence space, the set of phenotypes that it can produce changes at every step. Life is astonishing in its diversity, but an even deeper puzzle lies in the fact that only a tiny fraction of conceivable phenotypes have ever evolved, and those which have evolved are mostly limited to particular taxa[74–76,123]. Chance and selection are likely important factors in explaining the patchy distribution of phenotypes on Earth. But the very particular biology we observe today must also reflect the constantly changing potential of biological systems, as they vary and diverge, to generate new forms of life at every moment in time.

2.5 Methods

2.5.1 RE reporter strains

To measure binding of SR DBD to the 16 RE variants, we adapted a yeast GFP reporter system previously developed to measure binding to ERE and SRE, where GFP expression is well correlated with DNA affinity over a range of at least 2 M^{-2} ($r^2 = 0.74$)[43]. We engineered 16 yeast strains, each of which reports on binding of the DBD to one RE. We modified the yeast strain

CM997 (YPS1000 MATa ho::KMX)[124] to replace the *KMX* gene at the *HO* locus with a construct containing yeast-enhanced GFP downstream of a minimal *CYC1* promoter with an array of four palindromic RE sites (tcaAGNNCAcagTGNNCTtga), each separated by a 19-nt sequence, along with a *HygR* gene. To ensure a consistent dynamic range of fluorescence across strains, we made changes to two RE strains in the nucleotide sequences flanking the palindromes at sites that do not affect specificity[96,97] (see Appendix A for details). These constructs were transformed into yeast using the lithium acetate method[125] and selected for resistance to hygromycin and susceptibility to G418; integration was confirmed by Sanger sequencing.

To validate this reporter system, we measured fluorescence of each strain in the presence and absence of a DBD variant with universally high affinity to all REs (AncSR1+11P+GGKA)[17,67]. We used a low-copy yeast vector (pDBD) to express this DBD variant as a C-terminal fusion with an SV40 nuclear localization signal and a *S. cerevisiae* Gal4 activation domain (Gal4AD) under control of a pGAL1 promoter. We transformed this construct into each yeast strain using the lithium acetate method followed by G418 selection (50 μg/mL). Single colonies were inoculated in YPD+G418 and transferred to YPGal+G418 media for 6 hours to induce DBD expression. GFP fluorescence was measured on a BD LSRFortessa flow cytometer using a 488 nm laser with 505 nm long pass and 525/50 nm band pass filter. We used as the metric of fluorescence $\log_{10}(\text{GFP}/\text{FSC-A}^{1.5})$, which normalizes fluorescence to cell volume. All 16 strains showed DBD-dependent fluorescence across a similar dynamic range (Fig. A.1A–C).

2.5.2 AncSR1 and AncSR2 combinatorial library construction

We used as the wild-type protein sequences the maximum *a posteriori* AncSR1 and AncSR2

DBD sequences inferred from a maximum likelihood phylogeny of nuclear receptors[43].
We optimized codon usage for yeast and cloned the ancestral DBDs into the pDBD2.1 expression vector, which is modified from the pDBD vector[26,43] to express GFP at a level within the dynamic range of fluorescence for the wild type AncSR1:ERE and AncSR2:SRE complexes. A bidirectional pGAL1/GAL10 promoter simultaneously drives DBD and mCherry expression, which allowed us to monitor plasmid retention in yeast (Fig. A.1D).

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

Maxiprepped libraries (GenElute HP, Sigma-Aldrich) were transformed into the yeast reporter strains using an optimized yeast electroporation protocol (Appendix A). Transformation yields exceeded 10^7 cfu per library (50-fold coverage), estimated by dilution plating (Table A.2). Yeast libraries were flash-frozen in liquid N_2 in 200 OD_{600}-mL aliquots with 25% glycerol and stored at –80°C. Multiple transformant rates estimated from Sanger sequencing of individual colonies[126] were estimated to result in 0.03% or fewer cells with multiple plasmid copies at time

of sorting.

2.5.3 Cell sorting

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

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

Enriched cells from different libraries were pooled by GFP bin and grown in either 700 mL (GFP+) or 2 L (GFP–) of YPD+G418+chlor. Cultures were grown overnight at 225 rpm and 22–30°C, depending on the ratio of cells to media, until they were at least OD_{600} 3 but not yet saturated. 200 OD_{600}-mL 25% glycerol stocks were then made for both the GFP+ and GFP– cultures. 10 OD_{600}-mL of the GFP– culture was used for plasmid extraction using a previously described protocol[19].

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

2.5.4 Deep sequencing

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

Sequencing libraries were constructed from plasmids extracted from the enrichment sort

GFP– population and the four binned sort populations using two rounds of amplification. In the first round, the RH scanning and REBC regions of the DBD were amplified with primers that added a 6-nt barcode for bin and replicate identification (BRBC)[127]. For every 10 OD_{600}-mL of yeast used for plasmid extraction, 3 µL of plasmid template was used in a 10 µL Q5 PCR reaction (NEB). AncSR1- and AncSR2-specific primers were mixed proportionally to background-specific cell counts (estimated from flow cytometry) to minimize amplification bias. To introduce nucleotide diversity for improved cluster identification during Illumina sequencing, eight unique forward and reverse primer pairs were used per bin and background to encode frameshift diversity and attach read 1 primer sequences in both directions. PCR conditions included 52°C annealing for 13 cycles. Reactions were then pooled by bin/replicate and purified using the Zymo DNA Clean & Concentrator Kit. In the second round, half of the first-round product was amplified with primers to add Illumina P5 and P7 adapter sequences. PCR was performed in 50 µL Q5 reactions (NEB) per 10 µL round 1 product reaction at 68.4°C annealing for 12 cycles. The final product was size-selected on a 2% agarose gel, excised, purified using the Qiagen Gel Extraction Kit, and re-purified with the Zymo DNA Clean & Concentrator Kit.

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

2.5.5 Mean fluorescence estimation, data cleaning and validation

Sequencing reads were processed using a custom pipeline. We used *sickle* v1.33[128] to filter reads based on their quality: we kept reads with a Phred score ≥ 30 and a minimum length of 79 nucleotides. We then used *PEAR* v0.9.6[129] to merge the trimmed paired-end reads (minimum assembly length 100 nucleotides). Finally, we used Biopython toolkit v1.79[130] to demultiplex the assembled reads by DBD background, REBC, and BRBC. We only considered reads that mapped exactly to the DBD background and allowed reads with at most one mismatch in the REBC and one in the BRBC.

The mean fluorescence for protein:RE complexes observed in the binned sort data was estimated as previously described[43]. We first estimated the proportion of cells of each complex g in each bin b ($c_{g,b}$) from the proportion of reads in b that mapped to g. The mean fluorescence estimate F_g for each complex was then estimated by taking the weighted mean fluorescence across bins (mean fluorescence of each bin was measured during sorting), with weights $c_{g,b} / \sum_b c_{g,b}$.

We applied several filtering and correction steps to reduce global measurement error and normalize fluorescence estimates between replicates. First, complexes with fewer than 27 reads per replicate were removed to ensure >95% had a standard error (SE) of ≤ 0.1 (5% of the assay range; Fig. A.2A). Second, complexes observed in only one replicate were excluded. Third, batch effects were corrected by fitting I-splines to normalize fluorescence between replicates (Fig. A.2B). Finally, SE was recalculated and complexes with SE > 0.1 were removed (Fig. A.2C). The final dataset had a mean pairwise Pearson's $r^2 = 0.55$ across replicates. The poor correlation arises primarily because the vast majority of complexes are at the lower fluorescence bound, so r^2 is dominated by measurement noise; for variants with fluorescence above the lower

bound (roughly $F \geq -4.0$), r^2 improved to 0.92. Altogether, we obtained fluorescence estimates for 628,732 AncSR1 and 658,475 AncSR2 variants, covering 24.6% and 25.7% of possible variants, respectively (excluding nonsense variants).

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

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

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

2.5.6 Fluorescence inference for missing complexes

To predict the fluorescence of the remaining complexes for which we did not obtain experimental estimates, we fit a generalized linear model based on reference-free analysis (RFA)[47,52] to the experimental data. The model estimates a sigmoid function to capture the measurement bounds of the assay, plus additive and interaction effects (specific epistasis) for all amino acid states at the four variable sites in the DBD and all nucleotide states at the two variable sites in the RE. All possible intramolecular interactions up to third order amino acid interactions in the DBD and second order nucleotide interactions in the RE, and intermolecular interactions up to third order amino acid-by-second order nucleotide interactions were included. L2 regularization with 10-fold cross validation was used to reduce overfitting (Fig. A.3A; Appendix A). We fit separate RFA models for each ancestral background using the *glmnet* v4.1-6 R package[131]. Model fits to the observed data were $R^2 = 0.96$ for AncSR1 active complexes (0.31 all complexes) and $R^2 = 0.99$ for AncSR2 active complexes (0.88 all complexes) (Fig. A.3B). These models were used to predict fluorescence values for unobserved protein-RE complexes. We also used the fitted models to correct the predictions for complexes in one of the modified strains that had systematically lower fluorescence (Appendix A; Fig. A.3C, D).

2.5.7 Classification of functional complexes

We classified complexes as functional if their fluorescence was not significantly lower than the wild type complex, *i.e.* EGKA:ERE in the AncSR1 background and GSKV:SRE in the AncSR2 background. Complexes inferred as null from the enrichment sort were classified as nonfunctional. For complexes observed in the binned sort, we used a *t*-test to account for measurement error. For complexes with predicted fluorescence from the RFA models, we

performed a nonparametric bootstrap test using the distribution of model residuals concatenated over the ten cross-validation fits to account for model prediction error (Appendix A; Fig. A.3E). For both tests, we used a Benjamini-Hochberg FDR threshold of 0.25 to classify variants as nonfunctional if they were significantly less fluorescent than the wild type complex (Fig. A.3F). The low stringency of the FDR threshold was chosen to reduce the false positive rate for calling variants functional. The majority of complexes classified as functional in both backgrounds had fluorescence estimates obtained from the binned sort experiment (59.3% AncSR1, 75.4% AncSR2; Fig. A.3G).

2.5.8 Protein genotype networks

Following Maynard Smith's sequence space formalism[12], we built genotype networks consisting of all functional RH variants in each DBD background. RH genotypes are connected by an edge if they differ by a single amino acid mutation that can be produced via a single nucleotide mutation given the standard genetic code. Genotype networks for joint protein-DNA models follow a similar logic (Appendix A). We used the R package *igraph* v1.5.1[132] to build and analyze the genotype networks, and the software *gephi* v0.10.1[133] for network visualization. To identify clusters of densely connected genotypes within the networks, we used the cluster_edge_betweenness function from the R *igraph* package.

2.5.9 Model of evolution on GP maps

We modeled evolution on the genotype networks as an origin-fixation process under a strong selection-weak mutation regime[33,134] To isolate the effect of the GP map's structure on evolution, we considered a scenario in which all functional genotypes have equal fitness, so the fixation

probability is affected only by drift, and nonfunctional variants are removed by purifying selection. The relative probability $P(i,j)$ of substitution from protein genotype i to genotype j is therefore equal to the amino acid mutation rate μ_{ij}, normalized over all single-step neighbors of i in the network. We assumed that there are no biases in the nucleotide mutation process (*e.g.* transition vs. transversion rate), so μ_{ij} is affected only by unequal mutational access between amino acids imposed by the genetic code. To incorporate this effect, we scaled μ_{ij} by the number of possible nucleotide mutations that can change any nucleotide sequence that encodes i to any nucleotide sequence that encodes j:

$$\mu_{ij} = \eta_{ij}^{o^*} \times \prod_{o \neq o^*} c_o \quad (1)$$

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

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

genotypes:

$$\pi_{(k)} = \pi_{(0)} \times P^k \quad (2)$$

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

2.5.10 Effects of background substitutions

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

$$F(g)_{AncSR1} = \frac{1}{1 + \frac{K_d(g)}{[RE]}} \quad (3)$$

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

$$F(g)_{AncSR2} = \frac{1}{1+\alpha\left(\frac{K_d(g)}{[RE]}\right)} \quad (4)$$

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

$$F(g)_{AncSR2} = \frac{1}{1+\alpha\left(\frac{1-F(g)_{AncSR1}}{F(g)_{AncSR1}}\right)} \quad (5)$$

We fit this model to the AncSR1 and AncSR2 fluorescence data using orthogonal regression, which accounts for measurement error in both backgrounds. We used only complexes that had fluorescence measurements from the binned sort in both backgrounds, and whose fluorescence was significantly greater than that of nonsense variants in either background (n = 2,627). Fluorescence was normalized in each background to scale between the upper and lower bounds inferred from the RFA models. Confidence intervals (CI) were constructed by bootstrapping the data and refitting the model. The effect of the background substitutions was estimated to be α = 0.014 (95% CI: 0.010–0.014), corresponding to a 70-fold increase in affinity (95% CI: 70–99). We next considered a model where the background substitutions have a different effect on ERE affinity than they do on other REs. We modified the model such that α_1 represents the ERE-specific effect of the background substitutions and α_2 the effect on the other 15 REs. We fit this

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

2.5.11 Data availability

The datasets generated and analyzed during this study are available at https://doi.org/10.5061/dryad.18931zd7m

2.5.12 Code availability

Detailed scripts including how to use and analyze the data are available at www.github.com/JoeThorntonLab/RH-RE_scanning.

Chapter 3: Small-magnitude epistasis shapes a protein-DNA specificity landscape

3.1 Summary

Specific binding between biological molecules underpins many of life's functions, but the genetic architecture of these interactions—the rules by which variation in the sequence of two molecules generates variation in their binding—is poorly understood. I decomposed the genetic architecture of binding and specificity in a deep mutational scanning dataset of a transcription factor-DNA interface, in which all possible combinations of amino acid states at four protein sites and all possible nucleotide states at two DNA sites were systematically assayed for binding. I used reference-free analysis to estimate the main and epistatic effects of amino acid and nucleotide states in this dataset. Over 80% of genetic variation was attributable to main effects of amino acids on nonspecific binding and on specificity for individual DNA bases, implying that the overall genetic architecture is quite simple. Despite this, pairwise epistasis between amino acids and third-order epistasis between amino acids and nucleotides were essential for wild type levels of binding and specificity in most complexes, and contributed strongly to the effects of function-switching mutations. These results show that while epistatic effects tend to be small, their accumulation over many sequence sites strongly influence the landscape of intermolecular binding and specificity.

3.2 Introduction

Binding between biological macromolecules underpins many of life's most important functions, from the regulation of gene expression and protein function to the catalysis of complex biochemical reactions—like translation and mRNA splicing—by large assemblies of proteins

and nucleic acids. Understanding the genetic architecture of these interactions—the set of rules by which genetic sequence encodes intermolecular recognition via binding—is key to understanding how they diversify during evolution and how they can be engineered to carry out new functions[11].

Deep mutational scanning (DMS)[19] is a powerful technique for studying this problem because it generates large mutational datasets from which the genetic architecture of sequence-function relationships can be learned. Combinatorial designs that measure the effects of all possible combinations of sequence states across multiple sites[20,21,23–27] are particularly useful for understanding the extent to which epistasis—nonadditive interactions between sequence states across multiple sites in a molecule—contributes to molecular phenotype. While many studies have found evidence for extensive epistasis in molecular genotype-phenotype maps[11,20,21,23,24,45,48,49], recent advancements in computational methods used to infer epistasis have challenged these findings. In particular, Park et al.[47] found that by accounting for epistasis induced by measurement bounds and using a less biased regression approach termed reference-free analysis (RFA), the majority of phenotypic variance in existing combinatorial DMS datasets—including many for protein-protein or protein-DNA binding—can be described by the main (non-epistatic) effects of individual sequence states. Almost all of the remainder can be accounted for by pairwise interactions between pairs of residues; very little variance is attributed to higher-order interactions. These findings suggest that molecular sequence-function relationships, including those for intermolecular binding, are much simpler than previously thought, an encouraging prospect for our ability to interpret and predict the effects of mutations on molecular function.

While a number of DMS studies have characterized the genetic architecture of binding

between biological molecules[20–22,24,29,51,52,59,135,136], fewer have studied the architecture of intermolecular specificity[22,51,52]—an important feature of many interactions. Specific interactions require that one interaction partner (or a small number of them) be bound better than any other potential interaction partner. For instance, transcription factors must scan the entire genome and bind to specific DNA motifs to avoid off-target gene regulation, and many protein-protein interactions require distinguishing between paralogs. To study the genetic architecture of specificity, the impact of sequence variation on binding to all possible interaction partners must be measured. A few studies have accomplished this by assaying the effects of mutations in one molecule on binding to all existing members of a class of functionally or evolutionarily related potential partner molecules[22,51,52,122]. Mutation effects can then be partitioned into nonspecific effects, defined as the average effect across all assayed partners, and specificity effects, defined as the deviation between a mutation's nonspecific effect and its effect on binding to a particular partner[51,52]; specificity effects can also be thought of as intermolecular epistasis[17]. These studies have shown that specificity effects account for a minority of variance in binding across interaction partners[51,52]; this may be because the binding partners assayed tend to differ at only a few sequence states. However, one study found that protein-DNA specificity involved a large amount of intramolecular epistasis between amino acid mutations[52], suggesting that the genetic architecture of specificity may be more complex than that of general binding. Epistasis was also found in this study to be important for mutations that switch specificity, thus enabling functional diversification.

A limitation of the study design described above is that it only considers how mutations in one binding partner impact specificity for existing variation in the other partner. This strategy cannot decompose the effects of sequence variation in both binding partners to their residue-level

effects, because variants of the second partner typically differ from each other at multiple sites and sample only a subset of possible sequence states. Understanding specificity in terms of the intermolecular epistatic interactions between individual residues would not only provide more detailed information about the genetic architecture of specificity, but also give better insight into how mutations in both molecules shape the evolution of a coevolving complex[17].

In Chapter 2, I generated a dataset consisting of all 20^4 combinations of amino acid mutations at four sites in the DNA binding interface of an ancestral steroid receptor protein, measured for binding against all 4^2 combinations of nucleotide mutations at two sites in its cognate DNA response element. To my knowledge, this is the first experiment to combinatorially mutagenize sites on both sides of an intermolecular interface to all possible sequence states. In this chapter, I perform RFA on this dataset to analyze the effect on binding of all possible amino acid states in the protein and nucleotide sequence states in the DNA, as well as their intramolecular and intermolecular interactions. This allows me to gain insight into the complexity of interactions involved in protein-DNA binding and specificity, the effects of intra- and intermolecular epistasis on binding and specificity on individual genotypes and their accessibility via mutation, and the mechanisms that preference protein genotypes towards specificity for particular DNA variants.

3.3 Results

3.3.1 Inferring the genetic architecture of steroid receptor-DNA specificity

Steroid receptors are a family of metazoan transcription factors that regulate gene expression in response to hormone signaling. In vertebrates, two paralogous classes of steroid receptors—termed estrogen receptors and ketosteroid receptors—regulate distinct gene expression pathways

by binding specifically to two different DNA response elements (RE) motifs[96,137]. Protein-DNA specificity in this system is mediated by four phylogenetically variable amino acid sites in the recognition helix (RH) of the protein's DNA binding domain (DBD) and two nucleotide sites in the RE (Fig. 3.1A, B)[67,96,137]. To systematically probe the impact of sequence variation at these six sites on binding and specificity, in Chapter 2 I used a yeast-based GFP reporter assay to map the ability of all 160,000 possible combinations of amino acid mutations at the four variable RH sites to bind all 16 possible combinatorial mutants of the RE that differ at the two variable nucleotide sites (Fig. 3.1B). This experiment was performed in the genetic background of the AncSR2 protein, which is the common ancestor of all extant ketosteroid receptors[67].

The resulting genotype-phenotype map exhibited two notable features. First, binding was extremely rare—of the 160,000 assayed protein genotypes, only a tiny fraction (2,407 genotypes, 1.5%) bound at least one of the 16 REs at least as well as the wild-type AncSR2-RE complex (genotype GSKV:AA; Fig. 3.1A). Second, specific binding for some REs was encoded much more frequently than for others—of the protein genotypes that bound only a single RE as well as the wild type complex, the most common RE bound by far was AA (also known as SRE; Fig. 3.1C), which is also the RE bound specifically by the wild type AncSR2 protein. In general, specificity for REs that contain adenine at at least one of the two variable sites, especially at site 4, was more common than for other REs (Fig. 3.1C).

To understand the genetic basis behind these features of the genotype-phenotype map, I used RFA[47] to decompose the effects of individual amino acid and nucleotide states and their combinations on protein-DNA binding. RFA models phenotype y (in this case, GFP fluorescence) as a function of a genotype s's "genetic score" x, which is defined as the summed effects of individual sequence states and their epistatic interactions:

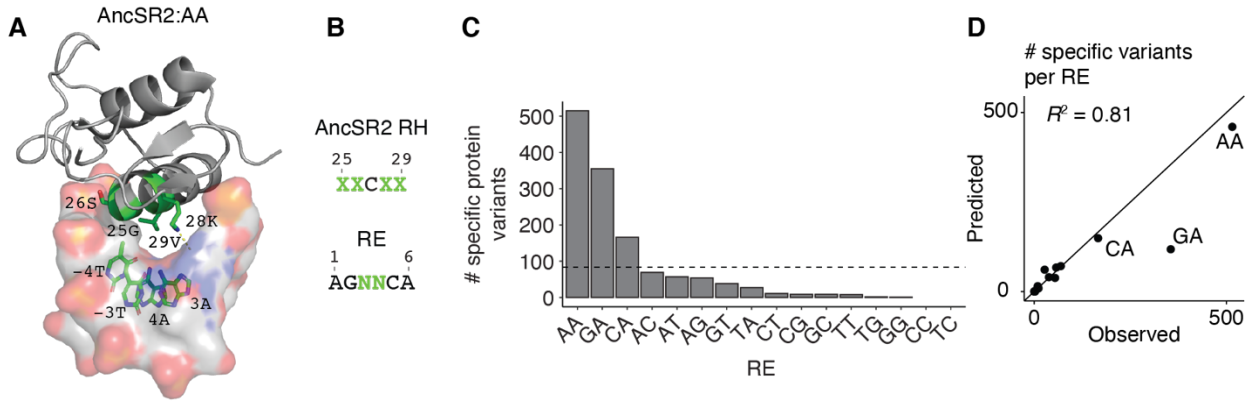

Figure 3.1. Specificity in the AncSR2-RE genotype-phenotype map. (**A**) Crystal structure of the AncSR2 DBD in complex with AA-RE (PDB 4OOR). Specificity-determining sites that were varied in the DMS experiment are colored in green, with amino acid side chains and DNA bases shown as sticks and labeled with their wild type sequence and position. The complementary thymine bases in the RE are also shown as sticks, with the entire RE shown as surface. A hydrogen bond between AA28K and the invariable NTG2 base (shown as surface) is shown as a yellow line. (**B**) Sites varied in the DMS experiment. Four sites in the DBD (green Xs) and two sites in the RE (green Ns) were combinatorially mutated to all possible amino acids and nucleotides, respectively. (**C**) Empirical distribution of specificity phenotypes in the AncSR2-RE GP map, as reported in Chapter 2. The dashed line shows the theoretical frequency of all 16 possible specificity phenotypes if all were equally frequent. (**D**) Comparison between the observed frequencies of specificity phenotypes shown in **C** and the predicted frequencies from the best-fit RFA model. The frequency of predicted GA specificity is much lower than observed because the observed frequencies were corrected for a strain-specific artifact that systematically reduced the fluorescence of protein variants on GA-RE; this artifact was not accounted for in the RFA model.

$$y(s) = g(x(s)) = g\left(\beta_0 + \sum_i \beta_i(s_i) + \sum_{i<j} \beta_{i,j}(s_i, s_j) + \sum_{i<j<k} \beta_{i,j,k}(s_i, s_j, s_k) + \cdots\right).$$

Here, i, j, k index the variable amino acid and nucleotide sites in the sequence, β_0 is the effect of the genetic background, β_i is the first-order (main effect) of the state at site i, $\beta_{i,j}$ is the second-order interaction (epistatic) effect of the states at sites i and j, and so on. Epistasis can occur intramolecularly between amino acid sites in the protein and nucleotide sites in the DNA, as well as intermolecularly between amino acid and nucleotide sites. A key feature of RFA is that the β

terms for each site and site combination are constrained to sum to zero across all possible states and state combinations; this property has been shown to minimize bias in effect estimates, especially in the face of missing and noisy data[47]. The model can theoretically include as many interaction orders as there are variable sites, but in practice is often truncated to a certain interaction order to reduce overfitting to measurement noise.

Another key feature of RFA is the function g, which captures nonspecific epistasis—a nonlinear mapping between an underlying additive trait and observed phenotype[44,45]. Nonspecific epistasis in the AncSR2-RE genotype-phenotype map arises from measurement bounds on GFP fluorescence, which has a likely biophysical basis in the thermodynamics of protein-DNA binding. I therefore used a sigmoid function $g = L + (U - L)/(1 + e^{-x})$ to model nonspecific epistasis, where L and U are globally fitted lower and upper fluorescence bounds, respectively[47].

I estimated main and epistatic effects of amino acid and nucleotide sequences states on fluorescence using L1-regularized regression with 10-fold cross-validation. To determine the maximum order of interaction terms to include, I fit a series of models truncated at various effect orders and evaluated overfitting using the out-of-sample predictions from cross-validation (Fig. B.1A). The best-fit model fit the data very well (R^2 = 0.88; Fig. B.1B), and accurately reproduced the distribution of RE specificity phenotypes seen in the full dataset (Fig. 3.1D; the much lower predicted number of GA-specific variants is due to a technical artifact of the yeast strains; see Methods, Section 3.5). The model included up to third-order intramolecular interactions in the DBD, second-order intramolecular interactions in the RE, and third-order intermolecular interactions between pairs of amino acid sites and single nucleotide sites and between single amino acid sites and pairs of nucleotide sites (Table 3.1). The vast majority of

these state combinations were observed in at least ten different genetic backgrounds (Fig. B.1C), meaning that their effects on observed fluorescence are likely accurately estimated. However, for some interaction orders (particularly third-order), a large proportion of state combinations were only measured at the lower bound of fluorescence (Fig. B.1C), meaning that their effects on binding may be masked by the limited dynamic range of the assay (Fig. B.1D). High-order epistatic effects on binding may thus be underestimated by the model.

3.3.2 Low-order effects are the primary determinants of binding and specificity

I first used the effects inferred by the model to determine the relative contribution of each effect order to AncSR2-RE binding. I rescaled effects so that the wild type complex has a genetic score of 1. I then calculated the variance in genetic score (genetic variance, $Var(\beta)$) explained by each effect term. RFA terms have a simple relationship to genetic variance: $Var(\beta)$ is proportional to the square of the effect, normalized by the number of possible states (or state combinations) at the site(s) affected by term β[47,52]. Higher-order terms explain a smaller proportion of genetic

Table 3.1. RFA model terms. Number of terms and degrees of freedom for each order of effect in the RFA model (not including global parameters L, U, and β_0). AA, amino acid state; NT, nucleotide state. Colons denote interactions between states at different sites. The number of terms per order corresponds to the number of possible states across all sites/site combinations. The degrees of freedom are always less than the number of terms for a given order because of the zero-mean constraint of RFA.

Effect order	# model terms	Degrees of freedom
AA	80	76
AA:AA	2,400	2,382
AA:AA:AA	32,000	31,972
NT	8	6
NT:NT	16	13
AA:NT	640	616
AA:AA:NT	19,200	19,116
AA:NT:NT	1,280	1,252
Total	55,624	55,433

variance than lower-order terms of the same magnitude, since they affect fewer genotypes. Similarly, amino acid effects contribute less to genetic variance than nucleotide effects of the same magnitude because each amino acid state is only present in $1/20^{th}$ of all genotypes, whereas each nucleotide state is present in $1/4^{th}$ of all genotypes.

I found that the vast majority of genetic variance in the AncSR2-RE map is explained by the additive (main) effects of amino acids on nonspecific binding (AA terms, 62.8%), and on specific binding to single nucleotide states (AA:NT terms, 18.5%) (Fig. 3.2A). AA effects are strongest at the two central amino acid positions (sites AA26 and AA28, Fig. 3.2A) and are dominated by a strong preference for residues A, G, and P at AA26, and K and R at AA28 (Fig. 3.2B). Negatively charged residues are strongly deleterious at all sites, likely due to electrostatic repulsion with the sugar-phosphate backbone. In contrast, AA:NT effects are strongest at the two outer amino acid positions (sites AA25 and AA29, Fig. 3.2A, C). The magnitude of the effect varies by amino acid and nucleotide site—AA:NT interactions tend to be strongest at site AA29, particularly in combination with site NT4 in the DNA; conversely, site AA25 tends to interact more strongly with site NT3 (Fig. 3.2A, C). The differential nonspecific and specific effect sizes among sites can be rationalized structurally using crystal structures of AncSR2 DBD variants in complex with RE variants[138]. Sites AA26 and AA28 interact with invariant atoms in the DNA—AA26 sits adjacent to the DNA backbone where small, nonpolar residues are likely to fit better, whereas K (and likely R) at site AA28 forms a hydrogen bond with the G base at site NT2 (Fig. 3.1A). On the other hand, sites AA25 and AA29 can form contacts with the variable DNA bases: site AA29 is positioned close to both bases on the sense strand, whereas site AA25 is close to the bases on the antisense strand (especially the NT–3 base; Fig. 3.2D). The majority of variation in

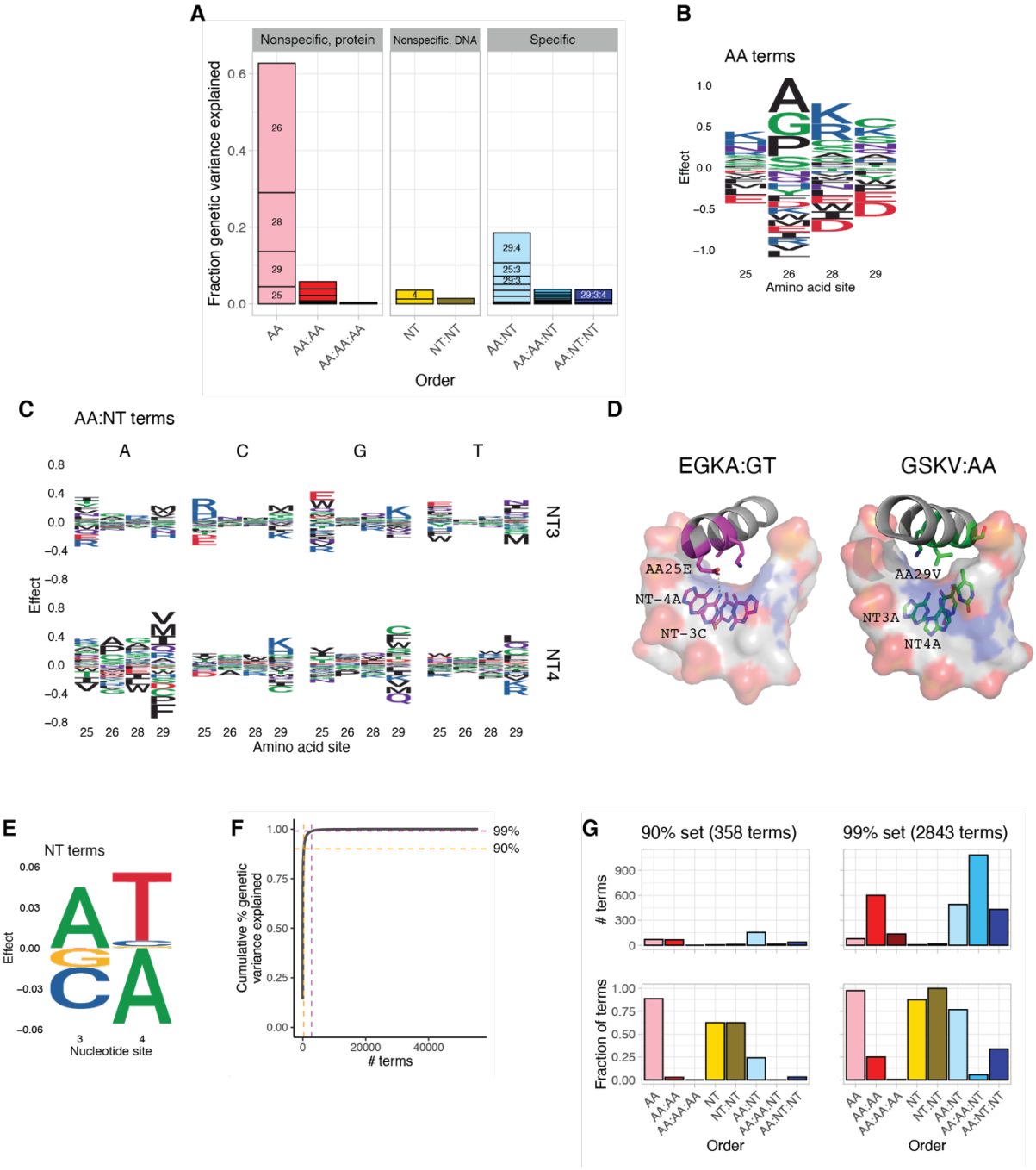

Figure 3.2. Effects of amino acid and nucleotide states on binding and specificity. (**A**) Bars show the fraction of total genetic variance of the model attributable to terms of each order. Stacked rectangles within each bar show the fraction of genetic variance attributable to each site or site combination, with site combinations that contribute >2% labeled. (**B**) Logo plot showing amino acid main binding effects. Letter height shows the effect of an amino acid residue. Residues are colored according to their biochemical properties: red, acidic; blue, basic; black, hydrophobic; purple, neutral, green, polar. (**C**) Amino acid-nucleotide specificity effects. Each logo plot shows the effects of amino acid residues on specific binding to a given DNA base (columns) at a given site in the DNA (rows). Colors are as in **B**. (**D**) Structural rationale for site-

specific specificity effects. (Left) Crystal structure of the AncSR2 DBD with the genotype EGKA at the four variable RH sites, in complex with GT-RE (PDB 4OND). Representation is as in **Fig. 3.1A**, but with only the RH shown for the DBD and the variable amino acid residues and DNA bases colored magenta. The AA25E residue is in close proximity to the variable bases on the antisense strand, and forms a hydrogen bond with the NT–3C base (dashed yellow line). (Right) Same as in **Fig. 3.1A**, but with only the RH shown for the DBD and rotated 180° so that the RH and RE are shown from the back. The AA29V residue is in close proximity to the variable bases on the sense strand. (**E**) Nucleotide main binding effects. Letters are colored according to nucleotide base. (**F**) Fraction of genetic variance in the full model captured with increasing numbers of model terms. Terms were ranked by the fraction of variance explained. Dashed lines show the 90% (orange) and 99% (pink) sets. (**G**) Bars show terms of each order included in the 90% (left) and 99% (right) sets, with stacked rectangles showing sites/site combinations. Top, number of terms; bottom, fraction out of all possible terms of each order.

protein-DNA binding and specificity can thus be described by rationalizable, site-specific interactions between individual amino acid residues and nucleotides.

Higher orders of interaction—both for nonspecific and specific binding—contribute a much smaller fraction of genetic variance in the model (Fig. 3.2A). Pairwise epistasis for nonspecific binding between amino acid residues (AA:AA terms) explain 5.8% of genetic variance, and third-order epistasis between amino acids (AA:AA:AA terms) just 0.4%. Higher-order interactions for specificity—between pairs of amino acids and single nucleotides (AA:AA:NT terms, Fig. B.2C), and between single amino acid residues and pairs of nucleotides (AA:NT:NT terms, Fig. B.2D)—explain similarly small amounts of genetic variance (3.8% and 3.7%, respectively; Fig. 3.2A). These epistatic interactions show site- and state-specific trends: interactions involving AA26 and other amino acid sites tend to be weaker, possibly because the AA26 side chain is oriented toward the side of the major groove, while those of other sites are oriented directly into the major groove where they can interact more easily (Fig. 3.1A). Strong interactions are also enriched for pairs of negatively charged amino acid residues, probably due to electrostatic repulsion (Fig. B.2A–C). AA:NT:NT effects are much stronger for site AA29 than for other amino acid sites (Fig. 3.2A, Fig. B.2D), probably because AA29 can directly contact both variable DNA bases simultaneously (Fig. 3.2D, right).

Nonspecific effects of nucleotide states are quite small (Fig. 3.2E, Fig. B.2E) and explain only a small fraction of the genetic variance of the model (3.6% for NT and 1.4% for NT:NT terms, Fig. 3.2A). Contrary to the high frequency of protein genotypes that encode specificity for adenine at site NT4 (Fig. 3.1C), the main effect of adenine at site NT4 has the strongest negative effect of all NT terms (Fig. 3.2E). NT:NT terms do not account for this discrepancy (Fig. B.2E). Nonspecific binding effects of nucleotide states therefore do not account for the biased encoding of specificity in the genotype-phenotype map.

To better understand how many terms are required to accurately describe protein-DNA binding, I ranked terms according to their variance explained, and asked how many are required to explain 90% of the genetic variance of the full model (the 90% set)[23,52]. This analysis revealed that the model is very sparse—only 358 out of the 55,624 possible terms (0.6%) were included in the 90% set (Fig. 3.2F, orange lines). Of these, almost all AA terms (71 out of 80) and most NT (5 out of 8) and NT:NT terms (10 out of 16) were included, as well as a substantial fraction (24%) of AA:NT terms; only a very small fraction (<4%) of terms of other orders were included (Fig. 3.2G, left). Explaining the vast majority of genetic variance in the data thus requires knowledge of only a very small number of terms, most of which are main effects of amino acid and nucleotide states and pairwise intermolecular interactions.

To understand which orders of effect are needed for even greater predictive accuracy, I asked which terms are needed to explain 99% of the genetic variance of the model. 2,843 terms were included in the 99% set—many more than in the 90% set, but still a small fraction of the number of terms in the full model (5.1%; Fig. 3.2F, pink lines). The fraction of AA:NT, AA:AA, and AA:NT:NT terms increased dramatically from the 90% set—roughly three quarters of all AA:NT terms and a quarter of all AA:AA and AA:NT:NT terms were now included (Fig. 3.2G,

bottom right). Only a small fraction of AA:AA:NT (6%) and AA:AA:AA terms (0.4%) were included; however, AA:AA:NT terms make up the largest category of terms in the 99% set because there are very many of them (Fig. 3.2G, top). AA:AA and AA:NT:NT terms also make up a substantial fraction of total terms in the 99% set. While the genetic architecture is sparse overall, a substantial number of higher-order interactions are still necessary for more precise knowledge of binding and specificity.

3.3.3 Epistasis and specificity are necessary for strong protein-DNA binding

TFs must bind with relatively high affinity to their cognate DNA motifs to regulate gene expression. The observation that the fraction of strongly bound complexes in the genotype-phenotype map is very small implies that strong binding is difficult to encode in the sequence space of the AncSR2-RE interface. To understand which genetic effects contribute to strongly bound complexes, I used the best-fit RFA model to predict the genetic score of every possible protein-DNA complex in the map, and defined bound complexes as those with genetic score at least as high as that of the wild type complex. I then examined the distribution of effects at each order and site combination for these genotypes to understand which effects contribute most to their genetic score.

For the vast majority of genotypes, neither nonspecific nor specific binding terms on their own are sufficient to generate wild type binding levels. Of the 4,250 complexes predicted to be bound, only 320 (7.5%) reach the wild type binding threshold with nonspecific terms alone, and no complex can do so with specificity terms alone (Fig. 3.3A). Nonspecific binding terms contribute most of the genetic score needed to become a bound complex (median 74%), but specificity terms also tend to contribute a substantial amount (median 39%).

Examining contributions across interaction orders, AA and AA:NT terms contribute the most to the genetic score of bound complexes, but higher-order terms both for binding and specificity also contribute (Fig. 3.3A). AA and AA:NT terms contribute 59% and 26%, respectively, of the score needed for binding. Higher-order contributions tend to be smaller, but are still positive on average for most interaction orders (median effect 13% for AA:AA terms, 11% for AA:AA:NT terms, and 4% for AA:NT:NT terms). These substantial epistatic contributions come from the summed effect of many very small interactions across site combinations (Fig. B.3A)—for example, the median per-site effect across all AA:AA:NT terms is less than 0.8%. An exception is an outsized contribution from AA:NT:NT terms involving amino acid site 29 (median effect 2%). Only AA:AA:AA, NT, and NT:NT terms have negligible contributions to bound complexes (Fig. 3.3A; median effect <2% per order).

To ask how terms of different orders affect the number and identity of bound complexes, I truncated the model at various orders by setting higher-order terms to zero and recalculated genetic scores. The results showed that specificity and epistasis are both necessary to create binders. The most truncated model, which only included amino acid and nucleotide main effects on nonspecific binding, did not produce any strongly bound complexes (Fig. 3.3B). Adding intramolecular epistasis created only a small number of strongly bound complexes, as did adding low-order specificity (AA:NT) terms. The combination of intramolecular epistasis and AA:NT terms created more than the addition of each subset of terms alone, but still only around half the number of strong complexes as in the full model. Only after adding higher-order specificity terms—particularly AA:AA:NT interactions—did the number of bound complexes approach the number in the full model. For all truncated models, the majority of bound complexes were also strongly bound in the full model, indicating that epistasis tends to create strong binding from

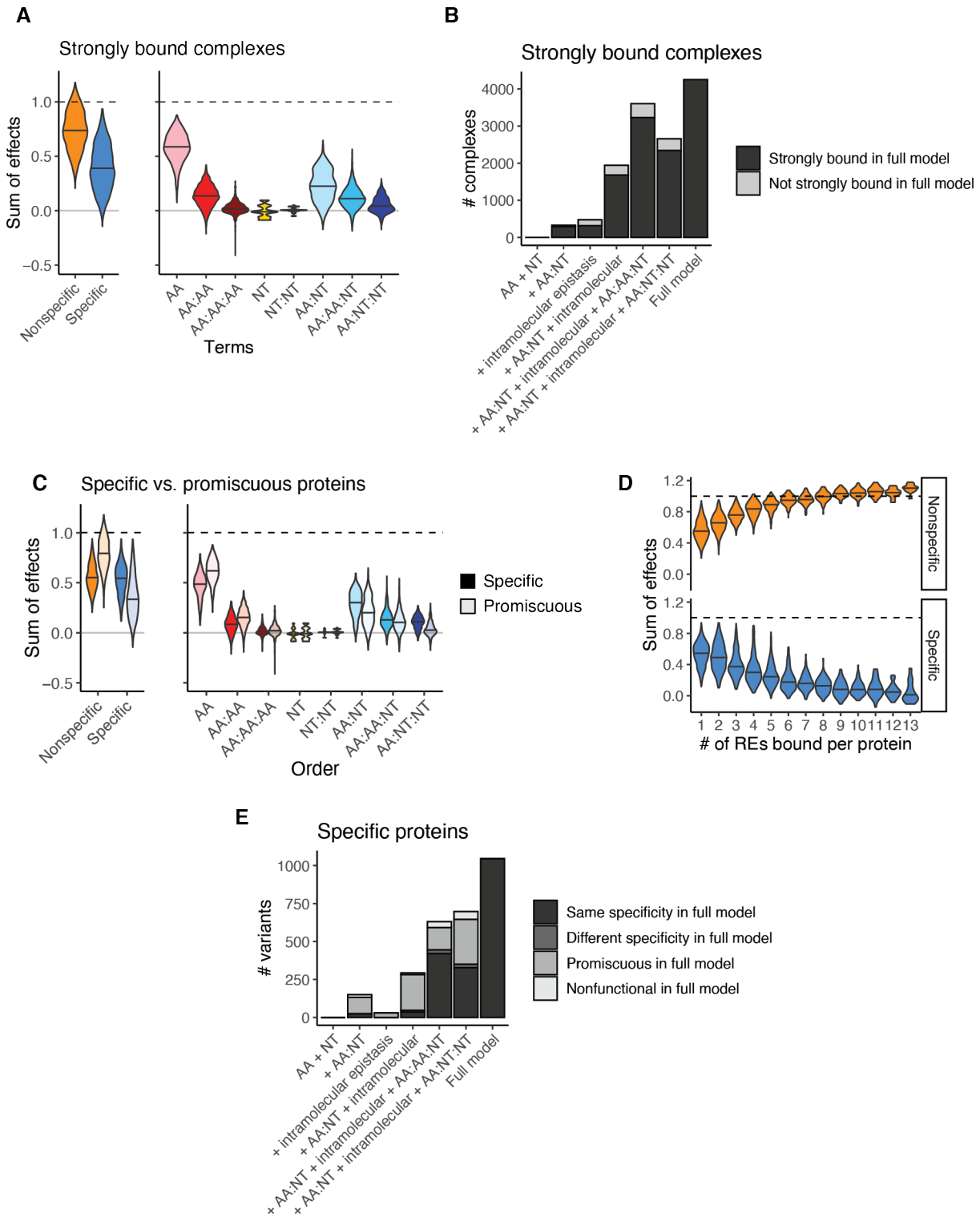

Figure 3.3. Genetic architecture of bound and specific variants. (**A**) Distribution of effects summed across nonspecific or specific binding terms (left) and for each order of interaction (right) for strongly bound complexes. Horizontal lines within each violin plot show the median; the dashed line shows the genetic score of the wild type complex (1). Violin plots are scaled to

have the same width within each panel. (**B**) Bars show the number of strongly bound complexes predicted for each truncated model. The leftmost bar corresponds to a base model with all except AA and NT terms set to zero; the x-axis labels on all other bars indicate terms added to this base model. Bar shading shows complexes that are strongly bound in the full model (dark) or not strongly bound in the full model (light). (**C**) Same as in **A**, but split by complexes involving specific (dark colors) and promiscuous (light colors) protein variants. (**D**) Distribution of effects summed across nonspecific (top) or specific (bottom) binding terms, plotted for complexes involving protein variants that bind a given number of RE variants (x-axis). The dashed horizontal line shows the genetic score of the wild type complex. (**E**) Bars show the number of specific protein variants predicted for each truncated model. Shading indicates the protein phenotype in the full model: darkest, variants with the specificity for the same RE in the full and truncated models; medium dark, variants with specificity for different REs in the full and truncated models; medium light, variants that are specific in the truncated model but promiscuous in the full model; lightest, variants that are specific in the truncated model but nonfunctional in the full model.

complexes that are weakly bound, rather than destroy binding for complexes with large positive contributions from lower-order effects. These analyses show that specificity and epistasis, while being of much smaller magnitude than main amino acid effects on nonspecific binding, are nonetheless necessary for creating wild type binding levels.

3.3.4 High-order intermolecular epistasis is necessary for specific binding

I next examined the contribution of different effect orders to RE-specific binding. I classified protein variants as specific if they bound to only one of the 16 REs, and promiscuous if they bound to multiple REs. I then compared the contribution of different model orders to the genetic score of complexes involving each of these classes of proteins.

The analysis showed that protein-DNA specificity involves a tradeoff between states that are beneficial for nonspecific binding and for specific binding. Complexes involving specific protein variants tend to have similar contributions from nonspecific and specific binding terms, whereas complexes involving promiscuous protein variants have much larger contributions from nonspecific than from specific binding terms (Fig. 3.3C, left). This suggests that having too high of a nonspecific binding score is detrimental for protein-DNA specificity. Indeed, nonspecific

binding scores tend to increase quantitatively with the number of RE variants bound per protein, and specific binding scores tend to decrease (Fig. 3.3D). Differences in nonspecific and specific binding scores between specific and promiscuous variants are not restricted to main effects, but instead come from multiple effect orders—most of the difference in specificity score comes from AA:NT and AA:NT:NT interactions, and for binding scores from AA and AA:AA terms (Fig. 3.3C, right; Fig. B.3B).

To test which orders of effect are required to generate specific binding, I predicted the number and identity of specific protein variants under the same truncated models as before and compared them to the identity of specific variants in the untruncated model. The results show main nonspecific binding and AA:NT terms are sufficient to generate some specific variants, but these are almost all promiscuous in the full model (Fig. 3.3.E). Only after higher-order intermolecular epistasis—AA:AA:NT or AA:NT:NT terms—are added does the identity of specific variants start to become more similar to the full model, although each category alone still misses around half of the specific variants of the full model (Fig. 3.3E). All of the truncated models predict a large number of variants to be specific that are promiscuous in the full model (Fig. 3.3E), implying that in the full model, the combined effect of multiple orders of epistasis often creates promiscuity by conferring binding on multiple REs. Together, these results suggest that specific binding requires a moderate nonspecific binding score and a select few contributions from higher-order intermolecular interactions to ensure that a protein variant only binds strongly to a single RE variant.

3.3.5 Mutations that alter binding and specificity use site-specific genetic mechanisms

During evolution, mutations can alter the binding and specificity of transcription factors

to their cognate DNA motifs. To understand the genetic mechanisms by which this occurs, I calculated the contribution of each model term to the total effect of single mutations that alter DNA binding or specificity by taking the difference between the effect of the starting state and the ending state for that term.

I first identified all single amino acid mutations cause the protein to go from unbound to bound on a given RE. There were 256,268 such mutations distributed relatively evenly across the four variable sites (Fig. 3.4A). Mutations that create binding at the central two sites tended to have larger effects on genetic score (Fig. 3.4B), likely due to the presence of large-effect AA terms those sites (Fig. 3.2B). Examining the contribution of nonspecific and specific binding terms at each of these four sites revealed that mutations tend to use different genetic mechanisms to create binding depending on the amino acid site. Mutations at the two central sites, especially site AA26, tend to have much larger contributions from nonspecific binding effects compared to specific effects (Fig. 3.4C). In contrast, mutations at site AA25 have a relatively equal contribution from specific and nonspecific binding terms, whereas mutations at AA29 tend to have the largest contribution from specificity terms. Further partitioning contributions by order of interaction shows that binding-creating mutations at different sites also differ in the importance of different interaction orders. Although mutations at sites AA26 and AA28 both get their largest contributions from nonspecific binding terms, most of that effect comes from AA effects at site AA26, whereas at site AA28 a large fraction comes from AA:AA terms (Fig. 3.4D). Mutations at site AA25 tend to have even larger contributions from AA:AA terms than from AA terms. These trends broadly reflect the distribution of effect sizes at these sites (Fig. 3.2, Fig. B.2). Mutations that create binding can therefore occur via a variety of genetic mechanisms—through enhancing nonspecific binding generally or in the context of other

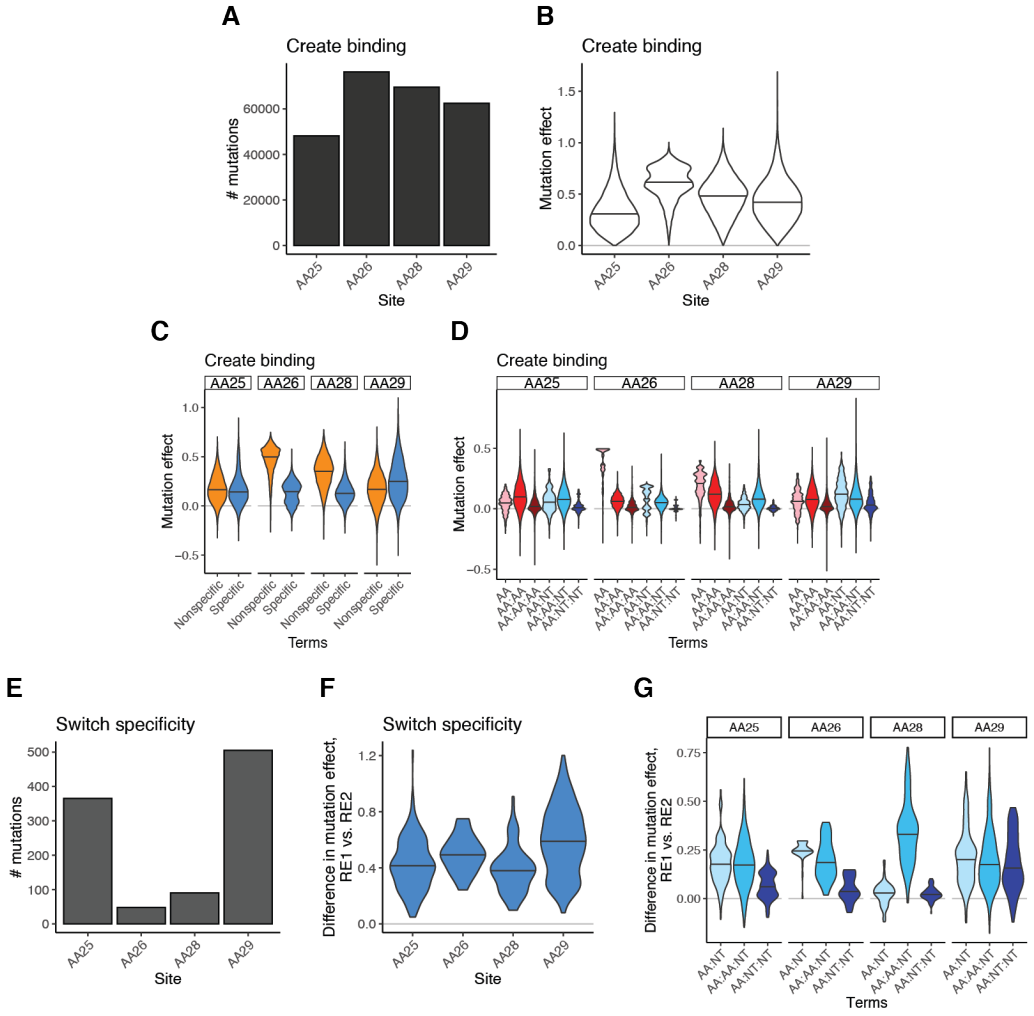

Figure 3.4. Genetic architecture of function-switching mutations. (**A**) Number of mutations per protein site that convert the protein from a nonbinder to a binder on a given RE. (**B**) Distribution of effect sizes per for mutations that create binding, split by the site of the mutation. (**C**) Contributions from nonspecific (orange) and specific (blue) binding terms for mutations that create binding, split by the site of the mutation. (**D**) Contributions from each order of effect for mutations that create binding, split by the site of the mutation. Note that NT and NT:NT terms do not contribute to mutation effects, since mutations are with respect to a single RE variant. (**E**) Number of mutations per protein site that change the protein's specificity from one RE to another. (**F**) Difference in mutation effect on the starting RE (RE1) and the ending RE (RE2) for mutations that switch specificity, split by the site of the mutation. Note that difference in effect is always positive because the starting complex (the starting protein genotype on RE1) is defined as having a lower genetic score than the ending complex (the ending protein genotype on RE2). Violin plots are colored blue since only specific binding terms contribute to differences in mutation effects on different REs. (**G**) Same as **F**, but split by the order of the effect.

existing amino acid states in the protein, or through enhancing specific binding to particular

DNA bases.

I next identified all single amino acid mutations that switch the specificity of the protein from one RE to another. These mutations were much rarer than mutations that create binding—only 1,008 in the entire sequence space. The majority (81%) cause specificity to change between REs that differ by a single DNA base. Specificity-switching mutations are highly enriched at sites AA25 and AA29 (Fig. 3.4E), the sites with the strongest AA:NT and AA:NT:NT effects (Fig. 3.2C, Fig. B.2D). To understand the genetic architecture of these mutations, I calculated the difference between the effect of each amino acid mutation on binding to the starting RE and to the ending RE, where the starting complex is defined as having the lower genetic score. Note that only specificity terms contribute to differences in amino acid mutation effects on different REs. Overall specificity effects do not differ strongly by amino acid site (Fig. 3.4F), but specificity-switching mutations at different sites tend to use different orders of intermolecular epistasis to achieve their effect (Fig. 3.4G). Specificity-switching mutations at sites AA25 and AA26 tend to use a combination of AA:NT and AA:AA:NT terms, with smaller contributions from AA:NT:NT terms, while mutations at AA29 have similar average contributions from all orders of effect. Mutations at AA28 get almost all of their specificity effects from AA:AA:NT terms, mostly from interactions involving AA29 and NT4 (Fig. B.4); they must therefore occur in the context of other amino acid states that contribute positively to these interactions. All in all, these results show that the mechanisms that are used to generate function-switching mutations depend on the architecture of genetic interactions that exist for a given site.

3.3.6 Binding and specificity effects are correlated

Specific binding for REs that contain adenine at the two variable sites, especially at site 4, are much more frequently encoded across the RH-RE sequence space than specificity for other REs,

with specificity for AA-RE the most frequent overall (Fig. 3.1C). This trend implies that there are features of the genetic architecture that make binding to adenine more favorable on average than binding to other DNA bases. Nonspecific nucleotide effects do not explain this trend—adenine at site NT3 does have a weak positive effect on binding, but adenine at site NT4 is the most negative of all NT terms (Fig. 3.2E), and pairwise interactions between nucleotide states do not make up for this effect (Fig. B.2E). Furthermore, NT and NT:NT terms explain a very small fraction of the genetic variance of the model overall (Fig. 3.2A). Nonspecific effects of nucleotide states on DBD binding therefore do not explain the trend towards adenine specificity seen in the GP map.

Another possible source of the bias toward adenine specificity is that states that are beneficial for nonspecific binding tend to be even better with adenine, especially at site NT4. The reverse could also be true: states that are deleterious for nonspecific binding are even worse for binding to adenine. This would cause variants with strong nonspecific binding scores to be more likely to bind to adenine than to other bases at the variable RE sites. To test this hypothesis, I computed the average change in the effect of amino acid states caused by interaction with each individual nucleotide state; this was done by taking the regression slope of AA:NT against AA terms for each individual nucleotide state. I found that nonspecific amino acid effects are amplified by 27 ± 13% (95% confidence interval (CI)) on adenine at site NT4 (NT4A), and weakened by other DNA bases at NT4 (Fig. 3.5A; note that due to the zero-mean property of RFA, the slopes must sum to zero across all nucleotide states at a site). The effect of NT4A is significant only for the two central amino acid sites (Fig. B.5A), which have the strongest nonspecific binding effects (Fig. 3.2B). A much weaker but still significant amplifying effect exists with adenine at site NT3 for the two central amino acid positions (Fig. B.5A), although the

effect is not significant when averaged across all four amino acid sites (Fig. 3.5A). Nonspecific binding effects at site AA26 and AA28 thus tend to be stronger on adenine and weaker on other bases.

To examine whether other orders of effect are also amplified or weakened in the context of adenine, I performed a similar analysis for the relationship between AA:AA:NT and AA:AA terms, and between AA:NT:NT and AA:NT terms. In both cases, lower-order effects were amplified in the context of adenine at either nucleotide site and weakened or unaffected in the context of other DNA bases. AA:AA interactions were 21 ± 3% (95% CI) stronger in the context of NT3A, and 19 ± 3% (95% CI) stronger in the context of NT4A (Fig. 3.5B). The strongest concordance was observed for interactions not involving site AA26 (Fig. B.5A), similar to the distribution of AA:AA and AA:AA:NT effect sizes (Fig. 3.2A, C). AA:NT interactions were also stronger when adenine was at the other nucleotide site, with AA:NT4 interactions 23 ± 6% (95% CI) stronger in the context of NT3A, and AA:NT3 terms 46 ± 10% (95% CI) stronger in the context of NT4A (Fig. 3.5C). The effect of NT4A on AA:NT3 interactions was strongest for site AA29—AA29:NT3 effects are on average almost doubled in the context of NT4A—whereas the effect of NT3A on AA:NT4 effects was more similar across amino acid sites (Fig. B.5C). Effects on of AA, AA:AA, and AA:NT states therefore tend to be systematically amplified in the context of adenine at one or both RE sites, consistent with the bias towards specificity for adenine REs among strongly bound protein variants.

To evaluate the extent to which these adenine-specific effect correlations contribute to the map's bias towards encoding adenine specificity, I shuffled effects between states at individual sites and site combinations for each order of specificity term to break the association between high and low-order effects, while preserving the overall effect size for each site combination and

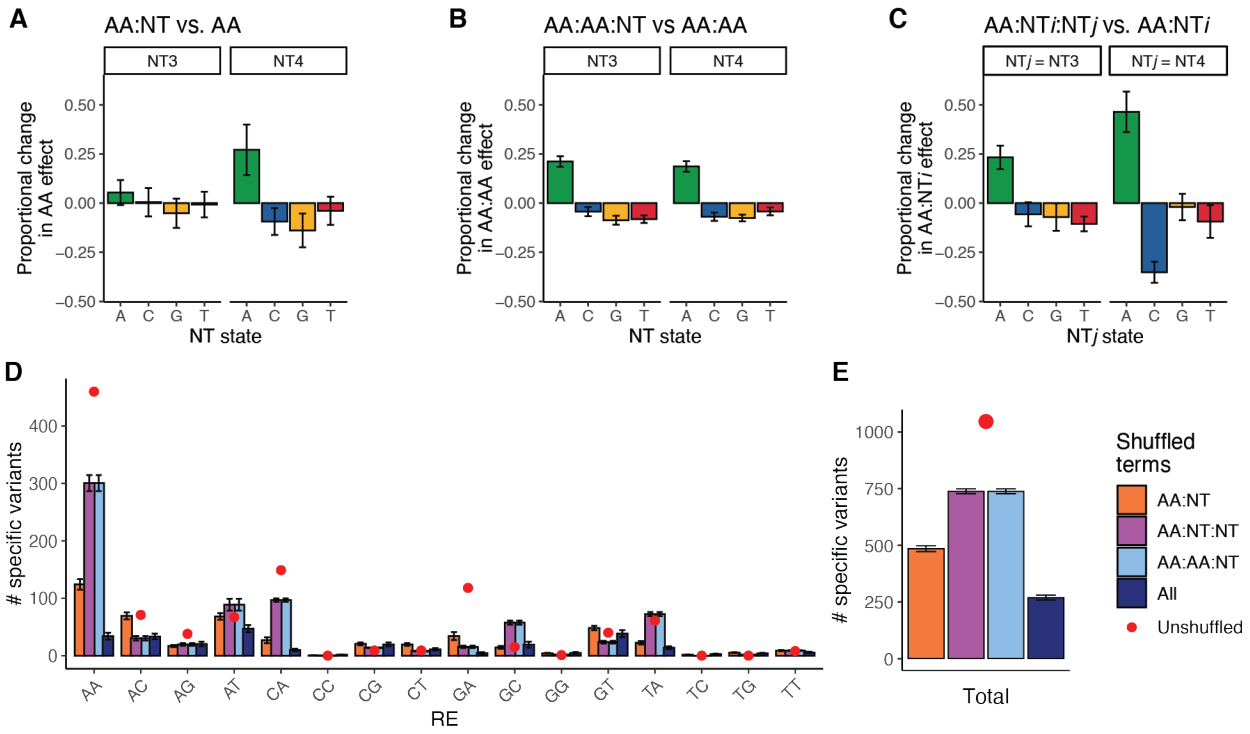

Figure 3.5. Adenine amplifies lower-order effects. (**A**) Bars show the mean proportional change in amino acid main effects on nonspecific binding caused by each single nucleotide state (x-axis) in the RE, calculated as the slope of AA:NT effects regressed against AA effects. Panels, nucleotide sites. Error bars, 95% CI of the mean. Colors, nucleotide states as in **Fig. 3.2E**. (**B**) Same as **A** but for the proportional change in pairwise amino acid effects on nonspecific binding caused by single nucleotide states. (**C**) Same as **A** but for the proportional change in amino acid-nucleotide interactions caused by the nucleotide at the other site. *i* indexes the nucleotide site involved in the AA:NT interaction; *j* indexes the other nucleotide site (panels). (**D**) Bars show the mean number of specific variants per RE predicted under $n = 100$ models in which the associations between specificity terms and sequence states have been randomly shuffled. Orange, shuffling AA:NT terms with respect to AA states; pink, shuffling AA:NT:NT terms with respect to AA:NT states; light blue, shuffling AA:AA:NT terms with respect to AA:AA states; dark blue, shuffling all of the above terms. Error bars, 95% CI of the mean. Red dots show the number of specific variants per RE under the unshuffled model. (**E**) Same as **D**, but showing the total number of specific variants across all REs.

nucleotide state. I then recomputed genetic scores for all complexes and asked how many variants bound specifically to each RE. Shuffling AA:NT terms strongly reduced the number of protein variants that bound specifically to REs with NT4A compared to the unshuffled model; the effect on REs with NT3A was much weaker (Fig. 3.5D). Similar results were obtained for the number of bound protein variants per RE (Fig. B.5D). In addition, the shuffled model produced

only about half as many specific protein variants and bound complexes as the full model (Fig. 3.5E; Fig. B.5E). Shuffling AA:AA:NT and AA:NT:NT terms had similar but much weaker effects (Fig. 3.5D, E; Fig. B.5D, E). When all orders of specific effects were shuffled together, the resulting GP map had a much more unbiased distribution of specific variants per RE (Fig. 3.5D), and only 35 ± 1% (95% CI) as many bound complexes as the unshuffled model (Fig. B.5E). The correlations between specificity and lower-order effect terms thus explain the majority of the map's bias toward adenine specificity, as well as most of the bound variants in the map.

While these results imply that the biased phenotype production propensity of the AncSR2-RE system is caused by a correlated genetic architecture of nonspecific and specific RE binding, this mechanism is difficult to rationalize structurally. The sites at which the correlation between AA and AA:NT4A terms are the strongest are AA26 and AA28 (Fig. B.5A), which do not directly contact A4 nor its complementary T–4 base (Fig. 3.1A). In general, correlations are strongest at sites/site combinations that have the strongest effects on binding (Fig. 3.2B, C; Fig. B.2A; Fig. B.5A–C). This observation suggests an alternative possible cause for the observed correlations—that they are an artifact of the fitting procedure caused by the limited dynamic range of the DMS assay, which causes complexes with observable fluorescence to be enriched for combinations of amino acids and nucleotides that are good for nonspecific binding. This could cause erroneous inference of positive epistasis between these states, since they tend to be observed together within the subset of highly fluorescent complexes. Further optimization of the RFA fitting procedure may be able to rule out this potential artifact.

3.4 Discussion

By decomposing the contributions of individual sequence states and their epistatic interactions in a combinatorial intermolecular genotype-phenotype map, I showed that the genetic architecture of protein-DNA binding and specificity is surprisingly simple. The vast majority of genetic variance in both nonspecific and specific binding was captured by the first-order effects of amino acid states on nonspecific binding, and by pairwise interactions between individual amino acid residues and DNA bases. This finding is in line with previous studies using RFA to decompose the genetic architecture of numerous molecular genotype-phenotype maps[47,139], adding to a growing body of evidence that sequence-function relationships are much less epistatic than previously thought.

The fact that the genetic architecture of protein-DNA binding in this system is dominated by low-order AA and AA:NT effects makes sense from a structural perspective. Intermolecular interactions are mediated by specific biochemical contacts between molecules—salt bridges, hydrogen bonds, hydrophobic packing, etc. An individual amino acid side chain can only make a limited number of these contacts with other atoms—either in the same molecule or in an interaction partner. In the case of the AncSR2-RE system, the variable amino acid side chains are all oriented out towards the DNA and not towards other residues within the protein, meaning that strong AA:AA interactions are unlikely, except in the case of electrostatic repulsion between long, charged side chains. Instead, they primarily interact structurally with the DNA—either with backbone atoms or invariable DNA bases as is primarily the case with AA26 and AA28, or with variable DNA bases as with AA25 and AA29. Only the AA29 residue is oriented in such a way as to be able to make strong specific contacts with both variable bases simultaneously, and as such is the only site with a high frequency of third-order interactions with pairs of nucleotide

bases. It is likely that if more sites in the protein and DNA had been mutated, other site-specific interactions between molecules would likely have been uncovered. Nevertheless, these findings suggest that the genetic architecture of protein-DNA interactions is, to a large degree, structurally rationalizable based on low-order effects, and not dominated by high-order epistasis as has been previously suggested[11,20,21,23,24,45,48,49].

Despite the fact that most genetic variance comes from low-order effects, however, moderately high-order epistatic terms are also clearly important for genetic architecture. AA:AA, AA:AA:NT, and AA:NT:NT effects tend to be smaller in magnitude (except for AA29:NT3:NT4 terms), perhaps arising from less direct interactions between amino acid residues and nucleotide bases. They thus have a small contribution to genetic variance, but the fact that there are very many of these possible interactions means that they make up a large proportion of terms needed to accurately describe binding. The sum of these small effects over many combinations of sites contributes substantial amounts to the genetic score of bound complexes and to the effects of mutations that change a protein's binding or specificity, to the point where they are necessary for the binding of most strong complexes. This is consistent with findings from a previous study that analyzed the genetic architecture of binding to two of the 16 REs studied here[52]. The fact that epistatic interactions have such a large effect on the number of bound and specific complexes is likely due to the fact that the genetic score of the wild type complex—the reference genotype for binding—has a very high genetic score compared to the vast majority of genetic variants in the map. To reach this level of binding therefore requires substantial contributions across many effect orders. It is also possible that the effect of epistasis on genotypes with lower genetic score is underestimated because epistasis cannot be estimated for state combinations that only occur at the lower bound of fluorescence. Nevertheless, these findings show that while AA and AA:NT

terms are sufficient to predict most of the variation in binding and specificity, knowledge of many small-magnitude interactions between combinations of indirectly interacting residues is necessary for finer scale phenotype predictions, and are especially important for generating wild type levels of function.

Mutations that alter a protein's phenotype—either creating binding to an RE or switching specificity between REs—used different genetic architectures depending on the mutated site. This reflects the fact that different amino acid sites engage in different sets of interactions with other residues in the protein or nucleotides in the DNA. Mutations that predominantly affect lower-order terms are less context-specific, and their effects may be more conserved during evolution than mutations that have larger effects on higher-order terms. These site-specific genetic architectures may explain the observation that mutations at different sites tend to vary in their effect conservation over evolutionary time. Notably, most mutations that change specificity have sizable contributions from higher-order intermolecular interactions (AA:AA:NT and AA:NT:NT terms); specificity switches therefore require the correct combinations of states at sites in the protein and DNA outside of the site that is actually mutated in the specificity switch. This is consistent with the very low frequency of specificity-switching mutations in this dataset; whether similar epistatic architectures apply to function-switching mutations in other systems is an interesting question for future investigation.

A surprising finding of this study was that the effects of sequence states on specific and nonspecific RE binding were correlated, which was sufficient to explain the observed bias towards adenine-specific protein genotypes in the genotype-phenotype map. Biased phenotype production is a widespread feature of many biological systems that can be an important driver of patterns of phenotypic diversity[32,34,84,101,110,140]; understanding its genetic basis is therefore key

for understanding how the genetic architecture of GP maps shapes evolutionary outcomes. It is still unclear whether the correlations found in this study reflect the true genetic architecture of the system, or whether they are an artifact of the model fitting procedure and limited dynamic range of the dataset. Resolving this question could have major implications for the generality of the observed patterns of phenotype bias—if the bias is primarily explained by correlations between specificity effects (intermolecular epistasis) and lower-order binding effects, then it is unlikely to be generalizable across different portions of sequence space and may instead be a property of the particular GP map characterized here. If instead the bias is caused not by epistasis but by some nucleotide states being better for binding than others on average across (NT effects), then the bias we observe here may be more generalizable across sequence space and perhaps even in other protein-DNA binding systems. Improved experimental designs and model fitting procedures can help to overcome this shortcoming in the future.

This study has implications for the evolution and engineering of intermolecular interactions. Epistasis introduces context dependence into the effects of mutations, making evolution more contingent[43,44,141–145] and interactions more difficult to predict and design[146–148]. If variation in binding and specificity are dominated by low-order genetic effects as seen in this study and others, then evolving and designing new interactions should be relatively straightforward. However, the relatively small amount of genetic variance contributed by epistasis between amino acid states on nonspecific and specific binding can still have subtle quantitative effects, which may be particularly consequential if very strong or precise levels of binding or specificity are required. While at a bird's eye view the genetic architecture of intermolecular interactions may be simple, knowledge of many subtle interactions may still be important for a detailed understanding of the evolution and function of molecular systems.

3.5 Methods

3.5.1 RFA model fitting

To infer the genetic architecture of protein-DNA specificity in the AncSR2-RE system, I applied RFA[47] to the combinatorial DMS dataset generated in Chapter 2, in which all combinations of amino acid states at four sites in the recognition helix were assayed for binding to all combinations of nucleotide states at two sites in the RE using a yeast-based GFP reporter system. Of the 2,560,000 possible protein-DNA complexes, 658,475 (25.7%) had high-confidence quantitative estimates of fluorescence obtained from cell sorting and sequencing (mean Pearson's $r^2 = 0.70$ across 3 replicates for all complexes, $r^2 = 0.95$ for complexes above the lower bound of fluorescence); I used this dataset as the training data for the RFA model.

In RFA, the "genetic score" of a variant is the sum across all states and state combinations contained within that variant; the genetic score is then transformed through some function to obtain the variant's phenotype. The effect of a state or state combination is defined as the mean genetic score of all genotypes containing that state (combination), relative to the genetic score predicted by the summed effects over all lower-order states and state combinations; the intercept is defined as the global mean genetic score. In the AncSR2-RE DMS dataset, there are four variable sites in the protein each with 20 possible amino acid states, and two variable sites in the RE each with 4 possible nucleotide states; epistasis can in theory occur between any combination of amino acid and/or nucleotide states. In practice, epistasis is observed to be mostly restricted to third-order or lower[47,52]; I therefore considered effects up to third-order intramolecular and third-order intermolecular epistasis (Table 3.1), as well as fourth-order AA:AA:NT:NT interactions. Almost all of these state/state combinations were observed at high frequency in the data (Fig. B.1C; 98.3% observed in at least 10 genetic backgrounds), so their

effects can be estimated from the data. I encoded the genotype of each complex in the dataset as a vector of 0s and 1s, where each element corresponds to a state or state combination in the model; 1 indicates that state is present in the complex and 0 that it is absent. Combining the genotype vectors for all complexes in the dataset forms the model matrix on which the RFA model is fit.

To account for nonspecific epistasis introduced by the fluorescence bounds of the yeast GFP reporter assay, I used a sigmoid function $g = L + (U - L)/(1 + e^{-x})$ to model the relationship between genetic score and fluorescence. Failure to account for measurement bounds can lead to highly inflated estimates of specific epistasis between sequence states[47], especially when a large proportion of variants are at the measurement bounds—as is the case in the AncSR2-RE dataset (Fig. B.1D). The sigmoid function also has the benefit of capturing the nonlinear relationship between effects of mutations on free energy of folding/DNA binding and occupancy of the DNA-bound state, which likely underlies the bounded range of GFP in the assay. I inferred the upper and lower bound parameters U and L of the sigmoid function by fitting a model with up to second-order intramolecular and third-order intermolecular epistasis using unregularized nonlinear least-squares regression. I inferred different lower bound parameters for each of the 16 different RE variants; this was to account for slight variation in the autofluorescence levels of the 16 yeast strains that each reported fluorescence on a different RE (Fig. B.1E). Failure to account for this variation in the lower bound by inferring a single L across all REs resulted in much larger effect size estimates for NT and NT:NT terms that reflected strain-specific autofluorescence levels (Fig. B.1F, G); these were also poorly correlated with effects inferred in the RE-specific L model (Fig. B.1H). Other model terms were systematically smaller in the RE-specific L model, but well correlated with those in the universal L model (Fig.

B.1H).

I next estimated the main and epistatic effects of amino acid and nucleotide states using L1-regularized regression implemented in the R package *glmnet* v.4.1-6[131]. The previously-fit L and U parameters were used to specify a link function, and 10-fold cross validation (CV) was used to obtain the regularization parameter that minimized out-of-sample mean squared error (MSE). To evaluate how many orders of epistatic effects were needed to accurately capture variation in fluorescence, I fit a series of models containing different orders of effects and evaluated their fit to the data using the distribution of out-of-sample MSE obtained from CV. I found that excluding AA:AA:AA terms significantly increased the mean MSE (>1 standard error) compared to models that included them, whereas excluding AA:AA:NT:NT terms did not significantly increase mean MSE (Fig. B.1A); I therefore chose to include a model with up to third-order intramolecular and intermolecular interactions but excluding fourth-order interactions as the best-fit RFA model (Table 3.1). This model fit the data very well, with $R^2 = 0.88$ (Fig. B.1B).

The above fitting procedure does not guarantee that effects at all sites and site combinations are zero-centered, a key tenet of RFA. To enforce the zero-mean constraint, I used a post hoc normalization procedure developed by Park et al.[47] Briefly, the zero-mean constraint can be enforced for terms of any given order k using the formula

$$\delta_{i_1,\ldots,i_k}(s_1,\ldots,s_k) = \beta_{i_1,\ldots,i_k}(s_1,\ldots,s_k) + \sum_{l=1}^{k}\left((-1)^l \sum_{\alpha_1<\cdots<\alpha_l\in\{1,\ldots,k\}} \overline{\beta_{i_1,\ldots,i_k}(s_1,\ldots,s_k)}_{i_{\alpha_1},\ldots,i_{\alpha_l}}\right).$$

This formula calculates the mean-centered effect term $\delta_{i_1,\ldots,i_k}(s_1,\ldots,s_k)$ corresponding to the

original estimated effect $\beta_{i_1,\ldots,i_k}(s_1,\ldots,s_k)$ by iteratively subtracting and adding the sum of effects averaged over every single state, every pair of states, etc. The notation $\overline{\beta_{i_1,\ldots,i_k}(s_1,\ldots,s_k)}_{j_1,\ldots,j_l}$ refers to the mean of terms $\beta_{i_1,\ldots,i_k}(s_1,\ldots,s_k)$ across sites $j_1,\ldots,j_l$. The terms corresponding to the highest epistatic order K in the model are normalized first; to ensure that the model still specifies the same genotype-phenotype relationship, lower-order terms are then re-estimated using regression to find a model of order $K-1$ that predicts genetic scores $y_K - z_K$, where y_K are the genetic scores predicted by the original model and z_K are the contributions to genetic score from normalized terms $\delta_{i_1,\ldots,i_K}(s_1,\ldots,s_K)$. The re-estimated terms of model order $K-1$ are then normalized using the same formula, and the procedure is repeated iteratively until all terms are normalized.

3.5.2 Effects of nucleotide states on binding and specificity

Using linear regression, I calculated the extent to which the effects of amino acid states on binding and specificity are amplified/weakened in the context of specific nucleotide states. For each of the eight nucleotide states (four at each site), I performed regression of AA:NT terms against AA terms, AA:AA:NT terms against AA:AA terms, and AA:NT:NT terms against AA:NT terms. Positive regression coefficients indicate that a given lower-order term is amplified in the context of a given nucleotide state, and negative coefficients that lower-order terms are weakened in the context of a given nucleotide state. To test whether nucleotide-specific amplification/weakening effects were specific to particular sites/site combinations, I performed the regression analysis separately for each possible site/site combination at the non-NT site (or for AA:NT site combinations at the other NT site for the case of AA:NT:NT terms). Significance was evaluated using 95% confidence intervals. Note that for a given nucleotide site, regression

coefficients must sum to zero due to the zero-mean constraint of RFA; amplification of effects in some nucleotide contexts must therefore be balanced by a weakening of effects in others.

To evaluate the extent to which nucleotide-specific amplification/weakening effects caused nonuniformity in the number of specific protein variants per RE, I performed a permutation test to break the observed correlations between specificity terms and lower-order terms. For each set of AA:NT, AA:AA:NT, and/or AA:NT:NT terms associated with a particular NT state, I shuffled the inferred effects with respect to the state(s) at the other site(s), such that the total variance for a given effect order and NT state was preserved, but the associations with states at the other site(s) were randomized. I shuffled effects for each site independently and used the same shuffling order for all terms involving the same NT site in order to preserve the zero-mean property. For example, for AA25:AA26:NT3 terms, I first shuffled effects relative to the state at site AA25 using the same order with respect to all AA26 and NT3 states, then shuffled effects relative to the state at site AA26 using the same order with respect to all AA25 and NT3 states; this keeps the sum of effects across all AA25 states, AA26 states, and AA25:AA26 state combinations constant for each NT3 state. I recalculated genetic scores for all variants under the shuffled models and counted the number of variants that were bound and specifically bound to each RE, as well as the total across all REs. I performed this procedure 100 times for each effect order and for the combination of all three effect orders.

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 Summary

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 Introduction

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.[149] 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.

4.3 Results

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[150,151]. Hadzipasic et al. also place microsporidians as the most basally branching eukaryotic lineage, despite strong evidence for their inclusion within Fungi[152]. 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.

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[153].

A more likely cause of the incongruence of Hadzipasic et al.'s tree with the species

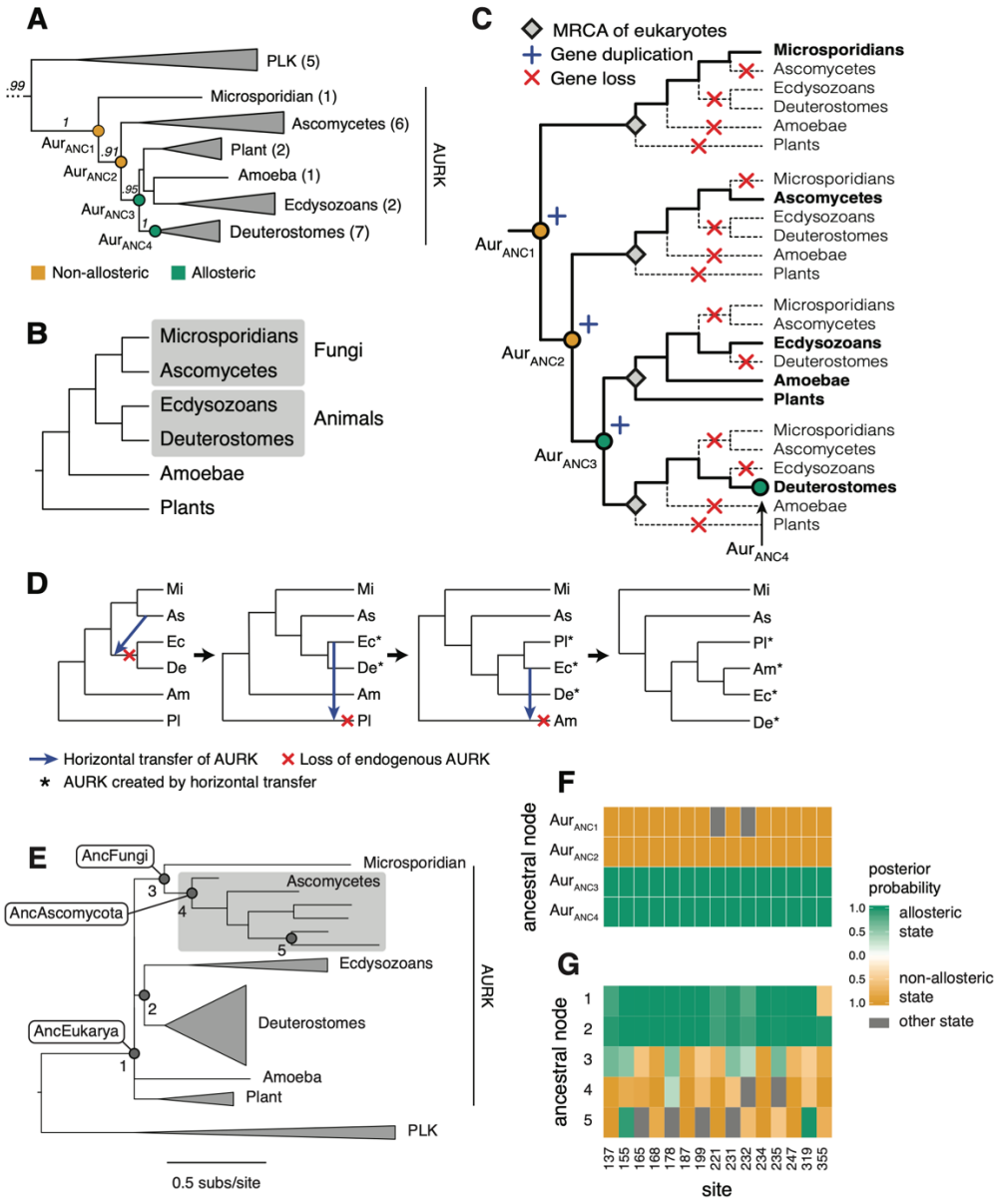

Figure 4.1. A plausible phylogeny reverses Hadzipasic *et al.*'s ancestral reconstructions. (**A**) The phylogeny of AURKs and their nearest outgroup (PLKs) used by Hadzipasic *et al.* Parentheses indicate the number of sequences in each clade. Circles mark the experimentally characterized ancestors, 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.

phylogeny is long branch attraction (LBA)[71]. 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[72]. Strong support for misplaced branches is consistent with systematic bias caused by LBA[72].

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.

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[154]), 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.

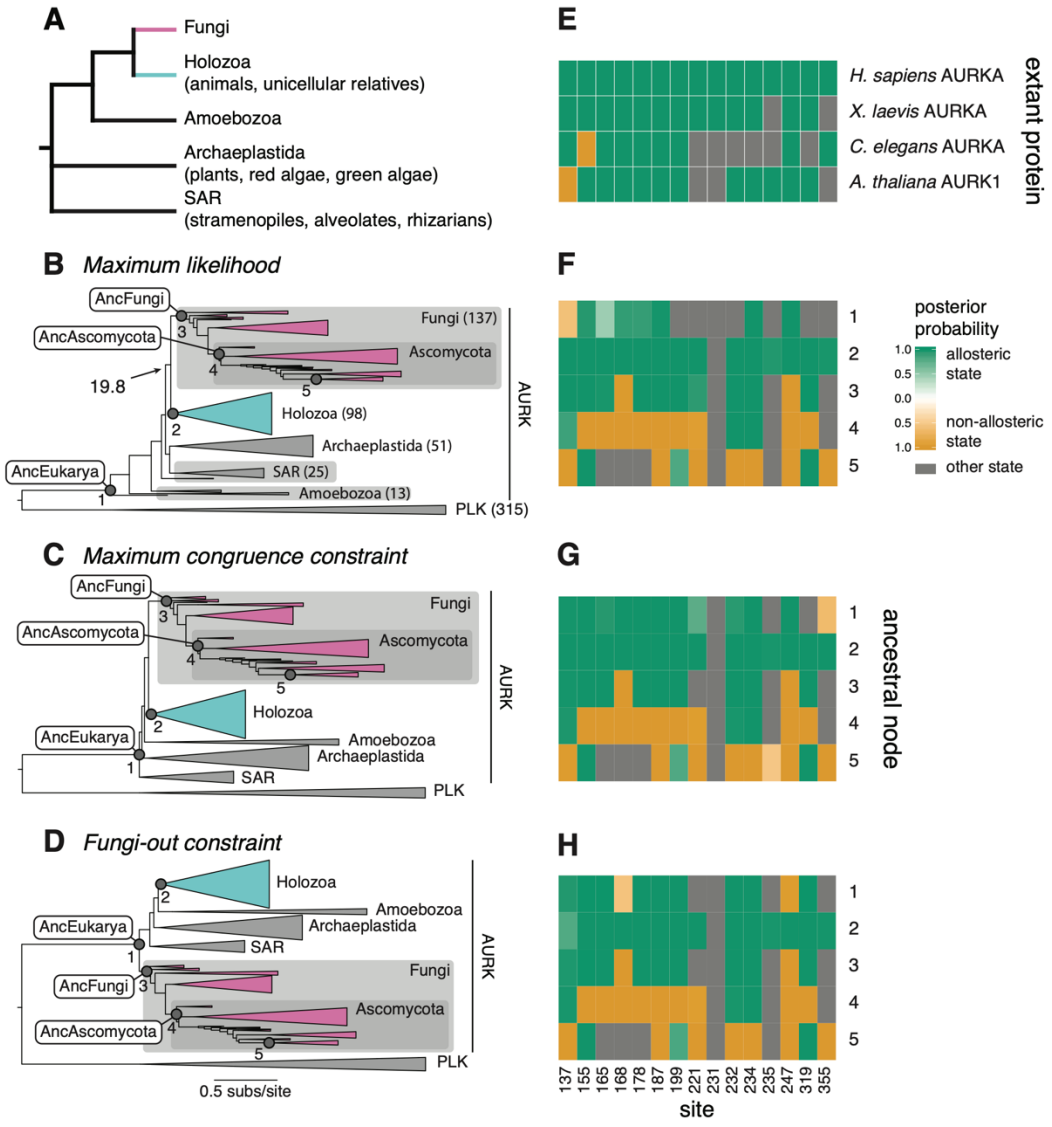

Figure 4.2. Improved sequence sampling reverses Hadzipasic *et al.*'s ancestral reconstructions. (**A**) The established phylogeny of the major eukaryotic groups. Polytomy, branching order not established. (**B**) Maximum likelihood phylogeny when AURK and PLK are densely sampled. Numbers in parentheses 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 ($p < 0.01$ by χ^2 approximation[155]). (**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–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.

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. 4.2E)[156–158]. 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. 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[73]. 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.

4.4 Discussion

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.

4.5 Methods

4.5.1 Ancestral sequence reconstruction using Hadzipasic et al. sequences under congruence constraint

We acquired the AURK and PLK sequences analyzed by Hadzipasic *et al*. We aligned them using MUSCLE (v3.8.425)[159], 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)[160] to infer the constrained ML phylogeny, imposing the constraint shown in Fig. 4.1B, and used PAML (v4.8)[161] 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).

4.5.2 Phylogenetics and ASR with improved sequence sampling

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[162]. 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)[163]. 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)[164]. 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)[165]. For the maximum congruence constraint analysis, we imposed the topological constraint shown in Fig. 4.2A 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. 4.2A, 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[166], 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

In this dissertation, I analyzed the structure and genetic architecture of two ancestral GP maps, and how biases in the production and accessibility of phenotypes in those maps shaped the phenotypes that evolved from them. I experimentally demonstrated the presence of production bias by mapping specificity for all possible substrates of a given class in the sequence space of a protein, and linked this bias to the lineage-specific distribution of phenotypes observed in nature. I also decomposed the main and epistatic effects of individual amino acids and nucleotides in this genotype-phenotype map to uncover the genetic architecture of intermolecular binding and specificity. I now link these findings to understand how the intrinsic structure of this biological system influenced its evolutionary history.

One of the key findings of my work is that the GP maps of the ancestral steroid receptor-RE binding interfaces are anisotropic, producing specificity phenotypes with unequal probability. I attempted to elucidate the genetic mechanism by which this anisotropy arises in the AncSR2-RE map by decomposing its genetic architecture, finding that it may have been caused by a systematic correlation between the effects of sequence states on binding and on adenine specificity. However, this finding may have been a statistical artifact, and the cause of the adenine specificity bias may instead have been that adenine tends to be easier to bind in general than the other three possible nucleotide states at the variable RE positions. The latter scenario would be a simpler genetic mechanism and would likely to be more generalizable across biological systems, since it seems likely that some sequence states may be better for intermolecular binding in general. For example, variation in protein-protein binding between leucine zipper transcription factors was found in a recent study to be mediated primarily by nonspecific variation in binding affinity between different substrate proteins, not by specific

interactions between variable residues across the interaction interface[51]. If this mechanism can be demonstrated to be true, it would provide support and a mechanistic basis for the theory that anisotropy is a universal feature of biological systems[167].

Anisotropy in systems where the phenotype is not defined by binding to a particular substrate is likely to result from different genetic mechanisms. In the case of RNA secondary structure, for instance, the phenotype is defined by the sites that form base pairs; anisotropic phenotype production in this case probably involves extensive intramolecular pairwise epistasis. For phenotypes that are formed by the interaction of proteins and/or their enzymatic products in regulatory circuits, differences in phenotype often arise from quantitative variation in the expression or function of each protein. Intergenic epistasis is thus likely to be important in these GP maps, although it would not necessarily arise from structural contacts. In the future, it would be interesting to extend the formalism of RFA or similar linear decomposition methods to these other types of systems to understand the genetic basis of anisotropy.

Another key property of the ancestral GP maps was that they were heterogeneous—the specificity phenotypes that were encoded were not uniformly distributed throughout sequence space. Heterogeneity can also be thought of as anisotropy on a more local level of sequence space. The observed heterogeneity had two properties: 1) genotypes close together in the map tended to encode the same phenotype, and 2) access to *new* phenotypes differed across the map. The first property may have its basis in the relative simplicity of the map's genetic architecture—changing a single amino acid residue in a map where specificity is dominated by AA:NT terms is less likely to change the specificity phenotype than in a map where amino acid residues are highly epistatic for specificity (AA:AA:NT terms), because in the former case positive AA:NT interactions with other amino acid sites are less likely to be impacted. The reason for the second

property might partially be found in the fact that most single amino acid mutations change specificity between REs that differ by only a single DNA base. This is probably easier than changing specificity for an RE that differs by two bases, since specificity for one of the bases can be maintained by amino acid residues other than the one that mutated. Other structural aspects of the GP map—such as the fact that the genetic code is more likely to produce mutations between amino acids with similar biochemical properties—likely also contribute to the heterogeneity of the map. The fact that similar properties have been observed in other GP maps suggests that these too are likely universal GP map features.

This study was limited in that it only considered variation in a few sites in both the protein and DNA. If the scope of genetic variation had been expanded to include more sites in both molecules, it is possible that additional specificity phenotypes could have been found, and that amino acid effects inferred to be nonspecific could turn out to be specific with respect to nucleotide variation at other sites (for instance, interactions between AA28 and the NT2 base). In the future, combining random experimental sampling of combinatorial GP maps with RFA to infer the effects of missing variants is likely to enable the characterization of much larger GP maps.

Overall, this study provides a comprehensive view of how the genetic structure of a protein-DNA binding system shaped its potential and actual evolution.

Appendix A: Supplementary Information for Chapter 2

A.1 Supplementary Methods

A.1.1 Dynamic range correction for RE reporter strains

14 of 16 strains showed DBD-dependent fluorescence across a similar dynamic range, with two fluorescence peaks in the presence of DBD—a GFP-negative peak (GFP–) corresponding to autofluorescence from cells that had spontaneously lost the DBD plasmid, and a GFP-positive peak (GFP+) from cells that retained the plasmid and expressed GFP—and a single GFP– peak in the absence of DBD (Fig. A.1A, B). With the CC RE strain, however, the GFP+ peak was absent (Fig. A.1A); with the GA RE, the GFP– peak was right-shifted, indicating high basal fluorescence even in the absence of DBD (Fig. A.1A, B). We repeated the transformation and obtained the same results. We hypothesized that inserting the CC and GA RE sites may have introduced cryptic yeast TF binding sites or caused chromatin remodeling that resulted in constitutive GFP repression and activation, respectively. The flank and spacer (FS) sequences surrounding the RE half sites do not affect SR binding specificity [97,168,169], so we reasoned that introducing mutations into these regions might preserve the ability of the strains to act as reporters for RE binding while disrupting any recognition sites for endogenous yeast proteins. We experimentally identified a set of FS mutations for the CC strain (aacAGCCCAaaaTGGGCTgtt) and another for the GA strain (ccaAGGACAatcTGTCCTtgg) that result in DBD-dependent GFP expression similar to the other RE strains (Fig. A.1C). We therefore used these two FS-modified strains along with the original strains for the remaining 14 RE variants for DMS experiments.

A.1.2 Details on replicate sorting, sequencing and processing

The binned sort was performed to yield 3 replicates per library. We sequenced and processed replicate 1 of the binned sort libraries before the other two replicates to assess data quality (see below). Analysis of this replicate revealed that the batch containing libraries with REBC 9-16 in the AncSR1 background contained inconsistent read counts—we expected about 1 read/cell, but these libraries contained very low or very high estimates of reads per cell (Table A.4). To fix this, we repeated the sorting procedure (enrichment and binned sorting) and used these data (replicate 4) to replace replicate 1 for these eight libraries (Table A.4).

Replicate 1 of the binned sort libraries was sequenced on a NextSeq High Output run. The remaining replicates (2, 3 and 4) were sequenced on a NovaSeq S1 run at the University of Chicago Genomics Facility. We obtained 1.27×10^8 and 2.1×10^9 read pairs for the NextSeq and NovaSeq runs, respectively. These encompassed the 32 experimental libraries, the isogenic controls, and the mini-libraries. We used *sickle v1.33*[128] to filter sequencing reads based on their quality: we kept reads with a Phred score ≥ 30 and a minimum length of 79 nucleotides. This procedure resulted in a cleaned dataset of 1.2×10^8 and 1.85×10^9 read pairs for the NextSeq and NovaSeq runs, respectively. We used *PEAR v0.9.6*[129] to merge the trimmed paired-end reads (minimum assembly length 100 nucleotides), resulting in 1.18×10^8 (>98%) and 1.75×10^9 (>94%) assembled reads, respectively.

We used Biopython toolkit *v1.79*[130] to demultiplex the assembled reads by DBD background, REBC, and BRBC. We only considered reads that mapped exactly to the DBD background and REBC, and allowed reads with at most one mismatch in the BRBC. We successfully mapped 1.01×10^8 (>85%) of the assembled NextSeq reads, and 1.65×10^9 (>94%) of the assembled NovaSeq reads. For downstream analyses, we combined the mapped reads from

NextSeq and NovaSeq.

We corrected the fluorescence estimates for batch effects between replicates 1–3. However, fluorescence estimates from replicate 4 were not corrected for batch effects due to the low number of active variants in the libraries.

A.1.3 Statistical classification of null variants from enrichment sort GFP– bin

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

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

$$p_m = Pr\left(c_m^{d+} < k \right) = F_{Binom}\left(k; c_m^d, f \right) \quad (3)$$

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

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

$$c_m^{d-} = \frac{r_m^{d-}}{r_l^{d-}} \times c_l^{d-} \quad (4)$$

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

$$c_m^{d+} = \frac{c_m^b}{c_l^b} \times c_l^{d+} \quad (5)$$

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

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

Using Equation 1 from the section 2.5 (Chapter 2 Methods) with our estimated parameter

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

A.1.4 Generalized linear model to predict fluorescence of missing variants

A remaining 42% of AncSR1 and 49% of AncSR2 variants were either unobserved or had insufficient read depth to be confidently assigned a phenotype. To predict the fluorescence of these variants, we used a type of generalized linear modeling approach called reference-free analysis (RFA)[47,52] to predict fluorescence based on the effects of sequence states inferred from empirically phenotyped variants. RFA is an unbiased method for inferring the phenotypic effects of genetic states and their interactions (*i.e.,* epistasis) in a combinatorial genotype space. Each genotype g of length n is represented as a vector of genetic states at each site $\left(g_1, g_2, \ldots, g_n \right)^T$. RFA relates the genetic states in g to a latent phenotype s, also referred to as the genetic score, through a linear combination of main and epistatic effects:

$$s(g) = e_0 + \sum_{i}^{n} e_i(g_i) + \sum_{i<j}^{n} e_{i,j}(g_i, g_j) + \ldots + \varepsilon \tag{6}$$

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

$$F(g) = L + \frac{U - L}{1 + e^{-s(g)}} + \varepsilon \quad (7)$$

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

We estimated separate RFA models for each ancestral DBD background. The model estimates the effects of all amino acid states at the 4 variable sites in the DBD and all nucleotide states at the 2 variable sites in the RE; it contains intramolecular interactions up to 3rd order amino acid interactions in the DBD and 2nd order nucleotide interactions in the RE, and intermolecular interactions up to 3rd order DBD-by-2nd order RE interactions. All variants with empirical fluorescence estimates from either the binned or enrichment sorts were used in training; for variants classified as null from the enrichment sorts, we used the mean fluorescence of nonsense variants from the same background. Models were fit in R using *glmnet* v4.1-6[131]. Because *glmnet* does not allow for estimation of parameters in the link function, we first fit an unregularized RFA model with up to 2nd order DBD-by-2nd order RE effects using nonlinear least squares regression to estimate the global *U* and *L* parameters. We then used these estimates to specify a link function for *glmnet*, which we used to fit the full model using L2-regularized regression. 10-fold cross validation (CV) was used to select the L2 penalty that minimized prediction error (Fig. A.3A). Our final models fit the data for active variants with $R^2 = 0.96$ (AncSR1) for and $R^2 = 0.99$ (AncSR2); for all variants, $R^2 = 0.31$ (AncSR1) and $R^2 = 0.88$ (AncSR2), because most variation in fluorescence for null variants is caused by measurement error (Fig. A.3B, C).

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

We used the model to correct for the observation that the GA strain has systematically lower fluorescence than expected given previous measurements of affinity for this RE and a panel of protein variants (Fig. A.3C)[67], presumably because the FS mutations we introduced reduce affinity (see section A.1.1). We estimated the magnitude of this effect by fitting the log-affinity-fluorescence relationship for these variants, including an effect of the RE, using the equation

$$F = L + \frac{U - L}{1 + e^{-\frac{\log(K_A) + d}{a}}} \tag{8}$$

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

A.1.5 Accuracy of predicted functional classification

To classify protein-RE variants with predicted fluorescence as functional or non-functional, we performed a nonparametric bootstrap test to address model prediction error: we concatenated residuals across the 10 RFA cross-validation models, then sampled residuals from within an interval of ±0.1 fluorescence units from the inferred fluorescence of each complex (Fig. A.3E). *p*-values were calculated as the proportion of bootstrap samples ($n = 250$) with fluorescence greater than or equal to that of the wild type complex.

To assess the accuracy of this functional classification, we analyzed the prediction error from the CV fits with the same regularization strength as the final models, which indicate how well the full model predicts unseen data. The AncSR2 CV models predicted the held-out data reasonably well, with mean $R^2 = 0.79$ for all variants and $R^2 = 0.75$ for active variants across the 10 CV fits (Fig. A.3E, right). However, prediction was considerably worse for the AncSR1 CV models, with mean $R^2 = 0.20$ for all variants and $R^2 = 0.07$ for active variants (Fig. A.3E, left). This can in large part be attributed to a bias towards underprediction observed in the AncSR1 CV models, which likely results from the fact that the vast majority of variants in this protein background are at the lower bound of fluorescence; genetic states that may be beneficial for binding in some genetic contexts are thus only seen at the lower bound, resulting in a negative bias in their inferred effects (Fig. A.3E). The AncSR2 CV models also exhibit a slight underprediction bias, but the effect is much less severe because there are many more active variants in this protein background. The prediction error and bias from the models are taken into account in later analyses.

A.1.6 Protein-RE genotype networks

To describe the effects of coevolution, we built joint protein-DNA genotype networks. We considered two functional protein-RE complexes as mutational neighbors if they differed by a single amino acid in the protein *or* a single nucleotide in the RE. For example, the genotypes EGKA:GT and EGKV:GT are neighbors by mutations in the protein, and the genotypes AAAI:AA and AAAI:TA are neighbors due to mutations in the RE. In this network, promiscuous genotypes (such as AAAI) are represented as two separate nodes, one for each complex. Network analyses and visualization were done as in the protein genotype networks. Although both the protein and the RE can substitute in this model, we calculated the length of evolutionary trajectories by including only changes in the protein's amino acid sequence, in order to fairly compare the length of these trajectories to those in the protein-only network. We therefore adjusted the total number of protein+DNA substitutions in any trajectory using the average fraction of all steps on the network that involve changes in the protein sequence. To estimate this fraction, we first computed for every protein-RE complex g the fraction of one-mutation neighbors that change the protein sequence ($f_{\mathrm{AA,g}}$). We ran a Markov chain starting from the set of all genotypes in the network, and at each step (k) we computed the fraction of all possible steps in the network that are amino acid changes ($N_{AA,k}$) as:

$$N_{AA,k} = \sum_{g} f_{AA,g} \times \pi_{(k)\,g} \tag{9}$$

where g iterates over all complexes and $\pi_{(k)}$ is the vector of probabilities of genotypes after k steps of the Markov chain. We took the average fraction across trajectory lengths and used this to scale the length of any evolutionary trajectory in the coevolution network as the expected

number of amino acid steps on that trajectory.

A.2 Supplementary Protocols

A.2.1 Combinatorial library assembly and bacterial transformation

This protocol aims to obtain ~100X coverage per barcoded DBD library (*i.e.* ~2×10^7 cfu). The protocol was repeated for each of the 32 barcoded libraries.

1. Perform second strand synthesis on 5 pmol of library oligo using Q5 DNA polymerase in a 10-µL PCR reaction. Denature at 95°C for 80 s, extend at 50°C for 30 s, anneal at 72°C for 3 min. 30 s.
2. Assemble the library oligo into the plasmid backbone using the BsaI-HFv2 Golden Gate Assembly Kit (NEB). For one 50-µL reaction, use 0.4 µL of double stranded library oligo (0.2 pmol) from the previous step and 0.1 pmol of pDBD2.1 AncSR1 or AncSR2 vector with BsaI cut sites. Perform 2 50-µL reactions per library. Incubate reactions at 37°C for 1 hr, then 50°C for 5 min. to inactivate the enzyme.
3. Pool the 50-µL reactions and purify using the Zymo Clean & Concentrator-5 kit. Elute in 8 µL ddH_2O and chill on ice.
4. Transform 4 µL of assembled pDBD2.1 library into Invitrogen ElectroMAX DH5α-E competent cells following the manufacturer's protocol. Electroporate at 1.7 kV in 1-mm cuvettes. Immediately add 1 mL of preheated SOC to the cuvette and transfer the mixture to a culture tube, using a P1000 tip inserted into a P20 tip.
5. Recover the cells for 1 hour at 37°C, 225 rpm.
6. Take 10 µL of the recovered culture and dilute with 1 mL LB. Take 10 µL of this culture and dilute again with 1 mL LB. Plate 100 µL of this culture on LB + 100 µg/mL carbenicillin.

This constitutes a 2×10^5-fold dilution.

7. Transfer the rest of the recovered cells to 25 mL LB + 100 μg/mL carbenicillin in a culture flask. Grow overnight at 37°C, 225 rpm.
8. The following morning, take 750 μL of the overnight culture to make 20% glycerol stocks. Maxiprep the remaining culture using the GenElute HP Plasmid DNA Maxiprep Kit. Elute DNA in 3 mL ddH_2O and concentrate DNA to 1.2 mL by ethanol precipitation. Measure concentration with NanDrop (should be ~500 ng/μL).
9. Count colonies on the serial dilution plate to estimate the number of colony forming units (cfu) obtained from transformation. Pick several colonies for Sanger sequencing to verify that library assembly was successful.

A.2.2 Yeast transformation

This protocol is for transforming one barcoded DBD library into a yeast RE reporter strain. We were able to transform three libraries at a time with two people working together; this is reflected in the media and buffer volumes. The protocol aims to obtain ~50X coverage per library (*i.e.* ~1×10^7 cfu).

Media and Buffers:

2L 2X YPD

- 20 g Yeast extract
- 40 g Peptone
- Add ddH_2O to 1.9 L in a 2 L flask and autoclave
- Add 100 mL filter-sterilized 40% dextrose

600 mL Buffer E (1M sorbitol, 1mM $CaCl_2$)

- 109.2 g sorbitol
- 0.087 g $CaCl_2$ [87 mg] (or 0.113 g $CaCl_2.2H_2O$ [113 mg])
- Bring to volume with ddH_2O
- Filter sterilize
- Store at 4°C

250 mL Buffer C (10X TE, 1M LiAc, 1M DTT), prepared fresh

- 2.5 mL 1 M Tris-Cl pH 8.0
- 500 uL 0.5 M EDTA pH 8.0
- 25 mL 1 M LiOAc
- 2.5 mL 1 M DTT
- 219.5 mL ddH_2O
- Filter sterilize into 50-mL conical tubes

600 mL recovery media (1M sorbitol in YPD)

- 109.26 g sorbitol
- Bring to 600mL with YPD, filter sterilize.

600 mL sterile water, chilled to 0°C

Day 1

1. Streak yeast glycerol stocks on a YPD plate.

Day 2 (at least 15h before day 3)

2. Inocubate 150 mL YPD with a single colony of each RE yeast strain in 1 L flask; incubate at

30°C, 225 rpm overnight.

Day 3

Preparation

Chill centrifuge to 4°C

3. Read overnight OD_{600}.
4. Aliquot 400 OD_{600}-mL in 50-mL conical tubes; spin 3 kg 2 min; discard supernatant.
5. Resuspend in 800 mL YPD in two 2.5 L flasks.
6. Incubate at 30°C, 225 rpm until $OD_{600} > 4$ (~4h)
7. Chill autoclaved water, Buffer E, 2-mm cuvettes, and prepare 250 mL fresh Buffer C.
8. Spin down cells in 8 50-mL conical tubes (two spins per tube); spin 3 kg 2 min; discard supernatant.
9. Resuspend and pool cells in 90 mL ice-cold ddH_2O in two 50-mL conicals. Spin 3 kg 2 min. Discard supernatant.
10. Wash once more with 45 mL ice-cold ddH_2O each tube.
11. Resuspend each in 45 mL ice-cold Buffer E; spin 3 kg 2 min; discard supernatant.
12. Resuspend both tubes in 160 mL Buffer C and add them to a 1 L culture flask. Incubate for 30 min at 30°C, 225 rpm.
13. Pellet the cells in two 50-mL conicals; decant supernatant. Wash once with ice-cold Buffer E (45 mL per tube); spin 3 kg 2 min. Decant supernatant.
14. Resuspend pellets in remaining volume of Buffer E; pool and add Buffer E for a final volume of 4.8 mL cells + DNA per library (add ~1.5 mL Buffer E per strain to resuspend). Resuspend using a 25 mL pipette and minimize cell loss on sides of pipette/tubes.

Electroporation

15. Mix 600 uL of plasmid library to tube with cells (at least 288 µg of DNA).
16. Add 171.2 mL of recovery media in an empty 1-L flask.
17. Aliquot 400 µL of cell/plasmid mixture in each chilled 2-mm electroporation cuvette (~12 cuvettes per library).
18. Electroporate at 2.5kV; immediately add 1 mL of recovery media at RT.
    - Add the recovery media gently and pipette up and down to bring the cells on the bottom of the cuvette to the surface.
19. Pool reactions together in a 1-L flask with an additional 171.2 mL of recovery media. Rinse each cuvette with an additional 1 mL of recovery media (final volume: 200 mL).
20. Incubate the transformed cells at 30°C, 225 rpm for 2 hr.
21. Make 10^{-4} and 10^{-5} serial dilution plates (YPD + 200 µg/mL G418) to estimate transformation yield.

Passage and storage

22. Aliquot recovery into 50-mL conicals.
23. Spin 3kg 2min room temperature; discard supernatant.
24. Resuspend to 400 mL YPD + 200 µg/mL G418 + 25 µg/mL chlor in a 2.5-L flask.
25. Incubate at 30°C 225 rpm until saturation, approximately 24 hr.

<u>Day 4</u>

26. Make 20 1-mL 25% glycerol stocks with 200 OD_{600}-mL of yeast; flash freeze in liquid N_2; store at −80°C.

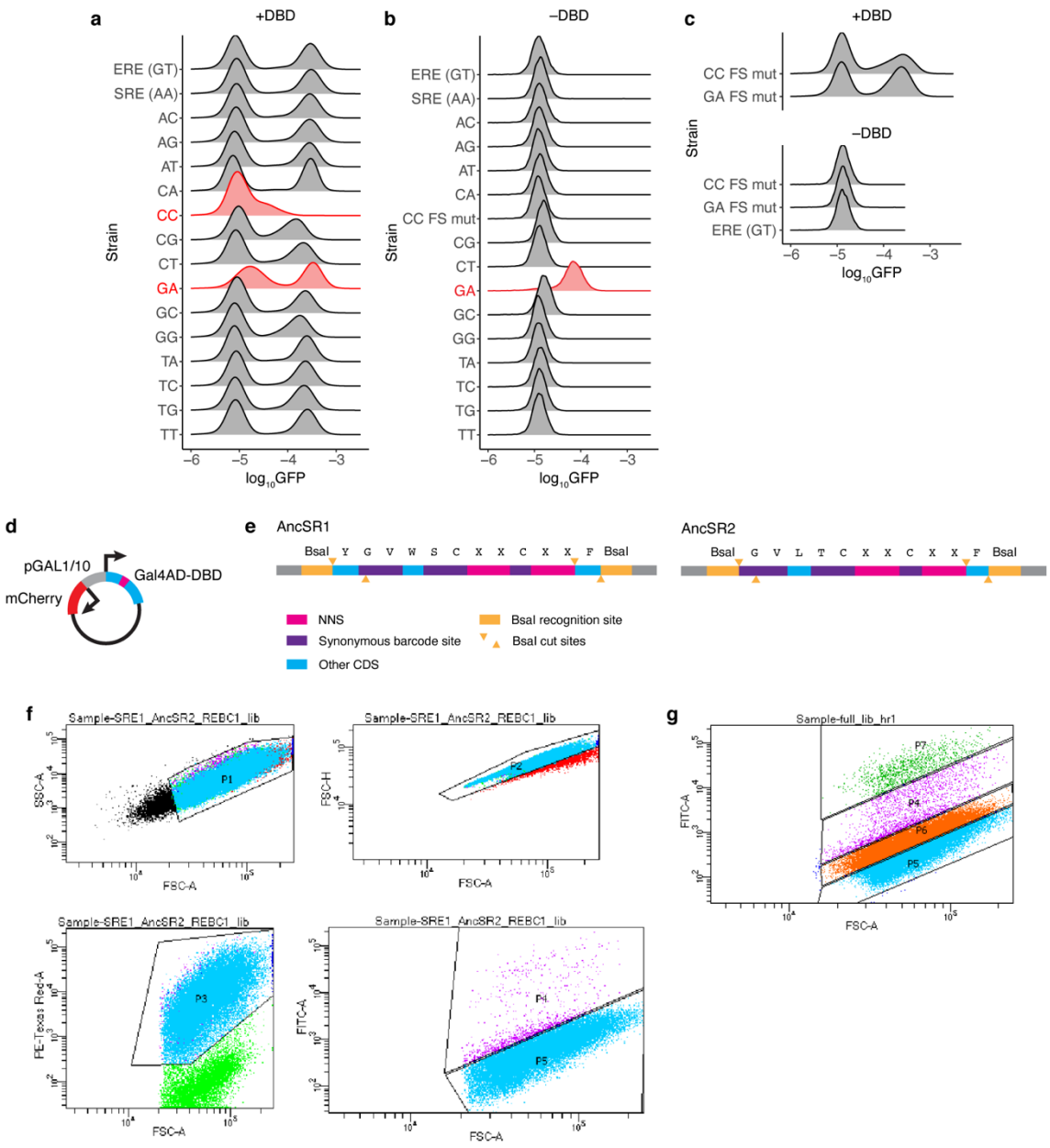

Figure A.1. DBD library construction and sorting. (**A**) Design of the DBD expression vector used for DMS. The SR DBD is fused to an N-terminal *S. cerevisiae* Gal4 Activation Domain. Its expression is under control of a bidirectional pGAL1/10 promoter, which simultaneously drives mCherry expression to select cells that maintain the plasmid during sorting. (**B**) Design of DBD library oligos. NNS codons (pink) were used to generate all possible combinations of amino acid mutations at the four RH scanning sites (marked as X in the amino acid sequence). For each background (AncSR1, left; AncSR2, right), we synthesized 16 libraries, each with a unique set of synonymous barcode mutations at five codons (purple, Table A.1), which allows each to be associated with one RE strain. BsaI sites (orange) were used for Golden Gate assembly into the pDBD2.1 backbone. (**C–E**) Validation of the RE reporter strains. GFP fluorescence was measured by flow cytometry in each strain in the presence (+DBD) or absence (–DBD) of a

universally high-affinity DBD variant (AncSR1+GGKA+11P[17]). In each row, the left peak corresponds to autofluorescence from cells that do not express GFP, either due to lack of DBD binding or loss of the DBD expression plasmid; the right peak corresponds to cells that are expressing GFP in response to DBD-RE binding. "FS mut" denotes strains with mutations in the flank/spacer regions of the RE that correct anomalous expression patterns shown in red (see section A.1). Red strains were not used in the final DMS experiment. Experiments were conducted on the same day within each panel. (**C**) Fluorescence in the presence of high-affinity DBD. (**D**) Fluorescence in the absence of DBD expression plasmid. (**E**) Fluorescence in the CC and GA FS mut strains, with the ERE strain included as a negative control. (**F–G**) Sorting gates used for DMS. (**F**) Enrichment sort gates. Homogeneous single cells were first selected by gating on FSC-A vs. SSC-A and FSC-A vs. FSC-H (top). Plasmid retention was then selected for by gating on mCherry expression (PE-Texas Red-A, bottom left). Finally, cells were sorted into GFP+ (P4) and GFP– (P5) populations (bottom right). The boundary between the GFP+ and GFP– gates was drawn to have a slope of 1.5 on a log-FSC-A vs. log-GFP (FITC) scale so that populations were sorted by GFP expression relative to cell volume. (**G**) Binned sort gates. Gates P1–P3 were drawn as in **C**. Cells were then sorted into four GFP bins, which were drawn to have roughly equal heights (P5–P7). The boundaries between GFP gates were again drawn to have a log-log slope of 1.5.

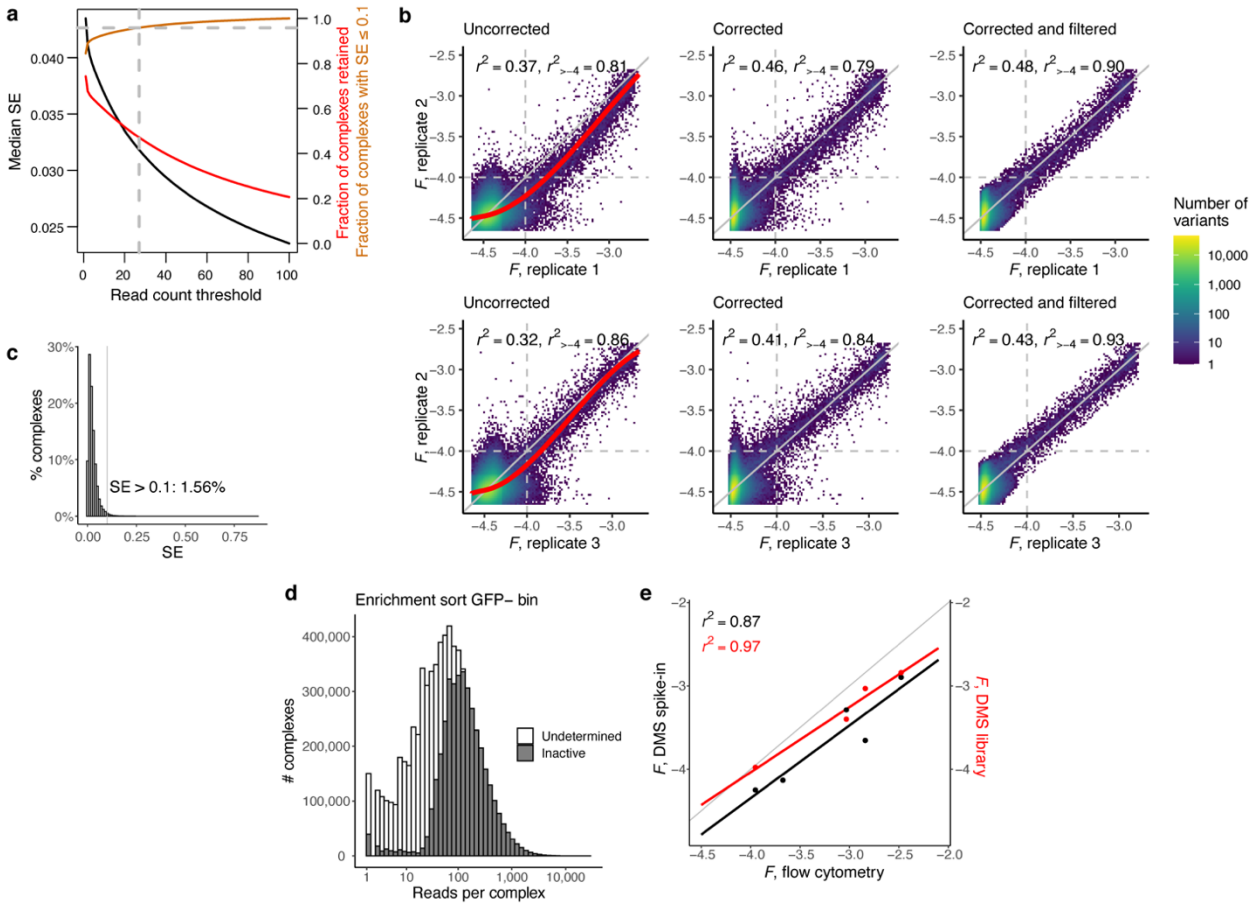

Figure A.2. DMS data cleaning. (**A**) Curves show characteristics of the binned sort dataset as a function of the read count threshold used to retain protein-RE complexes for further analysis (*x*-axis). Black, standard error of *F* (SE, left axis); red, complexes retained, expressed as a fraction of the number of complexes in the binned sort (right axis); gold, fraction of complexes retained that have SE ≤ 0.1 (right axis). We used a read count threshold of 27 (vertical dashed line), at which ≥95% of complexes have SE ≤ 0.1 (horizontal dashed line). (**B**) Correcting and filtering estimates of *F* from the binned sort. Left, correlation in *F* between replicates before correction. Pearson's r^2 is shown for all complexes, and for the subset of complexes with $F > -4$ in both replicates, which roughly corresponds to the boundary between active and inactive complexes (gray dotted lines). Red curves, I-splines fit using complexes with SE of $F < 0.1$. Center, correlation in *F* between replicates after correcting using the I-spline transforms. Right, correlation in *F* between replicates after filtering corrected variants for SE ≤ 0.1. (**C**) Distribution of SE across all complexes in the binned sort after the I-spline correction. Complexes with SE > 0.1 were discarded. (**D**) Read count distribution for complexes sequenced in the enrichment sort GFP– bin. Complexes were inferred to be inactive (gray) if they were not observed in the binned sort, but had high enough inferred cell count in the enrichment sort to have been detectable in the binned sort had they been at least minimally fluorescent (see section A.1). (**E**) Correlations between estimates of *F* from flow cytometry (*x*-axis) and DMS (*y*-axes). Left *y*-axis (black points) shows estimates from isogenic strains that were spiked into the DMS libraries prior to the binned sort. Right *y*-axis (red points) shows estimates from complexes that were encoded in the DMS libraries.

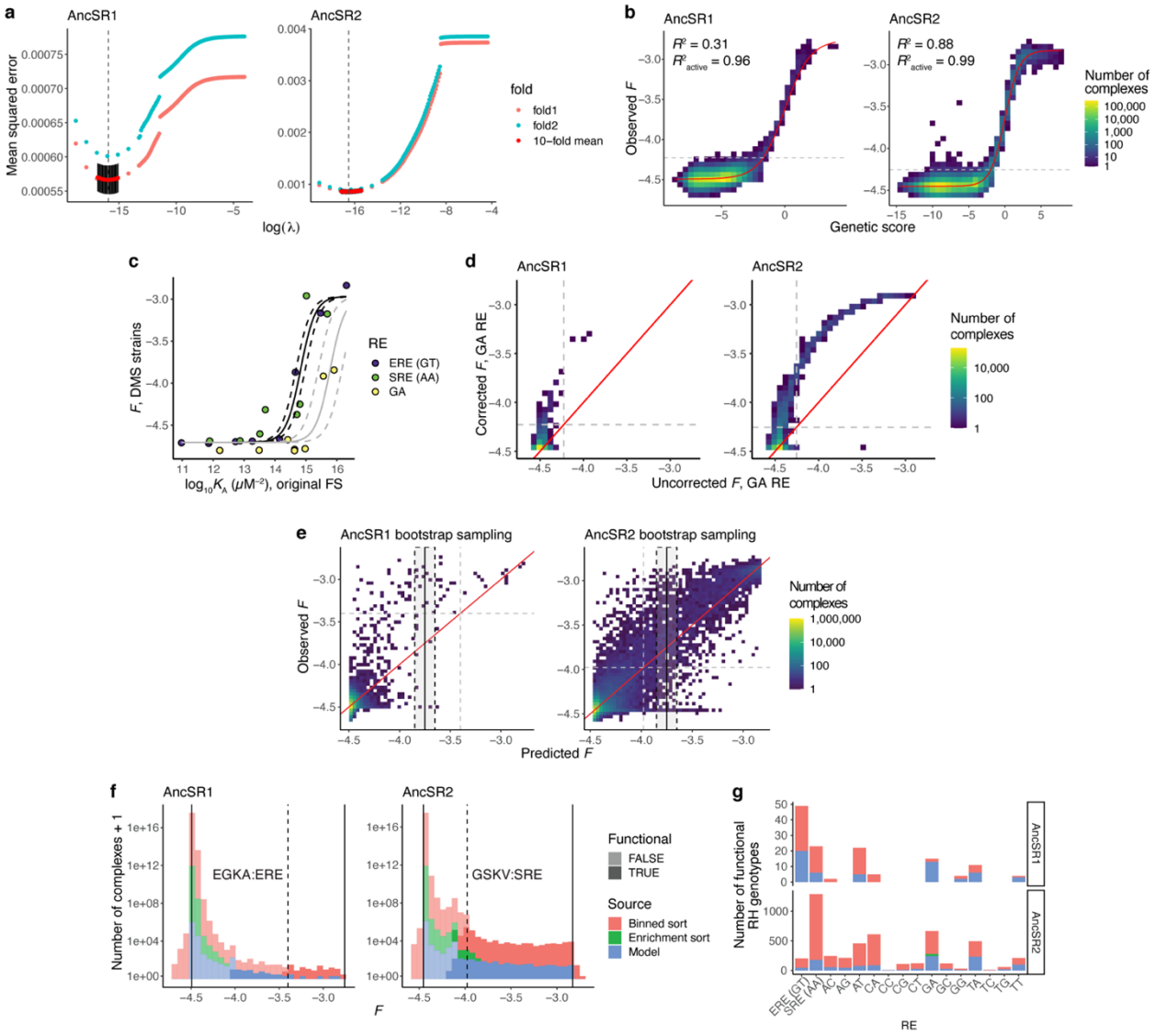

Figure A.3. Fluorescence imputation, GA fluorescence correction, and functional genotype classification. (**A**, **B**) A generalized linear model that predicts the fluorescence of each protein-RE complex from its sequence was fit to the data for each background, using L2 regularization to address overfitting. (**A**) Ten-fold cross-validation (CV) was used to identify the optimal L2 penalty parameter (λ). Red and black, mean and SE of the out-of-sample mean squared error (MSE) across the 10 folds. Initial range finding was performed using two folds (pink and cyan). Vertical line, λ that minimizes mean MSE. (**B**) Genetic score versus observed *F* for the regularized RFA models. Red line, best-fit nonspecific epistasis function. For display, the distribution was discretized; colors show the number of variants in the interval defined by each square. Coefficient of determination (R^2) is reported for all complexes and for the subset of active complexes (above the gray line). (**C**, **D**) Fluorescence correction for the GA strain. **(C)** Affinity (K_A) versus *F* for a panel of DBD variants measured on ERE, SRE, and GA. Affinities, measured by fluorescence anisotropy on the three REs, all with the original flank/spacer sequence, were previously reported[17,67]. *F* was measured by flow cytometry in the RE strains that were used for DMS, of which the ERE and SRE strains had the original flank/spacer sequence,

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

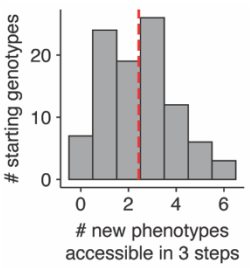

Figure A.4. Accessible new phenotypes after 3 substitution steps in the AncSR1 network. Bars show the distribution over every starting genotype in the AncSR1 main component. Dashed line, mean.

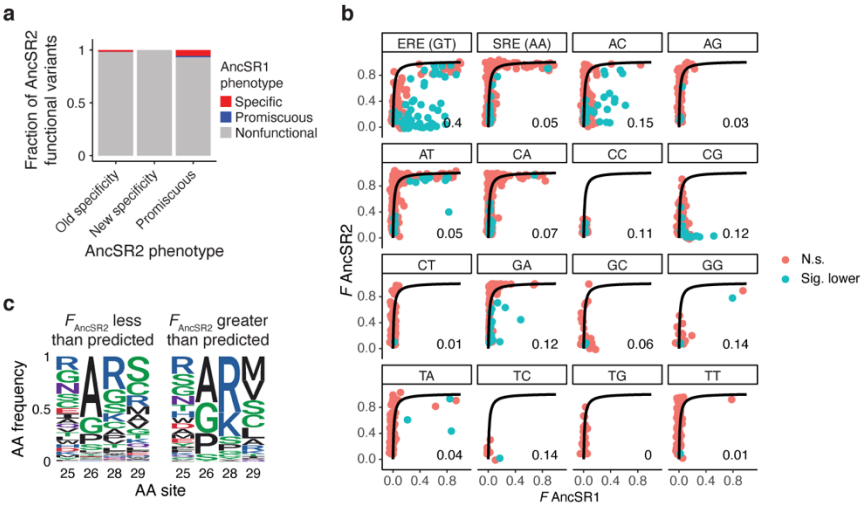

Figure A.5. Additional analyses for effects of background substitutions on DBD-RE affinity. (**A**) Changes in phenotype across the AncSR1-to-AncSR2 transition. Bars represent the set of protein variants in AncSR2 that have different classes of phenotypes: specificity phenotypes that were encoded in the AncSR1 map (old specificity), specificity phenotypes not encoded in the AncSR1 map (new specificity), or promiscuous in AncSR2. Colored sections show the fraction of variants in each class whose functional category in the AncSR1 background was specific, promiscuous, or nonfunctional. (**B**) Plots are the same as in Fig. 2.6A, but split into panels by RE. Blue points, protein-DNA complexes with significantly lower fluorescence in the AncSR2 background than predicted by the model; red, all other variants. Numbers at the bottom-right of each panel show the fraction of plotted variants with significantly lower than expected AncSR2 fluorescence. (**C**) Amino acid frequencies at the RH variable sites among all complexes that are significantly more (left) or less (right) fluorescent in the AncSR2 background than predicted by the ERE-specific model in Fig. 2.6E. To test for significance in **B** and **C**, we tested whether their Bonferroni-corrected 95% CI of fluorescence was outside of the 95% CI of the model in both the AncSR1 and AncSR2 backgrounds.

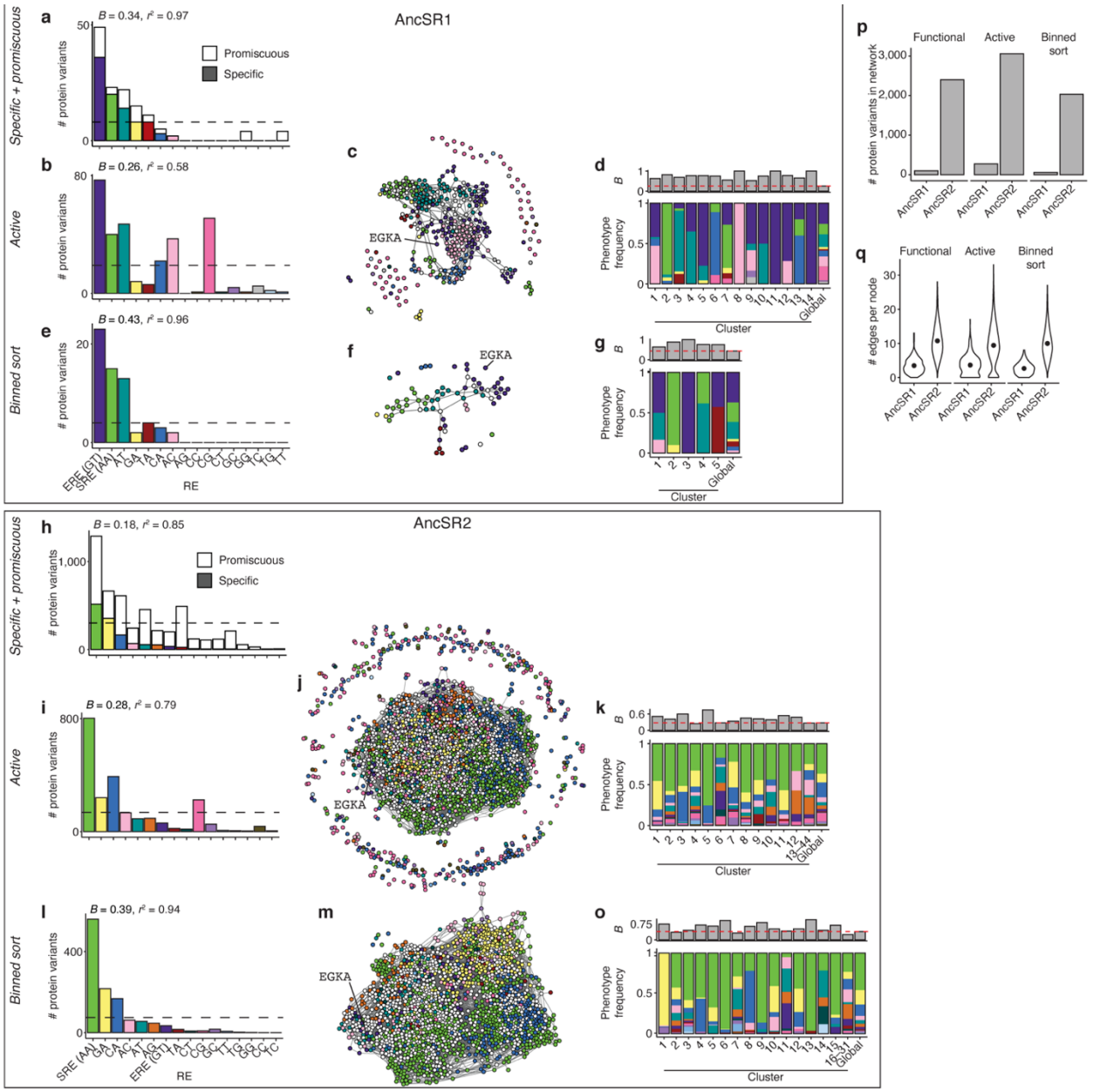

Figure A.6. Robustness to alternative phenotype assignment methods. (**A**) Global production distribution in the AncSR1 background, counting variants that bind specifically (colored bars) and promiscuously (white bars) to each RE. Dashed line shows the expected frequencies if the production distribution were isotropic. The bias, *B*, of the distribution and r^2 to the production distribution for specific variants (Fig. 2.2A) are reported. (**B**) Same as in **A**, with phenotypes calculated using data from variants with fluorescence significantly higher than that of nonsense variants (active variants). (**C**) Sequence space network for AncSR1 active variants. (**D**) Bottom: Frequencies of specificity phenotypes within each genotype cluster in the AncSR1 active variant networks; the global production distribution is shown for comparison. Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. (**E–G**) Same as in **B–D**, but with phenotypes calculated using only data from the binned sort experiment; protein-DNA complexes without experimental fluorescence measurements were assumed to have null fluorescence. **H–O**, Same as in **A–G**, but for the AncSR2 background. Note that the active

variant datasets are likely to be enriched for false positives due to misclassification of variants whose fluorescence is by chance slightly higher than the nonsense variant distribution. This may explain the high frequency of variants that do not share any mutational connections to other active variants. It may also explain the high frequency of CG-specific variants compared to the original classification scheme, since the CG yeast strain has a slightly higher null fluorescence level than most other strains (Fig. A.1C, D) and most CG-specific variants are unconnected in the active variant genotype networks. (**P**) Number of protein variants in each network under different methods of phenotype assignment. "Functional" indicates the original method used in the main text; note that this yields the same number of protein variants as the "specific + promiscuous" method. (**Q**) Number of edges per node in each network, with the original phenotype classification method (functional) shown for comparison.

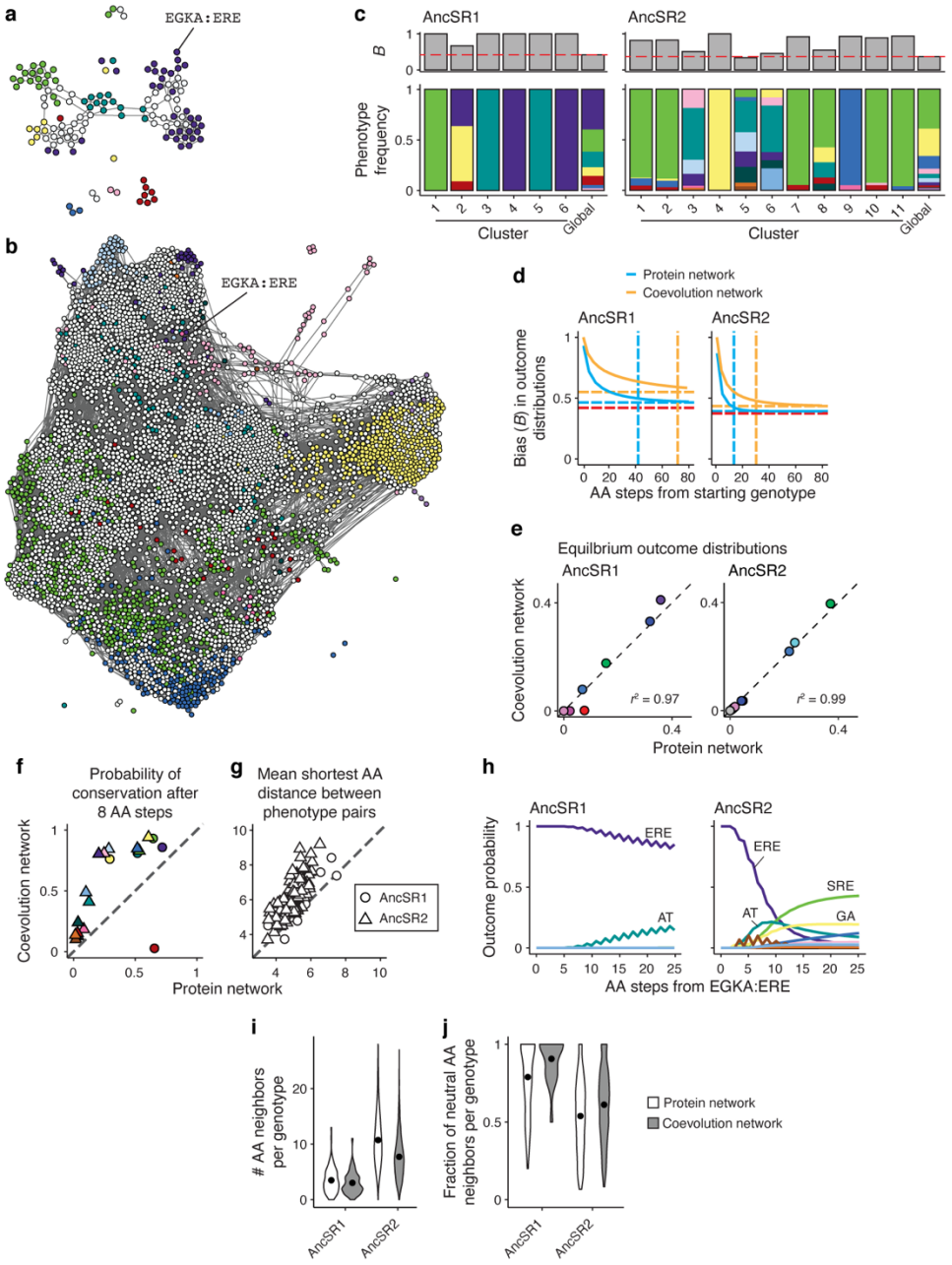

Figure A.7. Robustness to model of evolution using joint protein-DNA networks. (**A**) AncSR1 protein-DNA coevolution network. Nodes represent functional protein-RE complexes, colored by the RE specificity of the protein genotype; colors are as in Fig. 2.2B and 2.4B. Promiscuous protein genotypes are represented by multiple nodes, one for each RE it binds. Edges connect complexes that can be interconverted by a single nucleotide change in the RE or the coding sequence of the protein. (**B**) AncSR2 protein-DNA coevolution network. (**C**) Bottom: Frequencies of specificity phenotypes within each genotype cluster in the AncSR1 (left) and AncSR2 (right) coevolution networks; the global production distribution (right-most column) is shown for comparison. Top: strength of phenotype bias (*B*) in each cluster. Red line, *B* of global production distribution. (**D**) Bias (*B*) in evolutionary outcomes as a function of the length of evolutionary trajectories. Solid curves, mean *B* across starting genotypes in the protein (cyan) or

coevolution (orange) networks. Dashed horizontal lines, *B* of the equilibrium distribution in each network; dashed horizontal red line, global bias. Vertical dashed lines show the number of substitutions required for mean *B* to reach within 0.05 units of the equilibrium value within each type of network. The equilibrium distributions are more biased in the coevolution networks, and require more amino acid substitutions to be reached, because changes in protein genotype must occur between variants that can bind to the same RE sequence. (**E**) Comparison between equilibrium outcome distributions of the protein-only evolution and protein-DNA coevolution networks in each AncSR1 (left) and AncSR2 (right) backgrounds. Pearson's r^2 between the two distributions are shown. Dashed line, $y = x$. (**F**) Probability of conservation of each phenotype after 8 amino acid substitution steps in the protein vs. coevolution networks. (**G**) Mean shortest amino acid distance between all possible pairs of phenotypes in the coevolution vs. protein networks, calculated as in Fig. 2.2G. Circles, AncSR1 networks, triangles, AncSR2 networks. Dashed line, y = x. (**H**) Probability of evolving each specificity phenotype as a function of the number of amino acid substitutions away from EGKA:ERE in the AncSR1 (left) and AncSR2 (right) coevolution networks. In both backgrounds, conservation is more likely at short trajectory lengths than in the corresponding protein networks (Fig. 2.3G, 5F), but the relative likelihood of achieving each phenotypic outcome is similar. (**I**) Distribution of the number of neighbors per genotype with distinct RH sequences in each type of network. Dots, means of distributions. (**J**) Distribution of the fraction of neutral neighbors per node with distinct RH genotypes in each network. Dots, means of distributions.

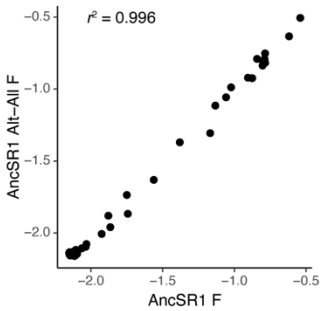

Figure A.8. Robustness of RH mutation effects to uncertainty in ancestral reconstruction. Effects on ERE binding of all possible single amino acid mutations at the four variable RH sites in the background of the maximum *a posteriori* (MAP) wild type AncSR1 protein (*x*-axis), and in the background of the AltAll wild type AncSR1 protein, which has the second-most likely amino acid state at all sites at which the posterior probability of the MAP state is less than 0.8 (*y*-axis)[43]. Pearson's r^2 is shown.

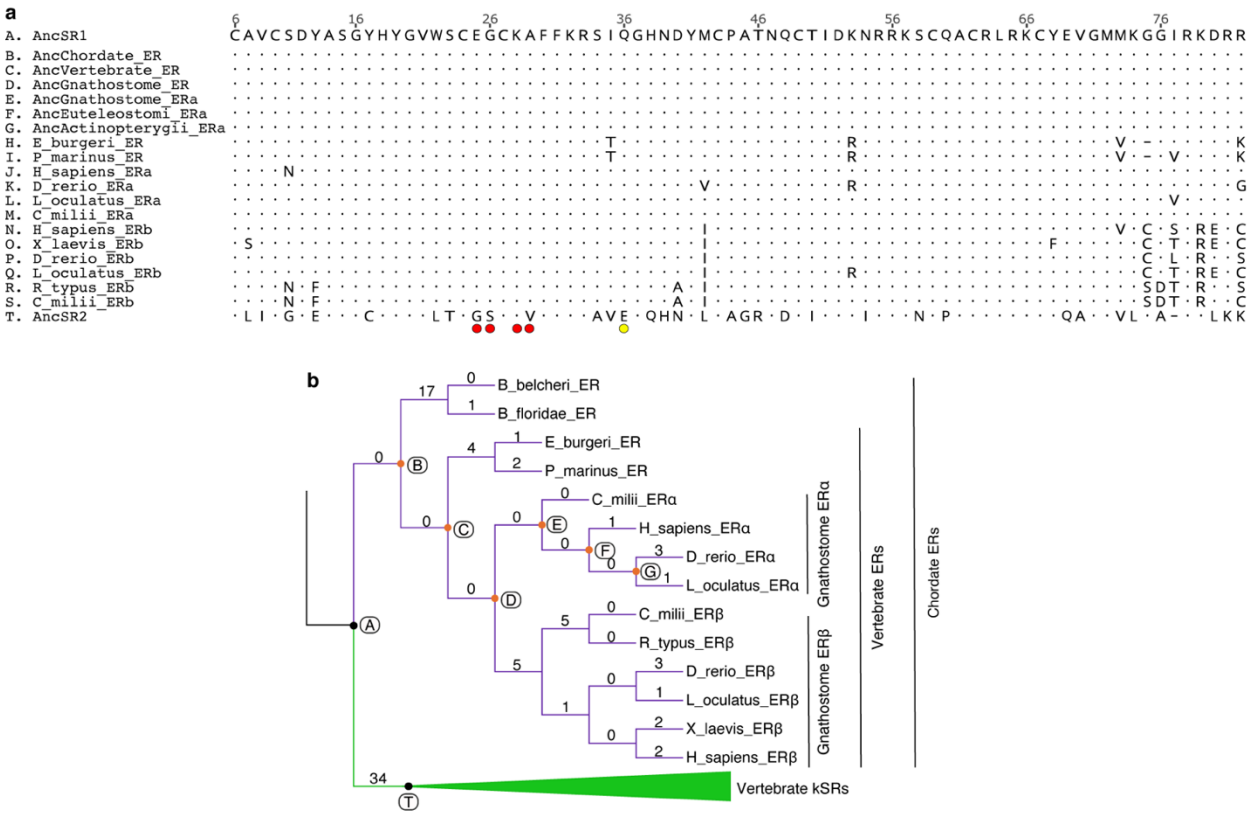

Figure A.9. Amino acid changes along the SR phylogeny. (**A**) Amino acid alignment of extant vertebrate ERs and the MAP protein sequences for key ancestral nodes in the SR phylogeny[43]. The AncSR1 sequence is used as the reference to indicate amino acid changes; dots, same amino acid state as that in AncSR1; dashes, gaps; red circles, variable sites in DMS experiment; yellow circle, historical substitution (q36E) that likely contributed to the shift in the direction of the global bias away from ERE. (**B**) Cladogram of SRs showing the number of substitutions that occurred along each branch. Letters, nodes shown in alignment in **A**; black nodes, AncSR1 and AncSR2; orange nodes, ancestral ER sequences identical to AncSR1. Branches and clades are colored according to their DNA specificity phenotype: purple, ERE-specificity; green, SRE-specificity.

Table A.1. Synonymous RE barcodes (REBCs). Sequences of synonymous mutations used to barcode the DBD libraries to associate them with RE strains. NNS, variable sites in the DMS experiment.

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

Table A.2. Library transformation and enrichment sort statistics. Library transformation yields, glycerol stock recovery rates, and number of cells sorted for enrichment sorts.

DBD back-ground	RE	REBC	Bacterial transfor-mation cfu (x10^7)	Yeast transfor-mation cfu (x10^7)	Top-up yeast transfor-mation cfu (x10^7)	Total yeast transfor-mation cfu (x10^7)	Enrich-ment sort batch	Glycerol stock recovery cfu (x10^7)	Enrich-ment sort GFP– cell count	Enrich-ment sort GFP+ cell count	Enrich-ment sort GFP+ proport ion
AncSR1	SRE	1	5.67	3.09		3.09	1	19.4	24375871	625635	0.025
AncSR1	GA	2	2.01	2.82		2.82	1	29.6	24800681	435760	0.017
AncSR1	ERE	3	2.38	2.58		2.58	1	25.4	24681039	405514	0.016
AncSR1	AC	4	2.64	1.73		1.73	1	10.6	24592905	528709	0.021
AncSR1	AG	5	2.07	3.25		3.25	1	25.8	24662980	408313	0.016
AncSR1	AT	6	3.39	3.8		3.8	1	25.8	24806774	472083	0.019
AncSR1	CA	7	1.56	2.07		2.07	1	14.4	24352862	726379	0.029
AncSR1	CC	8	1.08	2.9		2.9	1	27.8	24658946	448824	0.018
AncSR1	CG	9	1.66	0.85	1.1	1.95	6-May	18.4	24326839	858760	0.034
AncSR1	CT	10	2.37	2.34		2.34	6-May	25.6	24736293	379479	0.015
AncSR1	GC	11	1.97	2.26		2.26	6-May	26.4	24336612	744771	0.03
AncSR1	GG	12	1.38	1.32		1.32	6-May	33.6	24841745	285197	0.011
AncSR1	TA	13	1.86	0.69	1.29	1.98	6-May	10.4	24748789	329854	0.013
AncSR1	TC	14	1.1	2.16		2.16	6-May	20.6	24558731	587774	0.023
AncSR1	TG	15	1.22	0.91	1.01	1.92	6-Feb	15.6	24945619	348916	0.014
AncSR1	TT	16	3.07	2.54		2.54	6-Feb	16.4	25226917	437954	0.017
AncSR2	SRE	1	5.24	1.79		1.79	4	17.8	24220370	847039	0.034
AncSR2	GA	2	3.79	1.37		1.37	4	25	24825719	391276	0.016
AncSR2	ERE	3	5.1	1.5		1.5	4	17.4	24694083	425418	0.017
AncSR2	AC	4	3.54	0.7	0.87	1.57	4	9.8	24602087	533981	0.021

Table A.2 continued.

DBD back-ground	RE	REBC	Bacterial transfor-mation cfu (x10^7)	Yeast transfor-mation cfu (x10^7)	Top-up yeast transfor-mation cfu (x10^7)	Total yeast transfor-mation cfu (x10^7)	Enrich-ment sort batch	Glycerol stock recovery cfu (x10^7)	Enrich-ment sort GFP– cell count	Enrich-ment sort GFP+ cell count	Enrich-ment sort GFP+ proport ion
AncSR2	AG	5	4.6	0.43	0.9	1.33	4	8.8	24690832	434945	0.017
AncSR2	AT	6	4.3	1.6		1.6	4	1.6	24686331	421661	0.017
AncSR2	CA	7	2.2	1.17		1.17	4	1.6	24407887	777213	0.031
AncSR2	CC	8	2.8	1.35		1.35	4	2.8	24637616	409661	0.016
AncSR2	CG	9	6.9	1.98		1.98	3	18.6	24399903	903819	0.036
AncSR2	CT	10	4.5	1.16		1.16	3	4.6	24518676	796209	0.031
AncSR2	GC	11	5.4	0.97		0.97	3	21	24214644	940935	0.037
AncSR2	GG	12	5.1	2.52		2.52	3	18	24511230	679816	0.027
AncSR2	TA	13	1.46	1.08		1.08	3	8.2	24606769	673875	0.027
AncSR2	TC	14	12.9	1.3		1.3	3	18	24040623	1105703	0.044
AncSR2	TG	15	4.6	1.5		1.5	3	10.8	24441448	617178	0.025
AncSR2	TT	16	2.7	2.03		2.03	3	20.8	24619287	854539	0.034

Table A.3. Binned sort statistics. Glycerol stock recovery rates and number of cells sorted for binned sorts.

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

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

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

Appendix B: Supplementary Figures for Chapter 3

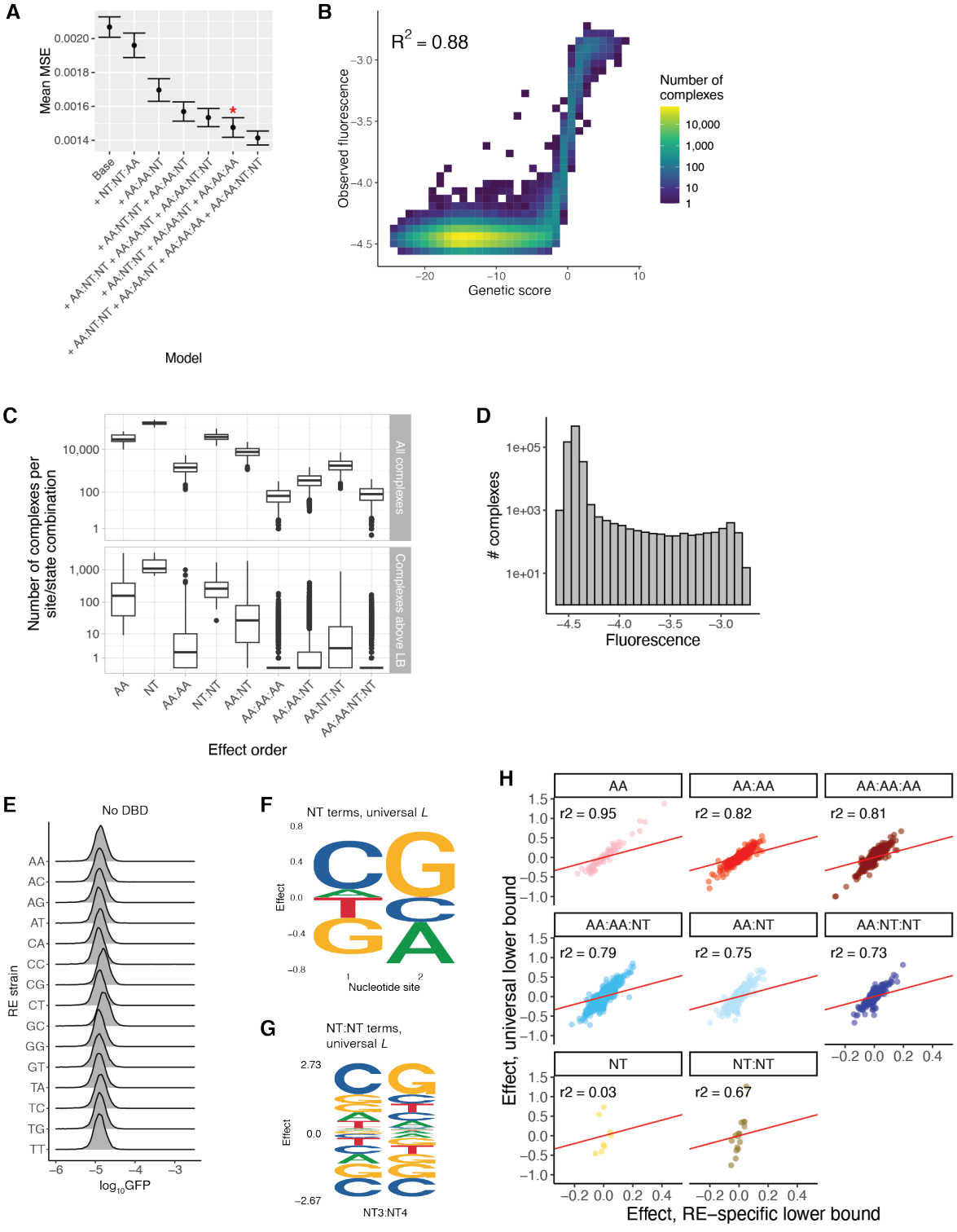

Figure B.1. Model selection and quality control. (**A**) Points show the mean out-of-sample mean squared error (MSE) from 10-fold cross-validation for RFA models truncated to various

orders. Error bars, standard error of the mean (SE). The "base" model includes terms for AA, AA:AA, NT, NT:NT, and AA:NT effects. Terms added to the base model are shown on the *x*-axis. Addition of AA:AA:NT, AA:NT:NT, and AA:AA:AA terms all decreased the mean MSE by greater than 1 SE compared to models that did not include them, while addition of AA:AA:NT:NT terms did not. The best-fit model was thus chosen to be the one without AA:AA:NT:NT terms (red asterisk). (**B**) Comparison of genetic score and observed fluorescence for the best-fit model. Both axes are binned for plotting; color shows the number of complexes in each bin. The R^2 of predicted versus observed fluorescence is shown. (**C**) Distribution of the number of genetic backgrounds in which state combinations of each order are observed in the training data. Top, distribution across all DBD-RE complexes in the training data. Bottom, distribution across complexes whose fluorescence is significantly greater than the lower bound of the assay. Points show outliers, defined as values that are beyond 1.5 times the interquartile range (box height) from the closest quartile. (**D**) Distribution of fluorescence among complexes in the training set. The *y*-axis is shown is log-transformed. (**E**) Distribution of autofluorescence (GFP expression in the absence of DBD) among the yeast reporter strains, measured by flow cytometry. The GA strain is not shown. (**F, G**) NT and NT:NT effects estimated from a model in which a single lower bound parameter was inferred for all 16 REs. Effects are scaled to the genetic score of the wild type complex. (**H**) Comparison of effects inferred using RE-specific lower bounds (*x*-axis) versus a universal lower bound parameter (*y*-axis), split by the order of effect. Pearson's r^2 for each order of effects is shown. Red line, $y = x$.

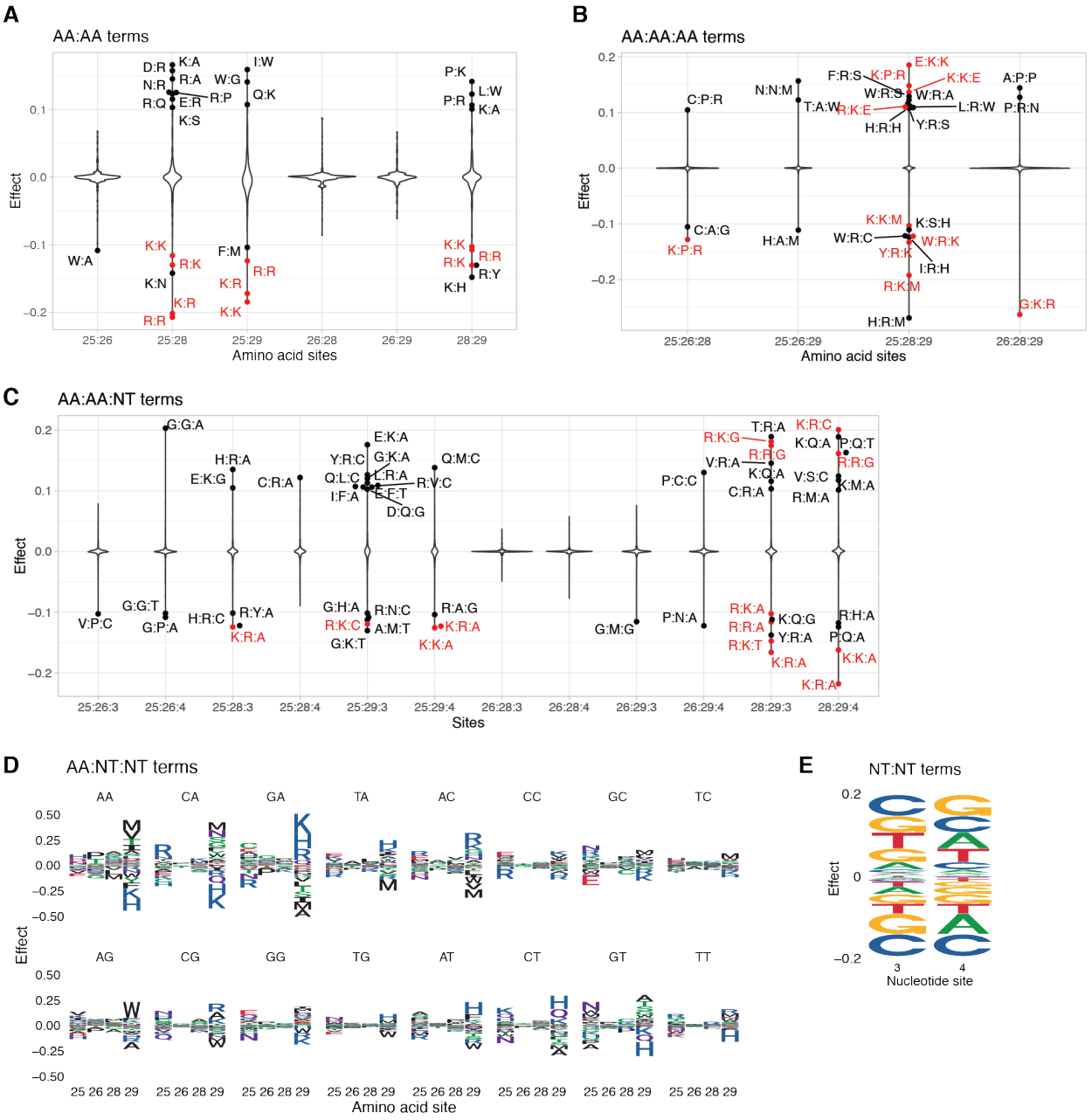

Figure B.2. Epistatic effects on binding and specificity. (**A**) Distribution of effect sizes for amino acid pairwise effects on binding, plotted by site combination. Effects larger than 0.1 in magnitude are shown as points and labeled by their amino acid states. Red, effects involving pairs of positively charged amino acids. Violin plots are scaled to have the same area for each site pair. (**B**) Same as **A** but for third-order amino acid epistatic effects on binding. Red, effects involving at least two positively charged amino acids. (**C**) Same as **A** but for amino acid pairwise epistatic effects on specific binding to single nucleotide states. Red, effects involving at least two positively charged amino acids. (**D**) Epistatic effects between single amino acid residues and pairs of nucleotide states. Each logo plot shows epistatic terms for each amino acid state to a pair of nucleotide states, with colors as in Fig. 3.2B and **C**. (**D**) Epistatic effects between pairs of nucleotide states on binding. Colors are as in Fig. 3.2E.

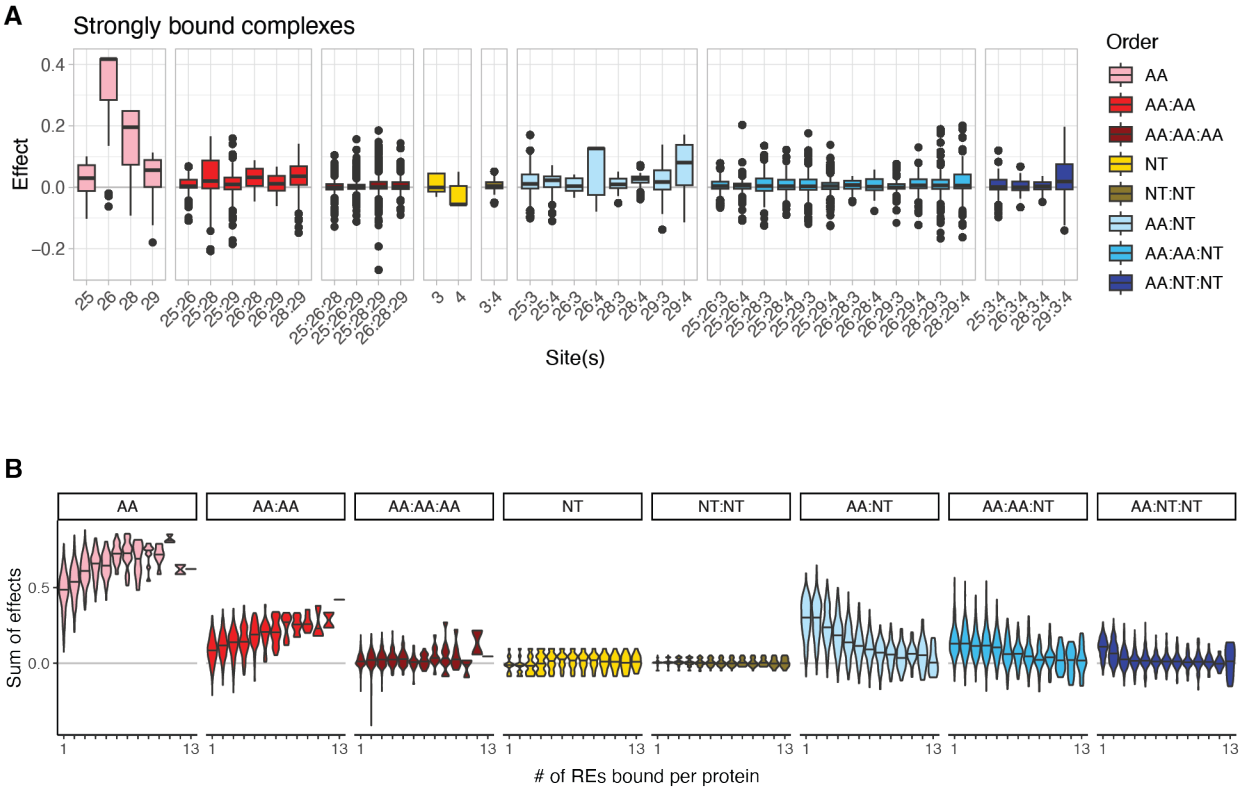

Figure B.3. Fine-grained genetic architecture of binding and specificity. (**A**) Distribution of effects at each order (colors) and site or site combination for strongly bound complexes. Points show outliers, defined as values that are beyond 1.5 times the interquartile range (box height) from the closest quartile. (**B**) Same as in Fig. 3.3D, but showing genetic score contributions split by effect order.

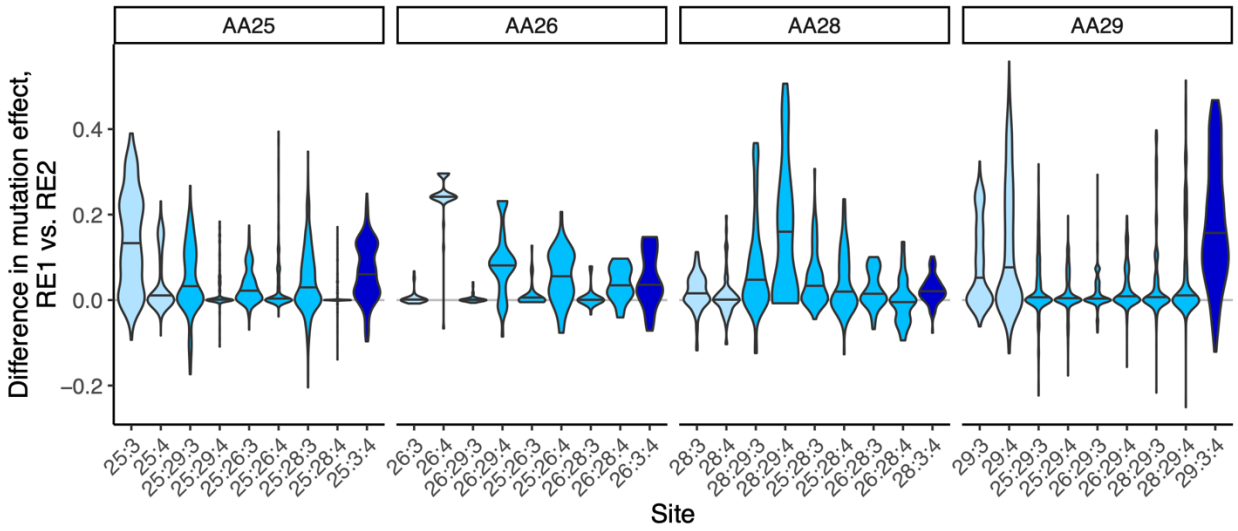

Figure B.4. Contributions of sites and site combinations to specificity-switching mutations. Difference in mutation effect on the starting RE (RE1) and the ending RE (RE2) for mutations that switch specificity, split by the site of the mutation (panels) and the sites that contribute to the effect (x-axis). Colors correspond to order of effect as in the main figures.

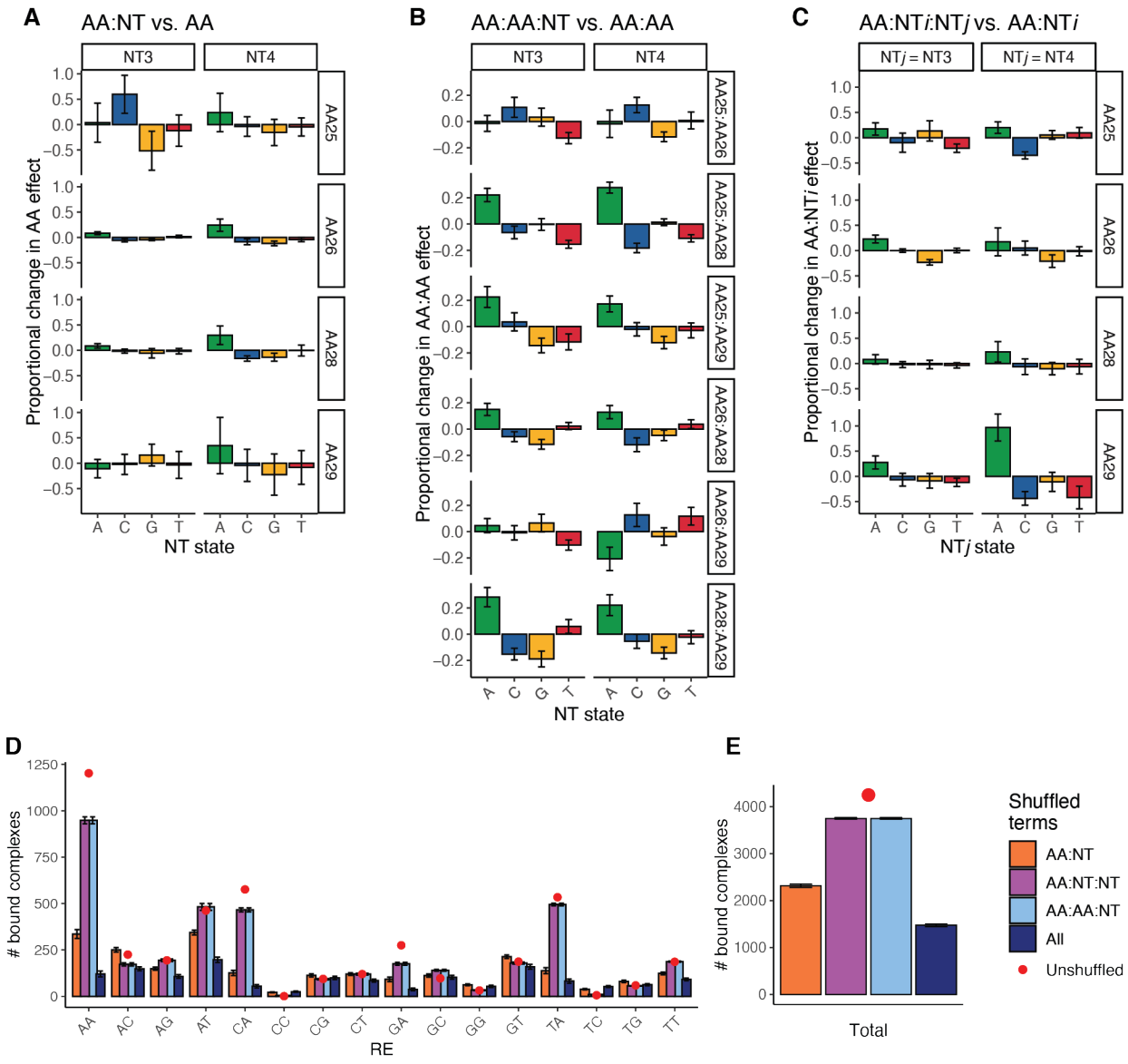

Figure B.5. Amino acid-specific adenine amplification effects and consequences for binding. (**A–C**) Same as Fig. 3.5A–C but split by amino acid site or site combination (panel rows). (**D, E**) Same as Fig. 3.5D, E but showing the number of bound complexes per RE (**D**) or in total across all REs (**E**).

References

1. Wright, S. The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution. in *Proceedings of the Sixth National Congress of Genetics* 356–366 (1932).
2. Rice, S. H. Phenotype Landscapes, Adaptive Landscapes, and the Evolution of Development. in *The Adaptive Landscape in Evolutionary Biology* 283–295 (Oxford University Press, Oxford, 2012).
3. Amundson, R. *The Changing Role of the Embryo in Evolutionary Thought: Roots of Evo-Devo*. (Cambridge University Press, Cambridge, 2005). doi:10.1017/CBO9781139164856.
4. Novick, R. *Structure and Function*. (Cambridge University Press, Cambridge, 2023). doi:10.1017/9781009028745.
5. Weinreich, D. M., Watson, R. A. & Chao, L. Perspective: Sign Epistasis and Genetic Constraint on Evolutionary Trajectories. *Evolution* **59**, 1165–1165 (2005).
6. Kingman, J. F. C. A Simple Model for the Balance between Selection and Mutation. *J. Appl. Probab.* **15**, 1–12 (1978).
7. Kauffman, S. A. & Weinberger, E. D. The *NK* model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.* **141**, 211–245 (1989).
8. Kauffman, S. & Levin, S. Towards a general theory of adaptive walks on rugged landscapes. *J. Theor. Biol.* **128**, 11–45 (1987).
9. Franke, J., Klözer, A., Visser, J. A. G. M. de & Krug, J. Evolutionary Accessibility of Mutational Pathways. *PLOS Comput. Biol.* **7**, e1002134 (2011).
10. Kondrashov, D. A. & Kondrashov, F. A. Topological features of rugged fitness landscapes in sequence space. *Trends Genet.* **31**, 24–33 (2015).

11. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23**, 700–707 (2013).
12. Maynard-Smith, J. Natural Selection and the Concept of a Protein Space. *Nature* **225**, 726–734 (1970).
13. Wagner, A. Neutralism and selectionism: A network-based reconciliation. *Nat. Rev. Genet.* **9**, 965–974 (2008).
14. Weinreich, D. M., Delaney, N. F., DePristo, M. A. & Hartl, D. L. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science* **312**, 111–114 (2006).
15. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative Epistasis Between Beneficial Mutations in an Evolving Bacterial Population. *Science* **332**, 1193–1196 (2011).
16. de Visser, J. A. G. M., Park, S. & Krug, J. Exploring the Effect of Sex on Empirical Fitness Landscapes. *Am. Nat.* **174**, S15–S30 (2009).
17. Anderson, D. W., McKeown, A. N. & Thornton, J. W. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its DNA binding sites. *eLife* **4**, e07864–e07864 (2015).
18. Hall, D. W., Agan, M. & Pope, S. C. Fitness Epistasis among 6 Biosynthetic Loci in the Budding Yeast Saccharomyces cerevisiae. *J. Hered.* **101**, S75–S84 (2010).
19. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protoc.* **9**, 2267–2284 (2014).
20. Phillips, A. M. *et al.* Binding affinity landscapes constrain the evolution of broadly neutralizing anti-influenza antibodies. *eLife* **10**, 1–40 (2021).

21. Phillips, A. M. *et al.* Hierarchical sequence-affinity landscapes shape the evolution of breadth in an anti-influenza receptor binding site antibody. *eLife* **12**, e83628 (2023).
22. Lite, T.-L. V. *et al.* Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *eLife* **9**, e60924 (2020).
23. Poelwijk, F. J., Socolich, M. & Ranganathan, R. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nat. Commun.* **10**, 4213–4213 (2019).
24. Moulana, A. *et al.* Compensatory epistasis maintains ACE2 affinity in SARS-CoV-2 Omicron BA.1. *Nat. Commun.* **13**, 7011 (2022).
25. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. *eLife* **5**, 1–21 (2016).
26. Starr, T. N., Picton, L. K. & Thornton, J. W. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature* **549**, 409–413 (2017).
27. Podgornaia, A. I. & Laub, M. T. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347**, 673–677 (2015).
28. Aguilar-Rodríguez, J., Payne, J. L. & Wagner, A. A thousand empirical adaptive landscapes and their navigability. *Nat. Ecol. Evol.* **1**, 0045–0045 (2017).
29. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *eLife* **7**, e32472–e32472 (2018).
30. Fisher, R. A. *The Genetical Theory Of Natural Selection*. (Oxford University Press, London, 1930).
31. Haldane, J. B. S. *The Causes of Evolution*. (Longmans, Green & Co., London, 1932).
32. Yampolsky, L. Y. & Stoltzfus, A. Bias in the introduction of variation as an orienting factor in evolution. *Evol. Dev.* **3**, 73–83 (2001).

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

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

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

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

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

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

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

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

41. Fontana, W. & Schuster, P. Continuity in Evolution: On the Nature of Transitions. *Science* **280**, 1451–1455 (1998).

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

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

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

45. Sailer, Z. R. & Harms, M. J. Detecting high-order epistasis in nonlinear genotype-phenotype maps. *Genetics* **205**, 1079–1088 (2017).

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

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

48. Buda, K., Miton, C. M. & Tokuriki, N. Pervasive epistasis exposes intramolecular networks in adaptive enzyme evolution. *Nat. Commun.* **14**, 8508 (2023).

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

50. Sailer, Z. R. & Harms, M. J. High-order epistasis shapes evolutionary trajectories. *PLOS Comput. Biol.* **13**, e1005541–e1005541 (2017).

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

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

53. Marks, D. S. *et al.* Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE* **6**, e28766 (2011).

54. Rollins, N. J. *et al.* Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).

55. Russ, W. P. *et al.* An evolution-based model for designing chorismate mutase enzymes. *Science* **369**, 440–445 (2020).

56. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 5869–5874 (2006).

57. Bloom, J. D. *et al.* Thermodynamic prediction of protein neutrality. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 606–611 (2005).

58. Dasmeh, P. & Serohijos, A. W. R. Estimating the contribution of folding stability to nonspecific epistasis in protein evolution. *Proteins Struct. Funct. Bioinforma.* **86**, 1242–1250 (2018).

59. Faure, A. J. *et al.* Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175–183 (2022).

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

61. Olson, C. A., Wu, N. C. & Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr. Biol.* **24**, 2643–2651 (2014).

62. Otwinowski, J. Biophysical Inference of Epistasis and the Effects of Mutations on Protein Stability and Function. *Mol. Biol. Evol.* **35**, 2345–2354 (2018).

63. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19**, 596–604 (2009).

64. Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci.* **108**, 9916–9921 (2011).

65. Zheng, J., Guo, N., Huang, Y., Guo, X. & Wagner, A. High temperature delays and low temperature accelerates evolution of a new protein phenotype. *Nat. Commun.* **15**, 2495 (2024).

66. Lunzer, M., Miller, S. P., Felsheim, R. & Dean, A. M. Evolution: The biochemical architecture of an ancient adaptive landscape. *Science* **310**, 499–501 (2005).

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

68. Rauscher, R. *et al.* Positive epistasis between disease-causing missense mutations and silent polymorphism with effect on mRNA translation velocity. *Proc. Natl. Acad. Sci.* **118**, e2010612118 (2021).

69. Zheng, J., Guo, N. & Wagner, A. Mistranslation reduces mutation load in evolving proteins through negative epistasis with DNA mutations. *Mol. Biol. Evol.* (2021) doi:10.1093/MOLBEV/MSAB206.

70. Schaerli, Y. *et al.* Synthetic circuits reveal how mechanisms of gene regulatory networks constrain evolution. *Mol Syst Biol* **14**, 8102–8102 (2018).

71. Philippe, H. *et al.* Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLOS Biol.* **9**, e1000602 (2011).

72. Brinkmann, H., van der Giezen, M., Zhou, Y., de Raucourt, G. P. & Philippe, H. An Empirical Assessment of Long-Branch Attraction Artefacts in Deep Eukaryotic Phylogenomics. *Syst. Biol.* **54**, 743–757 (2005).

73. Hedtke, S. M., Townsend, T. M. & Hillis, D. M. Resolution of Phylogenetic Conflict in Large Data Sets by Increased Taxon Sampling. *Syst. Biol.* **55**, 522–529 (2006).

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

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

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

77. Dawkins, R. *Climbing Mount Improbable*. (WW Norton & Company., 1996).

78. Grant, P. R. & Grant, B. R. *40 Years of Evolution: Darwin's Finches on Daphne Major Island*. (Princeton university press, Princeton, New Jersey, 2014). doi:10.5860/choice.52-0821.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

93. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **37**, D77–D82 (2009).

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

95. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**, 1173–1175 (2009).

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

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

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

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

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

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

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

103. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, 1983). doi:10.1016/B978-1-55938-802-3.50013-4.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

120. Aakre, C. D. *et al.* Evolving New Protein-Protein Interaction Specificity through Promiscuous Intermediates. *Cell* **163**, 594–606 (2015).

121. Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M. & Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* **445**, 383–386 (2007).

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

123. Lewontin, R. C. Four complications in understanding the evolutionary process. in *SFI Bulletin* vol. 18 (2003).

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

125. R.D. Gietz & R.A. Woods. Yeast Transformation by the LiAc/SS Carrier DNA/PEG Method. in *Yeast Protocol, W. Xiao, Ed.* 107–120 (Humana Press, Totowa, NJ, 2006).

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

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

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

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

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

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

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

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

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

135. Faure, A. J. *et al.* The genetic architecture of protein stability. *Nature* **634**, 995–1003 (2024).

136. Weng, C., Faure, A. J., Escobedo, A. & Lehner, B. The energetic and allosteric landscape for KRAS inhibition. *Nature* **626**, 643–652 (2024).

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

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

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

140. Maynard Smith, J. *et al.* Developmental Constraints and Evolution: A Perspective from the Mountain Lake Conference on Development and Evolution. *Source Q. Rev. Biol.* **60**, 265–287 (1985).

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

142. Blount, Z. D., Borland, C. Z. & Lenski, R. E. Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 7899–7906 (2008).

143. Harms, M. J. & Thornton, J. W. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* **512**, 203–207 (2014).

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

145. Jiménez, J. I., Xulvi-Brunet, R., Campbell, G. W., Turk-MacLeod, R. & Chen, I. A. Comprehensive experimental fitness landscape and evolutionary network for small RNA. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 14984–14989 (2013).

146. Miton, C. M. & Tokuriki, N. How mutational epistasis impairs predictability in protein evolution and design. *Protein Sci.* **25**, 1260–1272 (2016).

147. Lipsh-Sokolik, R. & Fleishman, S. J. Addressing epistasis in the design of protein function. *Proc. Natl. Acad. Sci.* **121**, e2314999121 (2024).

148. Chen, T. S. & Keating, A. E. Designing specific protein–protein interactions using computation, experimental library screening, or integrated methods. *Protein Sci.* **21**, 949–963 (2012).

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

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

151. del Campo, J. & Ruiz-Trillo, I. Environmental Survey Meta-analysis Reveals Hidden Diversity among Unicellular Opisthokonts. *Mol. Biol. Evol.* **30**, 802–805 (2013).

152. Keeling, P. J. & Fast, N. M. Microsporidia: Biology and Evolution of Highly Reduced Intracellular Parasites. *Annu. Rev. Microbiol.* **56**, 93–116 (2002).

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

154. Shimodaira, H. & Hasegawa, M. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference. *Mol. Biol. Evol.* **16**, 1114 (1999).

155. Anisimova, M. & Gascuel, O. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Syst. Biol.* **55**, 539–552 (2006).

156. Eyers, P. A., Erikson, E., Chen, L. G. & Maller, J. L. A Novel Mechanism for Activation of the Protein Kinase Aurora A. *Curr. Biol.* **13**, 691–697 (2003).

157. Tomaštíková, E. *et al.* TPX2 Protein of Arabidopsis Activates Aurora Kinase 1, But Not Aurora Kinase 3 In Vitro. *Plant Mol. Biol. Report.* **33**, 1988–1995 (2015).

158. Özlü, N. *et al.* An Essential Function of the C. elegans Ortholog of TPX2 Is to Localize Activated Aurora A Kinase to Mitotic Spindles. *Dev. Cell* **9**, 237–248 (2005).

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

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

161. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

162. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* gkaa892–gkaa892 (2020) doi:10.1093/nar/gkaa892.

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

164. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

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

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

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

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

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