INVESTIGATING THE MOLECULAR MECHANISMS OF EVOLUTIONARY NOVELTY

by

DAVID WILLIAM ANDERSON

A DISSERTATION

Presented to the Department of Biology
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

December 2014

DISSERTATION APPROVAL PAGE

Student: David William Anderson

Title: Investigating the Molecular Mechanisms of Evolutionary Novelty

This dissertation has been accepted and approved in partial fulfillment of the requirements for the Doctor of Philosophy degree in the Department of Biology by:

William Cresko	Chairperson
Joseph Thornton	Advisor
Patrick Phillips	Core Member
John Postlethwait	Core Member
J. Andrew Berglund	Institutional Representative

and

J. Andrew Berglund	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded December 2014

DISSERTATION ABSTRACT

David William Anderson

Doctor of Philosophy

Department of Biology

December 2014

Title: Investigating the Molecular Mechanisms of Evolutionary Novelty

Evolution is the descent with modification from common ancestors. Forms and functions diversify as a result of changes in genomic sequence that result in changing molecular functions performed by biological molecules such as proteins, RNA, or DNA. Not all genetic changes, however, result in a change in molecular function; highly distinct gene sequences may nonetheless produce similar functions. At the same time, there are some genetic changes that have a significant effect on molecular function, and sometimes highly similar gene sequences may nonetheless produce distinct functional molecules. In order to identify and understand the subsets of genetic changes that were responsible for novel functions we must apply the tools of molecular biology within an evolutionary framework in order to specifically characterize the functional differentiation of diversified genotypes and further to understand the molecular mechanisms that mediated their functional effects. This dissertation has sought to contribute to this work in three related ways: first, by analyzing the dominant approach used in molecular evolutionary research and outlining a program of research that would best yield insight into the mechanisms of evolutionary change; second, by examining the genetic, biochemical, and biophysical mechanisms that gave rise to a novel DNA-binding function in the steroid receptor transcription factors; and third, by functionally characterizing the sequence

space that separates the ancestral and derived DNA-binding function across that evolutionary transition. This body of work has sought to contribute to our general understanding of the principles that underlie the evolutionary process by characterizing the molecular mechanisms that were responsible for some of the interesting, diverse functions that evolution has produced. In doing so, it points towards some important potential general principles that guide evolutionary processes.

This dissertation includes published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: David William Anderson

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene, Oregon
McMaster University, Hamilton, Ontario, Canada

DEGREES AWARDED:

Doctor of Philosophy, Biology, 2014, University of Oregon
Master of Science, Biology, 2008, McMaster University
Bachelor of Arts and Science, Arts and Science, 2006, McMaster University

AREAS OF SPECIAL INTEREST:

Evolutionary Biology

Molecular Biology

PROFESSIONAL EXPERIENCE:

Graduate Teaching Fellow, Department of Biology, University of Oregon, 2008-2009

Teaching Assistant, Department of Biology, McMaster University, 2006-2008

GRANTS, AWARDS, AND HONORS:

American Heart Association Pre-doctoral fellowship, Understanding the Mechanisms of DNA-binding in the Steroid Receptors, University of Oregon, 2011-2013

Natural Sciences and Engineering Research Council (NSERC) Doctoral Scholarship, Studying the DNA-binding evolution in an important family of nuclear receptors, 2009-2012

IGERT-NSF fellowship, University of Oregon, 2009-2011.

Graduate Scholarship, McMaster University, 2006-2008.

PUBLICATIONS:

Alesia McKeown, Jamie Bridgham, Dave W Anderson, Michael Murphy, Eric Ortlund, Joseph W Thornton. (2014). Evolution of DNA Specificity in a Transcription Factor Family Produced a New Gene Regulatory Modules. *Cell*. 159, 58-68.

Adam Bewick, Dave W Anderson, Ben J Evans. (2011). Evolution of the closely-related, sex-related genes DM-W and DMRT1 in African clawed frogs. *Evolution*. 65, 698-712.

Dave W Anderson, Ben J Evans. (2009). Regulatory Evolution of a Duplicated Heterodimer Across Species and Tissues of Allopolyploid Clawed Frogs (Xenopus). *Journal of Molecular Evolution*. 68, 236-247.

Robi Banerjee, Ralph E Pudritz, Dave W Anderson. (2006). Supersonic turbulence, filamentary accretion and the rapid assembly of massive stars and discs. *Monthly Notices of the Royal Astronomical Society*. 373, 1091-1106.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere appreciation for my fellow scientists in both the Thornton lab and in the Institute of Ecology and Evolution at the University of Oregon. The work contained in this dissertation, as is often the case, has been the result of many collaborative experiences, both formal and informal, and could not have emerged without the positive, energetic, and intellectual milieu in which I have worked during my PhD.

In particular, I will single out my fellow graduate student Alesia McKeown; our collaborative projects were the high-point of my experience here, and while it may not have made it into this manuscript, your friendship was every bit as integral to my completing the long and arduous PhD process.

Additionally, I would like to acknowledge the other members of the Thornton lab; Jamie Bridgham, Paul Cziko, Geeta Eick, June Keay, Qinwen Liu, Carrie Olson-Manning, Lora Picton, Mo Siddiq, Tyler Starr, Aarti Venkat, and all of the others who have come through during my time here. The discussions, feedback, and general awesomeness of working with all of you have helped me become a better scientist.

Finally, I must acknowledge the exemplary support of my adviser, Joe Thornton, as well my committee members, Andy Berglund, Bill Cresko, Patrick Phillips, and John Postlethwait; I completed this work under many challenging circumstances, and each of you went above and beyond what was required of you in supporting me through it, whether it was by repeatedly flying across the country or by inviting me to your lab meetings and always making yourselves available for chats and feedback. I will always appreciate your support through this process.

The work in this dissertation was supported in part by scholarships from the Natural Sciences and Engineering Research Council (NSERC grant # 374293-2009) and the American Heart Association (AHA grant # 11PRE7510085).

For Krista and Little J

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

LIST OF TABLES

Table Page

LIST OF BOXES

CHAPTER I

INTRODUCTION

Evolution alters every level of biological systems

Evolution is the process of descent with modification from common ancestors; over time, it changes genetic content, which can cause altered functions for the molecules whose composition is encoded by that genetic content, which in turn can result in changing biological forms and functions. These multiple levels of evolutionary change – the content of genomes, the functions of molecules, and the higher order organismal forms and functions – are tightly interrelated but non-identical. On the one hand, genomes are the primary heritable material, and as such biological changes sustained by evolution must be in some way caused by an altered genetic sequence. On the other hand, important evolutionary forces like natural selection operate on the level of fitness, and thus occurs only when there are differences in organismal function and/or phenotype; in other words, there must be differentiated functions encoded by differentiated genomes for natural selection to act.

Many genetic changes do not change the functions of encoded molecules, and sometimes dramatically different sequences can result in common functions (Meyerguz et al., 2007); at the same time, some sequence changes can have a dramatic effect on molecular function, such that highly similar genes can have significantly different functions (Harms et al., 2013). As such, identifying genetic sequence differences, on their own, cannot uncover changes in molecular function. To do so, we must functionally characterize the specific genetic changes that produced novel functions in evolutionary

history. It is the goal of this dissertation to contribute to our understanding of how evolution has altered the functions of critical molecules, and the physical mechanisms that translated changes in genotype into changes in molecular function.

The rise of evolutionary genetics

Historically, evolutionary biologists explicitly studied differentiated physical traits. Darwin's original conception of descent with modification was based on the variations of organismal forms, such as the different beak morphologies of the famous Galapagos' finches (Darwin, 1859). Even Mendel's work, which gave birth to the field of genetics, was based on observing physical traits that happened to be directly caused by variation at a small number of loci, making the dynamics of inheriting alternate alleles traceable even without any understanding of what, exactly, comprised the physical substance of a "gene" (Mendel, 1866). Further development of evolutionary genetics during the "modern synthesis" era still occurred in the era before DNA was identified as the heritable material of life. As such, the mechanistic relationship between genotypes and phenotypes remained theoretical (Fisher, 1918; Wright, 1932; Haldane, 1933).

Since the discovery of DNA, and even more so since the development of gene-sequencing technology (Sanger and Coulson, 1975), evolutionary biology has taken on the characterization of genotypic change as a major objective. It is hard to overstate the importance of many of the insights that have come from this level of data – from the neutral theory (Kimura, 1968), to the idea that major phenotypic evolution can occur via genetic changes that act mainly to alter gene-expression levels (King and Wilson, 1975), to the refinement of otherwise ambiguous phylogenetic relationships (Koop et al., 1989),

to name only a few. The legacy of these deep insights can still be seen in the genetic focus of the majority of evolutionary biology, as a field.

Moving beyond the genome in molecular evolution

The suggestion that changes in gene regulatory machinery could be a major mechanism of evolutionary change was an important landmark for molecular evolution, suggesting that the relationship between phenotypic divergence and DNA sequence divergence is non-linear. Another landmark came with the formalization of research in evolutionary developmental biology, or "evo-devo", which argued that evolutionary biologists should be seeking to characterize the major developmental mechanisms that cause significant evolutionary changes (Goodman and Coughlin, 2000). Along these lines more recently are arguments in favor of using the classic tools of molecular biology within an evolutionary framework, in order to measure the functional effects of specific historical evolutionary changes, and further, to characterize the physical mechanisms that mediated these effects on function (Dean and Thornton, 2007; Harms and Thornton, 2013).

In this work, I am interested in understanding the molecular mechanisms that gave rise to important functions in evolutionary history. To this end, I have undertaken three separate but related projects. In general, these projects are united by a focus on characterizing the physical mechanisms that produced evolutionary innovation at the molecular level. Much of this work seeks to contribute to the growing catalog of evolutionary mechanisms, and when possible, it seeks to draw general principles that may guide evolutionary processes, and determine how evolution realizes biological novelty.

Content of this dissertation

Chapters II, III, and IV seek to contribute to the field of molecular evolution in distinct but related ways. Chapter II is a scholarly analysis of the methods and assumptions that dominate the field of molecular evolution. In particular, it discusses the ways in which the field currently studies evolutionary mechanisms, with particular discussion of the role that studies of adaptation play, or should play, in directing molecular evolutionary biological research. It uses many specific examples of work in the field in order to illustrate this analysis, and concludes with a proposed program of molecular evolutionary research that would best yield major insights into both the history of evolutionary change and the general principles that guide the evolutionary process. This work will be submitted for publication in the journal *Nature Reviews Genetics* with Joseph W. Thornton as co-author.

Chapter III is an example of the application of the proposed research program from chapter II. It studies the macromolecular evolution of DNA-binding function in the biomedically important transcription factor family of steroid hormone receptors, and resulted in a co-authored work with members of the Thornton and Ortlund labs. This work elucidates the genetic, biochemical, and biophysical mechanisms for novel DNA-specificity in the steroid receptors, and my contribution focused on using molecular dynamics simulations to identify the biophysical mechanisms for functional novelty in this system. It provides a novel mechanism whereby novel DNA-binding function is realized through a coordinated loss of specific positive interactions, as well as specific negative interactions, and the coincident gain of specific negative interactions, in order to

realize an overall shift in DNA-binding specificity that is critical to its developmental role in extant vertebrate species. This work has been published in the journal *Cell*.

Chapter IV characterizes in even further detail the transition of the Steroid Receptor DNA-binding function. By exhaustively characterizing the function of a combinatorially complete set of genotypes that separate the ancestral and derived genotypes, this work provides a detailed description and discussion of the mutational pathways that were available during the evolution of novel DNA-binding specificity. It includes a method for characterizing the genetic causes of differentiated function, and in doing so, allows for a quantitative description of the genotype sequence space that separates the ancestral and derived regulatory modules. Finally, it provides a partial general explanation for the physical mechanisms that caused the transition in binding function. This work was the product of extensive collaboration with Alesia McKeown in the Thornton lab, and we will submit it as co-first authors for publication in the journal *eLife*.

Using history to build general principles for evolutionary processes

Evolutionary biology is the study of the process of change in biological forms and functions over time. Articulating general principles that govern the process is important for understanding the historical legacy that gave rise to important biological forms and functions that now exist. At the same time, we can seek to characterize general evolutionary principles by studying the historical mechanisms that resulted in new biological forms and functions, and to draw general inferences on the basis of those studies. By examining the molecular mechanisms that produced specific novel biological

functions, we can begin to build up the general principles that guide evolution as it realizes new genotypes, functions, and phenotypes. This dissertation seeks to contribute to that important body of work.

Bridge to Chapter II

In Chapter I, I introduce the general scope of the work contained in this dissertation, which is focused on studying the molecular mechanisms that produce new evolutionary outcomes. In Chapter II, I will analyze the field of molecular evolution in more detail, with a focus on the preeminent program that guides contemporary research. This general analysis will conclude with a proposed program, of which chapters III and IV are examples.

CHAPTER II

BEYOND ADAPTATIONISM IN MOLECULAR EVOLUTION

This article will be submitted for publication with Joseph W. Thornton as co-author. I wrote the article and Dr. Thornton provided extensive comments and critiques.

INTRODUCTION

Adaptation and evolutionary biology

Evolution is change in the forms of living beings via descent from common ancestors. The goal of evolutionary biology in general should therefore be to provide historical explanations for how biological phenomena at any level -- from morphological and physiological characters to cellular signaling processes, the structure and function of proteins, or the content of genomes -- came to be, and to formulate a general understanding of the dynamics by which such systems evolve. Molecular evolutionary biology should be conceived broadly as the study of the molecular mechanisms and processes that underlie the evolution of biological phenomena.

There are two classes of evolutionary mechanisms by which organisms are modified over generations: First, population genetics processes that change the distribution of genotypes in a population from one generation to the next, such as mutation, natural selection, neutral drift, and migration; and second, molecular biological processes that transform the distribution of genotypes into a distribution of phenotypes upon which selection can act (Lewontin, 1974). The latter category comprises all the molecular, biochemical, developmental, and other processes by which changes in gene or

genome sequence produce changes in function, phenotype, and ultimately fitness (see Figure 1 – see Appendix A for Figures and Boxes from this chapter).

The former category refers to the population genetics forces that cause a biological character's frequency to change – and, potentially, to become fixed – within a population. To fully understand the evolution of a biological trait, one must establish the nature of both types of mechanisms by which it evolved. The ultimate goal of a unified research program in evolutionary biology should therefore be to specify the underlying mutations in genotype, the changes in biological function by which those mutations affected phenotype, the way(s) in which the phenotype affects fitness, and the population genetics processes that yielded the evolved genotype(s) that caused the discrete biological character to exist in the historical population. Among the population genetics processes, and of particular historical interest for evolutionary biologists, is natural selection leading to adaptation. The term "adaptation" was originally used in evolutionary biology to describe the ways in which organisms appear optimally "fit" for their specific environment (Darwin, 1859). One of the great strengths of Darwin's theory was its capacity to explain how adaptations could be realized through a process of natural selection, without the need to invoke supernatural design. To refer to a biological trait as an adaptation is therefore to make a specific claim about its history: Namely, that it became fixed within a progenitor population due to being the direct target of natural selection. An adaptive character is differentiated from non-adaptive characters that were fixed by other population genetics processes, such as neutral drift, physical linkage (in which a neutral or deleterious trait is caused by a genotype that is physically linked to another genotype that causes an adaptive trait), pleiotropy (in which a neutral or

deleterious trait is caused by the same genotype responsible for an adaptive trait), or random demographic events like population bottlenecks (Figure 2).

Establishing that a phenotypic character or genetic region contributed to adaptation is not a complete explanation of the historical evolutionary process. First, investigating the role of selection in driving the evolution of a gene or phenotype leaves unaddressed the roles of other population genetic factors that may also have contributed to its evolution (such as migration, inbreeding, sexual selection, demography, and linkage (Figure 1). More generally, however, a complete understanding of how and why a phenotype or locus evolved requires one to study the molecular, developmental, and physiological mechanisms that translated a particular genotype into a particular phenotype. Without studying the link between genotype and phenotype, one is left with a partial story at best. If one studies only the genotype, one cannot identify the biological consequences of genetic change or, in turn, understand why a genetic region may have increased fitness. If one studies only the phenotype, the genetic architecture of its evolution remains unknown, and key questions about evolutionary genetic processes must remain unanswered (Dean and Thornton, 2007).

The adaptationist program

Evolutionary biology's focus on adaptation is a product of its history. At the time of its formulation, Darwin's theory essentially acted to replace supernatural design as the source of biological diversity, and the adaptation of specific species to their specific environments. As such, research initially sought to demonstrate that evolution by natural selection could lead to adaptation. The concept of adaptation thus became nearly

synonymous with the concept of evolution. This equation was reinforced by the social debates that emerged as the theory of evolution by natural selection showed that adaptive phenotypes could no longer be considered evidence for a supernatural designer (Mayr, 1983).

At times, researchers were so focused upon the importance of adaptation in evolution that they failed to recognize that non-adaptive population genetics mechanisms for fixing novel characters even exist. As Theodosius Dobzhansky argued: "The basic postulate of the modern biological theory of evolution is that adaptation to the environment is the guiding force of evolutionary change…we have to suppose that most organs and functions of most organisms are, or at least were at the time when they were formed, in some way useful to their possessors. Nothing less than this is acceptable if the modern theory of evolution is sustained." (Dobzhansky, 1956) This became the major theme of Stephen J. Gould and Richard Lewontin's now classic paper: "The Spandrels of San Marco and the Panglossian Paradigm: a critique of the adaptationist programme". Gould and Lewontin argued that evolutionary biology was dominated by a "Panglossian" worldview (see Box 1), and that researchers took as a given the idea that if a trait exists in present day organisms, it must have been fixed because it acted as an adaptation in an historical population. The job of evolutionary biologists in the Panglossian adaptationist program was therefore to study extant species and provide explanations for why their biological traits acted as adaptations in their respective environments. Gould and Lewontin argued that the adaptive nature of a trait should not be assumed *a priori* because there are population genetics mechanisms that could allow non-adaptive traits to become fixed as well. Therefore, researchers must propose specific adaptive hypotheses

for the traits being studied that can then be experimentally tested (Gould and Lewontin, 1979).

Criticism of the Panglossian adaptationist program was supported by the neutral theory and studies that purported to show that the majority of fixed genetic differences in protein-sequences between species are functionally and selectively innocuous (Prakash et al., 1969; Ohta, 1976; Kimura, 1977). While evolutionary biologists broadly accepted the standard that adaptive hypotheses must be tested experimentally, they nonetheless held that adaptation comprises the most important aspects of evolutionary change. The focus on adaptation persisted as the program went from explaining how biological characters were adaptive to identifying the subset of evolutionary changes that were responsible for adaptation. As Ernst Mayr argued: "…showing that possession of the respective feature would be favored by selection…[is the] consideration which determines the approach of the evolutionist." (Mayr, 1983) The agenda of the adaptationist program became one of testing hypotheses about adaptive evolution and of identifying the subset of genetic changes and phenotypes that were responsible for adaptation. In other words, the ultimate purpose of studying evolution remained, in effect, one of studying adaptation.

Post-Panglossian adaptationism holds that demonstrating that a trait or a genetic region was involved in causing adaptation is both necessary and sufficient for a study to contribute to evolutionary biology. It is "post" Panglossian because it does not assume that all traits must have been adaptive. It remains, however, an adaptationist program because adaptation is its organizing principle. This means that research within this program must focus on the subset of evolutionary changes that were adaptive, as they are the most important and/or biologically relevant. It is common for research papers within

this program to open with a statement along the lines of: "Mapping adaptive trait loci (ATL) underlying ecological divergence is an essential step towards understanding the processes that generate phenotypic diversity." (Crawford and Nielsen, 2013), also (Nunes et al., 2010; Pritchard and Di Rienzo, 2010; Stapley et al., 2010). This central purpose has been widely embraced in evolutionary biology in general as gene-sequencing technology has improved and proliferated.

Historically, evolutionary biologists studied adaptation by characterizing the fitness advantage of a particular trait in a particular environment (Fretwell, 1969; Kaufman et al., 1977). More modern methods, however, rely primarily or entirely on DNA-sequence statistics to infer the genetic targets of natural selection, which are expected to have distinct "signatures" surrounding the specific genotype that contributed to a selectively favored function or phenotype. This approach is sometimes integrated with a study of the functional or phenotypic differences between species or populations, but in the vast majority of cases sequence-based evidence is presented either on its own or as the central component, and corroborating measurements of the functional or selective variation are left as future work still to be done (Sabeti et al., 2007; Fumagalli et al., 2011; Liu et al., 2014). This modern adaptationist program is manifest in several types of studies: Those that use sequence signatures to infer whether a specific gene or allele that was or is subject to historical positive selection (Evans et al., 2005; Mekel-Bobrov et al., 2005), those that compose a list of loci across the genome of a specific species that were subject to historical positive selection (Sabeti et al., 2007; Huerta-Sanchez et al., 2013), and those that seek to functionally characterize the molecular and developmental mechanisms of adaptation (Storz et al., 2009; Barrett and Hoekstra, 2011;

Kamberov et al., 2013). Studies from within this program are united by their exclusive focus on studying adaptation over non- or ambiguously-adaptive evolutionary change.

Our purpose is to argue two distinct but related points: First, that gene-sequence analyses identifying cases of adaptation in the genome is not sufficient to make deep or reliable insights into the evolutionary process; Second, that demonstrating adaptation is not necessary to meaningfully investigate evolutionary questions. In a post-adaptationist program, evolutionary research would not have adaptation as its organizing principle, and would instead focus its work on providing a complete explanation for how interesting traits and phenotypes came to exist. Our purpose is not to reproduce the adaptationist vs. neutralist debate, as we are not arguing that there is a preponderance of either adaptive or non-adaptive traits that arise through the evolutionary process. Rather, we argue that focusing on finding adaptations to the exclusion of studying non- or ambiguously-adaptive traits is an unproductive research agenda. A post-adaptationist program would involve characterizing all the mechanisms for the evolution of forms and functions, from the molecular and developmental mechanisms that translate alternative genotypes into phenotypes, to the population genetics processes that drove the relevant genotypes to fixation within ancestral populations. Adaptation is not the lens through which everything is judged; it is one component of the larger evolutionary process, and the goal is to understand that process as a whole.

IDENTIFYING ADAPTATION IN THE GENOME BY STATISTICAL ANALYSIS OF SEQUENCE DATA IS NOT SUFFICIENT

Sequence analysis is insufficient to demonstrate adaptation

The statistical analysis of gene-sequence data has become a hallmark of many modern studies of evolution. Gene sequence data has provided many important insights into phylogenetic reconstruction (Adoutte et al., 2000; Rokas et al., 2003; James et al., 2006) and the evolution of genome architecture (Postlethwait et al., 1998; Hokamp et al., 2003; Vandepoele et al., 2004). For the purposes of this discussion, however, we will limit our discussion to the analysis of gene-sequence data with the purpose of demonstrating a role in adaptation.

There are several different types of methods for statistically inferring a signature of adaptation in the genome (discussed in more detail in Box 2), which we will briefly outline here. One set of methods, sometimes referred to as "codon-based" analyses, identifies signatures of adaptation in the genome by examining the number or rate of non-synonymous changes (those that change the amino acid residue, the substitution rate abbreviated as "dN") versus synonymous changes (those that do not change the amino acid residue, the substitution rate abbreviated as "dS") between or across phylogenetically related genes. A second type of gene-sequence-based test uses population-level genetic variation to infer recent selective sweeps. Although analyzing gene sequences has become a common tool used to infer adaptation, they are challenged by the fact that many mechanistic factors, such as physical linkage, demographic fluctuations, and pleiotropy can result in non-adaptive traits and genotypes becoming

rapidly fixed (Figure 2). This means that there are many processes, some adaptive and some non-adaptive, that can produce an apparent signature of adaptation in the genome.

Natural selection acts on variation in biological function and phenotype. That variation, however, is rarely perfectly correlated with variation at a single nucleotide in the underlying causal genotypes. Thus, the way in which population genetics processes like natural selection and neutral drift alter the frequencies of different genotypes is depends on the molecular and developmental biological processes that translate alternative genotypes into alternative phenotypes (see Figure 1). For example, direct natural selection on a trait (for example, height) will change genotype frequencies differently if variation in that trait is caused by genetic variation at a single site versus if it is caused by genetic variation at many different sites (which may or may not be physically linked), and differently still if variation at any of those sites pleiotropically alter other traits, which may themselves be the target of natural selection. As such, it is possible for both non-adaptive phenotypes, as well as genotypes that do not directly contribute to an adaptive phenotype, to become fixed in populations, depending on the strength of adaptive and neutral population genetics processes (natural selection and demographic fluctuations, for example) and depending on the molecular and developmental mechanisms that translate alternative genotypes into alternative phenotypes (see Figure 2).

Statistical analyses of genetic variation seek to collapse these complex interrelationships between population genetics and molecular biological processes into a single metric. The problem, however, is that multiple different processes – some adaptive, some non-adaptive – can produce similar genetic signatures. For example, even

when there is a direct relationship between a single genetic variant and a differentiated phenotype, the causal genetic change is almost never clear from statistical analyses on their own due to the presence of physically linked variants in the genomic region identified (Figure 2C). Additionally, positively identified genetic changes may have become fixed not because they caused an adaptive phenotype, but because they interact epistatically with a genotype that did cause an adaptation, and were required in order to permit the causal genotype to be tolerated (Figure 2D) (Carroll et al., 2011; Gong et al., 2013). Such epistatically-linked variation would be driven to fixation by directional selection when the causal genotype emerged, despite not being directly responsible for the selected phenotype. Also, a non-adaptive phenotype may become fixed because it is caused by the same genotype responsible for a separate, adaptive phenotype (Figure 2E) (Barrett et al., 2008), which we refer to as pleiotropically-linked traits. Finally, both non-adaptive phenotypes and genotypes that do not cause an adaptive character can become fixed due to demographic events such as population bottlenecks (Figure 2F). In many systems, analyses based solely on genetic data cannot confidently distinguish between the alternate scenarios that can give rise to the same "adaptive" signatures in the genome, and even if they did, they cannot inform us on the phenotypic changes that arose in the adaptive process. Experimentally characterizing the functional and phenotypic effects of genetic variation is the best way to differentiate between these alternate scenarios.

There are also other fundamental issues with the most commonly used statistical methods for inferring adaptive genotypes, which have been extensively discussed by others but which we will briefly summarize here. For example, there is an explicit premise that underlies all codon-based tests: That synonymous changes are neutral in

terms of their effects on function and fitness. But it has been shown that synonymous mutations often have measurable effects on biological function by altering translational accuracy (Qian et al., 2012), folding properties (Drummond and Wilke, 2008), and transcription factor binding (Stergachis et al., 2013). Since both dN/dS model comparisons and the McDonald-Kreitman test rely on the assumption that synonymous changes are neutral, the finding that they can have functional (and therefore selective) consequences undermines the very foundation of these analyses. Additionally, both dN/dS and McDonald-Kreitman tests are challenged by potential error introduced by sequence-misalignment, which can bias these tests toward false positives (Markova-Raina and Petrov, 2011). There have also been several cases in which the conclusions drawn from codon-based methods have been directly contradicted by functional data (Yokoyama et al., 2008; Zhuang et al., 2009). Finally, while population-based tests (e.g. Fst, LD, etc.) sit on a somewhat firmer foundation, they nonetheless have potentially confounding factors, the most notable of which is demographic history. This has been well illustrated by the initial finding of adaptation in the *microcephalin* and *ASPM* loci in humans (Evans et al., 2005; Mekel-Bobrov et al., 2005) and the follow-up work that showed the supposed adaptive genotypes had no measurable phenotypic effect (Timpson et al., 2007), and further, that the geographic distribution of alternative genotypes matched that expected given the geographic history of human demographic expansion (Currat et al., 2006).

One way in which researchers have attempted to deal with this methodological uncertainty is to analyze genetic diversity data genome-wide. The argument goes that since the entire genome should have experienced the same demographic history, one can

overcome the confounding effects of demography because genomic regions that contain genotypes driven to fixation due to being the direct target of natural selection should exhibit significantly different patterns of genetic variation compared to rest of the genome. The implementation of this approach, however, often fails to recognize that while demographic history is equally likely to affect variation across the genome, its actual effect will play out independently at unlinked loci – in other words, the same demographic event can result in widely different reductions of genetic diversity at different loci even within the same genome. As such, one can have demographic processes that produce apparent outliers in terms of genetic diversity under purely neutral conditions (Excoffier et al., 2009), while the presence of purifying selection confounds this problem even further (Hermisson, 2009). Without a detailed understanding of demographic history it is not possible to say whether or not a particular pattern of genetic diversity, even an apparent outlier, was caused by neutral processes instead of natural selection and adaptation (Jensen and Rando, 2010). In some cases, the issue of demography can be overcome, for example when widespread parallel adaptation has occurred (Hohenlohe et al., 2012; Jones et al., 2012), but we are as yet unaware of a general approach for most species that has been demonstrably successful in doing so. And even given advances in our understanding of human demographic history, the methods we have at hand are not sufficient to make strong conclusions about adaptation in the human genome.

Genome-wide searches for signatures of selection have thus proved similarly unreliable. Hernandez et al. tested the assumption that classic selective sweeps, previously identified in the human genome (Akey, 2009), should exhibit patterns of

reduced genetic variation surrounding the genetic sites directly contributing to adaptation (which would be consistent with a scenario of recent selective sweeps). They analyzed these signatures in more detail, and found that they were not consistent with selective sweeps, and were instead consistent with background selection. Contrary to their expectations, they found that genetic differences that were non-synonymous and located within conserved non-coding gene regulatory regions did not exhibit a lower level of genetic variation surrounding them as compared to genetic differences that were synonymous or intergenic. The author's argument is that since the patterns of reduced variation are similar at both functionally relevant (i.e. non-synonymous and regulatory) and non-functionally relevant (synonymous and intergenic) sites, then this is most likely due to the effect of neutral genetic drift. They suggest that very little of the human genome has likely evolved under the classic selective sweep model expected for adaptation. Instead, the authors suggest that adaptation must have occurred predominantly via "soft" selective sweeps (i.e. by the fixation of a low-frequency allele that were previously segregating in the population for many generations prior to some environmental change that caused it to become favoured by natural selection). Another possibility is that whatever adaptation has happened over human evolution, it involved few selective sweeps of any time across the genome (Hernandez et al., 2011). Yet another possibility is that human evolution did not involve a large number of selective sweeps of any kind. In any case, what is clear is that applying these tests for signatures of adaptation genome-wide cannot be considered reliable, at least for human populations, where they are most commonly applied. Even given significant advances in our understanding of

human demographic history, the methods we have at hand remain insufficient to make strong conclusions about adaptation in the human genome.

Genetic signatures of adaptation should be interpreted with these issues of reliability in mind. At a minimum, adaptation cannot be confidently inferred unless genetic differences can be shown to have a functional or phenotypic effect in a way that matches some difference in environment, development, or lifestyle.

Sequence-based analysis alone provides limited insight into evolution

Even if the methods for inferring signatures of adaptation in the genome were reliable, however, demonstrating that a genotype or set of genotypes were fixed due to being the direct target of natural selection reveals little about the evolutionary process. The functional mechanisms of adaptation are typically left as either unknown or the subject of weak inference (Pavlidis et al., 2012). But simply knowing that a particular genetic region evolved under a model of adaptation does not reveal why it was adaptive. At best, this approach allows one to categorize a particular set of genetic changes as "adaptive" without elucidating anything about how it works, or why it would have been adaptive to a specific set of environmental, developmental, or life history conditions that differentiate a particular species or population. Efforts are often made to conclude something meaningful about how a set of genetic changes might work, sometimes by consulting the gene-ontology (GO) database; however, this approach is fraught with the potential for "just-so" storytelling (Gould and Lewontin, 1979; Pavlidis et al., 2012).

Critical questions remain unanswered by this approach: what phenotypes were produced? How did these genetic changes produce the different phenotypes? Was this a

result of many interacting genetic changes, or a few changes with independent effects? What was the nature of the evolutionary genetic process that allowed adaptation to occur? Was there epistasis? What was the effect size of individual mutations? Was there pleiotropy for any of these mutations that made this biological mechanism more or less likely than other means for increasing fitness? If a study only shows that a locus or allele was adaptive, then the details about what phenotypes were produced by those mutations, what about those phenotypes was adaptive, why those phenotypes lead to an increase in fitness, and how those phenotypes were produced biologically remain unknown. Did the mutations change gene regulation during development? Enzyme specificity? Metabolism? Did they rewire a genetic network, or did they change just one specific molecular function? Even 100% reliable statistical methods to identify adaptation in the genome cannot answer these questions on their own.

Despite the limitations for understanding functional evolution, studies that primarily or exclusively rely on the statistical analyses of gene sequences have become arguably the most prolific type of papers in contemporary evolutionary biology literature, including in top tier journals. For example, consider a recent study that analyzed gene-sequence diversity statistics comparing polar bears to other related species (Liu et al., 2014) – note that we could easily point to many analogous studies in the literature from 2014 alone (Cardona et al., 2014; Christiansen et al., 2014; Eichstaedt et al., 2014; Enard et al., 2014; Li et al., 2014; Steane et al., 2014; Udpa et al., 2014; Welch et al., 2014; Wuren et al., 2014). This study sought to construct a list of loci, with associated p-values that indicate the likelihood that the pattern of genetic variation could be caused by neutral drift under a likely demographic model. The argument is that those loci with a

statistically significant p-value have the signature of a selective sweep, which would be consistent with being the direct target of natural selection. In particular, the authors highlight a subset of the significant loci that comprise a number of genes whose GO entries show have been associated with fat metabolism and cardiovascular function in humans and mice. The authors argue that genetic changes at these loci (several of which encode what are expected to be premature stop-codons) have contributed to the polar bear's adaptation to a high-fat diet. But how did these genetic differences accomplish this proposed adaptation? Did they alter enzymes involved in lipid metabolism? Did they affect the expression levels for proteins that are involved in cardiovascular function? Did they cause a reorganization of a genetic network involved in either of these functions? And how? Why would premature stop-codons in these specific genes result in polar bear-like fat metabolism and cardiovascular function? If the signatures for adaptation are truly reliable, then this could be considered a first step toward understanding the evolutionary changes that have made polar bears an interesting and unique species. But it is not enough to leave all of these questions about the functional mechanisms to be done as future work. Deeper insight into the evolutionary process requires more than an unconfirmed signature in the genome (MacCallum and Hill, 2006; Pavlidis et al., 2012).

Understanding evolutionary change requires functional analysis

The mechanism of evolution is the change of genotype frequencies, but natural selection acts on phenotypes. To understand how evolution changed forms and functions, we must therefore study both genotypic and phenotypic change, and to elucidate the molecular mechanisms that connect the two. Similarly, in order to understand how

evolution has changed genotype frequencies and genome content, we must understand how and which alternative genotypes encode alternative phenotypes (which may be the target of natural selection), and how evolution has acted upon them. In order to accomplish this, we need to conduct functional biological analyses, by making direct measurements of the biological consequences of specific genetic changes. These measurements include things like specific protein functions, gene regulation, developmental changes, or differences in lifestyle between different species or populations. Meaningful insight has come from the adaptationist program, but only when researchers approach their study system with this broad perspective, extending their studies to the molecular mechanisms that underlie the evolution of forms and functions as a result of changing genotypes. For example, in a series of studies that examined at high-altitude adaptation in deer mouse populations, researchers began with the hypothesis that high-altitude populations are under selection for stronger hemoglobin-O_2 affinity as compared to lowland populations. Aerobic organisms must be able to survive the relatively low oxygen abundance at high-altitudes, which makes hemoglobin a particularly strong candidate for functional adaptation (Storz et al., 2007). They identified genetic variants in both α- and β-chain hemoglobin genes that were highly segregated between low-altitude and high-altitude populations – more so (though not dramatically) than the genome-wide average. In order to draw a strong conclusion regarding adaptation and these differentiated genetic regions, they directly tested the O_2-binding properties for hemoglobin encoded by these differentiated genotypes. They found that the high-altitude variant had greater O_2-binding affinity (Storz et al., 2009), which precisely matched their hypothesis for high-altitude adaptation. But it is worth emphasizing that the genetic

differentiation was not the definitive demonstration of adaptation; it was the clinal differentiation of hemoglobin function and its direct match with a strong biological hypothesis for why it would have been beneficial in different environments. Even more interestingly, they found extensive functional epistasis within and between the α- and β-chain hemoglobin loci for the genetic changes that exist between these differentiated functions, and which must have severely impacted the evolutionary pathway taken during the evolution of this trait (Natarajan et al., 2013). The most important evolutionary insights from this work come from its discovery of the detailed molecular and genetic mechanisms that resulted in different hemoglobin functions. While there is a good case that these changes were adaptive, this body of work would remain interesting and valuable even if the biological trait hadn't been fixed by natural selection.

Uncovering the detailed mechanistic basis for the evolution of an adaptive trait reveals significantly more about the evolutionary process. This is clear from a series of studies that looked at coat-colour adaptation between populations of *Peromyscus polionotus* field mice. These researchers hypothesized that divergent coat colouration was adaptive because of the different sand colouration in each population's local environment (Hoekstra et al., 2006). They established a specific functional hypothesis of environmental crypsis, whereby the mice are roughly color-matched to the sand, thus conferring a selective advantage via reduced predation. They then identified two genetic regions that showed significant association with the alternate coat-colour phenotypes, at *mc1r* and *agouti*, and they connected these genetic differences to differences in their gene expression that result in different coat colouration (Steiner et al., 2007). Finally, they demonstrated that genetic variation at the *mc1r* and *agouti* loci interact epistatically in

order to realize the full functional differences in gene expression and coat-colour between these populations. To directly test their adaptive hypothesis, they built plastecine models that were colour-matched with the habitats of these two populations, and then measured predation rates in the different environments (Vignieri et al., 2010). They found that colour-matched models experienced fewer predator attacks, convincingly demonstrating the adaptive advantage of coat-colour divergence between these populations. This body of work was successful by combining multiple levels of inference, from an environment-based hypothesis of adaptation, genetic differentiation that is associated with functional variation in such a way that matches the adaptive hypothesis, the interacting genetic mechanisms that cause the functional diversification to occur, and finally, a direct measurement of natural selection that acts on the differentiated phenotype. Without the extensive functional data, however, what could this study have shown? They would have failed to connect the differentiated genotypes to differences in phenotypes. Identifying a genetic signature in the absence of definitive knowledge about its functional consequences make for a very limited study of these species' evolution. On the other hand, however, how would the value of this work change if coat-colour did not show evidence of being adaptive? In that case, they would still have demonstrated the mechanistic basis of biologically interesting variation in function and phenotype. In other words, the thing that makes this study interesting has nothing to do with studying adaptation; it has everything to do with studying the underlying mechanisms of evolutionary change.

DEMONSTRATING ADAPTATION IS NOT NECESSARY

Insights into evolution come from both adaptations and non-adaptations

There persists a general focus in molecular evolution on the changes that were adaptive, with the attitude that they should be prioritized for study because they will provide greater insight into how species and populations became differentiated and unique. Many researchers make a claim along the lines of: "The genes that became fixed in our lineage as a result of positive selection are, after all, the ones that make us human." (MacCallum and Hill, 2006) Yet it is the traits that have differentiated – those that became uniquely fixed along the human lineage – and the underlying genetic changes that caused them to exist that make us uniquely human. This is true whether those traits were adaptive or not. Suppose that many genetic changes were required to realize a complex trait like the human capacity for learning and abstract reasoning, and further, suppose that at least some of those genetic changes were fixed by neutral or maladaptive processes: Would that mean that we should not be interested in understanding how they caused this trait to exist? Or that they are not among those genetic changes that "make us uniquely human"? If we only study adaptation, we ignore many other important traits that helped produce the diversity of biological forms and functions. To study evolution, we must study these non-adaptive contributions to diversity as well.

A more productive program of evolutionary research would study traits for their own innate biological interest, considering adaptive and non-adaptive components alike. Evolution involves all types of change, and mechanistic factors like pleiotropy, epistasis, and linkage, which govern the translation of genotypes into phenotypes, mean that both adaptive and non-adaptive phenotypic changes may be caused by tightly linked, or even

identical, genetic changes. As such, studying traits for their own sake, and uncovering the population genetics processes that led to the fixation of their causal genetic variants, can illuminate otherwise unknown aspects of important biological systems. To illustrate this point more specifically, we refer to a set of work that examined the phenotypic diversification of flower colour between populations of *Phlox drummondii*. In this case, the authors hypothesized that both colour intensity (light vs. dark) and colour hue (red vs. blue) were adaptive because they are highly differentiated between these populations (Hopkins and Rausher, 2011), and flower colour phenotypes like these have been shown to significantly influence pollinator behaviour in other systems. By conducting a detailed study of pollinator behaviour in mixed populations of these flowers, however, they showed that intensity is adaptive for pollination, while hue is not. Studying colour hue, which was not shown to be adaptive, as well as colour intensity, which was, revealed a fundamental aspect of system: Namely, that pollinators respond to differentiated color intensity while ignoring differentiated color hue (Hopkins and Rausher, 2012). This apparently non-adaptive detail revealed something fundamentally interesting about how the system evolved.

Similarly, consider work that studied the reduced armor plating in freshwater populations of three-spine stickleback. Oceanic and freshwater populations vary in the amount of hardened armor plating they develop along their flanks (Walker and Bell, 2000), and it has long been hypothesized that dramatically reduced armor plating was an adaptation for freshwater habitats (Reimchen, 1994). This group investigated the hypothesis that a specific genetic variant at the locus of *eda*, which had previously been established as a locus of major effect in a QTL study for armor plating development

(Cresko et al., 2004), was fixed as a result of its role in causing this potentially adaptive trait. In order to test the related hypotheses that reduced armor plating, and its associated *eda* variant, were adaptive, they built experimental freshwater ponds and stocked them with populations of three-spine stickleback that were polymorphic both at the *eda* locus and for the armor plating phenotype. They directly measured selection on both the phenotype and the *eda* genotype, demonstrating that natural selection in the freshwater environment favors the fixation of both reduced armor plating and the associated *eda* variant (Barrett et al., 2008). Even more interesting, however, was the finding that the direction of natural selection varied across the life history of the fish, and that during the juvenile stage (before armor plating has developed) selection acts against the reduced-armor-associated *eda* variant. This variant is still favored by natural selection over all, but only because the strength of positive selection during adulthood overcomes the negative selection during the juvenile stage. As such, one of the most interesting aspects of this work was that they uncovered a maladaptive pleiotropic function caused by the same *eda* variant that causes the reduced armor-plating adaptation in freshwater habitats (Figure 2C). The non-adaptive aspect of this study made it significantly more compelling by demonstrating that negative pleiotropic maladaptation accompanied this beneficial adaptive phenotype (Cresko, 2008). It revealed the underlying biological connection between the positively selected adult phenotype of reduced armor plating and the negatively selected juvenile developmental effect. In both the *phlox* and stickleback examples, studying the detailed function and effect of non-adaptive traits yielded a much richer and more thorough evolutionary study.

Limiting ourselves to studying adaptive genotypes and phenotypes would prevent us from understanding the evolution of many important biological systems and characteristics. This pertains to both non-adaptive evolutionary phenotypes, and to those phenotypes for which their adaptive nature cannot be determined. Research can explore fundamental questions of evolutionary processes by studying traits, and their genetic, molecular, and developmental causes, irrespective of whether they can be shown to have been adaptive. As an example, we highlight work that studied the functional evolution of the influenza A viral gene *nucleoprotein*. In this study, they measured the effect on influenza growth rates for each individual non-synonymous substitution that separates two extant sequences. Most substitutions were shown to be neutral. A few, however, were found to compromise growth (and thus, presumably, fitness) when introduced into the alternative extant genetic background. They then examined the effect of these apparently deleterious substitutions when introduced along with other changes that occurred at or near the same branch in the phylogenetic tree. This allowed them to identify several "permissive" changes that, while appearing neutral themselves, acted to epistatically alleviate the negative effect of the deleterious substitutions (Figure 2D). They also found that the deleterious substitutions appeared to significantly destabilize *nucleoprotein*'s protein product, while the permissive substitutions increased folding stability. These stabilizing substitutions are shown to be neutral when introduced in isolation, suggesting that beyond a certain minimum threshold, increased protein stability may not confer any selective advantage. But when they are introduced in conjunction with the deleterious substitutions, they reverse the deleterious effect on growth by keeping the overall protein above the stability threshold required for function (Gong et

al., 2013). By specifically identifying and characterizing non-adaptive substitutions, this work gleaned worthwhile insight into the importance of protein stability as a mechanism by which substitutions can act permissively to allow other, potentially adaptive, changes to become fixed.

Even complex phenotypes such as the underlying mechanisms of sexual differentiation cannot always be shown to have been adaptive or non-adaptive. But no one would argue that understanding the mechanisms of their evolution was unworthy of study. As an example, consider work that showed how a new gene became intercalated into the regulatory circuit that causes sexual reproduction in yeast species. This intercalation event allowed a novel signaling mechanism to be integrated into an existing regulatory circuit, which made starvation a cue for sexual reproduction (Booth et al., 2010). The regulatory changes themselves, however, were not shown to be either adaptive or non-adaptive, making the population genetics forces that led to its fixation unknown. If it could be shown to have been either adaptive or neutral, that would provide an interesting context for the acquisition of this trait. But the lack of information regarding the adaptive value for this intercalation event does not make these findings uninteresting, or this system unworthy of study.

Non-adaptive evolution can also play a critical role in functional and phenotypic evolution via processes like the Duplication-Degeneration-Complementation ("D-D-C") model of post-gene duplication evolution (Force et al., 1999). For example, a recent study demonstrated that the D-D-C model explains the functional evolution of the V-ATPase proton pump in fungi (Finnigan et al., 2012). Here, they demonstrated that two distinct protein components of this machine evolved from a common ancestral protein, which

was capable of fulfilling both of the extant protein's roles. Degenerative changes in the two descendant proteins occurred following the ancestral gene duplication event, but in such a way that the function of the ancestral protein was subdivided between the two descendants. In this study, they were able to identify the specific physical mechanisms that caused this diversification, and further, they demonstrated how increased biological complexity has evolved through neutral processes, with the overall function of the V-ATPase complex remaining the same throughout (Doolittle, 2012). Adaptation could not have driven these changes because there was no functional difference upon which directional selection could act. But the evolution of complex cellular machinery is a critically important area of study. Were we to accept the premise that adaptations comprise the subset of evolutionary changes that are worthy of study, then we would never uncover the details about how a complex system like this evolved, and our understanding of evolution as a whole would suffer because of it.

We could easily point to many other sets of work that appear unconcerned with demonstrating that the traits they are studying were adaptive, but which nonetheless made meaningful insights into the evolution of biologically interesting traits. Such work has uncovered details about the ways in which functional and phenotypic evolution are governed by epistasis (Bridgham et al., 2009), the relative importance of large- vs. small-effect substitutions in causing functional and phenotypic change (Shao et al., 2008; Harms et al., 2013), and the importance of changes in gene regulatory machinery (Gompel et al., 2005). These studies have been done without any attempt, through the use of sequence-signatures or fitness measurements, to make claims about adaptation. If the traits in question were shown to have (at least initially) been non-adaptive, however, they

would not be devalued in any way. Having additional information regarding the adaptive or non-adaptive value of these traits might supplement these findings in interesting ways, and provide insights into the underlying population genetics processes that led to their existence, but importantly, this is true whether they turn out to have been adaptive or not.

The adaptationist program says that adaptive traits are the most interesting, causing researchers to sometimes rely on weakly inferred roles in adaptation

The prioritization of adaptation has meant that demonstrating that a trait or gene has evolved adaptively is seen as a necessary component for any study to contribute to the field. Some researchers use genetic-signatures of adaptation to "supplement" other data regarding a particular gene's function that would otherwise be unable to support any claims, one way or another, about adaptation, thus satisfying the program's requirement. Such studies, however, do not actually test whether specific genetic variation (about which claims regarding adaptation are being made) cause measurable functional variation that would match such an adaptive hypothesis. For example, a recent study characterized the function of a captured retroviral *syncytin* protein that is required for placental fusion in humans and other related mammals (Cornelis et al., 2012). In the virus, this gene produces a protein necessary for the viral capsid to become fused to the target cell's membrane. *Syncitin* has been coopted by some mammals such that it similarly opens the cell membrane of some maternal and fetal cells in the placenta in order to allow them to become fused. This paper showed that this formerly-retroviral gene is required for proper placental development. Relevant to this discussion, however, they also performed a dN/dS analysis of *syncytin* gene sequences in order to argue that this gene has evolved

adaptively across the phylogeny of mammals that have diversified since it was initially captured. They did not perform functional assays to test how this function may have adapted differently between different species, nor did they hypothesize why *syncitin*'s protein function might be selected to function differently (the relevant species exhibit similar placental fusion phenotypes) in a way that would have been favoured by an "arms race" adaptive scenario (which is the case of adaptation detectable by dN/dS methods - Figure 2B) (Hughes, 2007).

The requirement that evolutionary biologists must study adaptations establishes a limiting lens that can blind us to rich insights that could otherwise emerge from scientific work. To illustrate this point, we refer to a set of papers investigating local adaptation between human populations (Sabeti et al., 2007; Kamberov et al., 2013; Tan et al., 2013). The requirement that evolutionary biologists must study adaptations establishes a limiting lens that can blind us to rich insights that could otherwise emerge from scientific work. To illustrate this point, we refer to a set of papers investigating local adaptation between human populations (Sabeti et al., 2007; Kamberov et al., 2013; Tan et al., 2013). These authors first implemented a statistical population genetics approach to examine genome-wide diversity data in human populations in order to construct a list of the loci that have mediated local adaptation. They initially identified a subset of loci across the genome, which exhibit the most differentiated patterns of genetic variation between human subpopulations and which also encode non-synonymous differences within protein-coding genes. The assumption underlying the conclusions from this statistical analysis is that such significant genetic differentiation implies a role in adaptation. One of the many such polymorphisms that were identified was located within the protein-coding region of

edar, a gene implicated in hair and eccrine gland development (among other phenotypes). Critically, however, a functional hypothesis for why variation in these traits would be involved in local adaptation between human populations remained unclear. They then characterized the functional effect of the *edar* variation in mouse models, and showed that variation in *edar* could explain some of the observed variation in human hair thickness and eccrine gland development between Han Chinese and non-Chinese populations (Kamberov et al., 2013).

The problem arises when the authors provide an adaptive explanation in order to support the gene-sequence signature that, as the underlying assumption holds, implies this genetic variant is involved in local adaptation – and which, according to the current adaptationist program, is also required if their mechanistic findings are to be considered valuable. "High humidity, especially in the summer [in ancient China], may have provided a seasonally selective advantage…Alternatively, another phenotype, such as mammary gland branching or fat pad size could have been adaptive…Reports of smaller breast size in East Asian women are notable in light of the effects of [the Han-associated *edar* variant] on fat pad size and the importance of breast morphology in human mate preference." The hypotheses put forth for why this phenotypic differentiation is locally adaptive are extremely weak: If heat and humidity made this *edar* variant the target of natural selection, why is it largely absent from India? And Papua New Guinea? Furthermore, even if there were a strong adaptive hypothesis for one trait, the finding that this individual genetic variant influences multiple traits pleiotropically would be interesting, whether those additional traits were adaptive or not.

These examples illustrate the dominant convention of post-Panglossian adaptationism, which is that an adaptive explanation is required in order to justify studying a particular trait. Contributions that are interesting and well supported are de-emphasized and contributions that are weakly supported take centre-stage. The overall result is that otherwise excellent bodies of work, that have made significant contributions to our understanding of the mechanistic basis of interesting traits, are forced to make weak claims about a role in adaptation.

CONCLUSIONS

Evolutionary biology would be a richer science if it embraced a post-adaptationist research program and broadened its scope to study all types of change

Evolution is the study of change and history, not a particular kind of change. The exclusive focus on adaptation is like an historian who studies only changes that made society "more advanced," or "wealthier," or "more democratic." This would lead to an impoverished understanding of history. It is the total nature of change, the role of different kinds of change within the overall picture, that makes the study of history and of evolution interesting, nuanced, and rich. By eschewing non-adaptive traits, the adaptationist program ignores many potentially interesting and important study systems, and more than that, restricts itself to a limited perspective on the evolutionary process as a whole.

What is the ideal study of adaptation? Our contention is that this is a counterproductive question. Adaptation cannot be extricated from broader evolutionary processes. Population genetics processes like natural selection and neutral drift will alter

the frequencies of genotypes differently depending on how those genotypes are translated into phenotypes. Linkage, pleiotropy, and epistasis mean that these processes will simultaneously affect adaptive and non-adaptive traits alike; to study adaptation thoroughly without studying non-adaptation is an impossible proposition. Studying the evolution of genotypes or phenotypes requires studying both, and the connections between them.

Rather than asking how to study adaptation on its own, we should seek to improve our study of evolution as a whole. The dichotomy of adaptations versus non-adaptations has been the governing lens for evolutionary study, with the goal of research being to discover what traits and genetic changes were adaptive, to separate them from non-adaptive changes, and to determine what proportion of evolutionary change has been adaptive. But in the modern era, this no longer remains a useful conceptual foundation for studying evolutionary processes. This is not a reproduction of the adaptationist/neutralist argument: we do not propose that the majority of evolutionary changes are either adaptive or non-adaptive. Instead, we argue that the future of evolutionary biology should transcend this dichotomy by answering deeper questions that are masked by the existing adaptationist program. Doing so will require a post-adaptationist program of research, wherein the field examines evolution at every possible level, from the molecular and developmental processes that translate genotypic variation into phenotypic variation, the nature and type of natural selection that operates on phenotypic variation, the demographic fluctuations and random genetic drift that change genotype frequencies between generations, and the ways in which processes like mutation, recombination, and assortment assemble new genotypes that allow novel phenotypes to be realized.

Evolution acts to change population genotype frequencies over time, with the result that gradual phenotypic change can accrue between different evolutionary lineages. The way that processes like natural selection, demographic fluctuations, and neutral drift alter genotype frequencies is a consequence of the ways in which those alternative genotypes result in different functions and phenotypes; in other words, evolutionary change is intimately connected to the macromolecular and developmental mechanisms that relate genotype space to phenotype space (Figure 1). As such, understanding how evolution has produced the diversity of forms and functions seen in the world requires that we understand the functional mechanisms that underlie evolutionary changes.

Studying the complex process that is evolution means studying adaptation and much more. Molecular evolutionary biologists should recognize that meaningful insights will come from studying all evolutionary change, whether it was adaptive or not, and whether or not it is known to have been so.

BRIDGE TO CHAPTER III

Chapter II proposed a research program in molecular evolution that is different from the current paradigm, and in Chapters III and IV I attempted to enact that agenda via a detailed analysis of the evolution of DNA-binding specificity in the biomedically important steroid hormone receptor family of transcription factors. In Chapter III, we identify the primary genetic, biochemical, and biophysical mechanisms that underlay the acquisition of novel DNA-binding specificity in this protein family, and draw conclusions about the general evolutionary process that may have guided this transition.

CHAPTER III

EVOLUTION OF NOVEL DNA SPECIFICITY IN A TRANSCRIPTION FACTOR FAMILY PRODUCED A NEW GENE REGULATORY MODULE

In this chapter, I performed molecular dynamics in silico experiments and conducted extensive biophysical analyses of those results. The purpose of these analyses was to provide an explanation for the biophysical mechanism that evolution utilized in order to realize novel DNA-binding specificity in the steroid hormone receptors. This work is a valuable contribution to the field of molecular evolution because it provides a thorough explanation of the genetic, biochemical, and biophysical mechanisms that produced the evolution of an important novel function. In particular, the biophysical mechanisms that I uncovered with my analysis showed that novel specificity was realized without making new, sequence-specific positive protein-DNA contacts, but rather by specifically excluding the ancestral DNA-targets; it is the first work to have demonstrated this type of physical mechanism for the historical evolution of novel DNA-specificity. Collaborators in the Thornton and Ortlund labs performed all other experiments and analyses. I performed all MD analyses, and interpreted those results in the context of other data with Joe Thornton and Alesia McKeown. Joe and Alesia were the primary writers of this paper, and I provided amendments, particularly for the sections of the work that most directly pertained to my analyses. This work has been published in the journal Cell.

INTRODUCTION

Transcription factor specificity and the evolution of gene regulatory networks

Development, homeostasis, and other complex biological functions depend upon the coordinated expression of networks of genes. Thousands of transcription factors (TFs) in eukaryotes play key regulatory roles in these networks, because their distinct affinities for DNA binding sites, for other proteins, and for small molecules allow them to specifically regulate the expression of unique sets of target genes in response to various hormones, kinases, and other upstream molecular stimuli. Most studies of the evolution of gene regulation have focused on how changes in *cis*-regulatory DNA can bring a new target gene under the influence of an existing TF (Wray, 2007; Carroll, 2008) or on changes in protein-protein interactions among TFs (Brayer et al., 2011; Lynch et al., 2011; Baker et al., 2012). Although TF specificity for DNA can and does evolve (Baker et al., 2011; Sayou et al., 2014), little is known concerning the molecular mechanisms and evolutionary dynamics by which such changes occur. In turn, it remains unclear how distinct gene regulatory modules – defined as a transcription factor, the molecular stimuli that regulate it, and the DNA target sequences it recognizes – emerge during evolution. If TFs are constrained by selection to conserve essential ancestral functions (Stern and Orgogozo, 2009), how can new regulatory modules ever arise? Do specific modules evolve by partitioning the activities of an ancestral TF that is promiscuous in its interactions with DNA targets and molecular stimuli (Sayou et al., 2014), or by acquiring entirely new interactions (Teichmann and Babu, 2004)? What is the genetic architecture of evolutionary transitions in TF specificity, and what kinds of biophysical mechanisms mediate these changes? Answering these questions requires dissecting evolutionary

transitions in TFs' capacity to interact specifically with DNA and molecular stimuli. Ancestral protein reconstruction, combined with detailed studies of protein function and biochemistry, has the potential to accomplish this goal (Harms and Thornton, 2010).

The knowledge gap concerning transcription factor evolution mirrors uncertainty about the physical mechanisms that determine TFs' specificity for their DNA targets. DNA recognition is usually thought to be determined by favorable interactions—especially hydrogen bonds but also van der Waals interactions—between a protein and its preferred DNA sequences (Garvie and Wolberger, 2001; Coulocheri et al., 2007; Rohs et al., 2010). Supporting this view, structural studies have established that positive interactions are typically present in high-affinity complexes of protein and DNA. Specificity, however, is determined by the distribution of affinities across DNA sequences, and it is unclear whether positive interactions sufficiently explain TFs' capacity to discriminate among targets. In principle, negative interactions that reduce affinity to non-target binding sites—such as steric clashes or the presence of unpaired polar atoms in a protein-DNA complex—could also contribute to specificity (von Hippel and Berg, 1986). Evaluating the role of negative interactions in determining specificity, however, requires analyzing not only high-affinity TF/DNA complexes but also poorly bound ones, which are vast in number and difficult to crystallize. We reasoned that by focusing on a major evolutionary transition in DNA specificity during the history of a family of related TFs, we could gain direct insight into the genetic and biophysical factors that cause differences in DNA recognition (Harms and Thornton, 2013).

Steroid receptors coordinate distinct gene regulatory modules

Steroid hormone receptors (SRs), a family of ligand-activated transcription factors, are a model for the evolution of TF specificity. SRs initiate the cascade of classic transcriptional responses to sex and adrenal steroid hormones in vertebrate physiology, reproduction, development, and behavior (Bentley, 1998). These proteins contain a conserved DNA-binding domain (DBD), which directly binds to DNA sequences in the vicinity of the target genes they regulate; they also contain a conserved ligand-binding domain (LBD), which binds hormonal ligands and then attracts coregulatory proteins, leading to ligand-regulated changes in gene expression (Kumar and Chambon, 1988; Beato et al., 1996; Bain et al., 2007). Additional poorly conserved N-terminal and hinge domains mediate other SR activities. All SRs bind as dimers to inverted palindromic DNA sequences consisting of two six-nucleotide half-sites separated by a variable three-nucleotide spacer (Figure 3A – see Appendix B for Figures from this chapter, (Beato, 1989; Umesono and Evans, 1989; Lundback et al., 1993; So et al., 2007; Welboren et al., 2009)).

There are two phylogenetic classes of SRs in vertebrates, which have distinct specificities for both DNA and hormonal ligands: the two SR classes therefore mediate distinct regulatory modules (Figure 3B). One class, the estrogen receptors (ERs), are activated by steroid hormones with aromatized A-rings (Eick et al., 2012) and bind preferentially to estrogen response elements (ERE, a palindrome of AGGTCA) (Welboren et al., 2009). The other class contains the receptors for the non-aromatized steroid hormones, including androgens, progestagens, glucocorticoids, and mineralocorticoids (AR, PR, GR, and MR; (Eick et al., 2012); this class of SR

preferentially binds to steroid response elements (SREs), including palindromes of AGAACA (SRE1) or AGGACA (SRE2) (Chusacultanachai et al., 1999; So et al., 2007). The two classes' DNA specificities are distinct: ERs bind poorly to and do not activate SREs, whereas members of the AR/PR/GR/MR group bind poorly to and do not activate ERE (Zilliacus et al., 1992). Although SRs can and do bind variants of these classic sequences (So et al., 2007; Welboren et al., 2009), the classical ERE and SRE sequences are physiologically relevant and have been the subject of extensive biochemical and structural analysis (Beato, 1989; Luisi et al., 1991; Zilliacus et al., 1992; Lundback et al., 1993; Schwabe et al., 1993).

Understanding the evolution of a TF-mediated regulatory module requires understanding the origin of the TF's interactions with both upstream stimuli and DNA targets. We recently reported on the mechanisms by which the two classes of SRs evolved their distinct specificities for aromatized or nonaromatized hormones (Eick et al., 2012; Harms et al., 2013). Here we use ancestral protein reconstruction (Thornton, 2004; Harms and Thornton, 2010; Harms and Thornton, 2013) to identify the genetic, biochemical, and biophysical mechanisms for the evolution of the distinct DNA specificity in the two classes of SRs. The results, together with previous findings on the evolution of SR ligand specificities, allow us to provide a detailed historical and mechanistic account for the evolution of a new regulatory module.

RESULTS

A discrete evolutionary transition in DNA specificity

To characterize the evolutionary trajectory of DNA recognition in the SRs, we first used ancestral protein reconstruction to infer the DBDs of the ancestral protein from which all SRs descend (AncSR1) and of the ancestor of all ARs, PRs, GRs, and MRs (AncSR2, Figure 3B). Both proteins predate the evolutionary emergence of vertebrates, more than 450 million years ago (Eick et al., 2012). We used maximum likelihood phylogenetics to infer the best-fit evolutionary model and phylogenetic tree for 213 SRs and related nuclear receptors from a wide variety of animal taxa using sequences of both the DBD and LBD (Fig. S1 – see Appendix C for supplemental materials for this chapter). We then inferred the maximum likelihood amino acid sequences of the DBD and the posterior probability distribution of amino acids at each sequence sites at the phylogenetic nodes corresponding to AncSR1 and AncSR2 (Fig. S1A-B). The vast majority of sites in the two sequences were reconstructed with little or no uncertainty; only 3 sites in AncSR2 and 12 in AncSR1 were reconstructed ambiguously, defined as having an alternate state with posterior probability >0.20 (Table S1).

The distinct specificities of extant SRs could have evolved by partitioning the activities of a promiscuous ancestor among descendants or by a discrete switch from ancestral to derived forms of specificity. To distinguish among these possibilities, we synthesized coding sequences for the inferred ancestral DBDs and characterized their functions and physical properties. We focused on the capacity to bind ERE, SRE1, and SRE2, because these classical REs differ only at two bases in the half-site and are completely distinct in their responses to the two classes of SR (Zilliacus et al., 1992).

Using a dual luciferase reporter assay in cultured cells (Figure 3C), we found that AncSR1 had DNA specificity like that of extant ERs, driving strong activation from ERE but exhibiting no expression above background from SREs. AncSR2, in contrast, specifically activated from both SREs but did not activate from ERE. These results are consistent with the strong sequence similarity between AncSR1 and extant ERs and between AncSR2 and the vertebrate ARs, PRs, GRs, and MRs (Figure 3B) and are further corroborated by the pattern of RE specificities across extant members of the SR family tree: because all known descendants of AncSR2 recognize SREs and all other family members and close outgroups bind ERE-like sequences, the most parsimonious expectation by far is SRE-specificity by AncSR2 and ERE-specificity by AncSR1 (Eick and Thornton, 2011), the most parsimonious expectation for AncSR1 is ERE-specificity.

Robustness to uncertainty

To determine whether the inferred functions of AncSR1 and AncSR2 are robust to uncertainty about the ancestral sequences, we synthesized reconstructions of each ancestor that contain every plausible alternate residue. These sequences represent the far edge of the "cloud" of plausible estimates of the true ancestral sequence and are different from the ML sequences at more residues than the expected number of errors in each ML reconstruction (Table S1). These alternative reconstructions therefore provide a conservative test of the robustness of inferences about the ancestral proteins' functions.

We synthesized and assayed these alternate reconstructions and found that the DNA specificities of the alternate reconstructions were nearly identical to those of the ML ancestors (Fig. S3.2A). Moreover, the sequences of extant SRs indicate that none of

the plausible alternative residues in AncSR1 or AncSR2 are sufficient to change DNA specificity (Table S2).

Taken together, these data indicate that the ancestral SR was ERE-specific, and recognition of SREs emerged via a discrete change in specificity during the interval between AncSR1 and AncSR2 (Figure 3B). This transition involved a complete loss of activation from the ancestrally preferred ERE and a wholesale gain of novel activation on SREs.

Thermodynamic basis for evolution of new DNA specificity

We next sought to understand the biochemical basis for this ancient change in DNA recognition by expressing and purifying ancestral proteins and characterizing their thermodynamics of binding to DNA. We used fluorescence polarization to determine the macroscropic binding affinity ($K_{A,mac}$) of each ancestral DBD for labeled DNA probes containing palindromic ERE or SREs. The relative affinities followed those in the activation assays, with AncSR1 showing strongly preferential binding to ERE and AncSR2 preferentially binding SREs (Figure 3D, Table S3). Both bound much more weakly to their non-target REs, with affinity apparently too low to activate reporter transcription. These data indicate that the evolutionary transition in the DBD's DNA specificity was due primarily to changes in DNA-binding affinity for the two classes of binding sites (see (Bain et al., 2012).

The macroscopic affinity of an SR dimer for a palindromic DNA sequence is determined by two components: the half-site binding affinity (K_1) of each monomer for its half-site and the binding cooperativity (ω) between half-sites, defined as the fold

excess of the macroscopic affinity beyond that expected if each monomer binds independently (Figure 3E, (Hard et al., 1990). To estimate these parameters, we performed fluorescence polarization binding experiments with both half-site and palindromic DNA constructs and globally fit the parameters of a two-monomer cooperative binding model to these data.

We found that AncSR1 binds ERE with high half-site affinity and low cooperativity. In contrast, AncSR2 displays much lower half-site affinity but greater cooperativity (Figure 3F-G, Table S3). AncSR2's novel RE specificity therefore evolved through a trade-off in the energetic mechanisms of binding: the protein's direct interactions with DNA became weaker as its specificity changed, but this effect was offset by an increase in cooperativity of binding. As a result, the derived DBD retained macroscopic DNA binding affinity for its favored targets similar to that of its ancestor, but for a new family of DNA sequences. These ancient changes in binding energetics persist to the present: human ERs, like AncSR1, bind DNA with high half-site affinity and low cooperativity, whereas human GR, like AncSR2, displays considerable cooperativity but lower half-site affinity (Hard et al., 1990; Alroy and Freedman, 1992).

Atomic structures of ancestral DBDs

To identify the causes of these evolutionary changes in DNA binding and recognition, we determined the crystal structures of AncSR1-DBD bound to ERE and of AncSR2-DBD bound to SRE1 at 1.5 and 2.7 Å, respectively (Figure 4, Table S4). Although their sequences are only 54% identical, AncSR1 and AncSR2 have very similar conformations (RMSD for protein backbone atoms = 0.82 Å). Each monomer buries a

recognition helix (RH) in the DNA major groove of one half-site and makes additional contacts to the DNA backbone; the monomers contact each other via a dimerization surface composed of an extended loop coordinated by a zinc atom (Luisi et al., 1991; Schwabe and Rhodes, 1991; Schwabe et al., 1993).

Despite these general similarities, there are several differences between the AncSR1 and AncSR2 structures. First, AncSR1's RH makes more hydrogen bonds to DNA than AncSR2 does (Figure 4B). Second, the loop that connects the RH to the dimerization surface is disordered in AncSR1 but adopts a resolved structure in AncSR2. Third, AncSR1 buries ~60% more of its surface area at the DNA interface than AncSR2 does, but AncSR2 buries ~40% more surface in its dimerization interface than AncSR1 (Figure 4C). These differences are consistent with AncSR1's greater affinity for DNA half-sites and AncSR2's greater cooperativity of dimeric binding.

Recognition helix substitutions are necessary but not sufficient for evolution of the derived function

We next sought to identify the evolutionary genetic changes that caused specificity to change between AncSR1 and AncSR2. We focused first on the recognition helix, because it makes the only direct contacts to bases in the DNA half-site. There are ten residues in the RH, but only three changed between AncSR1 and AncSR2—e25G, g26S, and a29V (Figure 5A, with lower and upper cases denoting ancestral and derived states, respectively). All three residues are strictly conserved in the AncSR1-like state in all ERs and the AncSR2-like state in all AR, PR, GR, and MRs (Figure S3A). This region

is also known to play an important role in the specificity of extant SRs (Alroy and Freedman, 1992; Zilliacus et al., 1992).

To test the hypothesis that these three substitutions were the main determinants of the evolutionary change in DNA specificity, we first reversed them to their ancestral state in AncSR2 (generating AncSR2+rh). As predicted, these changes are sufficient to restore the ancestral preference for ERE over SREs in a luciferase assay (Figure 5B). They do so by restoring the DBD's capacity to activate transcription from ERE while dramatically decreasing SRE activation.

We also determined the crystal structure of AncSR2+rh on ERE at 2.2 Å and found that reversing these three substitutions largely restores the ancestral protein-DNA interface (Figure S2B-C). The interactions of AncSR2+rh with ERE-specific nucleotides are almost identical to those made by AncSR1. Only a few minor differences are apparent in non-specific interactions to the DNA backbone and to nucleotides outside of the half-sites, presumably because of differences in crystallization conditions or protein sequence outside the RH. Taken together, these data indicate that the RH substitutions were the primary determinants of the evolutionary change in half-site specificity from ERE to SREs.

To determine whether the RH substitutions were also sufficient causes of the shift in specificity, we introduced the derived RH states into AncSR1 (Figure 5B). Surprisingly, activation was entirely abolished on all REs tested (Figure 5B). This result is robust to uncertainty about the ancestral sequence: introducing the RH substitutions – which are inferred unambiguously – into the reconstruction of AncSR1 containing all plausible alternative amino acids caused the same effect (Figure S2A). The lack of

activity is not due to differences in protein expression between AncSR1 and AncSR1+RH (Figure S2D), implying that the RH substitutions strongly compromise DBD function when introduced into AncSR1, rather than depleting protein in the cell. The derived RH states, however, are conserved in AncSR2 and all its descendants, all of which activate transcription. These data indicate that additional epistatic substitutions, which permitted the DBD to tolerate the RH substitutions must have also occurred during the AncSR1/AncSR2 interval.

Permissive substitutions outside the DNA interface were required for the evolution of new specificity

To identify these permissive substitutions, we divided the 35 other substitutions that occurred during the AncSR1/AncSR2 interval into 8 groups based on contiguity in the linear sequence and tertiary structure (Figure S3A). We tested the hypotheses that each group contained permissive substitutions by reverting it to the ancestral state in AncSR2: reversing a permissive substitution in the context of the derived RH should compromise function. We found that just three groups, containing a total of 16 amino acid replacements, significantly reduced activation when reversed, indicating that the derived states at these sites are necessary for full DBD function and therefore contribute to the permissive effect (Figure S3B, Table S5).

Using a series of forward and reverse genetic experiments testing the effects of the individual mutations within these groups, we ruled out a role for several substitutions and narrowed the set of permissive changes to 11 historical substitutions (11P) distributed among the three structural groups (Figure S4A-C, Table S5). When the derived residues

at these sites are introduced into the nonfunctional AncSR1+RH, they rescue activation and recapitulate the evolution of the derived DNA specificity (Figure 5 A-B). Their permissive effect is robust to uncertainty about the precise sequence of AncSR1 (Figure S2A). All three groups are necessary for the full permissive effect (Figure S4D, Table S5).

These substitutions are permissive in that they are required for the protein to tolerate the derived RH, but when introduced into AncSR1 they have no effect on specificity; rather, they enhance activation non-specifically on ERE and SREs alike (Figure 5B). Taken together, these data indicate that a large number of permissive mutations, which did not themselves affect specificity, were required for the specificity-switching substitutions to be tolerated.

The effect of these ancient permissive mutations persists to the present. We found that introducing the derived RH states from the human GR into human ERα results in a non-functional DBD, just as it did in AncSR1, consistent with the fact that the lineage leading to ERs branches from the rest of the SR phylogeny before AncSR2's permissive mutations occurred (Fig. S2E). Adding the 11P into the nonfunctional ERα+RH protein, however, rescued activation and yielded a DBD with preference for SREs. Conversely, the ancestral RH states can be introduced into human GR, where they dramatically increase activation on ERE, just as they do in AncSR2 (Fig. S2E; (Zilliacus et al., 1991; Alroy and Freedman, 1992). Taken together, these results indicate that the ancient RH and permissive substitutions provide a sufficient genetic explanation for the evolution of the distinct DNA specificities of the two major classes of extant SRs.

Evolution of specificity by negative protein-DNA interactions

Having identified the genetic changes that caused the evolution of AncSR2's new specificity, we sought to understand the biophysical mechanisms by which they did so. We first measured the effect of the RH substitutions on the energetics of sequence-specific DNA binding. We found that they improve the DBD's macroscopic binding preference for SREs by a factor of 30,000; this effect is caused by a 2,000-fold reduction in affinity for ERE and a 15-fold increase in SRE affinity (Figure 5C, Table S3). These effects are entirely attributable to changes in half-site binding affinity, as the RH substitutions do not affect cooperativity (Figure 5C).

To understand the atom-level mechanisms for the effects of the RH mutations, we compared crystal structures of the ancestral DBDs containing the ancestral or derived RH amino acids in complex with both ERE and SRE1; we also performed molecular dynamics (MD) simulations of AncSR1, AncSR1+RH, and AncSR2, each bound to ERE, SRE1 and SRE2. In principle, the evolutionary change in DNA specificity could have been caused by changes in positive interactions – hydrogen bonds or van der Waals attractions between protein and DNA atoms – or in negative interactions, such as electrostatic or steric clashes. If the change in specificity were solely due to changes in positive interactions, then the RH substitutions would reduce favorable interactions with ERE and increase favorable interactions with SREs.

Contrary to this prediction, we found that the RH substitutions primarily change negative interactions between the DBD and DNA binding sites, relieving clashes with SRE and establishing new ones with ERE. The ancestral RH does form more hydrogen bonds on ERE than on SREs, and the RH substitutions reduce the number of hydrogen

bonds to ERE (Figure 6A, Figure S5E); these observations are consistent with the view that positive interactions are the primary determinants of specificity. By removing hydrogen bond acceptors, however, these substitutions also establish negative polar interactions, leaving polar groups on ERE-specific bases unpaired and leading to penetration of transient solvent molecules into the protein-DNA interface (Figure S5A-D). The effect of these negative interactions is expected to be much stronger than the loss of the positive interactions: eliminating a protein-DNA hydrogen bond would reduce binding affinity only slightly, because the same number of total hydrogen bonds would form whether or not the protein and DNA are bound to each other or free in solvent. In contrast, leaving an unpaired polar atom at the protein-DNA interface results in more hydrogen bonds in the unbound than the bound state, leading to a much larger difference in energy between the bound and unbound states and a much more dramatic reduction in affinity (von Hippel and Berg, 1986).

The improvement in SRE binding also cannot be explained by an increase in SRE-specific positive interactions. The RH substitutions do not increase the total number of hydrogen bonds on SRE1 and actually reduce the number of hydrogen bonds on SRE2 (Figure 6A). They do so by eliminating or weakening hydrogen bonds formed by the ancestral protein to SREs without forming enough new hydrogen bonds to compensate. Although the derived RH does establish one novel hydrogen bond from derived residue Ser26 to the DNA backbone, this interaction actually forms more frequently on ERE than on SREs (Figure S5E). Overall, AncSR1+RH (like AncSR2) forms equal numbers of hydrogen bonds with ERE and SREs, indicating that hydrogen bonding does not explain the evolution of preference for SREs. As for van der Waals interactions, the RH

substitutions reduce the efficiency of packing on ERE, but they do not improve packing on SREs (Figure 6B). Taken together, these results indicate that changes in positive interactions—hydrogen bonds and van der Waals forces—do not explain AncSR2's increase in affinity or its preference for SREs.

If new SRE-specific positive interactions do not explain the increase in affinity for SREs caused by the RH substitutions, what mechanisms do mediate this effect? We found that the RH substitutions improve SRE affinity by relieving SRE-specific steric and electrostatic clashes with the ancestral RH. Crystal structures and MD simulations both show that the long sidechain of glu25 sterically clashes with T-4 and T-3 of SREs; these bases contain large methyl groups that protrude into the DNA major groove of SREs, but are absent from the corresponding bases in ERE (Figure 6C, Figure S6A-E). As a result of this clash, glu25 is forced to move away from the major groove of SREs and, in turn, to displace the conserved residue Lys28, which in high-affinity complexes forms hydrogen bonds to DNA bases that do not vary among REs (Figure 6D-E). As a result, Lys28 forms fewer hydrogen bonds on SREs compared to ERE (Figure 6F). Additionally, by pushing the negatively charged glu25 away from the bases in the center of the major groove, the SRE-protein interface is left with numerous unpaired hydrogen bond donors and acceptors, leading to water penetration into the interface with SREs (Figure S6F-H). The RH substitutions ameliorate this clash by replacing glu25 with the much smaller Gly, thus relieving the negative effect of the glu on SRE binding.

To test the hypothesis that removing glu25 improves SRE recognition by relieving negative interactions, we used site-directed mutagenesis to introduce e25G alone into AncSR1 containing the permissive mutations. We found, as predicted, that

SRE affinity and activation were enhanced, despite the fact that Gly25 makes no apparent favorable interactions with SREs (Figure 6G-H).

The other two RH substitutions preferentially reduce recognition of ERE, apparently by establishing additional ERE-specific negative interactions. When g26S and a29V are added to e25G, yielding the derived RH genotype, they reduce affinity and activation on all REs, but do so much more severely on ERE than SREs (Figure 6G-H). The mechanism for this effect is not obvious in the structures or simulations (Figure S6I-J), but it does not involve eliminating hydrogen bonds or van der Waals interactions with ERE: neither ancestral amino acid forms hydrogen bonds to ERE (Figure 6F), and they do not pack more efficiently against ERE than the derived amino acids do (Figure S6K).

Taken together, these data indicate that differences in sequence-specific positive interactions do not explain the switch in specificity caused by the RH substitutions. Rather, negative interactions that interfered with SRE binding in the ancestral state were lost, and new negative interactions that impair binding to ERE were gained (Figure 6I). The result was to transform the DBD's ancestral ERE-preference into AncSR2's derived SRE-preference. A secondary effect was to reduce affinity for the preferred DNA sequence and thus to require permissive substitutions for activation to be maintained.

Permissive substitutions non-specifically improve affinity for both the derived and ancestral REs

Permissive substitutions are often thought to act by increasing thermodynamic stability, allowing the protein to tolerate mutations that confer new functions but compromise stability (Bershtein et al., 2006; Gong et al., 2013). Using reversible

chemical denaturation, however, we found that the 11P substitutions do not increase stability, and the RH substitutions do not decrease stability (Figure 7A-B).

Because the RH substitutions radically reduce affinity for ERE and only weakly increase affinity for SREs – yielding a low-affinity receptor for both kinds of element – we hypothesized that the permissive substitutions might offset these effects by increasing affinity in a non-sequence specific manner. As predicted, introducing 11P into the ancestral background increases macroscopic binding affinity by increasing both cooperativity and half-site affinity on all REs (Figure 5C), indicating a tradeoff in the energetics of binding between the permissive and specificity-switching substitutions during evolution.

The crystal structures suggest that the permissive substitutions cause these effects by enhancing nonspecific protein-protein interactions at the dimerization interface and non-specific interactions with the DNA backbone and minor groove. Two of the permissive substitutions (v39H and v42L) may facilitate dimer formation, because they are located on the loop that links the RH to the dimerization surface (Figure 7A). In AncSR1, as in human ERα, the loop is unresolved, but it is fully resolved in complexes containing the derived state at these residues, including AncSR2, AncSR2+rh, and the human GR (Luisi et al., 1991). Using analytical ultracentrifugation, we found that the permissive substitutions do not measurably increase DBD dimerization in solution (Fig. 5C-D). We therefore propose that v39H and v42L contribute to cooperativity by stabilizing the dimerization interface in a DNA-dependent manner. Consistent with this view, this loop has been shown in extant SRs to undergo functionally relevant conformational changes when DNA is bound (Wikstrom et al., 1999; Meijsing et al.,

2009; Hopkins et al., 2012; Watson et al., 2013). The remaining permissive substitutions may enhance non-specific DNA binding because they are involved in contacts to the DNA backbone or other base-nonspecific interactions. Substitution w22L is adjacent to several backbone-contacting residues (Figure 7A), and the other permissive substitutions are in the C-terminal tail; although unresolved in our ancestral crystal structures, this region binds directly to the DNA backbone or minor groove just outside the core RE in other nuclear receptors (Nelson et al., 1999; Roemer et al., 2006; Meijsing et al., 2009; Helsen et al., 2012).

Taken together, our findings indicate that numerous permissive substitutions, which increased nonspecific affinity, were necessary for the affinity-reducing effects of the RH mutations to be tolerated. The evolving DBD therefore traversed sequence space extensively without changing its specificity, reaching regions relatively distant from AncSR1, before the transition to a new function via the RH substitutions could be completed. Selection for the derived specificity could not have driven this exploration; either neutral chance processes (such as drift and linkage) or selection for functions unrelated to specificity must therefore have played crucial roles in the evolution of AncSR2's DNA recognition mechanism.

DISCUSSION

Evolution of a new gene regulatory module

These results, together with our previous work on the evolution of the ancestral ligand binding domain, elucidate the mechanisms by which the distinct regulatory modules mediated by the two classes of extant SRs evolved from an ancestral module

mediated by a single TF. We recently reported that AncSR1's LBD also had ER-like functions, responding specifically to estrogens; after duplication of AncSR1, AncSR2 lost estrogen sensitivity entirely and gained activation by nonaromatized steroids (Eick et al., 2012; Harms et al., 2013); during this period, androgens and progestagens were already produced as intermediates in the synthesis of estrogens (Eick and Thornton, 2011). Our present findings therefore establish that during the interval after the duplication of AncSR1, both AncSR2's LBD and DBD both evolved entirely new specificities for upstream stimuli and downstream DNA targets (Figure 8A). The other protein lineage produced by this duplication, which led to the present-day estrogen receptors, maintained the specificity of the ancestral signaling module essentially unchanged for hundreds of millions of years.

By evolving distinctly new specificities in both domains after gene duplication, a new regulatory module was established without interfering with the functional specificity of the ancestral module. If one domain of AncSR2 had retained the ancestral specificity while the other evolved new interactions, the information conveyed by the ancestral signaling system would have been compromised by noise: ancestral targets would have been activated by additional stimuli, or the ancestral stimuli would have activated additional targets (Figure 8B). A similar effect would have ensued if the DBD and/or LBD became promiscuous (Figure 8C-D). Because the new specificities for hormone and DNA evolved during the same phylogenetic interval, we cannot determine which appeared first. It is possible that a promiscuous DBD arose as an evolutionary intermediate during the transition between the distinct RE-specificities of AncSR1 and AncSR2. If it did, however, it did so transiently, was abolished relatively rapidly, and left

no promiscuous descendants that persist in present-day species. Thus, the distinct AncSR2-mediated signaling module arose by establishing new functional connections and, just as importantly, by actively erasing the ancestral connections.

In both domains, just a few key mutations – three in the DBD and two in the LBD (Harms et al., 2013) – changed the protein's binding preferences by many orders of magnitude. These substitutions dramatically impaired interactions with the ancestral partner and, to a lesser extent, improved binding of the ancestral TF to the derived partner. In both domains, the biophysical mechanisms for this transition involved changes in negative determinants of specificity: the key mutations introduced unfavorable steric or electrostatic clashes with estrogens or ERE and removed clashes that in the ancestral state impaired binding to nonaromatized steroids and SREs (Harms et al., 2013). These data indicate that negative determinants of specificity – mechanisms that actively prevent binding to "non-target" partners – played key roles in the evolution of the new AncSR2-mediated regulatory module (Figure 8E).

Negative determinants of specificity: mutational constraints on TF evolution

AncSR2's new DNA specificity was conferred by a complex set of changes: three RH-mediated mutations that changed exclusionary interactions and a large number of permissive mutations that offset the affinity-reducing effects of the specificity-switching mutations. Why did evolution not utilize a simpler mechanism to cause the shift in specificity, such as gains and losses of positive interactions? We propose that differences in the abundance of mutational opportunities to establish negative vs. positive

mechanisms of specificity determined the evolutionary trajectory by which AncSR2's new mode of DNA recognition evolved.

As a protein evolves, it drifts through a "neutral network" of neighboring genotypes with similar functional outputs; it may cross into a network that encodes different functions, if one is accessible by mutation and compatible with selective constraints (Smith, 1970; Wagner, 2008). Biophysical considerations suggest that there may be few mutational opportunities to increase affinity in a sequence-specific fashion. Establishing a new sequence-specific positive interaction in the complex, heterogeneous interface with DNA would require introducing a side chain of fairly precise length, angle, volume, polarity, and charge to interact favorably with a feature of DNA that is unique to the target sequence, all without disrupting other aspects of the protein-DNA complex. In contrast, the requirements to establish a negative interaction via a steric or electrostatic clash are likely to be considerably less precise, as are those to abolish a hydrogen bond and thereby leave unpaired polar atoms in an interface. Thus, just as the integrated architecture of protein folds makes mutations that stabilize proteins more rare than those that destabilize them (Bloom et al., 2006), the biophysical architecture of protein-DNA interactions should make mutations that shift specificity by establishing new sequence-specific positive interactions much more rare than those that do so by reducing affinity for non-target sequences.

Evolutionary trajectories that utilize predominantly negative mechanisms to achieve specificity – like those during the evolution of AncSR2's DBD and LBD – should therefore be more likely to be realized than those that change specificity by establishing new, sequence-specific positive interactions. Consistent with this view,

directed evolution experiments that select for specific binding to a new DNA target typically reduce affinity (Rockah-Shmuel and Tawfik, 2012). Further, studies that select for binding without selecting for specificity usually increase affinity in a non-specific fashion (Cohen et al., 2004), indicating that increased affinity often evolves because of non-specific positive interactions, but specificity is realized largely through sequence-specific negative interactions.

Although they are more numerous, mutations that shift specificity by negative, exclusionary interactions would be eliminated by natural selection if they were to reduce affinity to a level below that required for target gene activation, as the RH substitutions do if introduced directly into AncSR1. The historical permissive mutations, by increasing cooperativity and nonspecific affinity, moved the evolving AncSR2 into a region of its neutral network in which the historical specificity-inducing mutations could be tolerated. This evolutionary dynamic is similar to that observed for permissive mutations that increase protein stability and therefore allow destabilizing mutations that confer new functions to be tolerated (Bloom et al., 2006). In the present case, however, the critical parameter is the binding affinity of a protein-DNA complex, rather than the stability of the protein fold. Because macroscopic binding affinity is determined by both half-site affinity and cooperativity, permissive mutations that enhance either parameter – or both, as is the case for the evolution of the SR DBD—could facilitate the evolution of new TF specificity and the rewiring of transcriptional circuits (Tuch et al., 2008; Li and Johnson, 2010).

Because of the limitations imposed by mutational opportunities and purifying selection, AncSR2 evolved distinct, high-affinity DNA binding using a mechanism that is

not the simplest or most elegant form imaginable for a TF-DNA complex. But it was the mechanism that happened to be available, given AncSR2's chance wanderings through sequence space and the constraints imposed by the physical architecture of SR proteins, DNA, and the interaction between them. That ancient, awkward mechanism persists to the present.

EXPERIMENTAL PROCEDURES

Ancestral sequences and posterior probability distributions for AncSR1 and AncSR2 DBDs were inferred using maximum-likelihood phylogenetics from an alignment of 213 peptide sequences of extant steroid and related receptors, the maximum likelihood gene family phylogeny, and the best-fit evolutionary model (JTT+G) (Eick et al., 2012). Complementary DNAs coding for these peptides were synthesized and subcloned and expressed as fusion constructs with the NFkB-activation domain in CV-1 cell line. Activation was measured using a dual luciferase assay in which firefly luciferase expression was driven by four copies of ERE or SRE. Variant proteins were generated using Quikchange mutagenesis and verified by sequencing. To measure the energetics of binding, tagged DBDs were expressed in *E. coli* and purified by affinity chromatography; we measured the change in fluorescence polarization of 6-FAM labeled double-stranded DNA oligos as protein concentration increased. Oligos containing a single half-site or a full palindromic element were assayed, and the data were globally fit to a two-site model with a cooperativity parameter to determine the half-site affinity and the cooperativity coefficient (the fold-increase in the K_A of dimeric binding compared to the expected value if the monomers bind independently (Hard et al., 1990)). To measure

protein stability we used circular dichroism to measure the reversible loss of secondary structure in increasing guanidinium chloride. Protein dimerization was assayed by sedimentation velocity analytical centrifugation. For crystallography, purified DBDs were crystallized in complex with palindromic DNA oligos and diffracted at the Advanced Photon Source; structures were determined using molecular replacement. Atomic coordinates were deposited as AncSR1:ERE (PDB 4OLN, 1.5 Å), AncSR2:SRE1 (4OOR, 2.7 Å), AncSR2+rh:ERE (4OND, 2.2 Å), and AncSR2+rh:SRE1, (4OV7, 2.4 Å). Molecular interactions were characterized with molecular dynamics simulations using Gromacs, TIP3P waters and AMBER FF03 parameters for protein and DNA. For each condition, three replicate 50 ns simulations were run, starting from crystal structures of ancestral proteins; historical mutations were introduced and energy minimized before MD simulation. For details, see Extended Experimental Procedures in Supplemental Information.

BRIDGE TO CHAPTER IV

In Chapter III, we uncover the genetic, biochemical, and biophysical mechanisms for novel DNA-binding specificity in the steroid receptors. In Chapter IV, we build upon that work by examining the transition in the finest detail possible: By measuring the binding function of all the genetic variants that directly separate the ancestral and derived regulatory modules, which are composed of the steroid receptor protein and its target response elements. By examining this transition at a finer scale of detail, we show that evolution must have proceeded via an intermediate protein with significant different

function from both the ancestral and derived states, suggesting some important general principles that may guide the evolution of new functions for molecular complexes.

CHAPTER IV

OF SPACE AND SPECIFICITY: MAPPING A FUNCTIONAL TRANSITION IN DNA-BINDING ACROSS THE STEROID RECEPTOR TRANSCRIPTION FACTOR FAMILY

In this chapter, I designed and implemented a statistical method to describe the genetic causes of functional variation for the binding affinity between steroid receptor and a library of possible response element targets. This analysis was applied to a large volume of binding affinity measurements that were made by Alesia McKeown. This statistical approach allowed us to describe important genetic interactions due to epistasis, as well as to identify the major first-order genetic drivers of binding function in the steroid receptor – response element system. This approach was critical for our description of the intervening genetic space that separates ancestral and derived functions across the evolutionary transition. Additionally, I performed molecular dynamics in silico experiments and analyses in order to complement this detailed genetic analysis of function. This allowed us to identify some of the important biophysical determinants of binding function in this system. This work provides a novel method of analyzing functional data for a library of alternate genotypes, and it showed the functions of a combinatorially complete set of genotypes that separate the ancestral and derived functions of this regulatory module. This allowed us to describe the possible evolutionary pathways that were available through that intervening genetic space under a few key evolutionary scenarios, and to infer some potential general principles that may have guided evolution as it realized functional novelty in this system. Alesia and I collaborated

extensively during all phases of this project, to such an extent that we are submitting the paper as co-first authors.

"The virtue of maps, they show what can be done with limited space, they foresee that everything can happen therein." -Jose Saramago

INTRODUCTION

Mapping functional sequence space using molecular cartography

Evolutionary biologists study how the evolutionary process changed genotypes and phenotypes, and thus led to the diverse forms and functions in the biological world. One aspect of the relationship between changing genotypes and the functions they encode is described by the classic metaphor of the "sequence space" (Smith, 1970), where the set of genotypes available to an evolving system is defined as those that are connected by single genetic mutations. Functional characterization of this sequence space requires a sort of molecular cartography, in which the tools of molecular biology and biochemistry are used to measure the functions for all the genotypes that were available to evolution. This molecular mapping reveals the connectivity of functional sequence space, where genotypes that encode viable functions are connected by single nucleotide changes, and uncovers potential mutational paths that result in the conservation of an ancestral function or lead to functional novelty (Smith, 1970; Stadler et al., 2001; Wagner, 2008).

Mapping the functions of genotypes across the sequence space that connects distinct functions results in the resolution of the evolutionary process that caused novel functions to arise. What sequence changes affected the function? What was the direction and magnitude of their effects? What were the characteristics of the intermediate

genotypes? To what extent are the functions across a given sequence space, and thus the pathways that traverse it, determined by epistatic interactions between genetic states at different sites (Fisher, 1918; Phillips, 2008)? Answering these questions is a necessary first step to understanding how specific biological systems evolved to their current form.

What functions existed across the sequence space of an evolving transcriptional module, and what are the physical interactions that caused them?

Many biological processes depend on the coordination of gene transcriptional modules, which we define as consisting of a *trans*-acting transcription factor (TF) and the *cis*-acting DNA response elements (REs) with which each TF interacts. The binding interaction between these two components of the regulatory module results in the targeted recruitment of additional cellular machinery and ultimately leads to the activation or repression of transcription for a nearby gene. Despite the central importance of these modules in development and homeostasis, the evolutionary processes and mechanisms by which they evolve are not clearly understood.

Some studies have attempted to characterize the relative contributions of *cis*- and *trans*-acting diversification in the evolution of regulatory networks. They have found that divergence in both *cis*-acting (Gompel et al., 2005) and *trans*-acting factors (Teichmann et al., 2010) can contribute to regulatory network evolution, though *cis*-acting diversification is more common (Carroll, 2005; Carroll, 2008; Wittkopp et al., 2008). However, in many cases (Landry et al., 2005), coincident changes in both *cis*- and *trans*-acting factors have maintained an ancestral connection, leading to overall conservation of regulatory function even when the module's components have undergone

diversification (Barriere et al., 2012). Therefore, characterizing the sequence space for an evolving transcriptional module should explicitly consider both interacting genetic loci: the TF, which can evolve by single step amino acid changes, and its set of high-affinity REs, which can also evolve by single nucleotide mutations. The functions across the sequence space for both of these loci are intimately related; substitutions in the protein may change the set of RE sequences with which it can have a regulatory interaction, and vice versa. Given the interconnected relationships of these molecular components, the evolvability of the system can only be determined by characterizing how genetic changes in the TF alter the high-affinity RE sequence space and how changes in the RE alters the accessible TF sequence space.

Mapping the functional sequence space across an evolutionary transition for a transcriptional module should therefore involve studying the mutations that were available to both the transcription factor and the RE. This would result in the resolution of key questions regarding transcriptional module evolution. Are there mutational pathways available to the transcription factor that results in the recognition of novel RE sequences, thereby contributing to transcriptional module diversification? What mutations are available to the RE that would result in conservation of a high-affinity interaction, and how are these dependent on transcription factor specificity? Are there mutational pathways that exist in the module's high-affinity network in which genetic changes in the *trans*-acting TF are compensated by changes in the *cis*-acting RE, thereby allowing both to change without ever compromising the module's ability to bind a critical gene target with high-affinity? To what extent is the evolution of novel function in the module dependent on promiscuous intermediates? Answering these questions would lend

insight into how changes in both the TF and the RE contribute to transcriptional module evolution and how each impact the module's evolvability.

Another goal in studying the sequence space across an evolutionary transition is to elucidate the biophysical interactions that translate different sets of genotypes into different functions. Based on the biophysical architecture of protein-DNA interacting systems, is it possible to describe the sequence space as a function of the same types of biophysical interactions across all RE sequences? If so, what are the physical determinants of TF-DNA interactions and how do they evolve to cause a novel binding function? Identifying these physical determinants would result in a mechanistic description of a regulatory module's evolving function, and could help us understand how this biophysical architecture gave rise to the system's available sequence space.

Steroid receptors are components of transcriptional modules and have evolved divergent specificities for distinct classes of DNA response elements

Steroid receptors (SRs) are an ideal model system for exploring the sequence space of an evolving transcriptional module. SRs are a class of ligand-activated transcription factors that regulate the physiological response to sex and adrenal hormones (Bentley, 1998). All SRs possess a highly conserved DNA-binding domain that binds cooperatively as dimers to a palindromic response element (RE) that consists of two six-nucleotide half-sites separated by a variable three-nucleotide linker (Bain et al., 2007). SRs group into two well-defined phylogenetic clades, each characterized by a distinct DNA-binding specificity (Figure 1A); estrogen receptors (ERs) bind to ERE, a palindrome of AGGTCA, while progestagen, androgen, mineralocorticoid and

glucocorticoid receptors (PAMGRs) bind to SREs, a palindrome of AGAACA (SRE1) and AGGACA (SRE2) (Beato, 1989; Umesono and Evans, 1989; Lundback et al., 1993; Welboren et al., 2009). Importantly, these REs differ only within the two middle positions in the half-site.

We previously reported on the historical mechanisms by which modern day SRs evolved their distinct DNA-binding specificities (McKeown et al., 2014). Using ancestral protein reconstruction, we resurrected the ancestor of all SRs (AncSR1) and the ancestor of all PAMGRs (AncSR2) and assayed their binding preference for ERE and SREs (Figure 1A). We found that AncSR1 was ER-like, preferentially binding to ERE, and that AncSR2 was PAMGR-like and preferentially bound to SREs. Of the 38 differences that occurred on the interval between AncSR1 and AncSR2, three substitutions were necessary and sufficient to cause a change in DNA-binding preference. These three substitutions (glu25GLY, gly26SER, ala29VAL; ancestral and derived states denoted by lower and upper case letters, respectively) occur in the 10-residue recognition helix (RH) that inserts into the DNA major groove and makes numerous polar contacts to DNA (Figure 1B). When introduced into the ancestral background, these three substitutions are sufficient to change the protein's specificity from preferring ERE to preferring SREs. The presence and effect of these three substitutions persist in modern day SR proteins.

To examine the contribution of all the sequence changes that occurred during this functional transition in DNA_binding specificity, we considered all genetic combinations of the three RH substitutions within the protein and in the middle two positions in the RE half-site. We chose to vary the two middle positions in the RE half-site because they are

the only nucleotides that differ between the two classes of REs and are therefore the most relevant for this transition. We aimed to functionally characterize the combinatorial set of RH protein intermediates existing within the sequence space along the transition from ERE-specificity to SRE-specificity, and to identify the physical interactions that produced these differentiated functions.

RESULTS

The derived RH changes DNA preference by exploiting a latent binding function

To describe the functional transition in binding affinity and specificity, we first characterized the binding functions of AncSR1 and AncSR1+RH. To determine binding preference, we rank-ordered the binding affinities for AncSR1 and AncSR1+RH to all 16 alternate REs and identified the highest affinity sequence (Figure 1C). As predicted, AncSR1 binds with highest affinity to ERE and AncSR1+RH binds with highest affinity to SREs. Relative to AncSR1's affinity for ERE, AncSR1+RH binds with much lower affinity to its preferred sequences. In accordance with our previous work (McKeown et al., 2014), these data indicate that the derived RH caused a switch in DNA-binding preference by greatly decreasing single-site affinity for the ancestrally preferred sequence without increasing affinity for SREs by an equivalent energy. This resulted in a protein with a novel DNA preference, but with much lower affinity for its preferred sequence.

In the rank-ordered affinity plots, ERE, SRE1 and SRE2 are all among the top 4 highest affinity REs for both AncSR1 and AncSR1+RH while the identity of the low-affinity sequences remains consistent between the ancestral and derived proteins (Figure 1C). These results indicate that evolution of new binding preference was due to changes

in the interactions with sequences that were historically bound with moderate affinity and did not require drastic changes in the interactions with other low-affinity sequences. These results imply that the derived preference for SREs arose via the exploitation of the ancestral protein's latent binding affinity for the derived proteins RE targets.

Despite this relatively simple re-ordering of the top four ancestral binding targets, the shift in binding energetics caused AncSR1 and AncSR1+RH to have very different occupancies across these 16 REs (Figure 1D). To determine the relative occupancy across different REs, we calculated the expected occupancy across all 16 REs in a competitive binding environment in which all REs are present in equal frequency. AncSR1's occupancy is dominated by REs with a G and T in positions 3 and 4, respectively, indicating its extremely strong preference and high specificity for ERE. AncSR1+RH prefers SRE nucleotides A or G in positions 3 and A in position 4. However, AncSR1+RH is much less specific, and has appreciable occupancies for REs with all other nucleotide states at both positions. Together, these data indicate that the derived RH caused a change in DNA-binding preference and a reduction in specificity, resulting in a protein that preferred a new sequence, but displayed far greater promiscuity.

Intermediate proteins were either promiscuous or low affinity

We next wanted to determine how each individual RH substitution contributed to a change in DNA preference and specificity. To investigate these contributions, we measured binding affinity to all 16 REs by all 6 intermediate protein sequences between AncSR1 and AncSR1+RH (Figure 2A). By comparing the affinity distributions for each

protein genotype, we were able to determine the individual effects of each amino acid substitution as well as the epistatic interactions between them.

To assess how the historical substitutions in the RH impacted the protein's DNA-binding function, we implemented a linear modeling approach to identify the genetic determinants that predict the free energy of binding. We generated two alternative linear models that use dependent variables that reflect the variation of the genotypes across the recognition helix. These dependent variables include both first-order effects of the individual independent sites and second-order effects that represent all two-way combinations. We applied two models to the data to minimize over-fitting and to minimize the potential for overestimating statistical effects as a result of type II error. The first model is constructed by optimizing the Akaike Information Criterion (AIC) score for a model that includes potential first- and second-order terms (for more detail see Materials and Methods). This approach aims to avoid overfitting error variation in the data by including extraneous statistical terms. The second linear model is a global model that includes all the terms identified with the AIC-optimized method, as well as any additional terms necessary to completely describe the total range of genetic variation. This ensures that statistical terms will not be excluded as a result of type II error, which can lead to the overestimation of the retained statistical terms. In the second model, all of these terms are optimized and retained regardless of whether they are found to be statistically significant (discussed further in Materials and Methods). These alternative models are designed to minimize over-fitting (the AIC-optimized model), and to minimize the potential of overestimating statistical effects as a result of type II error (the global model). The sign of the significant statistical effects were consistent in both

models (Table S7), and the effects that were significant in both models will be the focus of our discussion.

Considering the effects of the substitutions in the RH, we uncovered three first-order terms and two second-order epistatic terms (Figure 2B). The first-order terms represent the general effect of each substitution on binding affinity averaged across all 16 REs and all protein genotype backgrounds. We observed that glu25GLY increased binding affinity to all 16 REs, while gly26SER and ala29VAL decreased binding affinity to all 16 REs (Figure 2A). We also identified two second-order epistatic terms, which both acted to reduce average binding affinity beyond that expected for the average effects of each substitution individually (Figure 2B). These included an interaction between glu25 and gly26, as well as between SER26 and ala29. These results imply that the distribution of affinities across the space that separated the ancestral and derived transcriptional modules was shaped both by the individual positive and negative effects of protein substitutions as well as the interactions between them.

The effects of these first-order and epistatic terms result in protein intermediates across this transition that either bind all RE sequences with low-affinity or are promiscuous (Figure 2A). We defined low-affinity proteins as those that do not bind any RE sequences with an affinity that is above the average affinity across all proteins and REs. Three of the six intermediate protein genotypes (glu-gly-VAL, glu-SER-ala and glu-SER-VAL) were low-affinity proteins that did not bind with high affinity to any of the 16 REs (Figure 2A). Two intermediate protein genotypes (GLY-gly-ala and GLY-gly-VAL) were extremely promiscuous, binding with high-affinity to all or nearly all RE sequences. The remaining intermediate, GLY-SER-ala, was less promiscuous, but still

bound with high affinity to both ERE and SREs as well as one additional off-target RE. When mapped onto protein sequence space, these observations imply that the evolving protein was forced to sample either a low-affinity intermediate or promiscuous intermediate as it evolved its derived function (Figure 2C).

Ancestral and derived proteins have different genetic determinants of high-affinity in the RE

We next wanted to determine how the RH substitutions changed the protein's RE specificity. To do so, we used the same linear modeling approach to estimate the statistical effects of the state at positions 3 and 4 in the RE on binding affinity for each protein genotype. This analysis identified genetic states that were both positive determinants (i.e. genetic states that caused higher binding affinity) and negative determinants (i.e. genetic states that caused reduced binding affinity) of binding function. When we examine the distribution of affinities across all REs, we see that the positive determinants reflect the set of most highly occupied RE sequences for each protein genotype. Conversely, the significant negative determinants of affinity reflect the REs that remained in the tail of the distribution of affinities for each protein, thereby explaining variation between "bad" and "worse" binding affinities. We therefore chose to discuss the positive determinants because they are the genetic states that describe the set of highest-affinity RE targets. By applying this statistical framework to describe the map of high-affinity REs for each protein genotype, we were able to identify the nucleotide states that were generally preferred by each protein genotype, as well as any non-additive

epistatic interactions between states at the two RE positions that positively contributed to this preference.

As a whole, the derived RH changes the positive genetic determinants of affinity in the RE. For AncSR1, having G3 increases affinity regardless of the nucleotide state at position 4 (Figure 3), while REs with A3 also have greater than average binding affinity. We also observe an epistatic interaction between G3 and T4, which indicates that having these two states at positions 3 and 4 have a significantly greater-than-additive effect on affinity than would be predicted by the individual effect of G3. By contrast, AncSR1+RH has only one first-order term, with A4 increasing affinity, and no epistatic terms. This indicates that introduction of the derived RH drastically changed the RE genetic determinants of binding, eliminating all ancestral preference at site 3 and the epistasis between sites 3 and 4 and reorganizing the protein-DNA interface to only improve binding due to molecular information from nucleotides at position 4.

We next wanted to determine how the individual RH substitutions contributed to the change in the RE genetic determinants of binding. We quantified the positive genetic determinants of binding function within the RE for each protein genotype (Figure 3) and analyzed the effect that each RH substitution had on these determinants. The only substitution available to AncSR1 that avoids a low-affinity intermediate, glu25GLY, resulted in a protein that maintained two of the three ancestral genetic determinants for high affinity, losing the epistatic interaction between G3 and T4. The resulting protein therefore still binds preferentially to similar RE sequences as AncSR1, but with less specificity.

Once at the GLY-gly-ala genotype, the introduction of either possible second substitution (gly26SER or ala29VAL) further decreases the ancestral preference. However, only the ala29VAL substitution completely eliminates all the ancestral genetic determinants while simultaneously establishing the derived preference for A4. After the A4 effect is established, the final step from GLY-gly-VAL to GLY-SER-VAL maintains that effect. Going from GLY-gly-ala to GLY-SER-ala via the gly26SER substitution, we see that the ancestral G3 preference is maintained but the A3 preference in eliminated. Along this pathway, the final step from GLY-SER-ala to GLY-SER-VAL eliminates the final ancestral G3 preference while establishing the derived preference for A4. Both pathways (from GLY-gly-ala→GLY-gly-VAL→GLY-SER-VAL and GLY-gly-ala→GLY-SER-ala→GLY-SER-VAL) completely eliminate the ancestral preferences and decrease the promiscuity of the protein to realize the derived preference. These data indicate that the derived RH substitutions progressively re-ordered the genetic determinants of binding in the RE and each potential pathway had a step in which the last remaining ancestral preferences were eliminated while simultaneously establishing the derived preference.

The function of the evolving SR module is influenced by inter-molecular epistasis

We next wanted to understand how genetic variation across both the protein and the RE impacted binding affinity across the entire evolutionary transition. In particular, we were interested in any general effects of variation in the RE that improved binding on average across all protein backgrounds, as well as any epistatic interactions between the protein and the RE. We performed the same set of linear modeling analyses on the entire

dataset, but this time considered models that included interaction terms between genetic states in the protein and in the DNA. In addition to the same general protein effects discussed previously, this approach identified one positive first-order effect in the RE as well as six epistatic interactions between the protein and DNA that contributed to the change in positive determinants for binding in the RE across the evolutionary transition (Figure 4A). We identified a single positive first-order term indicating that A4 increased binding affinity averaged across all protein genotypes. This implies that preferential binding to A4 is an average effect across the transitional sequence space. Its absence from a sub-set of protein genotypes is due to the specific negative epistatic interactions with ancestral RH residues. In fact, all of the protein genotypes that lack an A4 determinant have at least one, if not both, ancestral states in the RH that produce this exclusionary epistasis (Figure 3).

We also identified six epistatic terms between the protein and the RE. These terms indicate the effects of specific individual amino acid states on binding to REs with specific nucleotide states that were preferred by either the ancestral or derived proteins. In particular, we identified 4 epistatic interactions between the protein and the RE that involved RE states that were positive genetic determinants for either ancestral or derived binding affinity (Figure 4A). First, we identified two positive epistatic interactions, between gly26 and G3, as well as between ala29 and G3. These effects imply that the ancestral gly26 and ala29 both specifically increase affinity for REs with G at position 3. Therefore, the gly26SER and ala29VAL substitutions contributed to the elimination of the ancestral preference for G3 by removing this interaction and decreasing affinity for ERE. Additionally, we identified negative epistatic interactions between glu25 and A4,

as well as between ala29 and A4. These negative effects imply that the ancestral glu25 and ala29 specifically reduced affinity for REs with A at position 4. Substitution of these ancestral residues for their derived states alleviated this negative effect and improved binding with the derived A4.

Together, these data indicate that the epistatic interactions between the ancestral residues and the preferred nucleotide states of the ancestral and derived proteins contributed to the ancestral specificity by (1) strongly favoring the ancestral nucleotide preferences and (2) excluding the derived nucleotide preference. Introduction of any of the derived RH substitutions eliminated these epistatic interactions between the protein and DNA. The elimination of these epistatic interactions removed the positive G3 effect, as well as the negative effect that specifically excluded A4. The removal of these specific exclusionary interactions revealed an average positive effect for A4, thereby resulting in the derived preference for A4.

Characterization of the sequence space across this transition reveals potential pathways to functional novelty

We next wanted to identify potential pathways through this space that would have resulted in the evolution of a high-affinity interaction with a novel RE. To identify these pathways, we characterized each protein's connected network of high-affinity RE targets. We defined this network as the interconnected set of RE sequences that were bound with high affinity and within 10-fold of the protein's highest affinity K_D. We reasoned that high-affinity REs that have large energetic differences relative to the preferred sequence would not successfully compete for TF binding and would thus have a low occupancy in

the cell, making them less likely to contribute a regulatory function. High-affinity REs with small energetic differences relative to the most preferred RE, however, would be expected to successfully compete and bind with appreciable occupancy. Describing the system in terms of the high-affinity RE network of each protein intermediate allows us to identify the mutational pathways – both in the protein and the RE – that would allow the evolving transcriptional module to realize a novel function or maintain a conserved ancestral interaction (Figure 4).

We observed two distinct mutational pathways in the TF by which high-affinity interactions with a novel RE could evolve (Figure 4). Novel high-affinity interactions were determined by identifying RE sequences that were not shared in the high-affinity networks for connected protein genotypes. We found that introduction of glu25GLY greatly increased the size of the high-affinity network, resulting in a highly promiscuous protein that bound to a set of 15 RE sequences, 13 of which are novel and completely distinct from the ancestral module. From the cloud of potential REs bound by GLY-gly-ala, there are differently sized subsets that are shared with the two potential subsequent intermediates, GLY-gly-VAL and GLY-SER-ala. Movement through GLY-gly-VAL further increases the set of high-affinity RE sequences from 15 to 16. Conversely, movement through GLY-SER-ala greatly decreases the high-affinity network, having only 4 potential high-affinity targets, two of which are shared with the ancestral module. The final step in both of these pathways is to diminish the number of RE targets in the protein's high-affinity network and eliminate those REs that are shared with the ancestral TF. This ultimately leads to a derived module with a set of novel high-affinity RE sequences that are completely distinct from those bound by the ancestor.

Identification of the connections between RE sequences that are shared between the high-affinity networks of TF genotypes also allowed us to identify the mutational pathways in the RE that would have maintained an ancestral high-affinity connection even upon TF divergence (Figure 4). We found multiple pathways through single-step nucleotide mutations in the RE that would have maintained an ancestral high-affinity interaction even as the protein diversified in its DNA-binding specificity. The presence of these high-affinity mutational pathways implies that the evolution of a novel binding function in a transcription factor may not always result in the establishment of novel network connections to previously unregulated *cis*- elements, but, through compensatory changes in ancestral *cis*- elements, may still maintain ancestral connections even upon diversification.

Novel specificity evolved by changing types of biophysical interactions

We next wanted to understand the underlying mechanisms that caused variation in binding affinity. To determine these mechanisms, we performed molecular dynamics (MD) simulations for AncSR1, AncSR1+RH and all intermediate protein genotypes, each bound to every one of the 16 DNA sequences. We then measured hydrogen bonding and packing at the protein-DNA interface, which are known to contribute to high-affinity interactions in this system (Garvie and Wolberger, 2001; Rohs et al., 2010; McKeown et al., 2014). For each protein, we used linear regression to analyze the statistical relationship between each biophysical parameter and the affinity for all 16 REs.

Hydrogen bonding and packing efficiency do not account for variation in binding affinity across all protein genotypes. Hydrogen bonding and binding affinity was

positively correlated for only 3 out of the 8 protein genotypes (Figure S2, Table 1), and explained only a small percentage of the variation in affinity for each. The strongest correlation was with AncSR1, in which hydrogen bonding accounted for 30% of the binding variation. Four of the protein genotypes showed no correlation between affinity and hydrogen bonding, and one showed a negative correlation. Differences in packing efficiency were correlated with binding affinity for only 3 protein sequences and explained at most 20% of the binding variation (Figure S2, Table 1). Further, hydrogen bonding and packing efficiency, together, explained only 8% of binding variation across all proteins. These data indicate that the number of hydrogen bonds and the extent of packing efficiency at the protein-DNA interface as predicted by MD simulations contribute to DNA binding affinity for some protein sequences, but these values are not global causes of binding affinity across protein sequences. Although hydrogen bonding and packing efficiency failed to predict most of the genetic effects observed in the binding data, the effects uncovered for AncSR1 and AncSR1+RH indicate that the change in specificity occurred by a change in the type of interaction that affects binding: the ancestral specificity was at least partially dependent on the number of hydrogen bonds formed between protein and DNA, while the derived specificity was more dependent on packing efficiency.

DISCUSSION

Novel DNA-binding function evolved by greatly reducing affinity for the ancestral targets while only slightly increasing affinity for derived targets

We found that novel DNA-specificity was largely realized by reducing affinity to ancestral targets and exploiting the existing ancestral affinity for specific sequences that ultimately became the derived targets. The derived RH caused small improvements in the binding affinity to the derived RE targets, but the main effect was to greatly decrease the binding to the ancestral RE targets. By dramatically reducing the protein's affinity to the ancestral targets without a comparable increase in the binding affinity to the derived targets, evolution resulted in a derived protein that bound a larger number of RE targets with similar affinity and thus had lower specificity. Similar evolutionary principles of latent functional exploitation have been observed in other systems (Bridgham et al., 2006; Khersonsky et al., 2006; Coyle et al., 2013), suggesting that it may be an important mechanism for evolutionary novelty.

The evolutionary transition in DNA specificity occurred by a change in the types of biophysical interactions at the protein-DNA interface

Novel DNA-specificity evolved by a change in the biophysical determinants of DNA-binding. The transition was from an ancestral mechanism dominated by hydrogen bonding to a derived mechanism that was more dependent on packing interactions at the protein-DNA interface. However, the ability of these interactions to explain overall variation in binding affinity of either of these complexes is fairly limited and fails to recover most differences in affinity across all protein intermediates.

We did not identify a single biophysical property that explains variation in binding across all proteins. Instead, DNA affinity and specificity appears to be determined by variation in biophysical interactions that are specific to each protein-DNA complex. For example, a specific steric clash between the ancestral residue at 25 and an A at position 3, which we described in previous work (McKeown et al., 2014), would not be a strong determinant of affinity for genotypes lacking the ancestral residue that clashes with this nucleotide. Similarly, differences in hydrogen bonding would not be expected to predict binding for protein constructs incapable of forming direct hydrogen bonds to DNA, such as the protein intermediate GLY-gly-ala. While the novel specificity of the derived protein likely evolved at least in part by establishing novel types of physical interactions and abolishing old ones, there remain many other physical interactions operating through specific mechanisms that are functionally relevant in this system, the determination of which is beyond the scope of this study.

A linear modeling approach resulted in a statistical description of the genetic determinants of binding-specificity

The linear modeling approach to describe the genetic determinants of binding function allowed us to quantitatively describe the evolution of binding affinity and specificity across this sequence space. Each of the three RH substitutions had large generic effects on binding affinity; one increased affinity and two decreased affinity across all REs tested. Although the signs of these effects were consistent across REs, the overall shift in preference occurred because the magnitude of each effect on affinity varied across the REs. glu25GLY increased affinity for SREs more than for ERE; the

other two substitutions caused a larger decrease in binding to ERE than to SREs. Thus, there was no single substitution that uniquely increased binding only to the derived targets, or uniquely decreased binding to the ancestral targets. We speculate that this is because such specific effects are difficult given the dense and heterogeneous properties of the biophysical architecture at the binding interface. Substitutions that specifically improve or specifically weaken interactions are likely more difficult to establish than those with a non-specific but differential effect, and would thus be expected to occur less frequently.

We also observed widespread epistasis within the protein, within the RE, and between the protein and the RE. In the case of SRs, intra-protein epistasis is likely to have limited the number of paths by which the novel function could have evolved. The negative intra-protein epistatic effects made it impossible to combine specific states and still maintain a high-affinity protein, likely constraining these mutational pathways, because the resultant proteins lack the ability to bind any REs with high affinity.

The existence of intra-RE epistasis greatly improves a system's specificity. These epistatic interactions result in a large difference between affinity for sequences with both of the interacting states and sequences that have only one. As such, an RE sequence with epistatically interacting states results in greater specificity because it can better compete for binding by a given TF relative to those whose binding is determined by only first-order effects.

By extending this analysis across macromolecules, we found that specific states in the protein differentially affected affinity for REs with specific nucleotide states, thereby leading to inter-molecular epistasis across interacting macromolecules. These differential

effects are the underlying genetic mechanisms that allowed substitutions in the protein to shift its DNA specificity; in the absence of inter-molecular epistasis, each protein substitution would have had a statistically equivalent effect across all REs, resulting in a protein that bound with a different absolute binding affinity but still preferred the same REs.

Inter-molecular epistasis implies that the effect of substitutions in each macromolecule is dependent on the other's genetic state. Depending on the genetic background of the protein, the RE may be able to drift through many single nucleotide mutations without detriment to the high-affinity interaction. Alternatively, a more specific protein will limit the number of genotypes available to the RE. The converse is also true: The identity of the RE may permit the protein to mutate to any of the derived residues without compromising the high-affinity interaction or may constrain the protein by permitting mutation to any derived residue. Depending on the functional constraints that exist for the system, these epistatic interactions could play a critical role in determining the evolutionary pathways that were available for the evolving SR module (Phillips, 2008).

The identification of such a diverse set of epistatic interactions within such a minimal system, encompassing only three amino acid substitutions in the protein and two variable nucleotide positions in the RE, is particularly noteworthy. This widespread epistasis suggests that evolution of larger, more complex molecular systems – and certainly whole genomes – should appreciate that non-additive epistatic interactions within and between interacting macromolecules are likely the norm rather than the exception (Breen et al., 2012).

Direct mutational pathways required the ancestral module to evolve through either a low-affinity or a promiscuous protein intermediate

All direct genetic pathways between the ancestral and derived proteins required passing through low-affinity or promiscuously binding intermediates. Based on available phylogenetic data, it is impossible to determine the exact mutational pathway taken by the evolving DBD, as none of these intermediate genotypes have persisted to the present. However, we can speculate on the potential evolutionary consequences, and therefore the plausibility, of taking each of these routes to the derived function.

After a gene-duplication, the redundancy of the second gene copy is thought to free it from functional constraint and allow it to sample genotypes that could potentially give rise to novel functions. If the duplicate were to sample a low affinity intermediate, however, it would be incapable of binding DNA sequences with an appreciable occupancy in a cellular environment, and would therefore be unlikely to maintain any regulatory function. The loss of regulatory interactions may be completely neutral; in this case, the evolving protein would be released from purifying selection and it would thus be expected to randomly sample its surrounding sequence space. While this would allow the evolving module to potentially traverse selectively-deleterious functional valleys that separate it from the derived state, the majority of these random mutations would be expected to further degrade the protein's binding function, potentially even compromising its structure (Guo et al., 2004; Lisewski, 2008). The increased rate of unconstrained mutation is expected to result in rapid degeneration and ultimately lead to pseudogenization (Fisher, 1935; Ohno, 1970; Lynch and Katju, 2004). This is true even

for a post-duplicate gene, as is the case with the evolving SR, as the duplicate would still need to evolve a new function-restoring mutation before accumulating additional non-functionalizing mutations (Haldane, 1933). This suggests that traversing through a low-affinity intermediate also made it more likely for pseudogenization. Given the presence of alternate pathways that would not have required a loss of purifying selection to evolve a novel DNA-binding function, these low-affinity pathways are unlikely to have been taken.

Evolving through a promiscuous protein intermediate would be expected to maintain the ancestral function, but would also have the potential for off-target effects, which could be deleterious. However, by expanding the number of possible DNA sequences that could be bound with high affinity, a promiscuous intermediate would greatly increase the evolvability of the RE. Subsequent substitutions could have then refined that promiscuity in order to ultimately realize the derived specificity. Additionally, a promiscuous protein would have been likely to maintain its ability to regulate gene targets *in vivo* and would have remained the subject of purifying selection, making it less likely than the low-affinity protein to have rapidly degraded into a pseudogene.

There is a significant body of evidence that supports the role of promiscuous intermediates in the evolution of novel specificity across diverse systems, including other transcription factors (Khersonsky et al., 2006; Howard et al., 2014; Sayou et al., 2014). Together with our data, this implies that traversing through a short-lived promiscuous intermediate may be the most likely pathway that the evolving protein took during its history.

Multiple pathways could have enabled the evolution of novel function without compromising high-affinity binding with an ancestral target

Given that REs can also evolve, it is possible that a change in transcription factor specificity could be compensated for by changes in the RE, ultimately resulting in the conservation of an ancestral connection. This scenario is of particular interest for understanding regulatory evolution, as it suggests that pathways may exist whereby the functions of TFs and REs can change even if the regulatory module is under strong purifying selection to maintain specific regulatory interactions (True and Haag, 2001). Further, such intermolecular compensation is thought to be an important source of genetic incompatibilities that drive speciation between recently diverged lineages (Haag and True, 2007; Barriere et al., 2012).

We determined that many pathways existed through this space by which single-step genetic mutations in both the protein and RE would have allowed the protein to maintain high-affinity binding with an ancestral gene target. By proceeding through a promiscuous protein intermediate, the RE high-affinity network was greatly increased, allowing the RE sequence of an ancestral target to freely mutate from an ancestral target to a derived target. As the module moved through this high-affinity network of genotypes, the promiscuous protein was refined by successive introduction of other derived residues in the protein, the realization of which was dependent on the RE first mutating from an ancestral RE target to a derived RE target. Given these interactions, the transcriptional module could have evolved by moving from one edge of this high-affinity network, through a densely connected region, until finally arriving at the derived

genotype on the other side. The movement of the module through this space was dependent on the evolution of both macromolecules, each step of which was contingent on the random mutations that have occurred in its interacting partner.

Mapping the functional sequence space reveals important details about how evolutionary novelty could have arisen

To reach a novel function, the protein had to proceed through at least one intermediate protein that was functionally distinct – either low-affinity or generally promiscuous – from both the ancestral and derived proteins. The functions of these alternate potential intermediates could not have been determined solely by looking at the beginning and end-points of the transition, but required characterization of the sequence space that separated them. By mapping the functional sequence space for this evolutionary transition in terms of both the protein and the RE, we uncovered a vast high-affinity network that would not have been discovered if only considering substitutions in either the protein or the RE in isolation. This implies that understanding the evolutionary pathways and processes that govern regulatory network evolution is best accomplished by studying *cis*- and *trans*-acting components in an integrated way. The evolvability of a transcriptional module – and certainly other multi-component systems – is a result of how changes in each of its interacting parts shape the function of the complex as a whole. Therefore, to understand the evolutionary potential of these systems, it is best to dissect genetic changes that extend across both interacting partners. By doing so, this work shows that it is possible for evolution to wander its way across the intervening sequence

space and, by altering each macromolecular component by single-step mutations, ultimately connecting functional spaces that might otherwise appear completely discrete.

BRIDGE TO CHAPTER V

Chapter IV examines the set of alternate functions that existed for the set of genotypes that encompasses an evolutionary transition in the steroid receptor transcriptional regulatory system, and shows that evolution likely proceeded either through a low-affinity or highly-promiscuous intermediate state n. This complements the characterization of the major genetic, biochemical, and biophysical mechanisms for that novel function from Chapter III. In Chapter V, I conclude by summarizing the major implications from the specific work in Chapters III and IV, and I place it in the context of the larger program of molecular evolutionary research that I presented, and for which I advocated, in Chapter II.

CHAPTER V

CONCLUSIONS

This dissertation has sought to contribute to the field of molecular evolution by examining the molecular mechanisms for the evolution of novel functions. It has done so in both general and specific ways, by analyzing the dominant program of research in contemporary molecular evolution and by conducting a set of detailed mechanistic studies of the evolution to novel DNA-binding specificity in the steroid hormone receptor family of transcription factors.

Uncovering the molecular mechanisms of evolutionary change

Evolution changes genetic content over time. Some of the genetic changes that that occur during evolution cause novel functions in macromolecules. This dissertation has sought to contribute to our understanding of how evolution has caused changes at the level of molecular function, and to draw out potential general features of that process.

Identifying these mechanisms requires applying the tools of molecular biology to the alternate genotypes that were realized during the evolution of novel functions. One way to accomplish this is to use ancestral sequence reconstruction in order to identify the specific set of genetic substitutions that produced a novel function. Once the specific genetic mechanisms for the evolutionary novelty are identified, it is possible to use further molecular biological methods in order to characterize the structural and dynamical features that mediated those genetic effects on function. By characterizing the genetic and biophysical mechanisms for specific evolutionary transitions, we can also begin to build a

body of knowledge from which to derive some of the general principles that describe how evolution changes these functions over time.

The value of studying all types of evolutionary transitions

In Chapter II, this dissertation raised the issue of molecular evolutionary biology studies being particularly focused on studying the evolution of adaptive traits, to the exclusion or de-emphasis of other traits that were of either non- or ambiguous adaptive value. It argues that this constitutes a modern adaptationist research program, in which the demonstration that a trait or a genetic region contributed to adaptation is both necessary and sufficient for a study to contribute to molecular evolutionary biology. An unfortunate by-product of this focus is that many studies emphasize relatively weak inferences of adaptation rather than stronger inferences into the mechanisms of evolutionary change. This chapter proposes that this adaptationist program should be replaced by a post-adaptationist program of molecular evolutionary research, in which researchers would recognize that major insights into evolution can come from characterizing the molecular mechanisms for all types of evolutionary change, whether it was adaptive or not, and whether or not it is possible to demonstrate its adaptive value at all.

Characterizing the molecular basis for novel DNA-binding function in the steroid receptor family

The steroid receptor family provides an excellent case study for characterizing the molecular mechanisms that underlie functional novelty. Chapter III has uncovered the

major genetic, biochemical, and biophysical mechanisms for novel DNA-binding function. In particular, my work performing molecular dynamics *in silico* experiments and analyses revealed the biophysical evolution of the steroid receptor-DNA complex, and showed that functional novelty arose because of a combination of lost, specific positive contacts, and the gain of specific, exclusionary contacts. The role of specific negative contacts suggests that evolution sometimes utilizes coarse, inelegant physical mechanisms to achieve new functions, possibly because these are the most easily accessible in the mutational sequence space that was available to evolution.

Chapter IV built upon these findings by dissecting the evolutionary transition in DNA-binding specificity to a much finer degree, characterizing the function for each point in the genotype-phenotype map that separates the ancestral and derived functions. With this work, we showed first of all that the genetic determinants of binding affinity included many epistatic interactions, within the protein, within the response element, and between the protein and the response element. Such widespread epistasis within such a minimal system has clear implications for the evolution of larger macromolecular complexes. Additionally, this work showed that the direct mutational pathways separating the ancestral and derived modules had to proceed through either a low-affinity or promiscuously binding intermediate protein. As a low-affinity intermediate is less likely to be functionally relevant, we consider the promiscuous pathway to have been more likely, which in light of other findings suggests that promiscuity may play an important general role in facilitating the evolution of novel protein functions. Further, this work demonstrated the potential for a sequence of substitutions in both the protein and

the RE that would have enabled an ancestral high-affinity interaction to be maintained throughout the evolution of novel protein function.

Overall, these two studies have characterized the evolution of novel DNA-binding specificity in the steroid receptor family in terms of the genetic, biochemical, and biophysical mechanisms, and further, characterized the direct mutational pathways that were available to both the evolving protein and the evolving DNA response elements. While the conclusions drawn from this work are based on this specific system, their implications for general evolutionary processes are significant, as they have shown how evolution acted to change the biochemical and biophysical interactions that caused a new, and critically important, molecular function to exist.

How molecular mechanisms shape the evolutionary process

By uncovering the molecular mechanisms for novel functions, we can begin to understand not only valuable mechanistic details about the systems we are studying, but we can also begin to induce general principles of evolution – such as the role of negative biophysical interactions in determining binding specificity, or the way that promiscuous intermediates mediate evolutionary change – that would otherwise remain hidden if we were to focus only on genotypic or phenotypic change in isolation. It is in the details of these mechanisms that we see what actually happened in evolutionary history to create the diversity of functions and phenotypes that exist. By studying the specific mechanisms that were used by evolution in the history of important gene families, we can begin to form general conclusions about how evolution works, and how such a process can yield

the great diversity of life forms and functions that we see in the world. This dissertation has sought to contribute to that body of work.

APPENDIX A

BOXES AND FIGURES FOR CHAPTER II

BOX 1

"Dr. Pangloss" as a symbol for the adaptationist program

In 1759, Voltaire published the novella *Candide: or, All for the Best*, a fictional work widely interpreted as a critique of Gottfried Leibniz's philosophy of optimism, which claimed that the world that exists must be the best world that could have existed (Leibniz, 1710) Voltaire's novella tells the story of the naïve protagonist "Candide" and his mentor "Dr. Pangloss", who was a satirical representation of Leibniz, characterized by his refrain: "All is for the best in the best of all possible worlds." (Voltaire, 1759).

Gould and Lewontin's seminal 1979 paper used Dr. Pangloss as a representation of what they argued was a flawed "adaptationist" program that dominated evolutionary biology research. They argued that many evolutionary biologists based their work on the assumption that all the biological traits that exist must have become fixed because they had some adaptive value. Gould and Lewontin found this worldview reminiscent of the "best of all possible world's" ideal represented by Dr. Pangloss in *Candide*, hence their label for the defining feature of that adaptationist program: "The Panglossian Paradigm."

BOX 2

Sequence-based statistical methods for inferring adaptation have become commonplace in evolutionary studies

There are several different analytical strategies that are commonly employed in order to infer the signature of an historical selective sweep in genomic data. While this is not an exhaustive discussion, we will highlight a few of the most common methods here.

First, codon-based tests, which compare the rate or number of non-synonymous and synonymous genetic changes, with the goal of identifying cases in which there is an excess of amino-acid switching changes. dN/dS tests compare the ratio of these rates on different lineages (and, sometimes, on different genes), with the expectation that adaptation will produce an inflated ratio (Hughes and Nei, 1989). Similarly, the McDonald-Kreitman test compares the relative number of non-synonymous and synonymous substitutions to the relative number of segregating intra-population non-synonymous and synonymous polymorphisms in one or a few species. If there are relatively more non-synonymous differences between species than one would expect given the relative number of both types of polymorphisms, then one infers that at least some of those non-synonymous differences must have contributed to adaptation (McDonald and Kreitman, 1991). Both of these tests search for cases of rapid evolutionary change in non-synonymous substitutions, which is consistent with an "arms race" scenario of molecular adaptation. They are not expected, however, to identify cases where adaptation is realized via a small number of genetic changes of large effect (Hughes, 2007).

Population-level tests calculate the fixation index (F_{st} – a measure of population differentiation at the genetic level), linkage disequilibrium (LD – a measure of the degree to which variation is correlated across a genetic distance), or a related statistic describing genetic variation that is often customized to particular demographic conditions. The goal of this set of methods is to identify genes or genetic regions that exhibit patterns of variation that could only have been produced by directional selection associated with adaptation (Evans et al., 2005; Mekel-Bobrov et al., 2005).

Figure 1 (next page). The path of evolutionary change from one generation to the next

Depicts the dynamic relationship between alternative genotypes, the consequent phenotypes that are the result of molecular and developmental biological processes, and the way in which population genetics processes can alter both genotype allele and phenotype frequencies over time (loosely based on figure 1 from (Lewontin, 1974)). Differently colored circles in genotype space indicate genetic variants, with associated phenotypes exhibiting a shared color in phenotype space. Shapes in phenotype space indicate phenotypic variation. The two spaces – genotypic and phenotypic – are connected but non-identical, and their dynamics are affected differently by natural selection. G_1 and P_1 represent the genotype and phenotype distributions for generation one, with the horizontal arrow representing all of the molecular processes (protein biochemistry, gene expression, cell biology, development, physiology) that translate genotypes into phenotypes. P_1' represents the phenotypes of the reproductively successful members of generation one, and G_1' represents their associated genotypes. G_2 and P_2 represent the genotype and phenotype distributions for generation two, which are the result of how population genetic forces (selection, drift, and migration, which translate P_1 into P_1'), as well as recombination, mutation, assortment, linkage, genetic drift, and genetic drive (which translate G_1' into G_2). The ways in which these population genetics forces affect the distributions of genotypes and phenotypes is significantly influenced by the molecular and developmental mechanisms that translate genotype into phenotype (horizontal arrows).

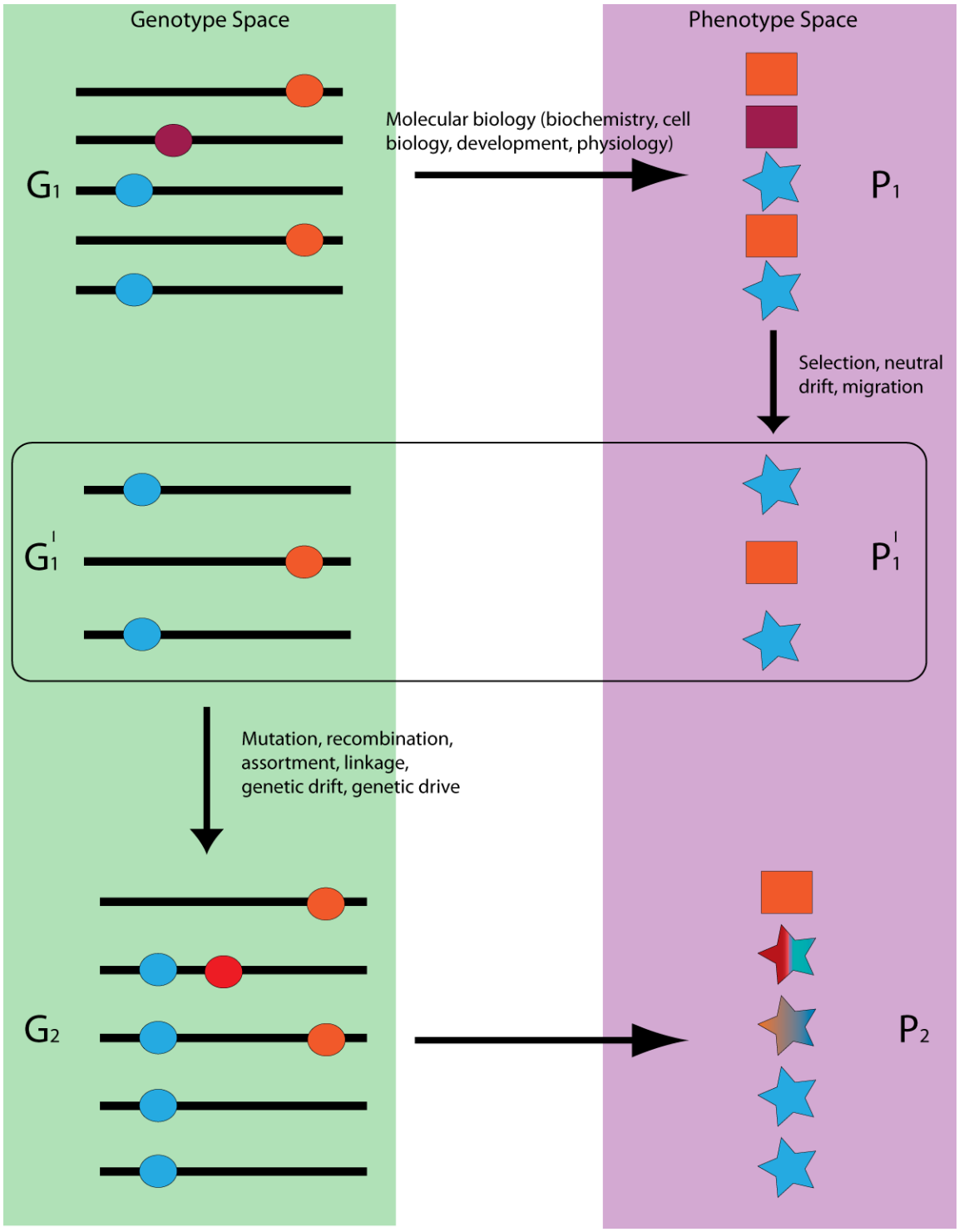
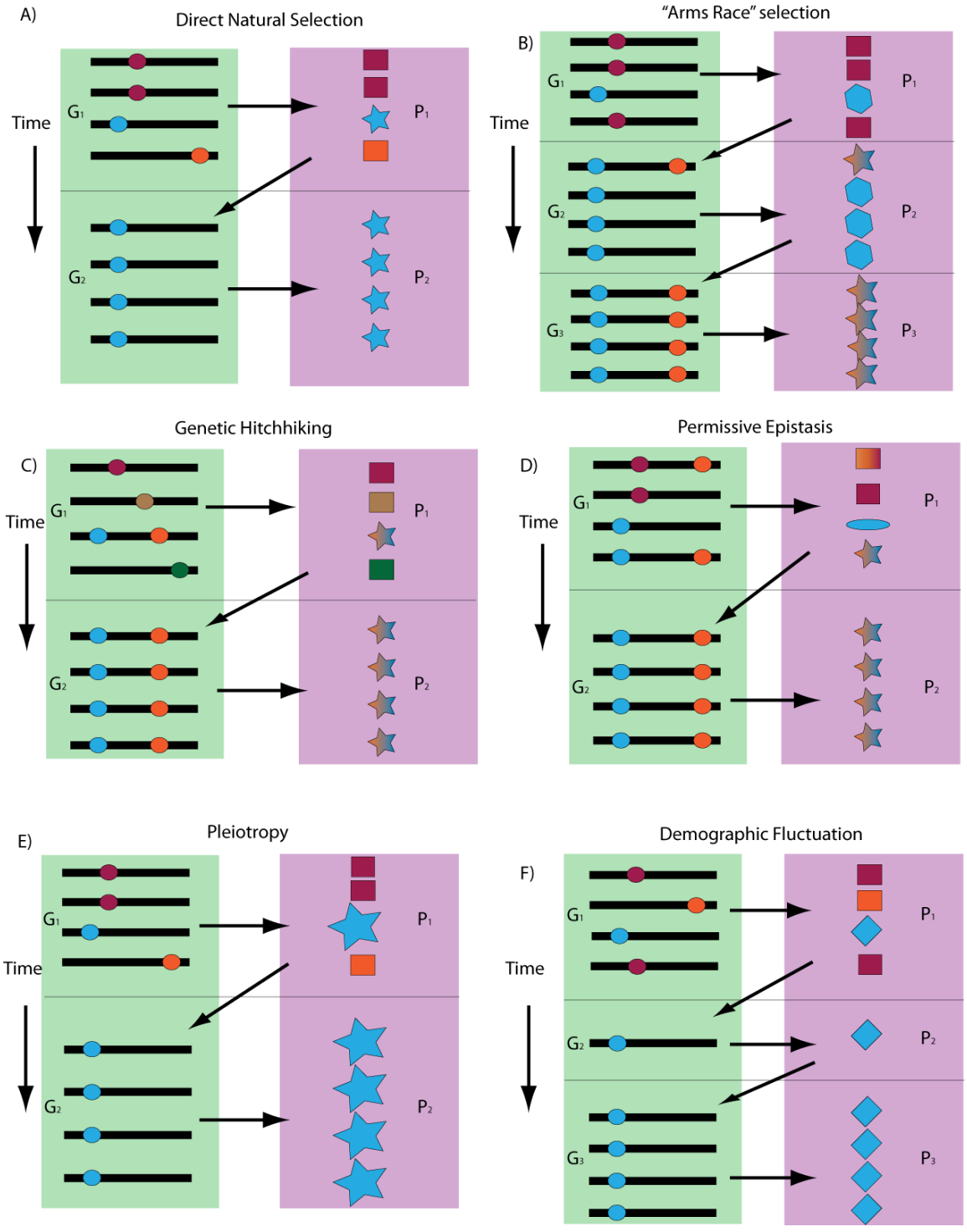

Genotype Space
Phenotype Space
Molecular biology (biochemistry, cell biology, development, physiology)
G1
P1
Selection, neutral drift, migration
G1'
P1'
Mutation, recombination, assortment, linkage, genetic drift, genetic drive
G2
P2

Figure 2 (next page). Alternative scenarios for the fixation of a low-frequency genotype and/or phenotype

Simplified schematics based on figure 1, in which the transition from generation one to generation two is collapsed into a single diagonal arrow. Stars represent positively-selected phenotypes, while other shapes are neutral in terms of fitness (unless otherwise indicated). A-F depict alternative scenarios that could result in the rapid fixation of a low-frequency genetic and/or phenotypic variant. A) An example of fixation due to direct natural selection, resulting from a one-to-one association between a specific genetic variant and a specific adaptive phenotypic variant. This will produce a positive adaptive signature using population-based metrics like F_{st} and LD, but will not be positively identified by codon-based methods. B) An example of protein domains involved in "arms race" dynamics where novel alleles are constantly favored by natural selection, such as the antigen-recognition domain of the major histone compatibility complex. In the first generation, the hexagon phenotype is positively selected, while in the second generation, the star phenotype is positively-selected. Population-based methods may positively identify the most recent variants if the timescale is very short. Codon-based methods should positively identify these genetic regions (Hughes, 2007). C) An example of fixation due to genetic linkage (also known as "genetic hitchhiking"). Natural selection favours the star phenotypes, thus fixing the causal blue genetic variant. The orange variant is fixed due to its physical linkage and the inability of recombination to dissociate the two. Population-based methods will positively identify the entire region, however they will be unable to conclusively identify the specific variant that contributed to adaptation. Codon-based methods will not positively identify either variant. D) An example of fixation due to one genetic variant (orange) acting as a neutral permissive change that does not directly alter function or phenotype, but which is required in order to tolerate an additional subsequent function-switching genetic variant (blue), which is inviable without the permissive change. Both genetic variants will be positively identified by population-based methods but not codon-based methods. E) An example of fixation due to direct selection on one trait (star) and the coincident fixation of another trait (increased size) caused by the same genetic variant (blue). In this case, both traits are fixed due to being caused by the same genetic variation, despite selection only favoring one of the traits. The genetic variant will be positively identified by population-based methods, but distinguishing between the adaptive and non-adaptive pleiotropic traits requires direct measurements of fitness (Barrett et al., 2008). Codon-based methods will not positively identify this variant. F) An example of how demographic fluctuations, like a severe population bottleneck, can result in the fixation of a non-adaptive genetic variant and its associated non-adaptive phenotype. Without a very confident accounting of demographic history (discussed in the text), population-based methods may identify this genetic variant as contributing to adaptation. Codon-based methods will not identify this region.

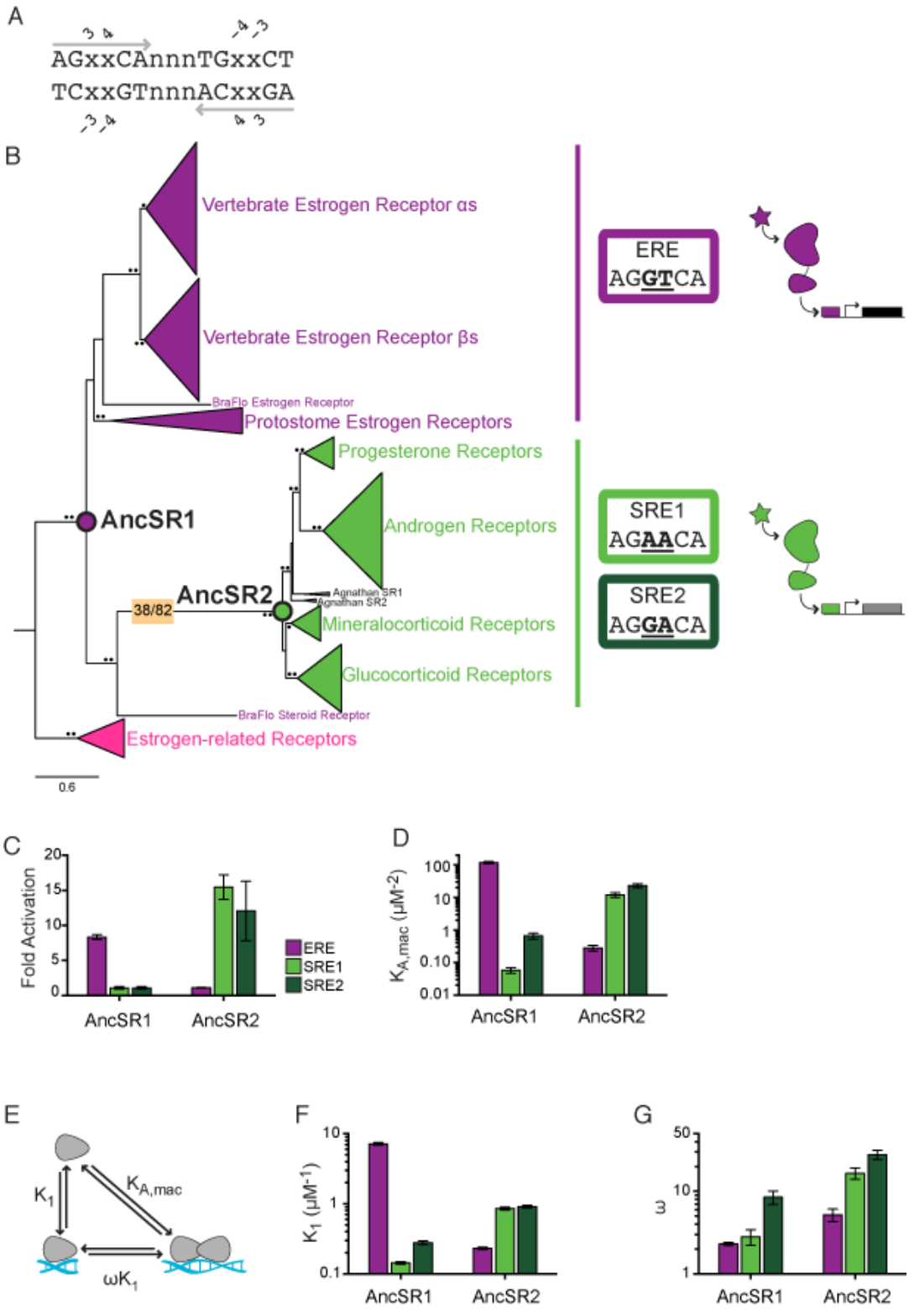

A)
Direct Natural Selection
Time
G1
G2
P1
P2
B)
"Arms Race" selection
Time
G1
G2
G3
P1
P2
P3
C)
Genetic Hitchhiking
Time
G1
G2
P1
P2
D)
Permissive Epistasis
Time
G1
G2
P1
P2
E)
Pleiotropy
Time
G1
G2
P1
P2
F)
Demographic Fluctuation
Time
G1
G2
G3
P1
P2
P3

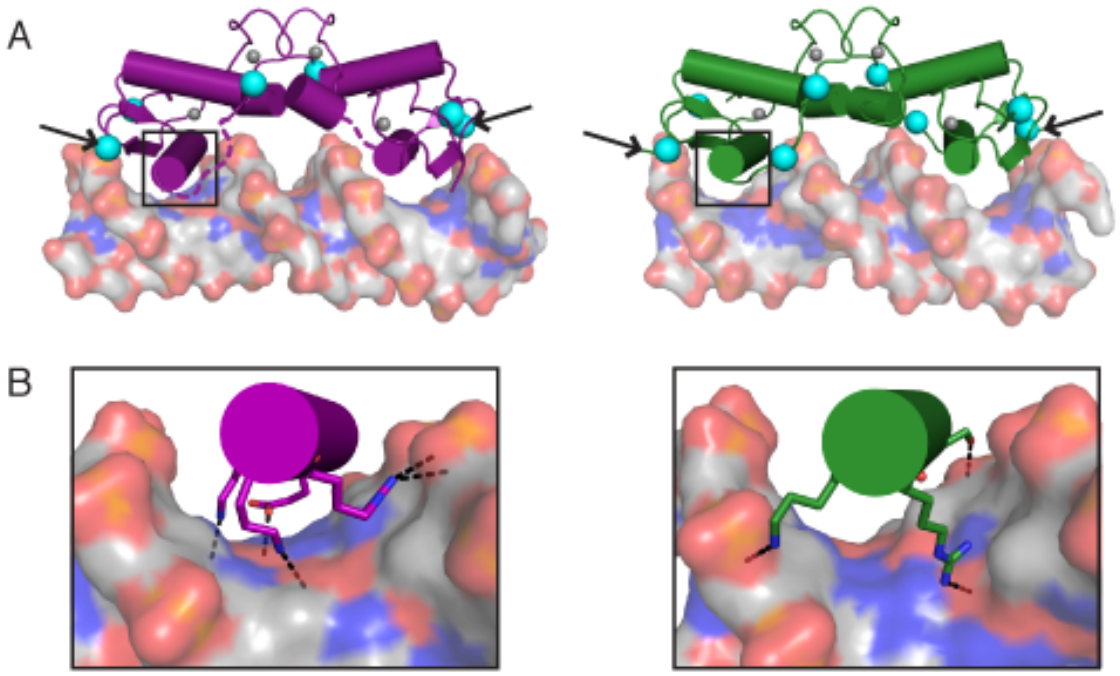
APPENDIX B

FIGURES AND TABLES FOR CHAPTER III

Figure 3 (next page). Evolution of novel specificity occurred via a discrete shift between AncSR1 and AncSR2

(A) Architecture of SR response elements. All SRs bind to an inverted palindrome of two half-sites (gray arrows) separated by variable bases (n). x, sites at which ERE and SREs differ. (B) SR phylogeny comprises two major clades, which have non-overlapping specificity for ligands (stars) and REs (boxes). Preferred half-sites for each clade are shown; bases that differ are underlined. Ancestral and extant receptors are colored by RE specificity (purple, ERE; green, SREs; blue, extended monomeric ERE). Orange box, evolution of specificity for SREs; number of substitutions on this branch and the total number of DBD residues are indicated. Nodal support is marked by the approximate likelihood ratio statistic: unlabeled, aLRS 1 to 10; *, aLRS 10 to 100; **, aLRS>100. Scale bar is in substitutions per site. (C) AncSR1 specifically activates reporter gene expression driven by ERE (purple bar), with no activation from SRE1 (light green) or SRE2 (dark green); AncSR2's specificity is distinct. Bar height indicates fold-activation relative to vector-only control. (D) Ancestral binding affinities reflect distinct specificities for ERE vs. SREs. Bars heights indicate the macroscopic affinity ($K_{A,mac}$) of binding to palindromic DNA response elements, measured using fluorescence polarization. Colors as in panel C. (E-G) The components of macroscopic binding affinity—affinity for a half-site (K_1) and cooperativity of binding (ω)—by AncSR1 and AncSR2, were estimated by measuring $K_{A,mac}$ on a full palindromic RE and K_1 on a half-site, then globally fitting the data to a model containing both parameters. Error bars show SEM of three experimental replicates. See Fig. S1; Tables S1-S3.


A
AGxxCAnnnTGxxCT
TCxxGTnnnACxxGA
B
Vertebrate Estrogen Receptor αs
Vertebrate Estrogen Receptor βs
BraFlo Estrogen Receptor
Protostome Estrogen Receptors
Progesterone Receptors
Androgen Receptors
AncSR1
AncSR2
38/82
Agnathan SR1
Agnathan SR2
Mineralocorticoid Receptors
Glucocorticoid Receptors
BraFlo Steroid Receptor
Estrogen-related Receptors
0.6
ERE
AGGTCA
SRE1
AGAACA
SRE2
AGGACA
C
Fold Activation
ERE
SRE1
SRE2
AncSR1
AncSR2
D
K_A,mac (μM-2)
AncSR1
AncSR2
E
K1
K_A,mac
ωK1
F
K1 (μM-1)
AncSR1
AncSR2
G
ω
AncSR1
AncSR2



C

	Buried Surface Area (Å^2)		
	Chain A::DNA	Chain B::DNA	Chain A::Chain B
AncSR1	617	610	401
AncSR2	405 (388)	343 (324)	565 (565)

Figure 4. Structures of ancestral proteins give insight into the molecular determinants of specificity. (A) X-ray crystal structures of AncSR1 bound to ERE (left); AncSR2 bound to SRE1 (right). Cartoon shows protein dimers; surface shows DNA. Black arrow, beginning of unresolved C-terminal tail. Dotted line, unresolved AncSR1 loop near dimerization interface. Cyan spheres, sites of permissive substitutions. Grey spheres, zinc atoms. (B) Enlarged view of recognition helix in the DNA major groove (black box in A). Sticks, side chains of RH residues making polar contacts with DNA. Dotted lines, hydrogen bonds and salt bridges from protein to DNA. (C) Buried solvent-inaccessible surfaces in Å^2 at the protein-DNA and protein-protein interfaces in the crystal structures for each protein chain. Parentheses, calculations when residues unresolved in the AncSR1 crystal structure are excluded. See Table S4.

Figure 5 (next page). Genetic basis for evolution of new DNA specificity
(A) AncSR1 and AncSR2 sequences. Substitutions between AncSR1 and AncSR2 are shown. Dots, conserved sites. ^, recognition helix (RH) and *, permissive substitutions. Grey box, RH. (B) Effect of RH and 11 permissive (11P) substitutions in luciferase reporter assays. Lower and upper case letters denote ancestral and derived states, respectively. Fold activation over vector-only control is shown, with SEM of three replicates. (C) RH substitutions shift half-site affinity among REs, and permissive substitutions non-specifically increase half-site affinity and cooperativity. The corners of the square represent genotypes of AncSR1, with or without RH and 11P substitutions. At each corner, circle color shows RE preference; numbers are the ratio of the K_{Amac} for binding to SRE1 (upper) or SRE2 (lower) versus ERE. Along each edge, vertical bar graphs show the effect of RH or permissive substitutions on the energy of association for the dimeric complex (grey background); contributions of effects on half-site binding (beige) and cooperativity (cyan) are shown. Bar color shows effects on binding to ERE (purple), SRE1 and SRE2 (light and dark green, respectively). Graphs in the square's center show the effect of 11P and RH combined. Mean ± SEM of three experimental replicates is shown. See Figures S2-S4; Tables S3 and S5.

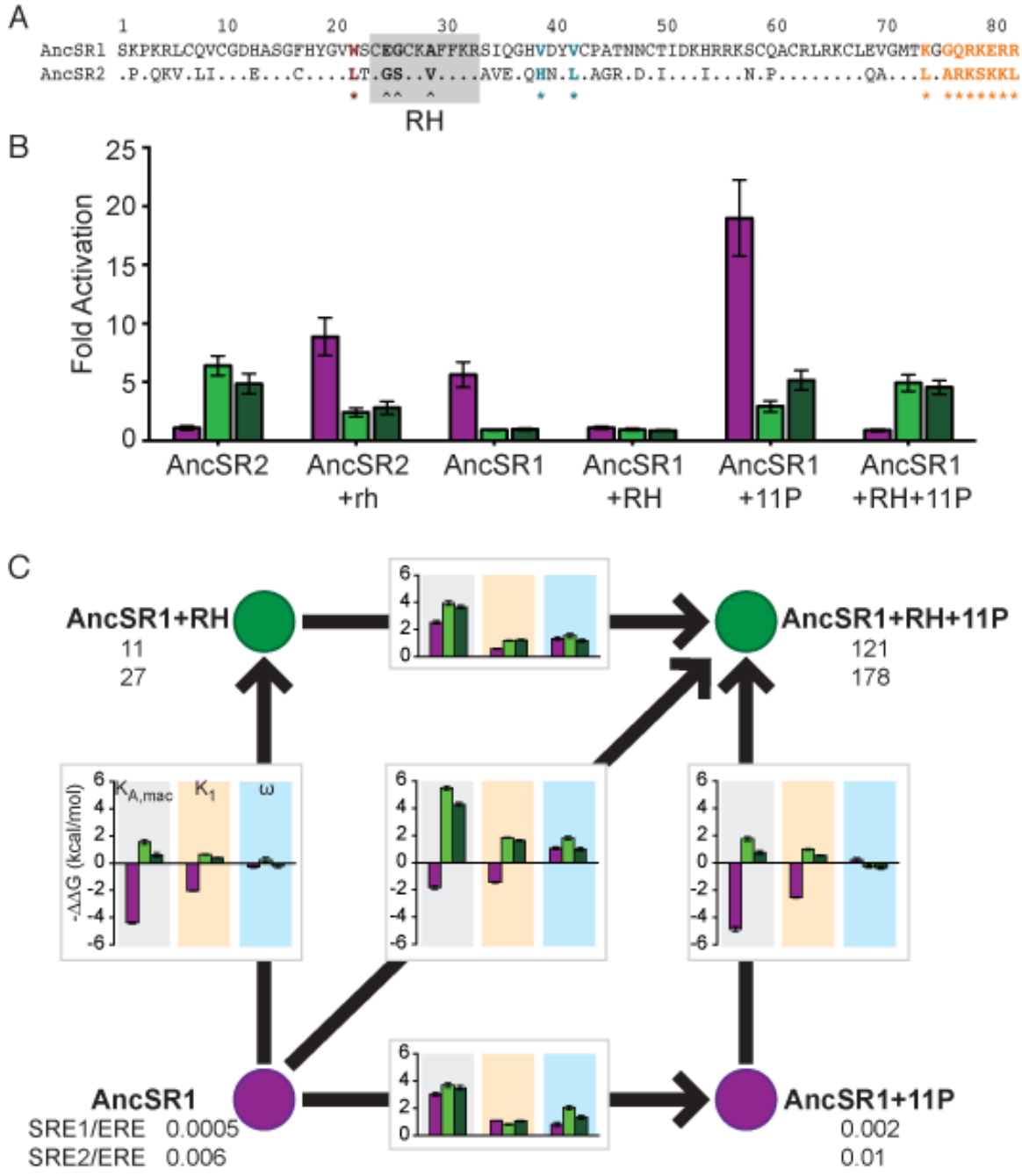

A
1 10 20 30 40 50 60 70 80
AncSR1 SKPKRLCQVCGDHASGFHYGVWSCEGCKAFFKRSIQGHVDYVCPATNNCTIDKHRRKSCQACRLRKCLEVGMTKGGQRKERR
AncSR2 .P.QKV.LI...E...C....LT.GS..V....AVE.QHN.L.AGR.D.I...I...N.P.....QA...L.ARKSKKL
RH
B
Fold Activation
25
20
15
10
5
0
AncSR2
AncSR2 +rh
AncSR1
AncSR1 +RH
AncSR1 +11P
AncSR1 +RH+11P
C
AncSR1+RH
11
27
AncSR1+RH+11P
121
178
-ΔΔG (kcal/mol)
K A,mac
K 1
ω
AncSR1
SRE1/ERE 0.0005
SRE2/ERE 0.006
AncSR1+11P
0.002
0.01

Figure 6 (next page). Recognition helix substitutions change DNA specificity by altering negative interactions

(A) In MD simulations, RH substitutions reduce hydrogen bonds to ERE but do not increase hydrogen bonds to SREs. Bars show mean number of direct hydrogen bonds from all 10 RH residues to DNA (Purple, ERE; light green, SRE1; dark green, SRE2), each sampled across three MD trajectories, with SEM. (B) RH substitutions reduce packing efficiency at the protein-DNA interface on ERE, but do not improve packing on SREs. Bars show the mean number of atoms in the 10 RH residues within 4.5 Å of a DNA atom. (C) Ancestral residue glu25 (sticks) shifts position due to steric clashes with T-4 and T-3 of SRE1. A representative sample frame from MD trajectories is shown for AncSR1 with ERE (purple) or SRE1 (green). DNA is shown as surface, with atoms in the variable bases -4 and -3 shown as lines; methyls of T-4 and T-3 are spheres. (D-F) Repositioning of glu25 by SREs causes Lys28 to shift, reducing hydrogen bonds to DNA. (D) The average position of these residues in MD trajectories of AncSR1 with various REs is shown when all atoms in the protein-DNA complex are aligned. Distance of lys28 from hydrogen bond acceptor G2 on ERE is shown in black. (E) Displacement of glu25 and lys28 of AncSR1 on SREs relative to their position on ERE. The mean positions of all atoms in each MD trajectory were calculated, the DNA atoms in these "mean structures" were aligned in pairs: bars shows the average distances from the atoms in complexes with SRE1 (dark green) or SRE2 (light green) to the corresponding atom in ERE were calculated. Purple bars, distances between pairs of atoms from independent ERE trajectories. Displacement toward the center of the palindrome was scored as positive, away as negative. Each bar shows the distance averaged across atoms in a residue and three pairs of trajectories with SEM. (F) Lys28 forms fewer hydrogen bonds to DNA on SREs than on ERE. Points show the mean number of hydrogen bonds formed by each RH residue to different REs, with SEM for three MD trajectories. (G,H) Effect of introducing e25G and other RH substitutions on half-site binding affinity (G) and transcriptional activation (H). See Figures S6-S7, and Table S3. (I) Summary of mechanisms by which ancestral RH excludes SREs. Ancestral glu25 and conserved residue Lys28 form hydrogen bonds (black dotted lines) with ERE bases. These side chains would sterically clash with methyl groups of SRE1 and SRE2, so they are repositioned and are unable to form hydrogen bonds to DNA, leaving unpaired donors (blue) and acceptors (red) at the DNA-RH interface. The RH substitutions resolve the steric clash and remove the unfulfilled donor on e25, increasing SRE affinity. See Figures S5-S6.

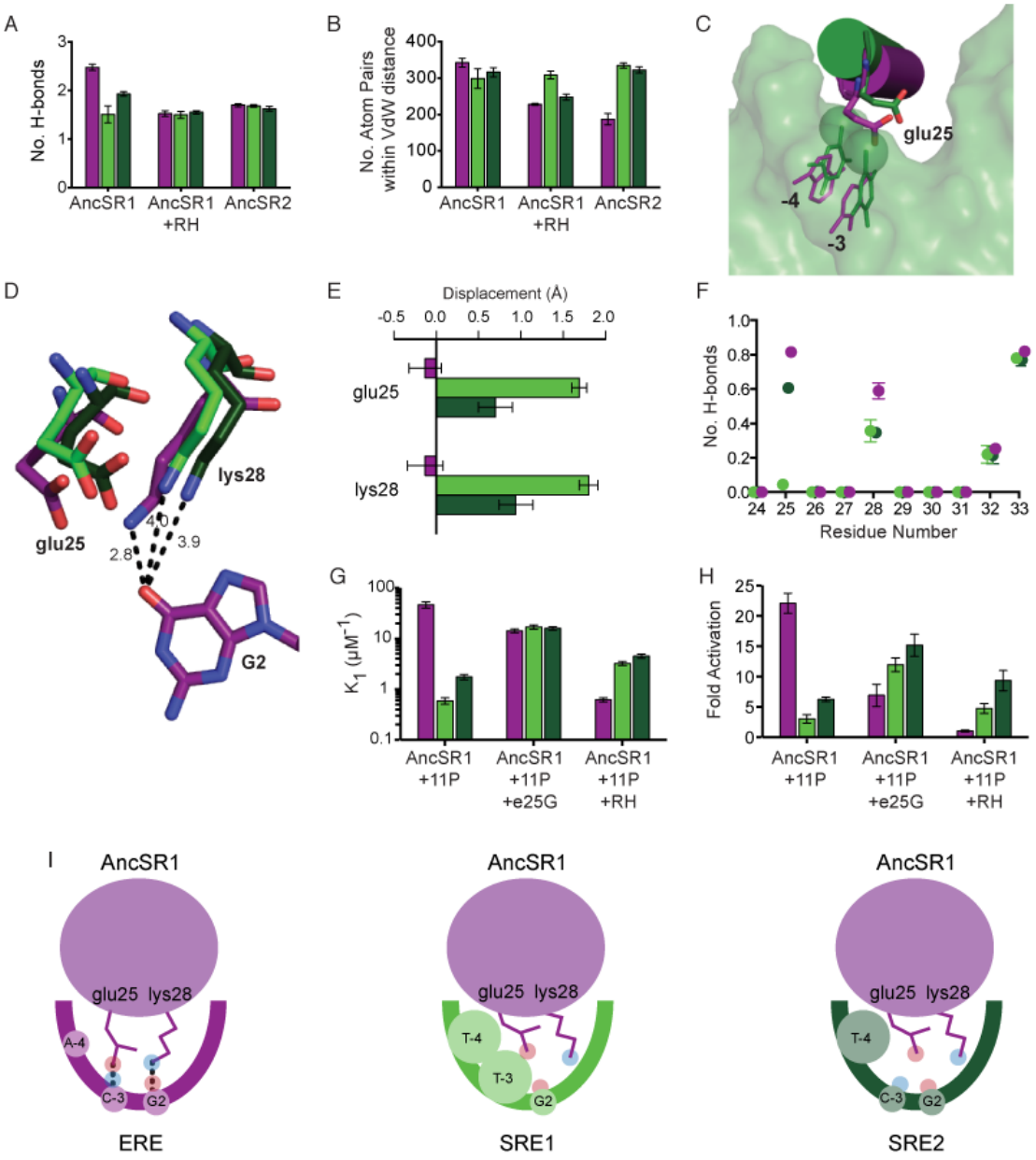
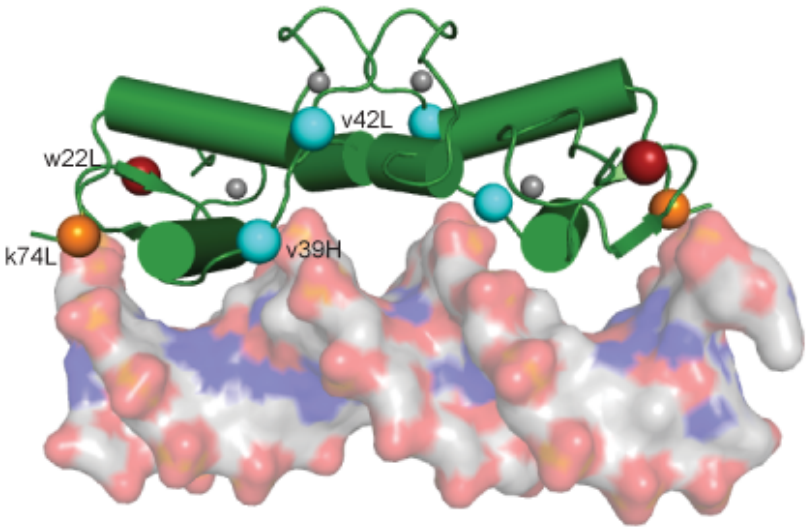

A
No. H-bonds
AncSR1
AncSR1 +RH
AncSR2
B
No. Atom Pairs within VdW distance
C
glu25
-4
-3
D
lys28
glu25
4.0
2.8
3.9
G2
E
Displacement (Å)
glu25
lys28
F
No. H-bonds
Residue Number
G
K_1 (μM^{-1})
AncSR1 +11P
AncSR1 +11P +e25G
AncSR1 +11P +RH
H
Fold Activation
I
AncSR1
glu25 lys28
A-4
C-3
G2
ERE
T-4
T-3
SRE1
SRE2

Figure 7 (next page). Permissive substitutions do not improve protein stability or dimerization in the absence of DNA
(A) Crystal structure of AncSR2 bound to SRE1. Sites of permissive substitutions are shown as Cα spheres; red, cyan, and orange indicate clustered groups of sites. Only one residue in the C-terminal group is shown). (B) Permissive substitutions (11P) do not increase protein stability. ΔG_{H2O}, calculated Gibbs free energy of chemically induced unfolding; m, slope of the unfolding transition; C_M, denaturant concentration at which 50% of protein is folded. (C,D) Permissive substitutions do not increase protein dimerization in the absence of DNA, measured by analytical ultracentrifugation. Distribution (C) and best-fit values (D) of sedimentation velocity coefficients ($S_{20,w}$) for AncSR1 (left) or AncSR1+11P (right) at 0.5 mM. The fraction of the total signal under the dominant peak (% total), the estimated molecular weight of that peak (MW) and the expected molecular weight of the monomeric protein (MW_{theo}) show that AncSR1 and AncSR2 are both predominantly monomeric. RMSD, root mean square deviation of the data from the model; f/f_0, total shape asymmetry. Signal at higher MW peaks may reflect aggregation due to high protein concentration.

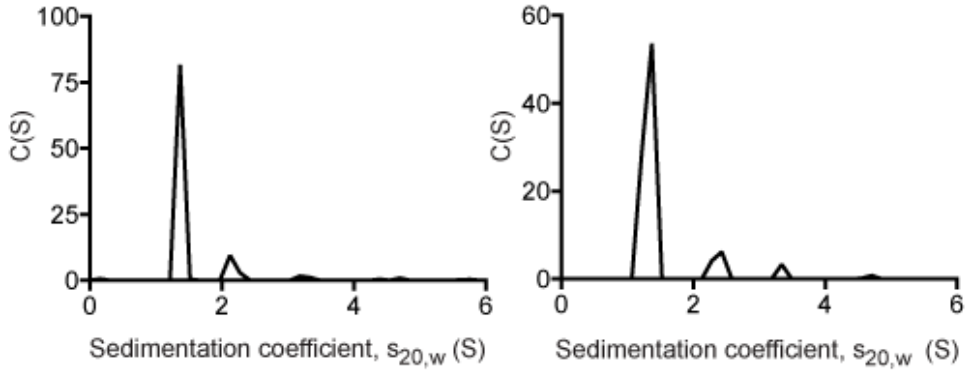
A



B

	ΔG_{H2O} (kcal/mol)	m (kcal/mol)	C_M (M)
AncSR1	-5.1 ± 0.1	1.53 ± 0.09	3.50 ± 0.2
AncSR1+11P	-5.1 ± 0.1	1.87 ± 0.01	2.73 ± 0.02
AncSR1+RH	-4.7 ± 0.1	1.65 ± 0.05	3.00 ± 0.04
AncSR1+RH+11P	-5.0 ± 0.1	1.99 ± 0.01	2.54 ± 0.00

C



D

	$s_{20,w}$	MW (kDa)	MW_{theo} (kDa)	% total	RMSD	f/f_0
AncSR1	1.364	10.5	9.62	83.7	0.013	1.3
AncSR1+11P	1.373	10.4	9.48	85	0.011	1.31

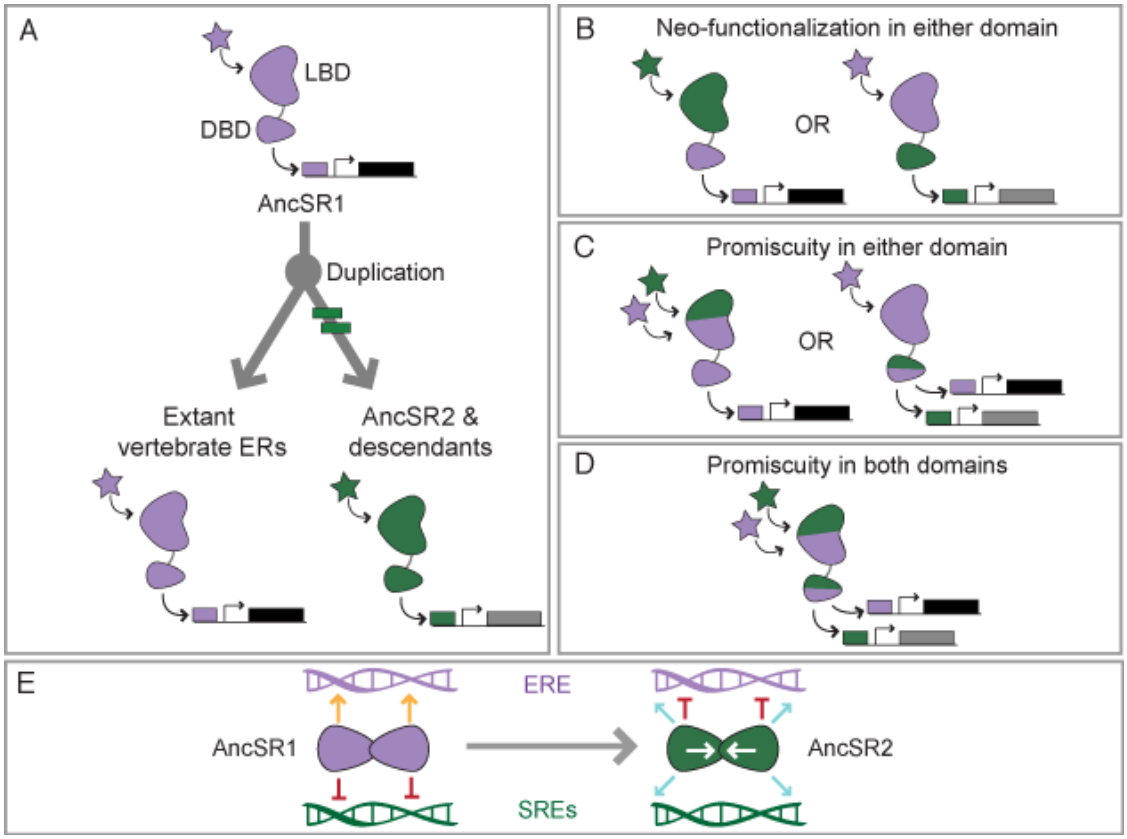


Figure 8. Evolution of a new regulatory module
(A) After duplication of AncSR1, the ancestral specificity for estrogens (purple stars) and ERE (purple box) was maintained to the present in the ER lineage. In the lineage leading to AncSR2, ancestral specificity for both DNA and hormone was lost, and novel sensitivity evolved for SREs (green box) and nonaromatized steroids (green star). A new set of target genes (light grey) was thus activated in response to different stimuli. Green hashes mark the branch on which these events occurred. (B-D) Other potential evoutionary trajectories for evolving new functions would interfere with the ancestral signaling network. (B) Evolution of new specificity for DNA or ligand would cause activation of old targets by new stimuli, or activation of new targets in response to ancestral stimuli. (C-D) Evolution of promiscuity in one or both domains would cause similar effects. (E) The shift in specificity from ERE (purple helices) to SREs (green helices) in AncSR2 involved losing favorable interactions (orange arrows) to ERE, losing unfavorabl negative interactions (red bars) to SRE, and gaining unfavorable interactions to ERE. Offsetting the loss of positive interactions in the DNA major groove, AncSR2 evolved favorable non-specific DNA contacts (blue arrows) and protein-protein interactions (white arrows in dimer interface) that increased cooperativity.

APPENDIX C

SUPPLEMENTAL INFORMATION FOR CHAPTER III

Figure S1 (next page). Inference of the ML steroid receptor phylogeny and reconstruction of AncSR1 and AncSR2 with high confidence; related to Figure 3
Tree is based on alignment of 213 steroid receptors and related sequences (Eick et al., 2012). Nodal support is indicated by likelihood ratio statistics and chi-squared values. Cyclostome sequences (cyan and red) were rearranged relative to the ML tree to minimize the number of gene duplication events. AncSR1 (purple) is the ancestor of all SRs and AncSR2 (green) is the ancestor of all PAMGRs. Ancestors were reconstructed with high confidence. Insets: Histograms for the distribution of posterior probabilities for (A) AncSR1 and (B) AncSR2. ERα/β- estrogen receptor α/β; PRs- progestagen receptors; ARs, androgen receptors; MRs, mineralocorticoid receptors; GRs, glucocorticoid receptors; ERRs, estrogen-related receptors; SF1, steroidogenic factor 1 receptors; RXR, retinoid X receptor; COUP-TFs, chicken ovalbumin upstream promoter transcription factors.

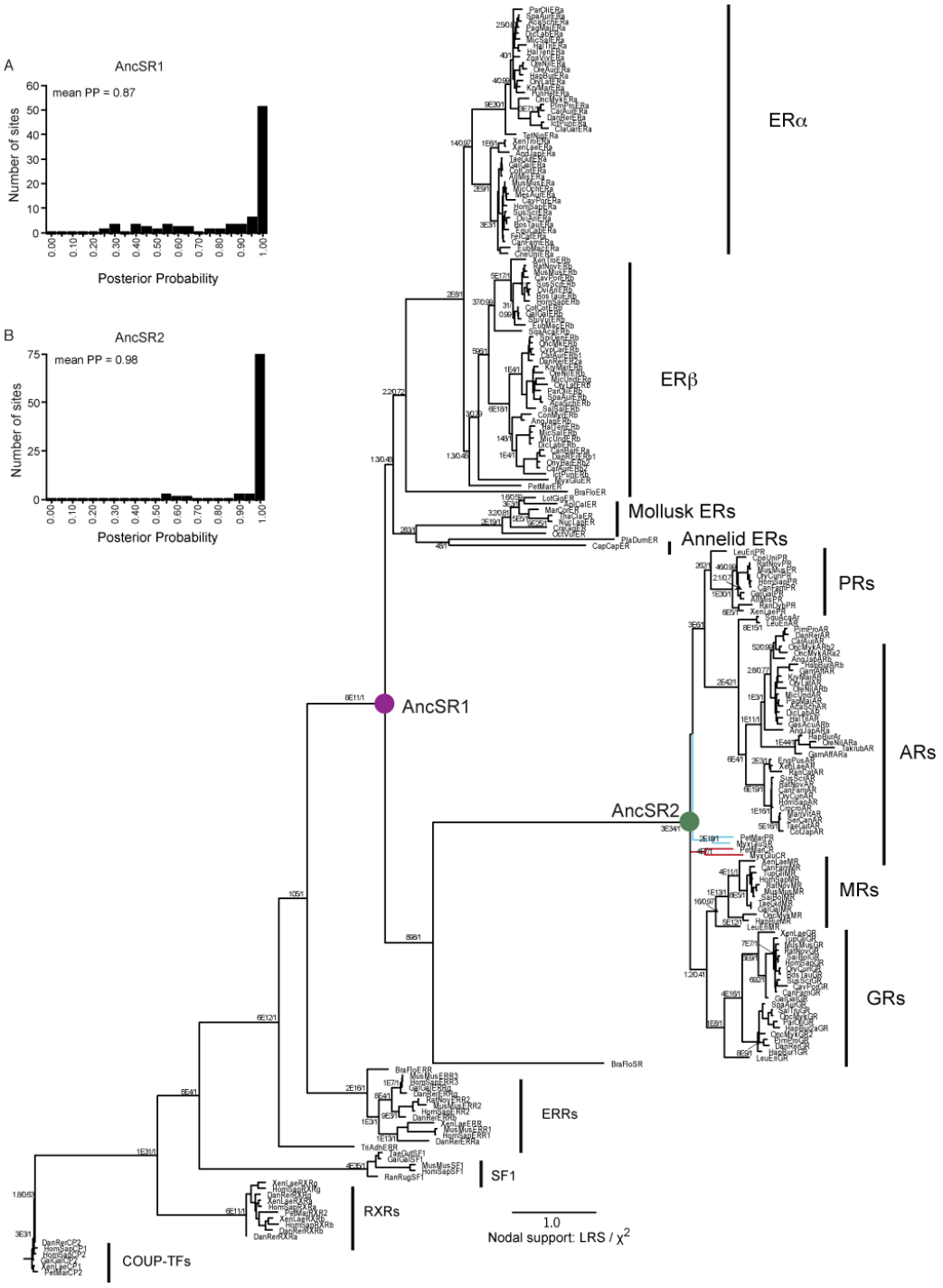
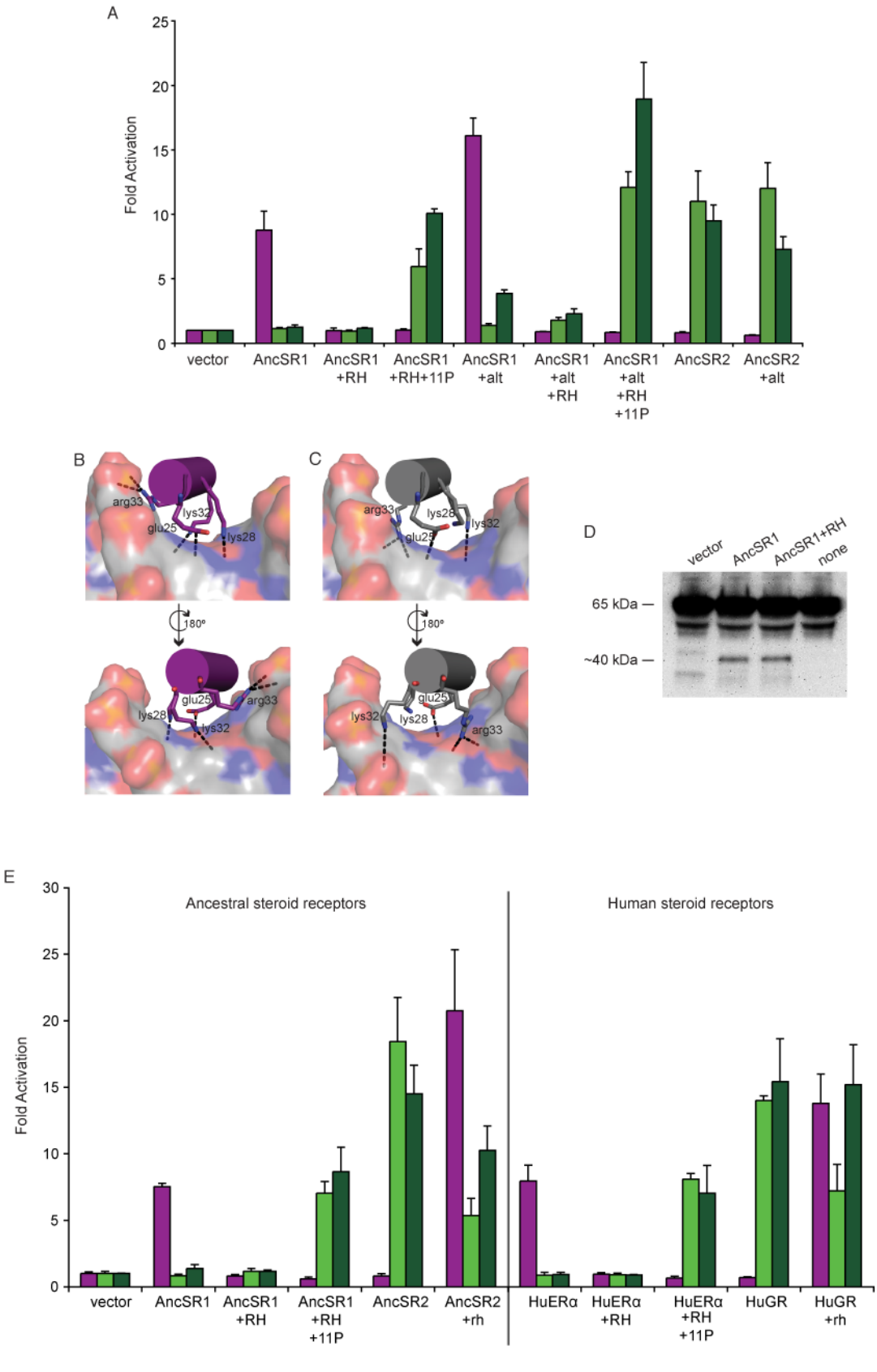

A
AncSR1
mean PP = 0.87
Number of sites
Posterior Probability
B
AncSR2
mean PP = 0.98
Number of sites
Posterior Probability
ERα
ERβ
Mollusk ERs
Annelid ERs
PRs
ARs
MRs
GRs
AncSR1
AncSR2
ERRs
SF1
RXRs
COUP-TFs
1.0
Nodal support: LRS / χ2

Figure S2 (next page). Functions of recognition helix and permissive substitutions identified using AncSR1 and AncSR2 are robust to uncertainty and their effects persist in present day human receptors; related to Figure 5

(A) Specificities of ancestors and intermediates are robust to uncertainty in the reconstruction. Reconstructions containing all alternate residues with posterior probability > 0.2 (+alt) have the same function as maximum likelihood ancestors. Derived groups of function-switching substitutions (RH, 11P) produce the same functional shifts in alternate states ancestors. (B-C) Reversal of the ancestral RH in the derived background nearly completely recapitulates the molecular interactions at the protein-DNA interface of the ancestral complex. Comparison of the protein-DNA interfaces of (B) AncSR1 bound to ERE and (C) AncSR2+rh bound to ERE. glu25 and lys28 have conserved hydrogen bonding partners. Favorably polar interactions between protein and DNA are drawn as dashed black lines. (D) The derived RH does not alter protein expression in the cell reporter assay. Western blot using NFκB antibody to detect the DBD+NFκB activation domain fusion construct shows: native full-length NFκB (~65 kDa) in non-transfected cells (none); truncated NFκB activation domain (band below 40kDa) in vector only control (vector); DBD-fusion protein (~40 kDa) in cells transfected with AncSR1 and AncSR1+RH, with no detectable differences between AncSR1 and AncSR1+RH. (E) Activation assays show that ancestors allowed for determination of residues important for observed DNA specificity of human steroid receptors. RH, recognition helix; 11P, 11 permissive substitutions; HuERα, human estrogen receptor α, HuGR, human glucocorticoid receptor. Lower-case letters, ancestral state; upper case, derived state. For all bar graphs: Purple, ERE; light green, SRE1; dark green, SRE2; error bars, ± SEM of three replicate experiments.


A
Fold Activation
25
20
15
10
5
0
vector
AncSR1
AncSR1 +RH
AncSR1 +RH+11P
AncSR1 +alt
AncSR1 +alt +RH
AncSR1 +alt +RH +11P
AncSR2
AncSR2 +alt
B
arg33
lys32
glu25
lys28
180°
glu25
arg33
lys28
lys32
C
arg33
lys28
lys32
glu25
180°
glu25
lys32
lys28
arg33
D
vector
AncSR1
AncSR1+RH
none
65 kDa —
~40 kDa —
E
Ancestral steroid receptors
Human steroid receptors
Fold Activation
30
25
20
15
10
5
0
vector
AncSR1
AncSR1 +RH
AncSR1 +RH +11P
AncSR2
AncSR2 +rh
HuERα
HuERα +RH
HuERα +RH +11P
HuGR
HuGR +rh

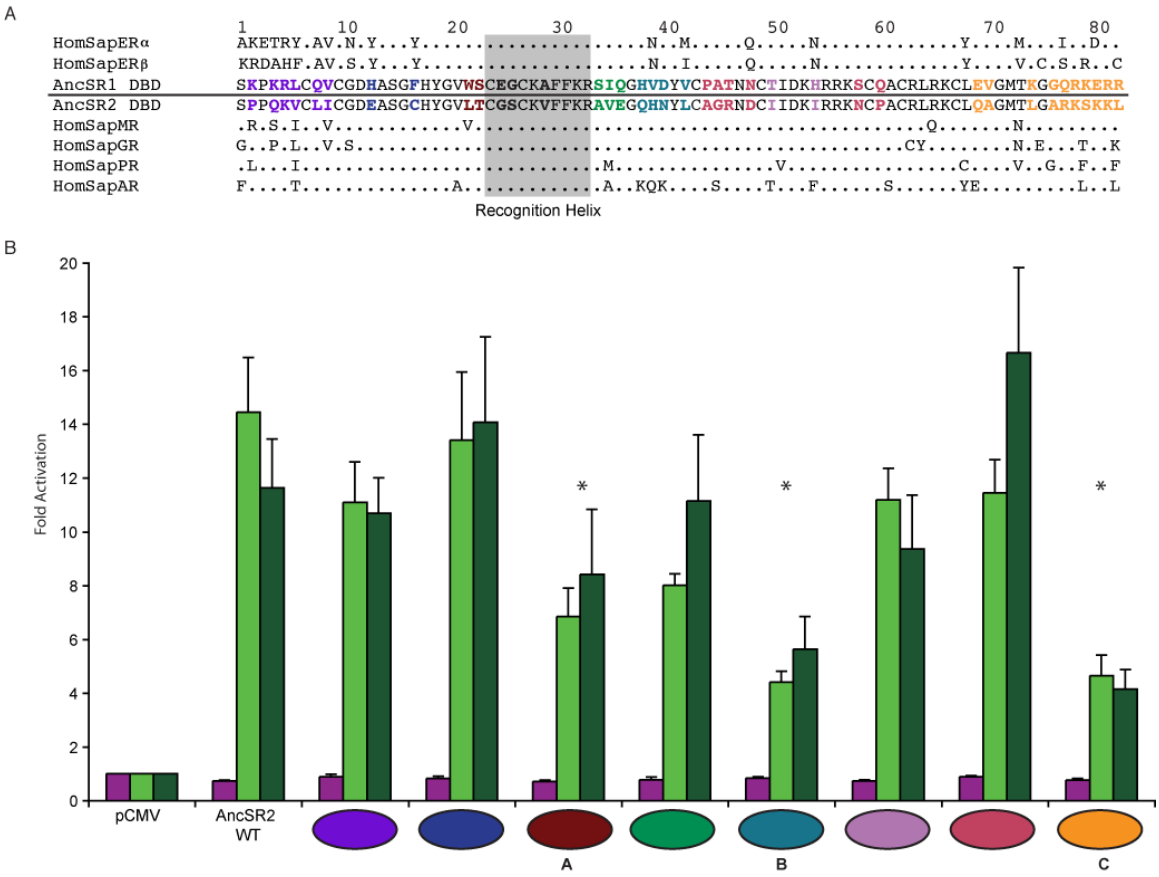


Figure S3: Three groups of permissive substitutions are required to support the derived specificity; related to Figure 5. (A) Alignment of ancestral and human DBDs shows amino acid differences; residues that are conserved between human DBDs and their closest ancestral sequence are indicated by '.' In addition to the RH substitutions, 35 substitutions occurred on the interval between AncSR1 and AncSR2. These substitutions were divided into 8 groups (indicated by color in the alignment) based on their contiguity in the linear sequence and tertiary structure. (B) Starting in AncSR2, each group was reverted to its ancestral state and assayed for cell reporter activation. A group containing permissive substitutions should result in a nonfunctional DBD when reverted to the ancestral state in the AncSR2 protein. Three groups (termed A, B and C, containing a total of 16 substitutions) had significantly reduced activation on SREs when reverted (indicated by *, $P<0.01$; see Table S5). Bar graph: Purple, ERE; light green, SRE1; dark green, SRE2. Error bars, ± SEM of three replicate experiments.

Figure S4 (next page). Three groups, totaling 11 substitutions, are sufficient for the protein to permit the derived RH; related to Figure 5

(A) Sequence alignment of AncSR1, potential permissive intermediates, and AncSR2. Colors indicate individual groups; 10 residues of the recognition helix are boxed gray. Recognition helix substitutions (^) and the narrowed set of permissives substitutions (*, referred to as 11P), are marked. (B) Sixteen substitutions, identified as supporting the derived RH by reversing groups of amino acids to their ancestral states in AncSR2 (see Supplemental Figure 3), were permissive for the derived function in AncSR1+RH (identified as AncSR1+RH+16P). These substitutions could be narrowed down to 13 and 11 without significant differences in function. (C) One of the two substitutions in group A (L22w) and two of the four members of group B (H39v, L42v) had statistically significant deleterious effects, indicating that necessary permissive substitutions occurred at these sites. Groups A and B could therefore be reduced to 1 and 2 substitutions respectively, narrowing the number of permissive substitutions to 13 (AncSR1+RH+13P). Two N-terminal members of group C (Q69e and A70v) could also be reversed, leaving a total of 11 substitutions that are sufficient to permit the derived RH (AncSR1+RH+11P). Decisive resolution of smaller set of permissive substitutions in group C is not possible because alignment of this region is ambiguous. Stars (*) indicate significant difference, $P<0.01$, from AncSR2 (see Supplemental Table 5). (D) All three groups of permissive substitutions are necessary for the fully permissive effect in cell reporter assays. For all bar graphs: Purple, ERE; light green, SRE1; dark green, SRE2. Values are average ± SEM of three replicate experiments.

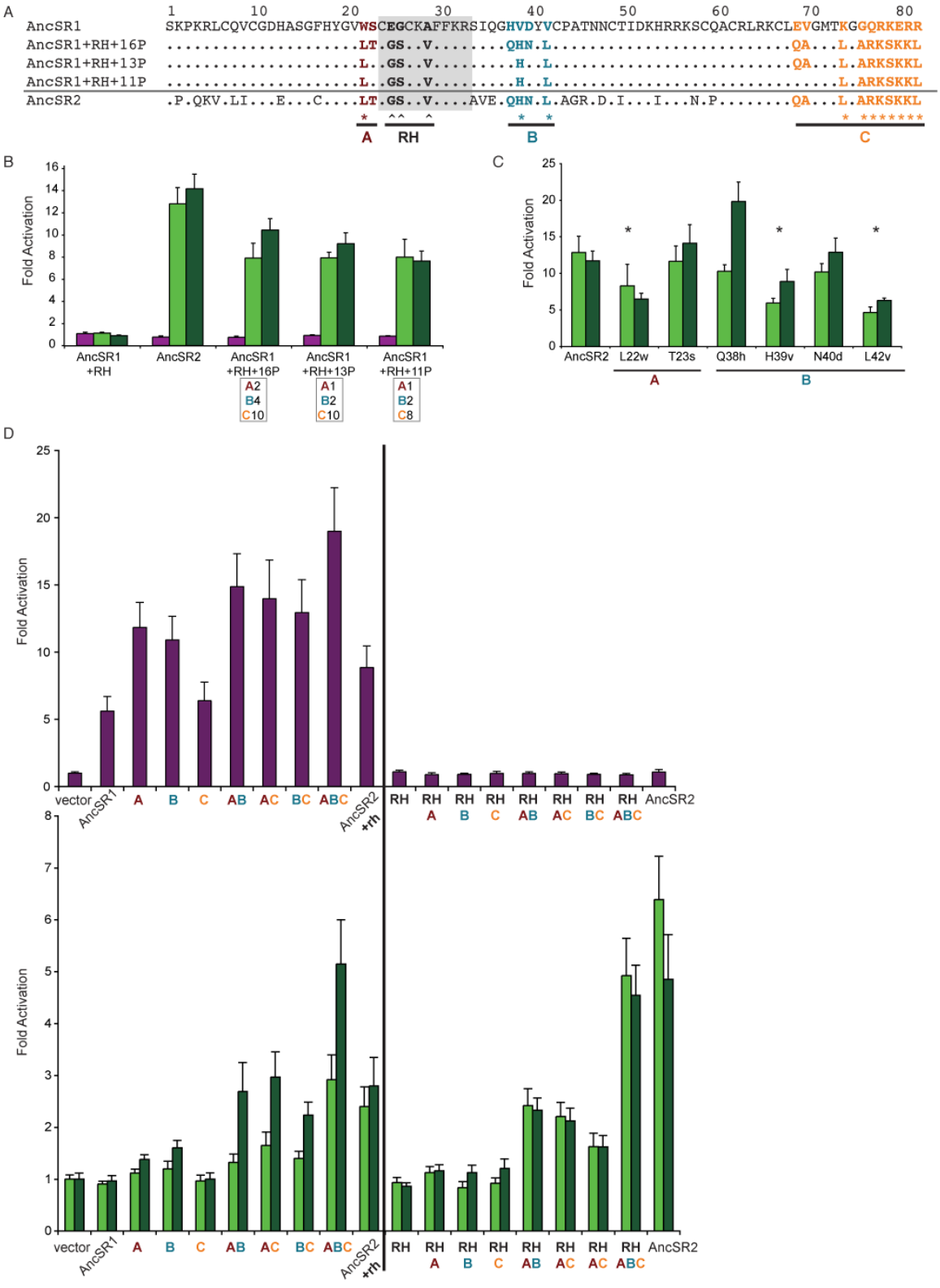
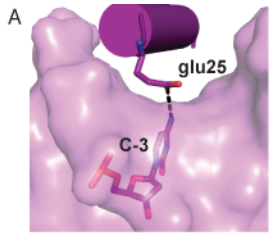
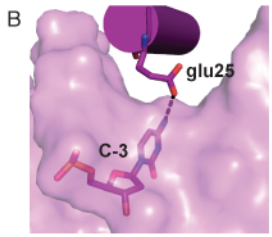
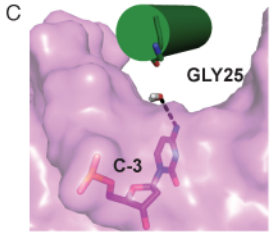
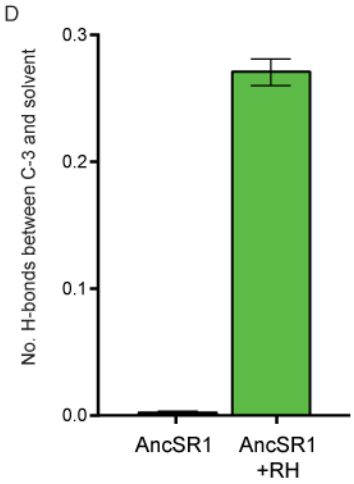

A
AncSR1 SKPKRLCQVCGDHASGFHYGVWSCEGCKAFFKRSIQGHVDYVCPATNNCTIDKHRRKSCQACRLRKCLEVGMTKGGQRKERR
AncSR1+RH+16P .....LT.GS..V.....QHN.L.....QA...L.ARKSKKL
AncSR1+RH+13P .....L..GS..V.....H..L.....QA...L.ARKSKKL
AncSR1+RH+11P .....L..GS..V.....H..L.....L.ARKSKKL
AncSR2 .P.QKV.LI...E...C....LT.GS..V....AVE.QHN.L.AGR.D.I...I...N.P.....QA...L.ARKSKKL
A
RH
B
C
B
Fold Activation
AncSR1 +RH
AncSR2
AncSR1 +RH+16P
A2 B4 C10
AncSR1 +RH+13P
A1 B2 C10
AncSR1 +RH+11P
A1 B2 C8
C
Fold Activation
AncSR2
L22w
T23s
Q38h
H39v
N40d
L42v
A
B
D
Fold Activation
vector
AncSR1
A
B
C
AB
AC
BC
ABC
AncSR2 +rh
RH
RH A
RH B
RH C
RH AB
RH AC
RH BC
RH ABC
AncSR2

Figure S5 (next page). The RH substitutions leave an unpaired hydrogen bond donor on ERE and yield no new SRE-specific hydrogen bonds; **related to Figure 6**
(A) In the crystal structure of AncSR1 bound to ERE, the ancestral glu25 accepts a hydrogen bond from C-3 of ERE. (B) This hydrogen bond also forms in MD simulations with AncSR1:ERE; a representative frame is shown. (C) The derived RH removes the hydrogen bond acceptor glu25, leaving C-3 unpaired; water molecules move into the interface and pair with C-3. A representative frame from AncSR1+RH:ERE simulation is shown. Potential hydrogen bonds between glu25 and water are dashed black lines. (D) Water penetration caused by RH substitutions. The average number of hydrogen bonds formed between C-3 base donor and solvent molecules in the presence of the ancestral (purple) and derived (green) RH; error is the SEM of three replicate MD simulations. (E) The RH substitutions do not increase hydrogen bonding on SREs. All hydrogen bonds from the RH residues to DNA in MD simulations were classified as homologous between complexes with and without the RH substitutions (involving the same donor and acceptor pair), unique to AncSR1 (not present in AncSR1+RH), or unique to AncSR1+RH (not present in AncSR1). Each hydrogen bond was weighted by its frequency of formation in each MD trajectory, and the average number of hydrogen bonds formed in each category across replicate trajectories was calculated. The RH substitutions eliminate some hydrogen bonds formed by AncSR1 to SREs and reduce the frequency of homologous bonds; they generate a single new hydrogen bond (from Ser26 to the protein backbone), which forms nonspecifically on all REs and is not sufficient to compensate for the loss of other hydrogen bonds.


A
glu25
C-3


B
glu25
C-3


C
GLY25
C-3


D
No. H-bonds between C-3 and solvent
0.3
0.2
0.1
0.0
AncSR1
AncSR1 +RH

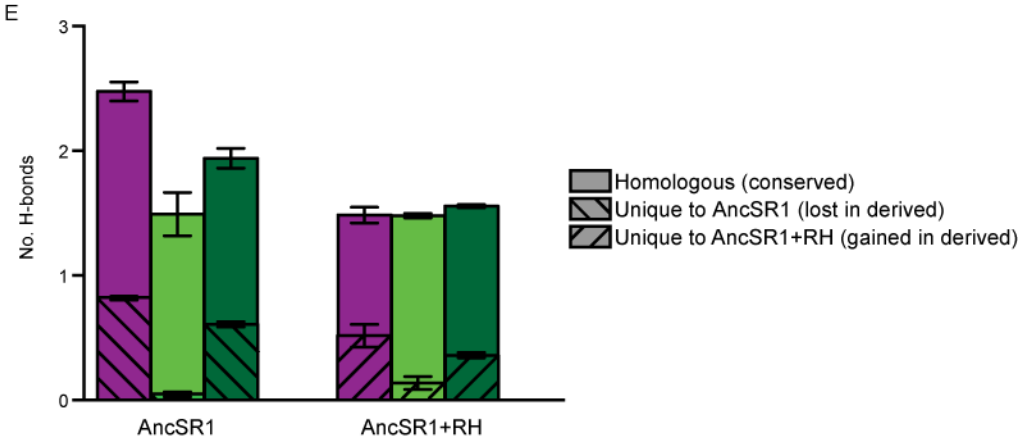

E
No. H-bonds
3
2
1
0
AncSR1
AncSR1+RH
Homologous (conserved)
Unique to AncSR1 (lost in derived)
Unique to AncSR1+RH (gained in derived)

Figure S6 (next page). The ancestral and derived RH exclude binding to non-target REs through negative interactions; related to Figure 6

(A) In MD simulations, SRE1-specific T-4 and T-3 add bulk into the DNA major groove relative to ERE. Overlay of the MD average positions of nucleotides -4 and -3 for ERE (purple) and SRE1 (green) when bound to AncSR1. Bulky methyls of T-4 and T-3 indicated by arrows. (B) Surface representation of ERE and SRE1 shows the more narrow major groove of SRE1 and the extra bulk of methyl groups of T-4 and T-3 (black arrows) fill in the major groove. Purple, ERE; green, SRE1. (C,D,E) In crystal structures, the steric interactions between glu25 and the SRE-specific T-4 forces glu25 to adopt an alternate conformation when bound to SRE1. (C) In the crystal structure of AncSR2+rh bound to ERE, the hydroxyl of glu25 points down into the major groove. When this crystal structure is aligned to the crystal structure of AncSR2+rh bound to SRE1, extra bulk is observed in the major groove of SRE1, but not in ERE. (D) If glu25 maintained the same conformation as when bound to ERE, it would sterically clash with the methyl of T-4 of SRE1. (E) In order to reduce this steric strain, glu25 adopts a different conformation when bound to SRE1. For crystal structure proteins: gray, AncSR2+rh bound to ERE; cyan, AncSR2+rh bound to SRE1. For DNA: purple, ERE; green, SRE1. (F,G,H) In MD simulations, the presence of unpaired electron acceptors on glu25 results in an influx of interfacial waters in the major groove when the ancestral RH is bound to SREs. (F) When AncSR1 is bound to ERE, glu25 makes hydrogen bonds with DNA and occasionally with solvent. (G) When AncSR1 is bound to SREs, glu25 is left unpaired, causing an influx of interfacial waters. Potential hydrogen bonds between glu25 and surrounding water molecules are dashed black lines. (H) glu25 is more solvent exposed when bound to SREs than when bound to ERE. For bar graphs: Purple, ERE; light green, SRE1; dark green, SRE2; values are average ± SEM for three replicate MD simulations. (I-K) The mechanisms for the sequence-specific negative effects of g26S and a29V are not obvious in crystal structures. Close-up of protein-DNA interactions for crystal structures of (I) AncSR1 bound to ERE and (J) AncSR2 bound to SRE1. The two RH substitutions, g26S and a29V, are shown as sticks; DNA is colored by element: N, blue, O, red; H, white; C, magenta (ERE) or green (SREs). (K) gly26 and ala29 do not pack preferentially on ERE. The number of DNA atom contacts within 4.5 Å of gly26 and ala29 were calculated for three replicate MD simulations of AncSR1 bound to ERE (purple), SRE1 (light green) and SRE2 (dark green); error bars are SEM.

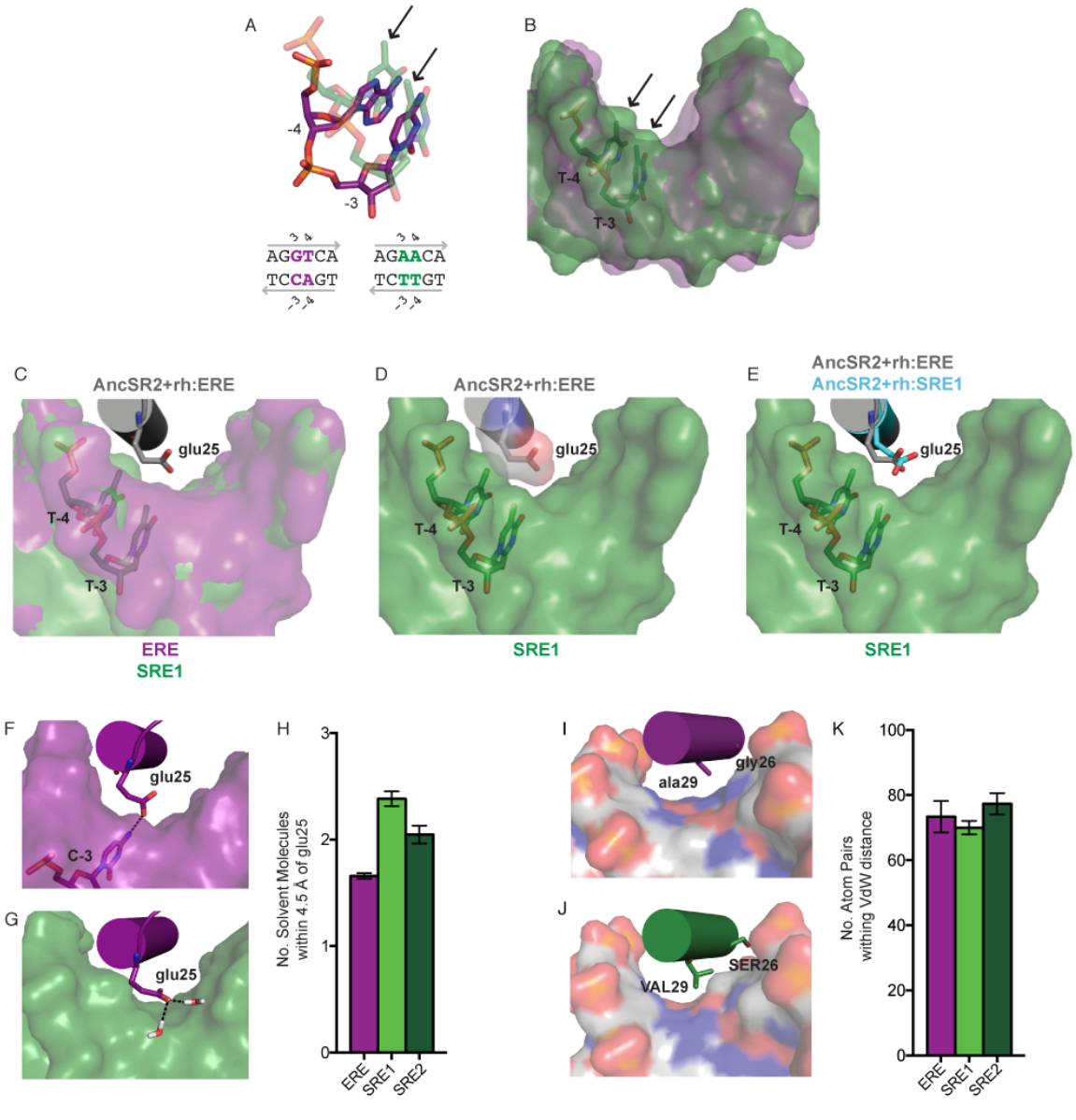

A
-4
-3
AGGTCA
TCCAGT
AGAACA
TCTTGT
B
T-4
T-3
C
AncSR2+rh:ERE
glu25
T-4
T-3
ERE
SRE1
D
AncSR2+rh:ERE
glu25
T-4
T-3
SRE1
E
AncSR2+rh:ERE
AncSR2+rh:SRE1
glu25
T-4
T-3
SRE1
F
glu25
C-3
G
glu25
H
No. Solvent Molecules within 4.5 Å of glu25
ERE
SRE1
SRE2
I
gly26
ala29
J
SER26
VAL29
K
No. Atom Pairs withing VdW distance
ERE
SRE1
SRE2

Table S1: Posterior probabilities for each amino acid residue of AncSR1 and AncSR2 DBDs; related to Figure 3. Alternate states and their posterior probabilities are shown. Plausible alternate states with PP>0.2, highlighted in green, were included in the alternate reconstructions (AncSR1+alt and AncSR2+alt) in Figure S2. For both the maximum likelihood and alternate reconstruction containing all plausible alternate states, the mean posterior probability across sites is shown, as is the expected number of errors in the sequence, calculated as one minus the posterior probability of the incorporated state at each site, summed over all sites.

AncSR1					*AncSR2*				
Position	ML state	Posterior probability	Alternate state	Posterior probability	Position	ML state	Posterior probability	Alternate state	Posterior probability
1	S	0.399	T	0.235	1	S	0.911	A	0.057
2	K	0.228	R	0.218	2	P	0.962	S	0.029
3	P	0.282	A	0.115	3	P	1		
4	K	0.61	T	0.182	4	Q	0.984	H	0.016
5	R	0.891	Q	0.053	5	K	1		
6	L	0.572	F	0.2	6	V	0.603	I	0.35
7	C	1			7	C	1		
8	Q	0.305	A	0.298	8	L	1		
9	V	0.999	I	0.001	9	I	0.992	V	0.008
10	C	1			10	C	1		
11	G	0.796	S	0.124	11	G	0.982	S	0.017
12	D	1			12	D	1		
13	H	0.534	N	0.125	13	E	1		
14	A	1			14	A	1		
15	S	1			15	S	1		
16	G	1			16	G	1		
17	F	0.936	Y	0.061	17	C	1		
18	H	1			18	H	1		
19	Y	1			19	Y	1		

20	G	1			20	G	1		
21	V	1			21	V	1		
22	W	0.669	L	0.177	22	L	0.999	I	0.001
23	S	0.998	A	0.001	23	T	1		
24	C	1			24	C	1		
25	E	1			25	G	1		
26	G	1			26	S	1		
27	C	1			27	C	1		
28	K	1			28	K	1		
29	A	1			29	V	1		
30	F	1			30	F	1		
31	F	1			31	F	1		
32	K	1			32	K	1		
33	R	1			33	R	1		
34	S	0.844	A	0.079	34	A	1		
35	I	0.994	V	0.005	35	V	0.929	I	0.07
36	Q	0.999			36	E	1		
37	G	0.999			37	G	1		
38	H	0.396	P	0.222	38	Q	1		
39	V	0.549	I	0.22	39	H	1		
40	D	0.899	E	0.06	40	N	1		
41	Y	1			41	Y	1		
42	V	0.727	I	0.19	42	L	1		
43	C	1			43	C	1		
44	P	1			44	A	1		
45	A	1			45	G	1		
46	T	0.968	N	0.025	46	R	1		
47	N	1			47	N	1		
48	N	0.933	D	0.025	48	D	1		
49	C	1			49	C	1		
50	T	0.934	I	0.018	50	I	1		

51	I	1			51	I	1			
52	D	1			52	D	1			
53	K	0.983	R	0.017	53	K	1			
54	H	0.584	R	0.305	54	I	1			
55	R	1			55	R	1			
56	R	1			56	R	1			
57	K	1			57	K	1			
58	S	0.994	N	0.006	58	N	1			
59	C	1			59	C	1			
60	Q	0.999	P	0.001	60	P	1			
61	A	1			61	A	1			
62	C	1			62	C	1			
63	R	1			63	R	1			
64	L	0.854	F	0.145	64	L	1			
65	R	0.957	K	0.03	65	R	1			
66	K	1			66	K	1			
67	C	1			67	C	1			
68	L	0.666	F	0.277	68	L	0.655	I	0.179	
69	E	0.909	D	0.04	69	Q	1			
70	V	0.997	I	0.002	70	A	1			
71	G	1			71	G	1			
72	M	1			72	M	1			
73	T	0.422	M	0.346	73	T	0.534	V	0.365	
74	K	0.95	R	0.046	74	L	1			
75	G	0.836	E	0.14	75	G	1			
76	G	0.991	S	0.005	76	A	1			
77	Q	0.286	R	0.244	77	R	1			
78	R	0.998	K	0.002	78	K	1			
79	K	0.459	R	0.313	79	S	0.549	L	0.412	
80	E	0.497	D	0.492	80	K	1			
81	R	0.991	K	0.009	81	K	1			

82	**R**	**0.437**	**K**	**0.36**	**82**	**L**	**0.912**	M	0.033
Mean PP (ML)		**0.88**					**.98**		
Mean PP (Alt-all)				**0.86**					**0.97**
Expected errors (ML)		**10.2**					**2.0**		
Expected errors (Alt-all)				**11.8**					**2.5**

Table S2: SRs in which plausible alternate ancestral amino acids are found; related to Figure 1. For ambiguously reconstructed sites in AncSR1 (top) and AncSR2 (bottom), the ML and next-most-likely (alternate) state are shown. X denotes that the alternate state is present in one or more extant members of the clade. Clades containing members known to recognize ERE-like sequences are shown in purple; those that recognize SRE-like sequences are shown in green. Asterisk denotes that lamprey and hagfish co-orthologs have been placed in these groups. Plausible alternate reconstructions are defined as having posterior probability > 0.20.

AncSR1 site	ML state	Alternate state	Vertebrate ERs	Protostome ERs	Cephalochordate ER	ERRs	Cephalochordate SR	SR	ARs	PRs*	GRs	MRs*
1	S	T	X			X				X	X	X
2	K	R	X									X
8	Q	A	X					X				
38	H	P		X								
39	V	I				X						
54	H	R				X		X				
68	L	F	X						X			
73	T	M	X	X		X						X
77	Q	R			X				X	X	X	X
79	K	R	X	X								
80	E	D	X	X		X						
82	R	K	X	X				X				

AncSR2 site	ML state	Alternate state	Vertebrate ERs	Protostome ERs	Cephalochordate ER	ERRs	Cephalochordate SR	SR	ARs	PRs	GRs	MRs
6	V	I								X	X	X
73	T	V	X							X		
79	S	L				X			X	X	X	X

Table S3: Macroscopic binding affinity ($K_{A,mac}$), half-site affinity (K_1) and cooperativity (ω) were calculated for each protein construct using fluorescence polarization assays; related to Figure 3, and Figures 5-6. Values were calculated by a global fit of half-site and palindromic-site binding data using a two-site cooperative binding model.

	ERE		SRE1		SRE2	
$K_{A,mac}$ (μM^{-2})	**Mean**	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**
AncSR1	118.57	0.14	0.06	0.28	0.66	0.30
AncSR2	0.28	0.26	12.15	0.25	23.28	0.22
AncSR2+rh	3.18	0.25	0.09	0.21	0.43	0.17
AncSR1+RH	0.07	0.17	0.81	0.31	1.88	0.24
AncSR1+11P	20243.35	0.30	32.19	0.33	257.10	0.28
AncSR1+RH+11P	5.27	0.25	637.24	0.23	936.77	0.22
K_1 (μM^{-1})	**Mean**	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**
AncSR1	7.18	0.29	0.14	0.01	0.28	0.02
AncSR2	0.23	0.01	0.86	0.04	0.91	0.04
AncSR2+rh	0.44	0.02	0.04	0.00	0.08	0.00
AncSR1+RH	0.22	0.01	0.43	0.02	0.55	0.02
AncSR1+11P	46.33	1.70	0.59	0.04	1.75	0.10
AncSR1+RH+11P	0.62	0.03	3.23	0.15	4.50	0.21
AncSR1+11P+Gga	16.11	1.28	3.71	0.04	7.66	0.28
ω	**Mean**	**SEM**	**Mean**	**SEM**	**Mean**	**SEM**
AncSR1	2.30	0.14	2.84	0.55	8.52	1.58
AncSR2	5.25	0.94	16.53	2.54	27.89	3.66
AncSR2+rh	16.41	2.47	59.51	7.30	68.93	6.76
AncSR1+RH	1.40	0.15	4.36	0.91	6.20	0.91
AncSR1+11P	9.43	2.12	92.37	17.60	83.57	13.60
AncSR1+RH+11P	13.72	2.14	61.08	8.11	46.26	5.97

Table S4: Crystal structure refinement statistics (molecular replacement); related to Figure 4 and Experimental Procedures for Chapter III.

	AncSR1:ERE	AncSR2:SRE1	AncSR2+rh:ERE	AncSR2+rh:SRE
Data Collection				
Space group	C2	$P2_1$	$P2_1$	$P2_1$
Cell dimensions				
a (Å)	97.2	47.5	48.3	47.8
b (Å)	36.4	80.4	79.8	80.5
c (Å)	90.9	116.6	116.8	115.9
α (°)	90.0	90.0	90.0	90.0
β (°)	90.0	90.0	90.0	90.0
γ (°)	121.6	96.7	96.8	96.4
Resolution range (Å)	41.40-1.50 (1.53-1.50)*	30.60-2.70 (2.80-2.70)*	37.60-2.25 (2.33-2.25)*	29.12-2.35 (2.43-2.35)*
R_{sym} (%)	10.2 (29.6)	8.20 (35.6)	9.70 (78.6)	15.4 (57.2)
I / sI	32.5 (2.7)	19.8 (2.4)	13.1 (2.1)	3.4 (2.0)
Completeness (%)	83.3 (32.7)	97.9 (82.3)	99.2 (95.4)	97.7 (89.3)
Redundancy	3.7 (1.9)	3.5 (2.5)	3.7 (3.5)	3.3 (2.2)
Refinement				
Wilson B-factor	15.8	46.7	44.9	66.6
Resolution (Å)	1.50**	2.7	2.25	2.35
No. reflections	36436	23265	41533	34761
R_{work} / R_{free} (%)	17.5 (20.6)	19.1 (23.2)	18.6 (21.6)	19.87 (23.1)
No. atoms	2155	3685	3771	3688
Macromolecules	1852	3624	3631	3666
Water	298	53	132	22
B-factors	30.4	51.1	55.4	81.6
Macromolecules	29.1	51.3	55.6	81.7
Water	38.5	39.8	52.9	69.1
R.m.s.				

deviations				
Bond lengths (Å)	0.006	0.004	0.004	0.006
Bond angles (°)	0.97	0.77	0.67	0.93

*Data collected from a single crystal; values in parentheses are for highest resolution shell.

**After molecular replacement, all data was used in refinement since its inclusion improved map quality with no detrimental impact on model quality.

Table S5: T-tests to identify permissive substitutions; related to Figure 5 and Experimental Procedures for Chapter III. Statistical analysis of results shown in Figure S3B, and Figure S5C. *, genotypes that are significantly different from AncSR2 after Bonferroni correction.

Genotype	Mean Fold Activation of SRE1 and SRE2	P-value
AncSR2	13.28	--
Purple	11.27	0.209
Blue	13.10	0.85
Red (A)	8.15	1e-4*
Green	10.01	0.016
Teal (B)	5.14	2e-7*
Lavender	11.01	0.039
Pink	14.67	0.691
Orange (C)	4.00	2e-7*
AncSR2	12.35	--
Red (A)	4.30	2e-7*
L22w	7.05	6e-4*
T23s	12.74	0.776
Teal (B)	3.16	9e-9*
Q38h	14.51	0.121
H39v	7.25	3e-4*
N40d	11.39	0.493
L42v	5.39	1e-6
Orange (C)	3.32	2e-8
Q69e, A70v	5.92	5e-5

Table S6: Custom terms used in molecular dynamics simulations. (A) Zn-Cys interactions terms. (B) Partial charges for Cys and Zn atoms within each zinc finger; related to Experimental Procedures.

A

Atoms	**Interaction**	**Value**	**Reference**
Zn	VDW	σ = 1.10 Å ε = 0.0125 kcal/mol	Hoops, Anderson and Merz 1991
S-Zn	length	2.26 Å	Lin and Wang, 2010
	energy	92.8 kcal/mol	
Zn-S-CT	angle	104.90°	
	energy	75.2 kcal/mol	
S-Zn-S	angle	129.12°	
	energy	21.6 kcal/mol	
CT-S-Zn-S	dihedral	0°	
	energy	0 kcal/mol	

B

Atom	**Partial Charge**	**AMBER atom type**
N	-0.41570	N
H	0.27190	H
CA	-0.01819	CT
HA	-0.03191	H1
CB	0.36673	CT
HB1	-0.07039	H1
HB2	-0.07039	H1
SG	-0.84046	S
C	0.59730	C
O	-0.56790	O
Zn	1.11604	Zx*

*custom atom type

EXTENDED EXPERIMENTAL PROCEDURES

Phylogenetics and ancestral sequence reconstruction Annotated protein sequences for nuclear receptors were downloaded from UniPROTKB/TrEMBL, GenBank, the JGI genome browser, and Ensemble (Eick et al., 2012). To reconstruct the DBD of both AncSR1 and AncSR2, 213 steroid and related receptor sequences (both DNA binding and ligand binding domains with hinge removed) were aligned using the Multiple Sequence Alignment by Log-Expectation (MUSCLE) program (Edgar, 2004). The alignment was checked to ensure alignment of the nuclear receptor AF-2 domain and manually edited to remove lineage-specific indels. The ML phylogeny was inferred from the alignment using PHYML v2.4.5 (Guindon et al., 2010) and the Jones-Taylor-Thornton model with gamma-distributed among-site rate variation and empirical state frequencies, which was the best-fit evolutionary model selected using the Akaike Information Criterion implemented in PROTTEST software. Statistical support for each node was evaluated by obtaining the approximate likelihood ratio (the likelihood of the best tree with the node divided by the likelihood of the best tree without the node) and chi-squared confidence statistic derived from that ratio (Anisimova and Gascuel, 2006). AncSR1 and AncSR2 DBDs were reconstructed by the maximum likelihood method (Yang et al., 1995) on a single-branch rearrangement of the ML phylogeny that requires fewer gene duplications and losses to explain the distribution of SRs in agnathans and jawed vertebrates using Lazarus software (Hanson-Smith et al., 2010), assuming a free eight-category gamma distribution of among-site rate variation and the Jones-Taylor-Thornton protein model. Average probabilities were calculated across all DBD sites.

Luciferase reporter activation assay

DBDs of both ancestral and human receptors were cloned into the mammalian expression vector pCMV-AD (Stratagene), and fused in-frame with the NF-κB activation domain. Response element plasmids were modified versions of the plasmid pGL3-4(EREc38), gift from C. Klinge (Tyulmenkov et al., 2000), which contains 4 copies of the estrogen receptor recognition sequence upstream of a luciferase reporter gene. All other response elements were designed to replace each ERE half site (AGGTCA) with the alternate half-site. For example SRE1-luc was made by introducing the AGAACA half sites. These alternate response elements were synthesized by Blue Heron Biotechnology and then cloned into the pGL3-4(EREc38) plasmid.

These plasmids were then transfected into CV-1 cells (ATCC cat#CCL-70), which were restarted from frozen stocks of early passages frequently, as follows. A mix containing: 20ng of DBD plasmid, 20 ng response element containing luciferase reporter plasmid, 2ng of phRLtK plasmid for normalization, and 80 ng PUC19 plasmid (filler DNA) complexed with Lipofectamine and Plus reagents (Life Technologies) was added to each well of a 96 well plate, incubated for 4 hours and the transfection mixture was replaced with charcoal stripped DMEM supplemented with 10% fetal bovine serum. The ratio of DBD to reporter plasmid was optimized to ensure that activation was in the linear range for both high and low activation constructs. After 24 hours, luciferase production was measured using the Dual-Glo luciferase kit (Promega). Mutants were generated using site-directed mutagenesis (QuikChange Lightening, Stratagene), and all clones were verified by sequencing (Genewiz, Inc).

Statistical analysis of reporter activation assays

To determine which amino acids were required to permit the RH substitutions we designed experiments to be analyzed statistically using analysis of variance (ANOVA). Dual-luciferase reporter assays were performed using AncSR2 "wild-type" and mutant genotypes in which historical substitutions were reversed to the ancestral states on ERE, SRE1, and SRE2. Each condition was assayed in triplicate, and each experiment was performed independently three times. A Shapiro-Wilk W test found no evidence for deviation from normality, so we used a fully factorial ANOVA to analyze the effects of RE and genotype on activation. Activation of ERE was significantly different from both SRE1 and SRE2 (p=0.0007 and 0.005, respectively, using an all pairs Tukey-Kramer HSD), but there was no significant difference between activation of SRE1 and SRE2 (p=0.95). The ANOVA indicated a significant effect of mutant genotypes on activation (p<0.0001), so we performed t-tests to identify mutant genotypes with significant effects on activation of the SREs (combined) relative to the wild-type AncSR2 control. Mutations with p<0.01 were considered to be significantly different.

Western blots

CV-1 cells were grown in 6 well plates, transfected with DBD containing plasmids, and grown for 40 hours. Cells were lysed in RIPA buffer containing protease inhibitors (Santa Cruz Biotechnology, Inc cat #sc-24948), and proteins were quantitated using Bio-Rad protein assay (cat#500-0006). Twenty μg of protein was separated on a 12% acrylamide gel and transferred to PVDF membrane (Bio-Rad cat# 162-0175). Ancestral proteins were visualized by western blot using an antibody against the fused

NF-κB activation domain, diluted 1:500 (Santa Cruz Biotechnology, cat# sc-372) and goat-anti-rabbit HRP conjugated secondary diluted 1:10,000 (sc-2004), with Luminol chemiluminescent reagent [Santa Cruz (sc-2048)].

Protein purification

DBDs were cloned into the pETMALc-H10T vector (Pryor and Leiting, 1997) (a gift from John Sondek, UNC-Chapel Hill) C-terminal to a cassette containing a 6xHis tag, maltose binding protein (MBP) and a TEV protease cleavage site. DBDs were expressed in BL21(DE3)pLysS Rosetta cells. Protein expression was induced by addition of 1 mM IPTG at A_{600} of 0.8-1.2. After induction, cells were grown overnight at 15°C. Cells were harvested via centrifugation and frozen at -10°C overnight. Cells were lysed using B-PER® Protein Extraction Reagent Kit (ThermoScientific).

Lysate was loaded onto a pre-equilibrated 5 mL HisTrap HP column (GE) and eluted with a linear imidazole gradient (25 mM to 1 M) in 25 mM sodium phosphate and 100 mM NaCl buffer [pH 7.6]. The DBD was cleaved from the MBP-His fusion using TEV protease in dialysis buffer consisting of 25 mM sodium phosphate, 150 mM NaCl, 2 mM βME and 10% glycerol [pH 8.0]. The cleavage products were loaded onto a 5 mL HiPrep SP FF cation exchange column (GE) and eluted with a linear NaCl gradient (150 mM to 1 M) in 25 mM sodium phosphate buffer [pH 8.0]. DBDs were further purified on a Superdex™200 10/300 GL size exclusion column (GE) with 10 mM Tris [pH 7.6], 100 mM NaCl, 2 mM βME, 5% glycerol. Protein purity was assayed after each purification by visualization on a 12% SDS-PAGE gel stained with Bio-Safe™ Coomassie G-250 stain (Bio-Rad).

Fluorescence polarization (FP) binding assay

DNA constructs were ordered from Eurofins Operon as HPLC-purified single stranded oligos with the forward strand labeled at the 5'-end with 6-FAM. Sequences of forward and reverse strands, respectively, are as follows: ERE-half – CCAGGTCAGAG, CTCTGACCTGG; SRE1-half – CCAGAACAGAG, CTCTGTTCTGG; SRE2-half – CCAGGACAGAG, CTCTGTCCTGG; ERE-full – CCAGGTCAGAGTGACCTGA, TCAGGTCACTCTGACCTGG; SRE1-full – CCAGAACAGAGTGTTCTGA, TCAGAACACTCTGTTCTGG; SRE2-full – CCAGGACAGAGTGTCCTGA, TCAGGACACTCTGTCCTGG. Forward and reverse strands were re-suspended in duplex buffer (30 mM Hepes [pH 8.0], 100 mM potassium acetate) to a concentration of 100 μM. Equimolar quantities of complementary forward and reverse strands were combined and placed in a 95°C water bath for 10 minutes then slowly cooled to room temperature. The double stranded product was diluted to 5 μM in water.

Purified DBD was buffer exchanged using Illustra NAP-25 columns into 20 mM Tris [pH 7.6], 130 mM NaCl and 5% glycerol. A range of DBD concentrations was titrated in triplicate onto a black, NBS-coated 384 well plate (Corning 3575). Labeled DNA was added to each well to achieve a final concentration of 5 nM in 91μL total volume. Sample FP was read using a Perkin Elmer Victor X5, exciting at 495nm and measuring emission polarization at 520nm.

To determine K_1 and ω with high confidence, we performed two experiments for each protein-DNA pair. We measured binding to a half-site RE and to a palindromic RE

and applied a global fit, based on the model by Hard and colleagues (Hard et al., 1990), to both data sets to calculate K_1 and ω simultaneously.

Protein denaturation

Purified DBD was buffer exchanged into 10 mM sodium phosphate [pH 7.6], 25 mM NaCl, 2 mM BME. The reversible, two-state unfolding reaction was followed by measuring the loss of secondary structure using circular dichroism signal at 222nm as a function of increasing concentration of 8 M guanidinium chloride in 10 mM sodium phosphate [pH 7.6], 25 mM NaCl and 2 mM BME. The resulting data was fit to the model previously described by Pace and Scholtz (Pace and Scholtz, 1997).

Sedimentation velocity

Sedimentation experiments were performed on a Beckman ProteomeLab XL-I. Purified DBDs were dialyzed against a buffer containing 20mM Tris [pH 7.6] and 100mM NaCl. DBDs were concentrated to 0.5 mM and sedimented at 20°C using a rotor speed of 60,000 rpm for 10 hours. Sedimentation coefficients were calculated by measuring sample interference. The distribution of sedimentation coefficients was calculated using every 5^{th} scan of the first 190 scans in SedFit. Partial specific volumes were calculated using the method previously described by Arakawa (Arakawa, 1986).

Crystal structure determination

Reagents

Chemicals were purchased from Sigma, Fisher or HyClone. DNA oligos used for binding and crystallization were synthesized by Integrated DNA Technologies (Coralville, Iowa).

Protein Expression and Purification

The fusion proteins were expressed in BL21(DE3) pLysS cells using standard methods and purified using affinity chromatography (Ni Sepharose 6 Fast Flow, GE) in the presence of 1 M NaCl to remove non-specifically associated DNA. For crystallization the fusion tags were cleaved via TEV protease and constructs were re-purified using affinity chromatography. The protein variants were further purified via size-exclusion chromatography into 300 mM NaCl, 20 mM Tris-HCl [pH 7.4], 5% (v/v) glycerol, and concentrated to 1-3 mg ml^{-1} before flash freezing in liquid nitrogen and storage at -80 ^{0}C.

Crystallization and Structure Determination

Crystals of AncSR1 in complex with a 19-bp blunt ended duplex DNA canonical ERE (5'-CCAGGTCAGAGTGACCTGA-3') were grown by hanging-drop vapor diffusion at 20°C from solutions containing equal volumes of the 1:1.2 protein:DNA complex in the following crystallant: 12% PEG 3350, 100 mM ammonium acetate, 100 mM bis-Tris buffer (pH 5.5). Crystallization experiments were microseeded with a 1:100 dilution of crushed crystals of the same protein:DNA construct grown at a higher

concentration of PEG 3350 and 75 mM ammonium acetate. Crystals were cryoprotected in crystallant containing 30% PEG 3350, 150 mM ammonium acetate and 50 mM bis-Tris (pH 5.5) and were flash cooled in liquid N_2. Data to a resolution of 1.7 Å were collected at 100 K with a MAR 225 CCD detector at the SER-CAT 22 BM beamline at the Advanced Photon Source and were processed and scaled with HKL2000 (Otwinowski and Minor, 1997). Phases were determined with the Phaser-MR program from the Phenix software suite (Adams et al., 2010) using the structure of the human ER DNA binding domain (pdb code 1HCQ - 82% sequence identity over 81 equivalent residues (Schwabe et al., 1993)) as the search model. Model building and refinement was carried out with Phenix's Refine program (version dev-1627) (Adams et al., 2010). The final model contains one dimer of the AncSR1 DBD, 19 base pairs of dsDNA, four zinc atoms, 298 water molecules, 1 sodium atom, and exhibits good geometry as indicated by Procheck (Laskowski et al., 1993). 98% of the residues are within favored Ramachandran space with no outliers.

Crystals of AncSR2 in complex with a 19-bp overhang duplex DNA canonical SRE1 (5'-CCAGAACAGAGTGTTCTG-3', 5'-TCAGAACACTCTGTTCTG-3') were grown by hanging-drop vapor diffusion at 20°C from solutions containing equal volumes of the 1:1.2 protein:DNA complex in the following crystallant: 20% PEG 3350, 50 mM ammonium acetate, 100 mM bis-Tris (pH 5.5). Crystals were cryoprotected in crystallant containing 30% PEG 3350, 150 mM ammonium acetate and 50 mM bis-Tris pH 5.5 and were flash cooled in liquid N_2. Data to a resolution of 2.7 Å were collected at 100 K with a MAR 225 CCD detector at the SER-CAT 22BM beamline at the Advanced Photon Source and were processed and scaled with HKL2000. Phases were determined with the

Phaser-MR program from the Phenix software suite using the structure of the rat glucocorticoid receptor DBD (86% sequence identity over 84 equivalent residues, using PDB ID: 3G99 (Meijsing et al., 2009)) as the search model. Model building and refinement were carried out in Phenix (version dev-1627). The final model contains two dimers of the AncSR2 DBD, 18 base pairs of dsDNA, eight zinc atoms, 53 water molecules, and exhibits good geometry as indicated by Procheck. 95% of the residues are within favored Ramachandran space with no outliers.

Crystals of the AncSR2+rh variants in complex with blunt end ERE and SRE1 DNA identical to that used for the AncSR1 and AncSR2 complexes, respectively, were grown via hanging-drop vapor diffusion at 20°C from solutions containing 1:1.2 protein:DNA in the following crystallant: 14-20% PEG 3350, 100 mM NH_4Acetate, 100 mM bis-Tris (pH 5.5) with 2:1 and 4:1 respective protein-DNA solution: reservoir drop ratios. Crystals were cryoprotected in crystallant containing 20% PEG 3350, 10% glycerol and 100 mM bis-Tris pH 5.5 and were flash cooled in liquid N_2. Data to a resolution of 2.25 and 2.37 Å, for AncSR2+rh:ERE and AncSR2+rh:SRE1 respectively, were collected at 100 K with a MAR 300 CCD detector at the SER-CAT 22ID beamline at the Advanced Photon Source and were processed and scaled with HKL2000. Phases were determined with the Phaser-MR program from the Phenix software suite using the structure of AncSR2-rh as a search model. Model building and refinement were carried out with Phenix's Refine program. The final models contain two dimers of the AncSR2+rh, 19 and 18 base pairs of dsDNA (for the ERE and SRE1, respectively), eight zinc atoms and 132 and 22 water molecules, respectively. 97 and 95% of the residues are

within favored Ramachandran space with no outliers for the AncSR2+rh:ERE and AncSR2+rh:SRE1 complexes, respectively.

Protein Data Bank

The atomic coordinates and structure factors have been deposited in the RCSB Protein Data Bank, www.pdb.org with the following PDB ID codes: 4OLN for AncSR1:ERE, 4OOR for AncSR2:SRE1, 4OND for AncSR2+rh:ERE, and 4OV7 for AncSR2+rh:SRE1.

Molecular dynamics simulations

The crystal structures of AncSR1 and AncSR2 bound to their response elements were used as the starting point for all simulations. Historical substitutions and changes to the DNA response element sequences were introduced in silico (Emsley and Cowtan, 2004). Each system was solvated in a cubic box with a 10 Å margin, then neutralized and brought to 150 mM ionic strength with sodium and chloride ions. This was followed by energy minimization to remove clashes, assignment of initial velocities from a Maxwell distribution, and 1 ns of solvent equilibration in which the positions of heavy protein and DNA atoms were restrained. Production runs were 50 ns, with the initial 10 ns excluded as burn-in. The trajectory time step was 2 fs, and final analyses were performed on frames taken every 12.5 ps.

We used TIP3P waters and the AMBER FF03 parameters for protein and DNA, as implemented in GROMACS 4.5.5 (Duan et al., 2003). The zinc fingers were treated with a recently derived bonded potential for Cys-Zn interactions (Table S6A) (Lin and

Wang, 2010). Zinc finger partial charges were derived using the RED III.4 pipeline (Table S6B) (Dupradeau et al., 2010). We extracted a tetrahedral Cys_4 zinc finger from a 0.9 Å crystal structure (Iwase et al., 2011), optimized its geometry with an explicit quantum mechanical calculation using the 6-31G** basis set (Schuchardt et al., 2007), then derived partial charges using RESP (Dupradeau et al., 2010). All quantum mechanical calculations were performed using the FIREFLY implementation of GAMESS (Schmidt et al., 1993; Granovsky and GAMESS, 2009). We verified that the zinc fingers maintained their tetrahedral geometry over the course of the simulations.

Simulations were performed in the NTP ensemble at 300K, 1 bar. All bonds were treated as constraints and fixed using LINCS (Hess et al., 1997). Electrostatics were treated with the Particle Mesh Ewald model (Darden et al., 1993), using an FFT spacing of 12 Å, interpolation order of 4, tolerance of 1e-5, and a Coulomb cutoff of 9 Å. van der Waals forces were treated with a simple cutoff at 9 Å. We used velocity rescaled temperature coupling with a τ of 0.1 ps and Berendsen pressure coupling with a τ of 0.5 ps and a compressibility of 4.5e-5 bar^{-1}. Analyses were performed using VMD 1.9.1 (Humphrey et al., 1996)—with its built-in TCL scripting utility—as well as a set of in-house Python and R scripts.

APPENDIX D

FIGURES AND TABLES FOR CHAPTER IV

Figure 9 (next page). The derived RH causes a switch in DNA-binding preference and specificity

(A) SR receptors group into two well-defined clades based on their DNA-binding specificity. Phylogenetic relationships of extant receptors are shown with the DNA-binding specificity of each receptor indicated by color; purple, ERE and green, SRE. Reconstructed ancestors are also indicated by a circle and colored by RE specificity. The preferred RE half-site sequence is shown to the right with differences underlined and in bold. SRE-specificity evolved on the interval between AncSR1 and AncSR2, indicated by a gray box.

(B) Crystal structure of dimeric AncSR1 bound to palindromic RE full-site. Recognition of DNA occurs by insertion of the recognition helix (RH) into the DNA major groove of each DNA half-site. The three RH substitutions capable of switching DNA binding preference are indicated with Cα as spheres; glu25GLY is orange, gly26SER is cyan and ala29VAL is green. Protein is shown in cartoon; DNA is shown as surface and colored by atom (gray, carbon; blue, nitrogen; red, oxygen; orange, phosphate).

(C) AncSR1 binds with highest affinity to ERE; AncSR1+RH binds with highest affinity to SREs. Rank-ordered single-site DNA binding energies for AncSR1 (top) and AncSR1+RH (bottom). ERE, SRE1 and SRE2 are indicated by purple, light green and dark green bars, respectively. Data points are for three independent replicates; mean and SEM are shown with lines. Identity of the RH residues are indicated; lower case and upper case letters denote the ancestral and derived amino acid states, respectively.

(D) AncSR1 has greatest preference for G3T4; AncSR1+RH has highest preference for G3A4 and A3A4. Binding motifs display nucleotide preference for AncSR1 (top) and AncSR1+RH (bottom). Bar height indicates fractional occupancy of DNA sequences with a given nucleotide state at each position. The total binding energy of each protein construct was calculated by summation of the binding energies across all 16 RE sequences and is indicated to the right of the bar graphs.

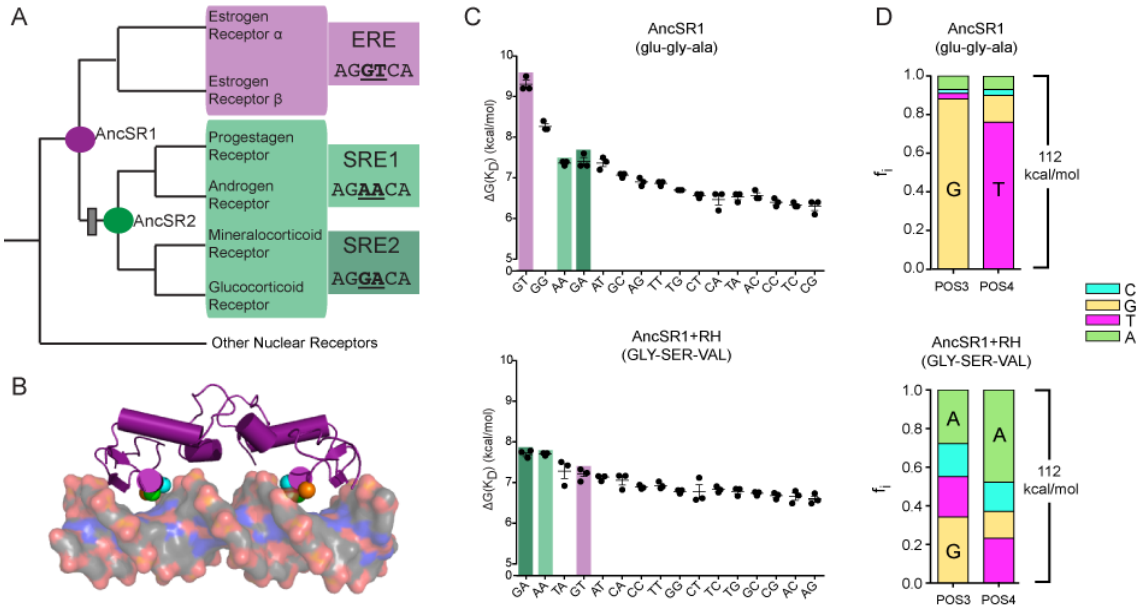
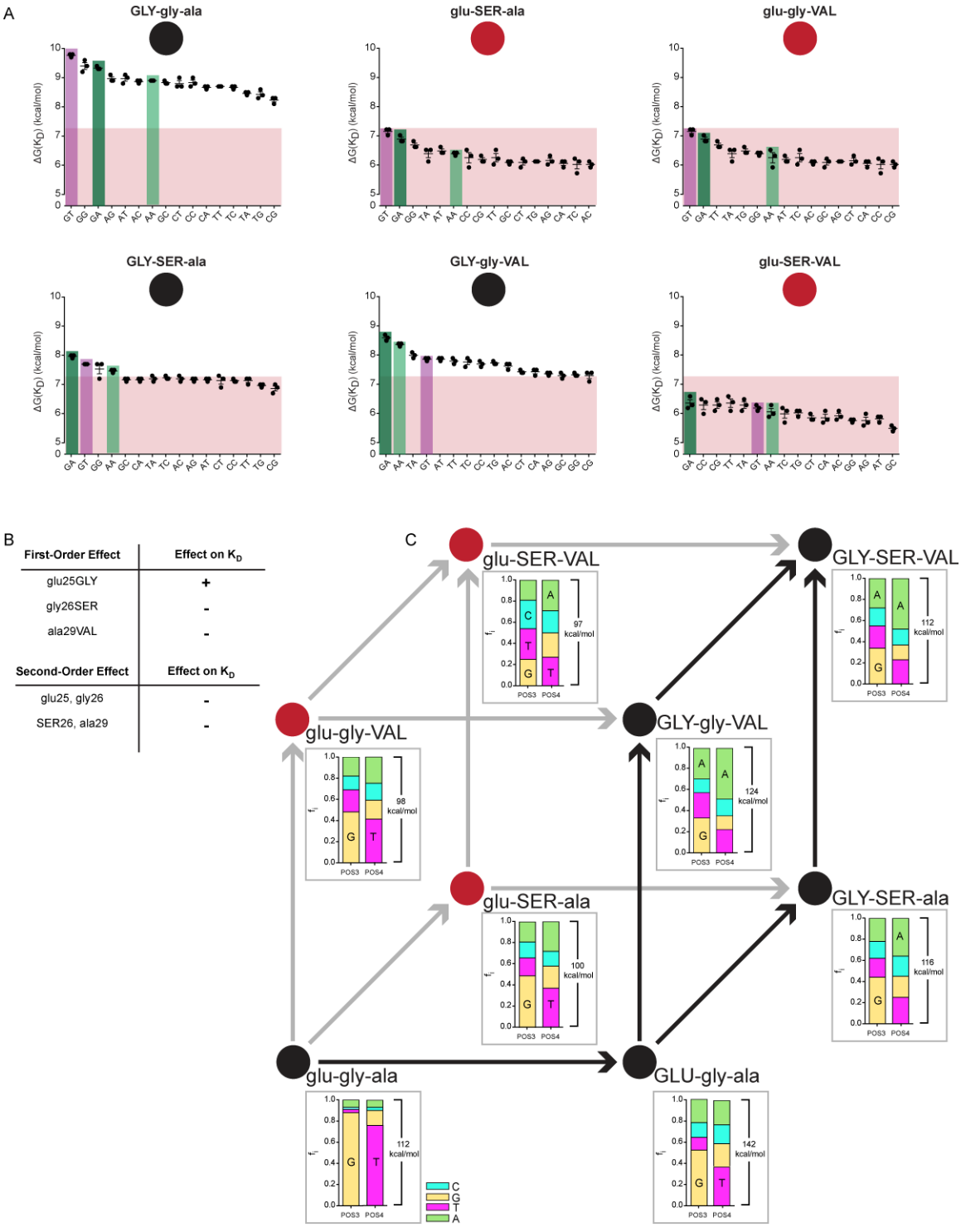

A
Estrogen Receptor α
Estrogen Receptor β
ERE
AGGTCA
AncSR1
Progestagen Receptor
Androgen Receptor
SRE1
AGAACA
AncSR2
Mineralocorticoid Receptor
Glucocorticoid Receptor
SRE2
AGGACA
Other Nuclear Receptors
B
C
AncSR1
(glu-gly-ala)
ΔG(K_D) (kcal/mol)
GT GG AA GA AT GC AG TT TG CT CA TA AC CC TC CG
AncSR1+RH
(GLY-SER-VAL)
GA AA TA GT AT CA CC TT GG CT TC TG GC CG AC AG
D
AncSR1
(glu-gly-ala)
f_i
G
T
POS3 POS4
112 kcal/mol
C
G
T
A
AncSR1+RH
(GLY-SER-VAL)
A
A
G
POS3 POS4
112 kcal/mol

Figure 10 (next page). Functional characterization of all protein intermediates allows for a complete mapping of the functional sequence space between AncSR1 and AncSR1+RH

(A) Ranked binding energies for all possible protein intermediates. ERE, SRE1 and SRE2 are shown with purple, light green and dark green bars, respectively. The low-affinity cut-off, defined by the mean of all binding measurements across all protein sequences, is shown as a red box. Data points are for three independent replicates; mean and SEM are shown with lines. Lower case and upper case letters denote the ancestral and derived amino acid states, respectively.

(B) Statistically significant first and second-order effects of the derived substitutions on binding affinity determined by linear modeling. $^{\Phi}$ indicates effect to increase $\Delta G(K_D)$, while – indicates effect to decrease $\Delta G(K_D)$.

(C) Only two mutational pathways were available to the evolving protein that allowed for evolution of the derived phenotype without passing through a low-affinity intermediate. Vertices of the cube represent unique combinations of RH residues. Low-affinity constructs, defined as not binding to a single sequence with an affinity above the mean binding affinity, are indicated by a red circle. High-affinity constructs are black circles. Bar plots at each vertex represents the fractional occupancy for each protein sequence. Arrows connecting vertices represent single genetic mutations. Accessible mutations that do not result in a low-affinity intermediate are black arrows; mutations that lead from or result in a low-affinity intermediate are gray. Lower case and upper case letters denote the ancestral and derived states, respectively.


A
GLY-gly-ala
glu-SER-ala
glu-gly-VAL
GLY-SER-ala
GLY-gly-VAL
glu-SER-VAL
ΔG(KD) (kcal/mol)
B
First-Order Effect
Effect on KD
glu25GLY
+
gly26SER
-
ala29VAL
-
Second-Order Effect
Effect on KD
glu25, gly26
-
SER26, ala29
-
C
glu-SER-VAL
GLY-SER-VAL
glu-gly-VAL
GLY-gly-VAL
glu-SER-ala
GLY-SER-ala
glu-gly-ala
GLU-gly-ala
97 kcal/mol
112 kcal/mol
98 kcal/mol
124 kcal/mol
100 kcal/mol
116 kcal/mol
112 kcal/mol
142 kcal/mol
POS3
POS4
C
G
T
A

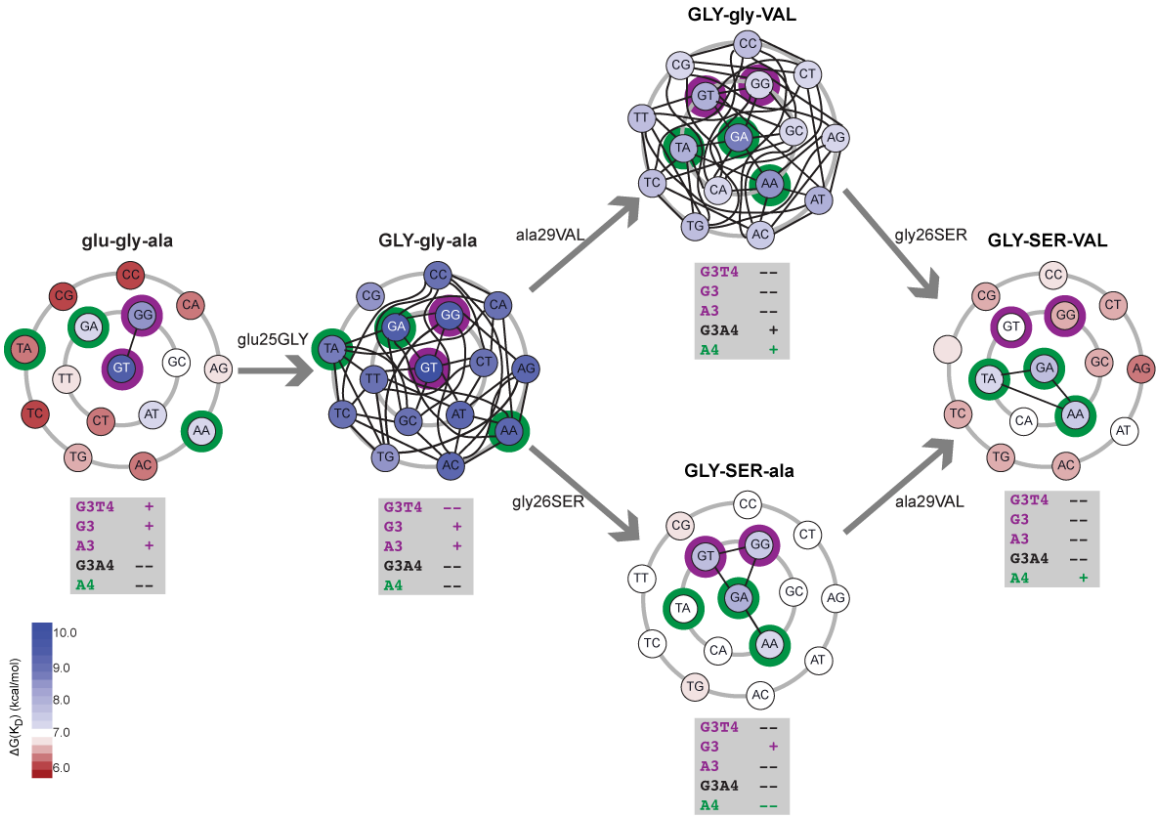


Figure 11. Protein promiscuity increases the size of the high-affinity RE sequence space

Maps of the RE sequence space for each high-affinity protein sequence. RE sequences are colored based on their binding affinity: blue, binding affinity greater than the mean binding affinity; white, mean binding affinity of 7.1kcal/mol; red, binding affinity less than mean binding affinity. Ancestrally preferred sequences are outlined in purple; sequences preferred by the derived protein are outlined in green. An RE sequence is defined as accessible if (1) it has binding affinity greater than 7.1kcal/mol and (2) has a binding affinity that is within 10-fold of the highest affinity RE sequence for each protein sequence. Single genetic mutations between accessible REs is shown as a black line. Both possible protein mutational pathways that do not pass through a low-affinity intermediate are shown. As the protein becomes more promiscuous, the accessible RE sequence space becomes less constrained, resulting in a much larger accessible RE network. Nucleotide preferences, determined by linear modeling, for each protein sequence is shown in the gray box; + indicates effect to increase affinity, while -- indicates that it is a non-significant effect. Ancestral preferences are colored purple. Derived preferences are colored green. Preferences that are neither ancestral nor derived are colored black. Lower case and upper case letters denote the ancestral and derived amino acid states, respectively.

Figure 12 (next page). Mapping the functional sequence space of the SR transcriptional module allows for identification of all accessible mutational pathways available for both the protein and RE during the evolution of novel DNA specificity

(A) The functional sequence space of the SR transcriptional module is characterized by inter protein-RE epistasis. Reported is the sole positive first-order RE effect, as well as the epistatic effects between a given protein residue and RE nucleotide state. Effects are indicated by +, increasing $\Delta G(K_D)$ and –, decreasing $\Delta G(K_D)$.

(B) Map of the functional sequence space for the evolving SR transcriptional module. The vertices of the cube represent all possible genetic combinations of ancestral and derived RH residues; edges of the cube represent single genetic mutations in the protein. Lower case and upper case letters denote the ancestral and derived amino acid states, respectively. The function of the protein is expressed by the accessible RE sequence space available to an evolving RE sequence while still maintaining regulation by the specific protein sequence. RE sequences are colored according to binding affinity:blue, binding affinity greater than 7.1kcal/mol; white, binding affinity equal to 7.1kcal/mol; red, binding affinity less than 7.1kcal/mol. Black connections between RE sequences within a given protein construct represent high-affinity nodes within the RE sequence space for that protein. Green connections between RE sequences that occur between protein sequences represents possible genetic changes within the protein that would still result in regulation of the connected RE sequences. Together, these data give a complete account for the evolvability of the system by describing all possible protein and RE mutations available to the evolving transcriptional module.

A

Second-Order Effect	Effect on K_D
A4	+
glu25, A4	-
ala29, A4	-
gly26, G3	+
ala29, G3	+
gly26, C3	-
ala29, T3	-

B

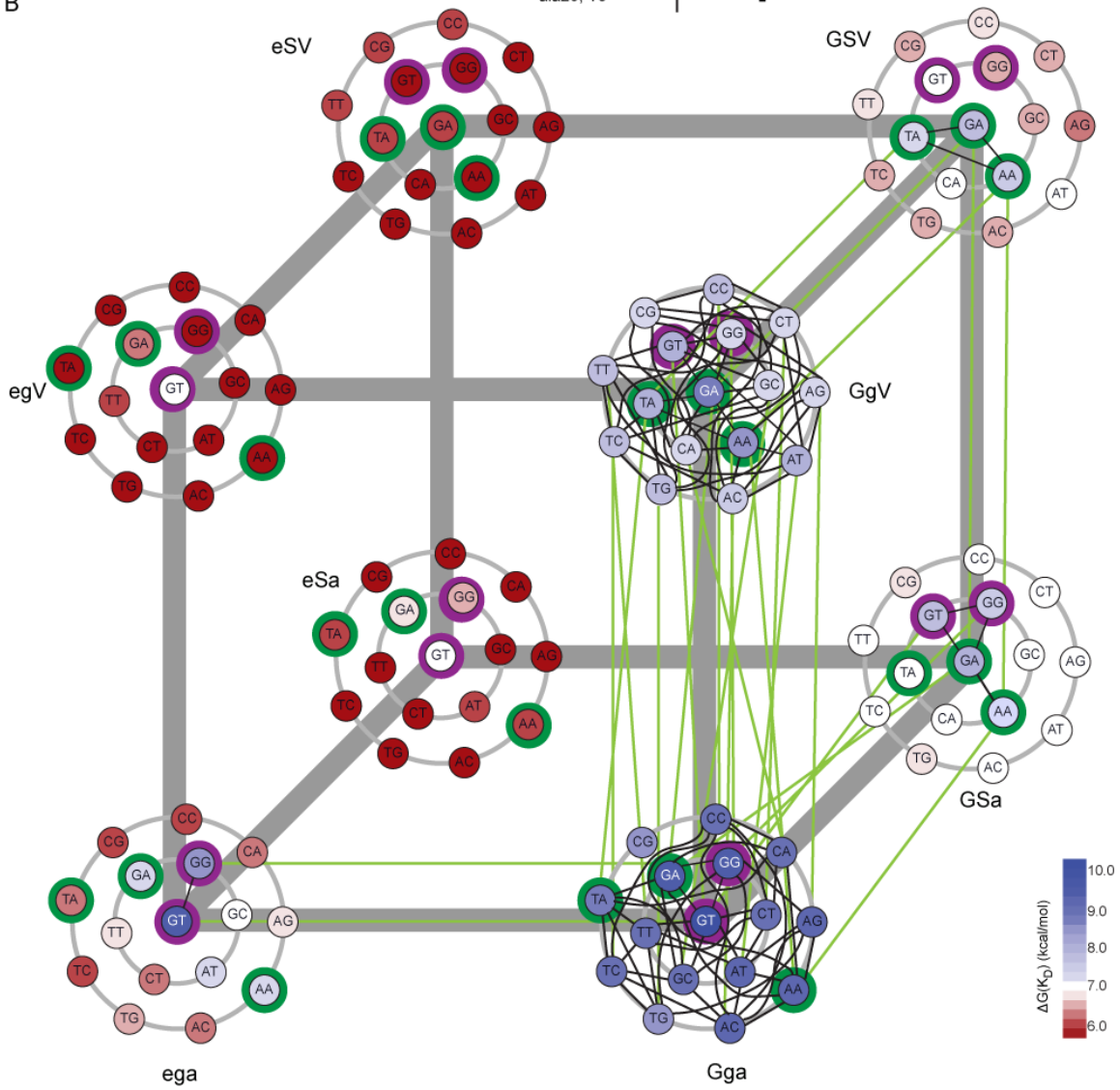


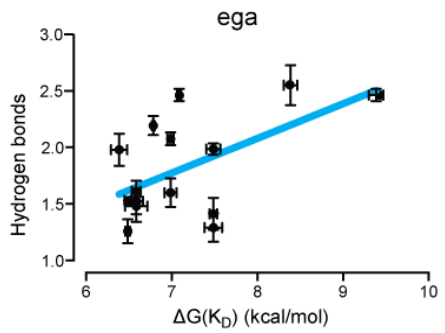
Table 1. Hydrogen bonding and packing efficiency are insufficient to explain variation in binding affinity across the transition from AncSR1 to AncSR1+RH. Correlation coefficients for hydrogen bonding versus binding and packing efficiency versus packing. Positive correlations are colored blue. Negative correlations are colored pink. Insignificant correlations are white.

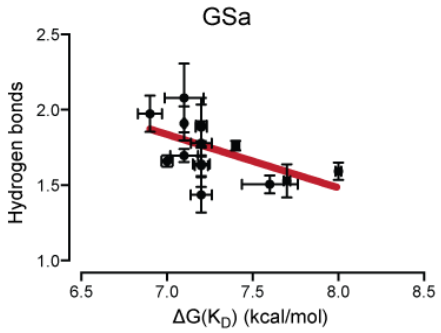
	Hydrogen bonding vs binding			Packing vs binding		
Protein sequence	Correlation	P-value	R^2	Correlation	P-value	R^2
ega	positive	< 0.001	0.2857	positive	0.0060	0.2264
Gga	NS	0.0941	0.0598	NS	0.7410	0.0024
eSa	positive	0.0052	0.1576	positive	0.0011	0.2082
egV	positive	0.0015	0.1991	NS	0.0772	0.0663
GSa	negative	0.0071	0.1474	NS	0.1708	0.0413
GgV	NS	0.9298	0.0002	NS	0.6531	0.0044
eSV	NS	0.7272	0.0027	NS	0.1589	0.0427
GSV	NS	0.327	0.0214	positive	0.0075	0.1455

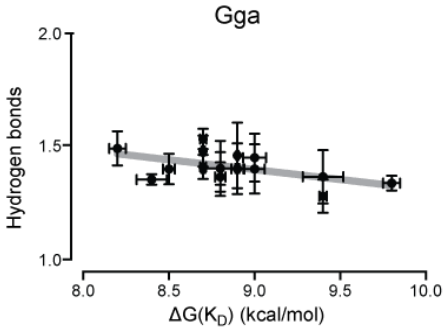
APPENDIX E

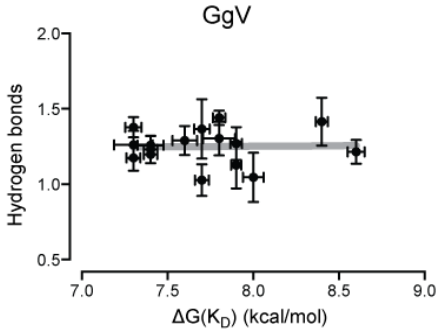
SUPPLEMENTAL INFORMATION FOR CHAPTER IV

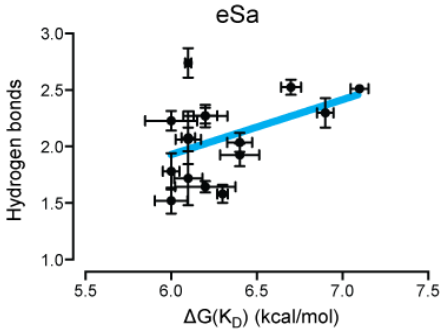
Figure S7 (next page). Hydrogen bonding is insufficient to account for variation in binding affinity across the transition from AncSR1 to AncSR1+RH
Linear modeling of hydrogen bonding data versus binding affinity. Hydrogen bonding has a positive correlation (blue line) with binding affinity for three protein sequences and a negative correlation (red line) with binding affinity for one protein sequence. The remaining 4 protein sequences show no significant correlation (gray line) between hydrogen bonding and binding affinity. For statistics, see Table 1.

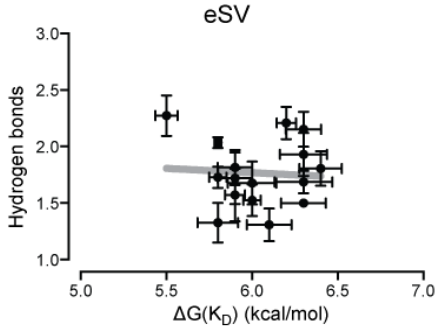

ega
Hydrogen bonds
ΔG(K_D) (kcal/mol)

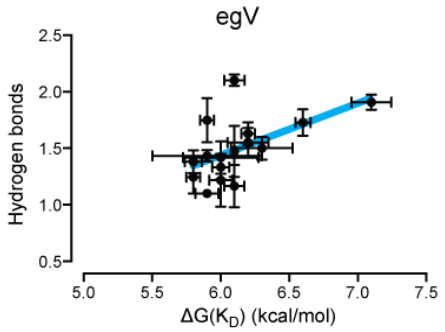

GSa
Hydrogen bonds
ΔG(K_D) (kcal/mol)


Gga
Hydrogen bonds
ΔG(K_D) (kcal/mol)


GgV
Hydrogen bonds
ΔG(K_D) (kcal/mol)


eSa
Hydrogen bonds
ΔG(K_D) (kcal/mol)


eSV
Hydrogen bonds
ΔG(K_D) (kcal/mol)


egV
Hydrogen bonds
ΔG(K_D) (kcal/mol)

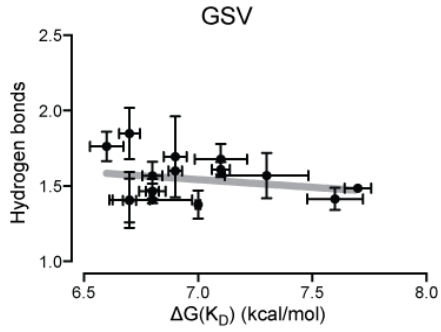
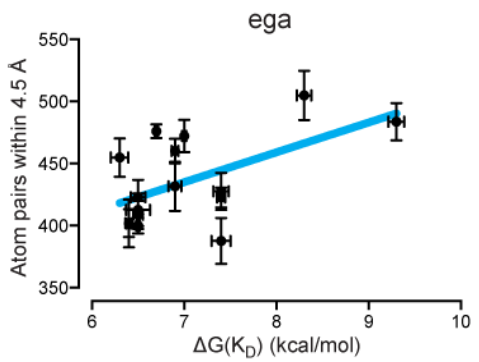
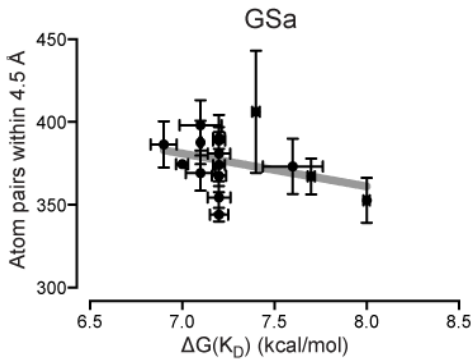
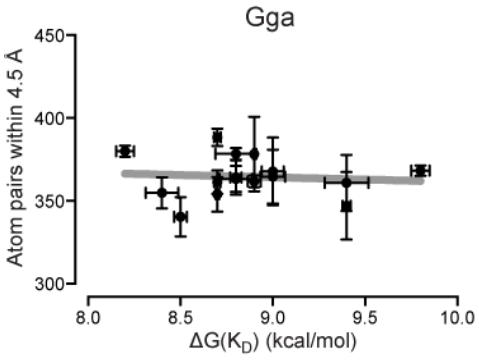
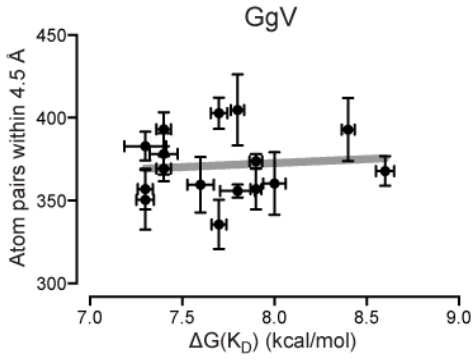

GSV
Hydrogen bonds
ΔG(K_D) (kcal/mol)

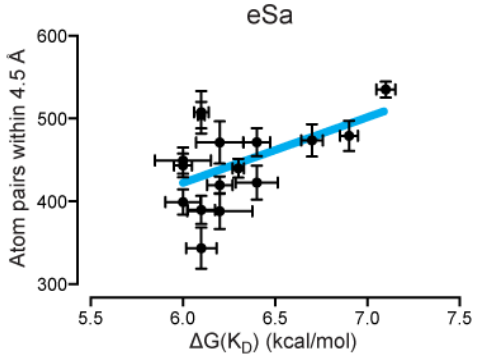
Figure S8 (next page). Packing efficiency is insufficient to account for variation in binding affinity across the transition from AncSR1 to AncSR1+RH
Linear modeling of packing efficiency data versus binding affinity. Packing efficiency has a positive correlation (blue line) with binding affinity for three protein sequences. The remaining protein sequences do not have a significant correlation (gray line) between packing efficiency and binding affinity. For statistics, see Table 1.

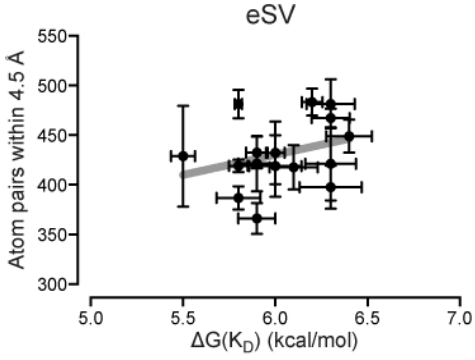

ega
Atom pairs within 4.5 Å
550
500
450
400
350
6
7
8
9
10
ΔG(K_D) (kcal/mol)

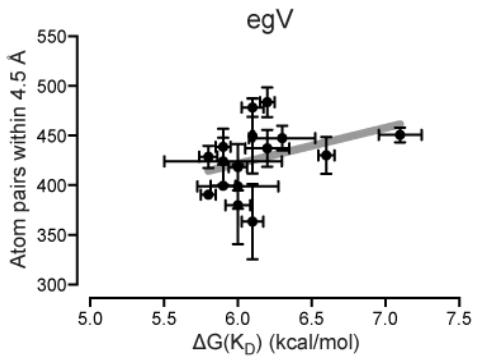

GSa
Atom pairs within 4.5 Å
450
400
350
300
6.5
7.0
7.5
8.0
8.5
ΔG(K_D) (kcal/mol)


Gga
Atom pairs within 4.5 Å
450
400
350
300
8.0
8.5
9.0
9.5
10.0
ΔG(K_D) (kcal/mol)


GgV
Atom pairs within 4.5 Å
450
400
350
300
7.0
7.5
8.0
8.5
9.0
ΔG(K_D) (kcal/mol)


eSa
Atom pairs within 4.5 Å
600
500
400
300
5.5
6.0
6.5
7.0
7.5
ΔG(K_D) (kcal/mol)


eSV
Atom pairs within 4.5 Å
550
500
450
400
350
300
5.0
5.5
6.0
6.5
7.0
ΔG(K_D) (kcal/mol)


egV
Atom pairs within 4.5 Å
550
500
450
400
350
300
5.0
5.5
6.0
6.5
7.0
7.5
ΔG(K_D) (kcal/mol)

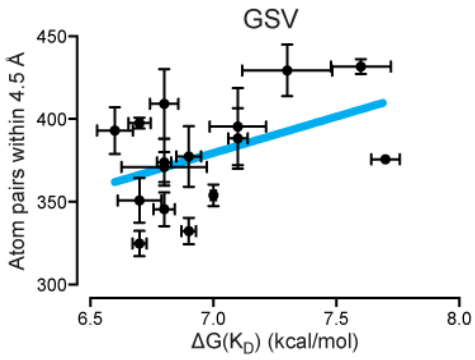

GSV
Atom pairs within 4.5 Å
450
400
350
300
6.5
7.0
7.5
8.0
ΔG(K_D) (kcal/mol)

Table S7: Significant first and second order terms from AIC-optimized and global linear models

Optimized statistical coefficients from both an AIC-optimized and a global linear model as described in the materials and methods. Table includes terms that were statistically significant in either model when applied across protein genotypes, across RE genotypes and across both protein and RE genotypes. Significance assessed with multiple testing correction (false-discovery rate of 5%). All significant coefficient effects act in the same direction to either increase or decrease binding affinity for both linear modeling approaches. * indicates terms significant in the AIC-optimized model but not in the global model, while Φ indicates terms significant in the global model but not the AIC-optimized model. N/A indicates absence from AIC-optimized model.

General protein effects				
	AIC-optimized Model		Global Model	
Genetic Term	Effect (Fold Affinity)	p	Effect (Fold Affinity)	p
glu25GLY	5.08	7.1e-42	4.56	3.3e-27
gly26SER	0.292	4.3e-17	0.262	3.3e-22
ala29VAL	0.142	1.6e-32	0.158	4.0e-37
glu25, gly26	0.222	2.8e-15	0.275	1.8e-16
SER26, ala29	0.225	5.3e-15	0.280	4.0e-16

Protein-specific RE effects				
	glu-gly-ala			
	AIC-optimized Model		Global Model	
Genetic Term	Effect (Fold Affinity)	p	Effect (Fold Affinity)	p
A3	2.02	4.6e-4	2.28	2.3e-5
G3	5.39	2.1e-10	14.0	1.0e-16
C4	0.436	1.0e-5	0.421	1.1e-5
G3, T4	13.7	2.7e-10	4.40	4.8e-7
G3, A4 Φ	N/A	N/A	0.300	1.5e-5
G3, C4 Φ	N/A	N/A	0.224	4.0e-7
	GLY-gly-ala			
	AIC-optimized Model		Global Model	
Genetic Term	Effect (Fold Affinity)	p	Effect (Fold Affinity)	p
A3	1.85	3.9e-10	1.61	2.5e-3
C3*	1.37	4.7e-4	1.21	1.9e-1
G3	4.38	4.8e-18	5.20	1.0e-12
G3, T4*	1.64	1.5e-3	1.25	2.8e-1
C3, G4	0.433	7.9e-7	0.561	8.5e-3
G3, C4	0.309	1.5e-3	0.240	8.0e-8

GLY-gly-VAL

Genetic Term	AIC-optimized Model Effect (Fold Affinity)	AIC-optimized Model p	Global Model Effect (Fold Affinity)	Global Model p
C3	0.560	6.3e-5	0.540	1.0e-4
G3 $^{\Phi}$	N/A	N/A	0.494	1.6e-5
A4*	1.76	2.3e-4	1.43	1.5e-2
G4*	0.477	8.5e-6	0.901	4.6e-1
A3, G4 $^{\Phi}$	N/A	N/A	0.582	6.7e-4
C3, A4*	0.497	3.3e-3	0.701	8.0e-2
G3, A4	2.27	7.5e-4	5.29	1.1e-9
G3, C4 $^{\Phi}$	N/A	N/A	0.347	3.6e-5

GLY-SER-ala

Genetic Term	AIC-optimized Model Effect (Fold Affinity)	AIC-optimized Model p	Global Model Effect (Fold Affinity)	Global Model p
G3	2.86	4.2e-15	2.70	3.8e-7
A4*	1.28	1.2e-3	1.18	3.1e-1
G4*	0.770	7.0e-4	0.790	1.4e-1
G3, C4	0.372	7.5e-9	0.347	3.6e-5

GLY-SER-VAL

Genetic Term	AIC-optimized Model Effect (Fold Affinity)	AIC-optimized Model p	Global Model Effect (Fold Affinity)	Global Model p
A4	1.97	2.3e-5	1.83	4.2e-3
C4*	0.641	3.3e-3	0.849	4.1e-1
G4*	0.523	3.3e-4	0.785	2.3e-1

Across Protein and RE

Genetic Term	AIC-optimized Model Effect (Fold Affinity)	AIC-optimized Model p	Global Model Effect (Fold Affinity)	Global Model p
glu25GLY	4.53	2.3e-67	3.53	3.4e-21
gly26SER	0.267	5.2e-39	0.238	3.9e-26
ala29VAL	0.179	5.4e-68	0.204	9.4e-31
C3*	0.710	8.8e-5	0.674	1.2e-2
G3 $^{\Phi}$	N/A	N/A	0.655	6.8e-3
A4	1.81	1.9e-8	2.37	5.7e-8
C4*	0.699	3.4e-5	0.781	1.1e-1
G4*	0.721	5.8e-6	0.892	4.6e-1
glu25, GLY26	0.275	1.9e-36	0.275	1.3e-41
SER26, ala29	0.280	1.0e-35	0.280	8.3e-41
glu25, A4	0.630	1.6e-5	0.545	4.2e-7

gly26, C3*	0.702	1.5e-3	0.815	8.2e-2
gly26, G3	1.54	1.1e-5	1.82	5.6e-7
ala29, A3 Φ	N/A	N/A	1.71	7.0e-6
ala29, G3	2.32	2.2e-16	3.45	1.1e-22
ala29, T3*	0.720	1.2e-4	0.950	7.0e-1
ala29, A4	0.636	2.4e-5	0.556	9.4e-7
G3, T4	1.72	7.0e-6	2.00	3.8e-5
C3, C4	1.79	8.4e-6	1.73	1.0e-3
G3, C4	0.459	1.1e-9	0.502	4.2e-5

Table S8: All first and second order terms from global linear models
Data was fit to a global model as described in the materials and methods. Table includes all terms when applied across protein genotypes, across RE genotypes, and across both protein and RE genotypes, as well as their optimized coefficient (effect) and associated p-value.

General protein effects		
Genetic Term	**Effect (Fold Affinity)**	**p**
glu25GLY	4.56	3.3e-27
gly26SER	0.262	3.3e-22
ala29VAL	0.158	4.0e-37
glu25, gly26	0.275	1.8e-16
glu25, SER26	0.805	1.5e-1
glu25, ala29	1.00	1.0e0
glu25, VAL29	1.00	1.0e0
GLY25, gly26	1.00	1.0e0
GLY25, SER26	1.00	1.0e0
GLY25, ala29	1.00	1.0e0
GLY25, VAL29	1.00	1.0e0
gly26, ala29	1.00	1.0e0
SER26, ala29	0.280	4.0e-16
gly26, VAL29	1.00	1.0e0
SER26, VAL29	1.00	1.0e0

Protein-specific RE effects		
glu-gly-ala		
Genetic Term	**Effect (Fold Affinity)**	**p**
A3	2.28	2.3e-5
C3	0.560	1.5e-3
G3	14.0	1.0e-16
T3	1.00	1.0e0
A4	0.563	1.6e-3
C4	0.421	1.1e-5
G4	0.726	6.4e-2
T4	1.00	1.0e0
G3, T4	4.40	4.8e-7
A3, A4	1.76	2.2e-2
A3, C4	0.591	3.3e-2
A3, G4	0.598	3.7e-2
A3, T4	1.00	1.0e0
C3, A4	1.62	4.9e-2
C3, C4	1.89	1.1e-2
C3, G4	0.878	5.9e-1
C3, T4	1.00	1.0e0

G3, A4	0.300	1.5e-5
G3, C4	0.224	4.0e-7
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0

GLY-gly-ala

Genetic Term	Effect (Fold Affinity)	p
A3	1.61	2.5e-3
G3	5.20	1.0e-12
C3	1.21	1.9e-1
T3	1.00	1.0e0
A4	0.716	2.9e-2
G4	0.671	1.0e-2
C4	0.983	9.1e-1
T4	1.00	1.0e0
G3, T4	1.25	2.8e-1
A3, A4	1.31	2.0e-1
A3, C4	0.907	6.4e-1
A3, G4	1.46	7.6e-2
A3, T4	1.00	1.0e0
C3, A4	1.13	5.7e-1
C3, C4	1.11	6.2e-1
C3, G4	0.561	8.5e-3
C3, T4	1.00	1.0e0
G3, A4	0.835	3.9e-1
G3, C4	0.240	8.0e-8
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0

GLY-gly-VAL

Genetic Term	Effect (Fold Affinity)	p
A3	1.14	3.5e-1
C3	0.540	1.0e-4
G3	0.494	1.6e-5
T3	1.00	1.0e0
A4	1.43	1.5e-2
C4	0.927	5.9e-1
G4	0.901	4.6e-1
T4	1.00	1.0e0

G3, T4	2.32	1.5e-4
A3, A4	1.33	2.1e-1
A3, C4	0.838	4.3e-1
A3, G4	0.477	6.7e-4
A3, T4	1.00	1.0e0
C3, A4	0.701	8.0e-2
C3, C4	1.68	1.3e-2
C3, G4	0.848	4.1e-1
C3, T4	1.00	1.0e0
G3, A4	5.29	1.1e-9
G3, C4	0.347	3.6e-5
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0

GLY-SER-ala

Genetic Term	**Effect (Fold Affinity)**	**p**
A3	1.50	5.0e-1
C3	0.963	8.1e-1
G3	2.70	3.8e-7
T3	1.00	1.0e0
A4	1.18	3.1e-1
C4	1.16	3.4e-1
G4	0.790	1.4e-1
T4	1.00	1.0e0
G3, T4	1.02	9.2e-1
A3, A4	1.33	2.1e-1
A3, C4	0.838	4.3e-1
A3, G4	1.23	3.6e-1
A3, T4	1.00	1.0e0
C3, A4	0.938	7.7e-1
C3, C4	0.934	7.6e-1
C3, G4	0.860	5.0e-1
C3, T4	1.00	1.0e0
G3, A4	1.33	2.1e-1
G3, C4	0.347	3.6e-5
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0

GLY-SER-VAL

Genetic Term	Effect (Fold Affinity)	p
A3	1.52	4.0e-2
C3	0.770	1.9e-1
G3	1.00	1.0e0
T3	1.00	1.0e0
A4	1.83	4.2e-3
C4	0.849	4.1e-1
G4	0.785	2.3e-1
T4	1.00	1.0e0
G3, T4	1.83	3.7e-2
A3, A4	1.34	3.0e-1
A3, C4	0.489	1.5e-2
A3, G4	0.479	1.2e-2
A3, T4	1.00	1.0e0
C3, A4	0.878	6.4e-1
C3, C4	1.43	2.0e-1
C3, G4	1.08	7.8e-1
C3, T4	1.00	1.0e0
G3, A4	0.481	1.3e-2
G3, C4	0.850	5.6e-1
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0

Global model effects across protein and RE

Genetic Term	Effect (Fold Affinity)	p
glu25GLY	3.53	3.4e-21
gly26SER	0.238	3.9e-26
ala29VAL	0.204	9.4e-31
A3	0.854	3.1e-1
C3	0.674	1.2e-2
G3	0.655	6.8e-3
T3	1.00	1.0e0
A4	2.37	5.7e-8
C4	0.781	1.1e-1
G4	0.892	4.6e-1
T4	1.00	1.0e0
glu25, gly26	0.275	1.3e-41
glu25, ala29	0.805	9.4e-3
SER26, ala29	0.280	8.3e-41
glu25, ala29	1.00	1.0e0
glu25, VAL29	1.00	1.0e0

GLY25, gly26	1.00	1.0e0
GLY25, SER26	1.00	1.0e0
GLY25, ala29	1.00	1.0e0
GLY25, VAL29	1.00	1.0e0
gly26, ala29	1.00	1.0e0
gly26, VAL29	1.00	1.0e0
SER26, VAL29	1.00	1.0e0
G3, T4	2.00	3.8e-5
A3, A4	1.30	1.2e-1
A3, C4	0.856	3.5e-1
A3, G4	0.827	2.5e-1
A3, T4	1.00	1.0e0
C3, A4	1.01	9.4e-1
C3, C4	1.73	1.0e-3
C3, G4	1.15	4.1e-1
C3, T4	1.00	1.0e0
G3, A4	1.38	5.5e-2
G3, C4	0.502	4.2e-5
G3, G4	1.00	1.0e0
T3, A4	1.00	1.0e0
T3, C4	1.00	1.0e0
T3, G4	1.00	1.0e0
T3, T4	1.00	1.0e0
glu25, A3	1.30	2.8e-2
glu25, C3	0.950	6.6e-1
glu25, G3	1.30	2.8e-2
glu25, T3	1.00	1.0e0
glu25, A4	0.545	4.2e-7
glu25, C4	0.653	3.4e-4
glu25, G4	0.991	9.4e-1
glu25, T4	1.00	1.0e0
GLY25, A3	1.00	1.0e0
GLY25, C3	1.00	1.0e0
GLY25, G3	1.00	1.0e0
GLY25, T3	1.00	1.0e0
GLY25, A4	1.00	1.0e0
GLY25, C4	1.00	1.0e0
GLY25, G4	1.00	1.0e0
GLY25, T4	1.00	1.0e0
gly26, A3	0.814	2.0e-2
gly26, G3	1.82	5.6e-7
gly26, C3	0.815	8.2e-2
gly26, T3	1.00	1.0e0
gly26, A4	0.618	5.3e-5
gly26, C4	0.733	8.6e-3
gly26, G4	0.773	2.9e-2

gly26, T4	1.00	1.0e0
SER26, A3	1.00	1.0e0
SER26, C3	1.00	1.0e0
SER26, G3	1.00	1.0e0
SER26, T3	1.00	1.0e0
SER26, A4	1.00	1.0e0
SER26, C4	1.00	1.0e0
SER26, G4	1.00	1.0e0
SER26, T4	1.00	1.0e0
ala29, A3	1.71	7.0e-6
ala29, C3	1.00	1.0e0
ala29, G3	3.45	1.1e-22
ala29, T3	0.950	7.0e-1
ala29, A4	0.556	9.4e-7
ala29, C4	0.706	3.2e-3
ala29, G4	0.945	6.3e-1
ala29, T4	1.00	1.0e0
VAL29, A3	1.00	1.0e0
VAL29, C3	1.00	1.0e0
VAL29, G3	1.00	1.0e0
VAL29, T3	1.00	1.0e0
VAL29, A4	1.00	1.0e0
VAL29, C4	1.00	1.0e0
VAL29, G4	1.00	1.0e0
VAL29, T4	1.00	1.0e0

MATERIALS AND METHODS

Protein purification

See Appendix C.

Fluorescence polarization (FP) binding assay

See Appendix C.

Linear Modeling the Genetic Determinants of Binding Affinity

To identify the genetic determinants of binding affinity, we implemented two alternative linear modeling approaches. We designed our models with an approach similar to that previously developed by others (Guenther et al., 2013). We built regression models that explain ΔG as a function of the genetic states at the three amino acid residues identified in the protein recognition helix or at the two middle positions in the response-element half-site. Linear coefficients were computed using ordinary least squares (OLS) regression with the open-source statistical package R (http://www.r-project.org/).

In the first linear model, we sought to identify the genetic factors that best explain the variation in binding affinity without over-fitting error variation as a result of including extraneous statistical parameters. We constructed our null model by regressing the $\log(K_a)$ (which is directly proportional to ΔG) measured for each genotype on the individual first-order identities at each genetic position. Each variable is 1 if the respective genetic state is at a given position, and 0 otherwise. For example, glu25 is 1 if there is a glu at position 25, and 0 in all other cases. An example of a null model is as follows:

$$\log(K_a) = C_0 + C_1(G3) + C_2(A3) + C_3(C4)$$

Where C0 is the y-intercept,C_1, C_2 and C_3 are coefficients of the effect for each respective variable. To identify cases of second-order epistatic interactions, we introduced one at a time all possible interaction terms for every two-way combination of genotypes at the variable sites being considered. These interaction terms take the same form as the first-order terms, but they are composed of identities at two sites. For example, G3T4 if 1 is the third position is a G and the fourth position is a T, and will be 0 otherwise. An example of an epistatic model is as follows:

$$\log(K_a) = C_0 + C_1(G3) + C_2(A3) + C_3(C4) + C_4(G3T4)$$

Where the additional variable's effect size is determined by its coefficient, C_4. This model has an extra explanatory variable compared to the null model, and we determine whether each potential second-order interaction term should be considered further via a likelihood ratio test. We also assessed the p-value for each variable, correcting for false-discovery rate of 5%; any terms that failed to reach this threshold were not considered further for this model. Finally, we construct a model that includes all statistically significant first- and second-order terms, and that model is pared down using stepwise regression (Bendel and Afifi, 1977). This final step removes any redundant first- or second-order terms, producing a final minimal model that best explains overall variation in the data, and includes only the terms reflecting genetic variation that provide the best explanatory power for the measured variation in ΔG. Overall, this approach identifies a linear model with optimized Akaike Information Criterion (AIC) score, thereby minimizing the potential for over fitting the data with excess variables.

While the AIC-optimized model effectively identifies the statistical terms with the greatest explanatory power, we wanted to ensure that our conclusions did not arise

because of overestimation of significant parameters that could be a result of failing to include non-significant terms in the model (i.e. type II error). This could inappropriately increase the amount of variation being explained by the terms we identified as significant in the AIC-optimized model. In order to assess this, we constructed a global linear model in which ΔG was modeled against all first- and second-order terms, including both the significant ones we identified in the AIC-optimized models, as well as any additional non-significant terms needed to complete the full span of possible genetic variation (Table S2). Statistical significance of terms was assessed by correcting for multiple testing (false-discovery rate of 5%). All terms were optimized and retained in the model whether they were statistically significant or not. In order to ensure that our conclusions are robust to both potential over-fitting and to overestimating effects due to type II error, we therefore limited our discussion in the text to statistical terms that were significant for both AIC-optimized and global linear models.

Molecular dynamics simulations

See Appendix C.

REFERENCES CITED

Adams, PD, PV Afonine, G Bunkoczi, VB Chen, IW Davis, N Echols, JJ Headd, LW Hung, GJ Kapral, RW Grosse-Kunstleve, AJ McCoy, NW Moriarty, R Oeffner, RJ Read, DC Richardson, JS Richardson, TC Terwilliger, and PH Zwart. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* 66, no. Pt 2: 213-221.

Adoutte, A, G Balavoine, N Lartillot, O Lespinet, B Prud'homme, and R de Rosa. 2000. The new animal phylogeny: reliability and implications. *Proc Natl Acad Sci U S A* 97, no. 9: 4453-4456.

Akey, JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* 19, no. 5: 711-722.

Alroy, I, and LP Freedman. 1992. DNA binding analysis of glucocorticoid receptor specificity mutants. *Nucleic Acids Res* 20, no. 5: 1045-1052.

Anisimova, M, and O Gascuel. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* 55, no. 4: 539-552.

Arakawa, T. 1986. Calculation of the partial specific volumes of proteins in concentrated salt, sugar, and amino acid solutions. *J Biochem* 100, no. 6: 1471-1475.

Bain, DL, AF Heneghan, KD Connaghan-Jones, and MT Miura. 2007. Nuclear receptor structure: implications for function. *Annu Rev Physiol* 69, 201-220.

Bain, DL, Q Yang, KD Connaghan, JP Robblee, MT Miura, GD Degala, JR Lambert, and NK Maluf. 2012. Glucocorticoid receptor-DNA interactions: binding energetics are the primary determinant of sequence-specific transcriptional activity. *J Mol Biol* 422, no. 1: 18-32.

Baker, CR, LN Booth, TR Sorrells, and AD Johnson. 2012. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell* 151, no. 1: 80-95.

Baker, CR, BB Tuch, and AD Johnson. 2011. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc Natl Acad Sci U S A* 108, no. 18: 7493-7498.

Barrett, RD, and HE Hoekstra. 2011. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* 12, no. 11: 767-780.

Barrett, RD, SM Rogers, and D Schluter. 2008. Natural selection on a major armor gene in threespine stickleback. *Science* 322, no. 5899: 255-257.

Barriere, A, KL Gordon, and I Ruvinsky. 2012. Coevolution within and between regulatory loci can preserve promoter function despite evolutionary rate acceleration. *PLoS Genet* 8, no. 9: e1002961.

Beato, M. 1989. Gene regulation by steroid hormones. *Cell* 56, no. 3: 335-344.

Beato, M, S Chavez, and M Truss. 1996. Transcriptional regulation by steroid hormones. *Steroids* 61, no. 4: 240-251.

Bendel, Robert B, and Abdelmonem A Afifi. 1977. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical Association* 72, no. 357: 46-53.

Bentley, Peter John. 1998. *Comparative vertebrate endocrinology*. Cambridge University Press.

Bershtein, S, M Segal, R Bekerman, N Tokuriki, and DS Tawfik. 2006. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444, no. 7121: 929-932.

Bloom, JD, ST Labthavikul, CR Otey, and FH Arnold. 2006. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* 103, no. 15: 5869-5874.

Booth, LN, BB Tuch, and AD Johnson. 2010. Intercalation of a new tier of transcription regulation into an ancient circuit. *Nature* 468, no. 7326: 959-963.

Brayer, KJ, VJ Lynch, and GP Wagner. 2011. Evolution of a derived protein-protein interaction between HoxA11 and Foxo1a in mammals caused by changes in intramolecular regulation. *Proc Natl Acad Sci U S A* 108, no. 32: E414-E420.

Breen, MS, C Kemena, PK Vlasov, C Notredame, and FA Kondrashov. 2012. Epistasis as the primary factor in molecular evolution. *Nature* 490, no. 7421: 535-538.

Bridgham, JT, SM Carroll, and JW Thornton. 2006. Evolution of hormone-receptor complexity by molecular exploitation. *Science* 312, no. 5770: 97-101.

Bridgham, JT, EA Ortlund, and JW Thornton. 2009. An epistatic ratchet constrains the direction of glucocorticoid receptor evolution. *Nature* 461, no. 7263: 515-519.

Cardona, A, L Pagani, T Antao, DJ Lawson, CA Eichstaedt, B Yngvadottir, MT Shwe, J Wee, IG Romero, S Raj, M Metspalu, R Villems, E Willerslev, C Tyler-Smith, BA Malyarchuk, MV Derenko, and T Kivisild. 2014. Genome-wide analysis of cold adaptation in indigenous Siberian populations. *PLoS One* 9, no. 5: e98076.

Carroll, SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* 3, no. 7: e245.

Carroll, SB. 2008. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, no. 1: 25-36.

Carroll, SM, EA Ortlund, and JW Thornton. 2011. Mechanisms for the evolution of a derived function in the ancestral glucocorticoid receptor. *PLoS Genet* 7, no. 6: e1002117.

Christiansen, MT, RS Kaas, RR Chaudhuri, MA Holmes, H Hasman, and FM Aarestrup. 2014. Genome-wide high-throughput screening to investigate essential genes involved in methicillin-resistant Staphylococcus aureus Sequence Type 398 survival. *PLoS One* 9, no. 2: e89018.

Chusacultanachai, S, KA Glenn, AO Rodriguez, EK Read, JF Gardner, BS Katzenellenbogen, and DJ Shapiro. 1999. Analysis of estrogen response element binding by genetically selected steroid receptor DNA binding domain mutants exhibiting altered specificity and enhanced affinity. *J Biol Chem* 274, no. 33: 23591-23598.

Cohen, HM, DS Tawfik, and AD Griffiths. 2004. Altering the sequence specificity of HaeIII methyltransferase by directed evolution using in vitro compartmentalization. *Protein Eng Des Sel* 17, no. 1: 3-11.

Cornelis, G, O Heidmann, S Bernard-Stoecklin, K Reynaud, G Veron, B Mulot, A Dupressoir, and T Heidmann. 2012. Ancestral capture of syncytin-Car1, a fusogenic endogenous retroviral envelope gene involved in placentation and conserved in Carnivora. *Proc Natl Acad Sci U S A* 109, no. 7: E432-E441.

Coulocheri, SA, DG Pigis, KA Papavassiliou, and AG Papavassiliou. 2007. Hydrogen bonds in protein-DNA complexes: where geometry meets plasticity. *Biochimie* 89, no. 11: 1291-1303.

Coyle, SM, J Flores, and WA Lim. 2013. Exploitation of latent allostery enables the evolution of new modes of MAP kinase regulation. *Cell* 154, no. 4: 875-887.

Crawford, JE, and R Nielsen. 2013. Detecting adaptive trait loci in nonmodel systems: divergence or admixture mapping? *Mol Ecol* 22, no. 24: 6131-6148.

Cresko, WA. 2008. Evolution. Armor development and fitness. *Science* 322, no. 5899: 204-206.

Cresko, WA, A Amores, C Wilson, J Murphy, M Currey, P Phillips, MA Bell, CB Kimmel, and JH Postlethwait. 2004. Parallel genetic basis for repeated evolution of armor loss in Alaskan threespine stickleback populations. *Proc Natl Acad Sci U S A* 101, no. 16: 6050-6055.

Currat, M, L Excoffier, W Maddison, SP Otto, N Ray, MC Whitlock, and S Yeaman. 2006. Comment on "Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens" and "Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans". *Science* 313, no. 5784: 172; author reply 172.

Darden, Tom, Darrin York, and Lee Pedersen. 1993. Particle mesh Ewald: An N□ log (N) method for Ewald sums in large systems. *The Journal of chemical physics* 98, no. 12: 10089-10092.

Darwin, Charles. 1859. *On the origin of species by means of natural selection*. London: J. Murray.

Dean, AM, and JW Thornton. 2007. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet* 8, no. 9: 675-688.

Dobzhansky, Theodosius. 1956. What is an adaptive trait? *American Naturalist* 337-347.

Doolittle, WF. 2012. Evolutionary biology: A ratchet for protein complexity. *Nature* 481, no. 7381: 270-271.

Drummond, DA, and CO Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134, no. 2: 341-352.

Duan, Y, C Wu, S Chowdhury, MC Lee, G Xiong, W Zhang, R Yang, P Cieplak, R Luo, T Lee, J Caldwell, J Wang, and P Kollman. 2003. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J Comput Chem* 24, no. 16: 1999-2012.

Dupradeau, FY, A Pigache, T Zaffran, C Savineau, R Lelong, N Grivel, D Lelong, W Rosanski, and P Cieplak. 2010. The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys Chem Chem Phys* 12, no. 28: 7821-7839.
Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, no. 5: 1792-1797.

Eichstaedt, CA, T Antao, L Pagani, A Cardona, T Kivisild, and M Mormina. 2014. The Andean adaptive toolkit to counteract high altitude maladaptation: genome-wide and phenotypic analysis of the Collas. *PLoS One* 9, no. 3: e93314.

Eick, GN, JK Colucci, MJ Harms, EA Ortlund, and JW Thornton. 2012. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet* 8, no. 11: e1003072.

Eick, GN, and JW Thornton. 2011. Evolution of steroid receptors from an estrogen-sensitive ancestral receptor. *Mol Cell Endocrinol* 334, no. 1-2: 31-38.

Emsley, P, and K Cowtan. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr D Biol Crystallogr* 60, no. Pt 12 Pt 1: 2126-2132.

Enard, D, PW Messer, and DA Petrov. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res* 24, no. 6: 885-895.

Evans, PD, SL Gilbert, N Mekel-Bobrov, EJ Vallender, JR Anderson, LM Vaez-Azizi, SA Tishkoff, RR Hudson, and BT Lahn. 2005. Microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science* 309, no. 5741: 1717-1720.

Excoffier, L, T Hofer, and M Foll. 2009. Detecting loci under selection in a hierarchically structured population. *Heredity (Edinb)* 103, no. 4: 285-298.

Finnigan, GC, V Hanson-Smith, TH Stevens, and JW Thornton. 2012. Evolution of increased complexity in a molecular machine. *Nature* 481, no. 7381: 360-364.

Fisher, RA. 1918. The correlation of relatives on the assumption of Mendelian inheritance. *Proc. Roy. Soc. Edinburgh*

Fisher, RA. 1935. The Sheltering of Lethals. *The American Naturalist* 69, 446-455.

Force, A, M Lynch, FB Pickett, A Amores, YL Yan, and J Postlethwait. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151, no. 4: 1531-1545.

Fretwell, Stephen. 1969. Ecotypic variation in the non-breeding season in migratory populations: a study of tarsal length in some Fringillidae. *Evolution* 406-420.

Fumagalli, M, M Sironi, U Pozzoli, A Ferrer-Admetlla, L Pattini, and R Nielsen. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet* 7, no. 11: e1002355.

Garvie, CW, and C Wolberger. 2001. Recognition of specific DNA sequences. *Mol Cell* 8, no. 5: 937-946.

Gompel, N, B Prud'homme, PJ Wittkopp, VA Kassner, and SB Carroll. 2005. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature* 433, no. 7025: 481-487.

Gong, LI, MA Suchard, and JD Bloom. 2013. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* 2, e00631.

Goodman, CS, and BC Coughlin. 2000. Introduction. The evolution of evo-devo biology. *Proc Natl Acad Sci U S A* 97, no. 9: 4424-4425.

Gould, SJ, and RC Lewontin. 1979. The spandrels of San Marco and the Panglossian paradigm: a critique of the adaptationist programme. *Proc R Soc Lond B Biol Sci* 205, no. 1161: 581-598.

Granovsky, AA. 2009. Firefly version 8. http://classic.chem.msu.su/gran/firefly/index.html

Guenther, UP, LE Yandek, CN Niland, FE Campbell, D Anderson, VE Anderson, ME Harris, and E Jankowsky. 2013. Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature* 502, no. 7471: 385-388.

Guindon, S, JF Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59, no. 3: 307-321.

Guo, HH, J Choe, and LA Loeb. 2004. Protein tolerance to random amino acid change. *Proc Natl Acad Sci U S A* 101, no. 25: 9205-9210.

Haag, ES, and JR True. 2007. Evolution and development: anchors away! *Curr Biol* 17, no. 5: R172-R174.

Haldane, JBS. 1933. The Part Played by Recurrent Mutation in Evolution. *The American Naturalist* 67, 5-19.

Hanson-Smith, V, B Kolaczkowski, and JW Thornton. 2010. Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol* 27, no. 9: 1988-1999.

Hard, T, K Dahlman, J Carlstedt-Duke, JA Gustafsson, and R Rigler. 1990. Cooperativity and specificity in the interactions between DNA and the glucocorticoid receptor DNA-binding domain. *Biochemistry* 29, no. 22: 5358-5364.

Harms, MJ, GN Eick, D Goswami, JK Colucci, PR Griffin, EA Ortlund, and JW Thornton. 2013. Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proc Natl Acad Sci U S A* 110, no. 28: 11475-11480.

Harms, MJ, and JW Thornton. 2010. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* 20, no. 3: 360-366.

Harms, MJ, and JW Thornton. 2013. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet* 14, no. 8: 559-571.

Helsen, C, S Kerkhofs, L Clinckemalie, L Spans, M Laurent, S Boonen, D Vanderschueren, and F Claessens. 2012. Structural basis for nuclear hormone receptor DNA binding. *Mol Cell Endocrinol* 348, no. 2: 411-417.

Hermisson, J. 2009. Who believes in whole-genome scans for selection? *Heredity (Edinb)* 103, no. 4: 283-284.

Hernandez, RD, JL Kelley, E Elyashiv, SC Melton, A Auton, G McVean, G Sella, and M Przeworski. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331, no. 6019: 920-924.

Hess, B, H Bekker, HJC Berendsen, and J Fraaije. 1997. LINCS: a linear constraint solver for comparative protein modeling. *J. Comput. Chem* 18, 1463-1472.

Hoekstra, HE, RJ Hirschmann, RA Bundey, PA Insel, and JP Crossland. 2006. A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science* 313, no. 5783: 101-104.

Hohenlohe, PA, S Bassham, M Currey, and WA Cresko. 2012. Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philos Trans R Soc Lond B Biol Sci* 367, no. 1587: 395-408.

Hokamp, K, A McLysaght, and KH Wolfe. 2003. The 2R hypothesis and the human genome sequence. *J Struct Funct Genomics* 3, no. 1-4: 95-110.

Hopkins, R, DA Levin, and MD Rausher. 2012. Molecular signatures of selection on reproductive character displacement of flower color in Phlox drummondii. *Evolution* 66, no. 2: 469-485.

Hopkins, R, and MD Rausher. 2011. Identification of two genes causing reinforcement in the Texas wildflower Phlox drummondii. *Nature* 469, no. 7330: 411-414.

Hopkins, R, and MD Rausher. 2012. Pollinator-mediated selection on flower color allele drives reinforcement. *Science* 335, no. 6072: 1090-1092.

Howard, C, V Hanson-Smith, KJ Kennedy, CJ Miller, HJ Lou, AD Johnson, B Turk, and LJ Holt. 2014. Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. *Elife* 3,

Huerta-Sanchez, E, M Degiorgio, L Pagani, A Tarekegn, R Ekong, T Antao, A Cardona, HE Montgomery, GL Cavalleri, PA Robbins, ME Weale, N Bradman, E Bekele, T Kivisild, C Tyler-Smith, and R Nielsen. 2013. Genetic signatures reveal high-altitude adaptation in a set of ethiopian populations. *Mol Biol Evol* 30, no. 8: 1877-1888.

Hughes, AL. 2007. Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity (Edinb)* 99, no. 4: 364-373.

Hughes, AL, and M Nei. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc Natl Acad Sci U S A* 86, no. 3: 958-962.

Humphrey, William, Andrew Dalke, and Klaus Schulten. 1996. VMD: visual molecular dynamics. *Journal of molecular graphics* 14, no. 1: 33-38.

Iwase, S, B Xiang, S Ghosh, T Ren, PW Lewis, JC Cochrane, CD Allis, DJ Picketts, DJ Patel, H Li, and Y Shi. 2011. ATRX ADD domain links an atypical histone methylation recognition mechanism to human mental-retardation syndrome. *Nat Struct Mol Biol* 18, no. 7: 769-776.

James, TY, F Kauff, CL Schoch, PB Matheny, V Hofstetter, CJ Cox, G Celio, C Gueidan, E Fraker, J Miadlikowska, HT Lumbsch, A Rauhut, V Reeb, AE Arnold, A Amtoft, JE Stajich, K Hosaka, GH Sung, D Johnson, B O'Rourke, M Crockett, M Binder, JM Curtis, JC Slot, Z Wang, AW Wilson, A Schussler, JE Longcore, K O'Donnell, S Mozley-Standridge, D Porter, PM Letcher, MJ Powell, JW Taylor, MM White, GW Griffith, DR Davies, RA Humber, JB Morton, J Sugiyama, AY Rossman, JD Rogers, DH Pfister, D Hewitt, K Hansen, S Hambleton, RA Shoemaker, J Kohlmeyer, B Volkmann-Kohlmeyer, RA Spotts, M Serdani, PW Crous, KW Hughes, K Matsuura, E Langer, G Langer, WA Untereiner, R Lucking, B Budel, DM Geiser, A Aptroot, P Diederich, I Schmitt, M Schultz, R Yahr, DS Hibbett, F Lutzoni, DJ McLaughlin, JW Spatafora, and R Vilgalys. 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* 443, no. 7113: 818-822.

Jensen, JD, and OJ Rando. 2010. Recent evidence for pervasive adaptation targeting gene expression attributable to population size change. *Proc Natl Acad Sci U S A* 107, no. 27: E109-10; author reply 111.

Jones, FC, MG Grabherr, YF Chan, P Russell, E Mauceli, J Johnson, R Swofford, M Pirun, MC Zody, S White, E Birney, S Searle, J Schmutz, J Grimwood, MC Dickson, RM Myers, CT Miller, BR Summers, AK Knecht, SD Brady, H Zhang, AA Pollen, T Howes, C Amemiya, J Baldwin, T Bloom, DB Jaffe, R Nicol, J Wilkinson, ES Lander, F Di Palma, K Lindblad-Toh, and DM Kingsley. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* 484, no. 7392: 55-61.

Kamberov, YG, S Wang, J Tan, P Gerbault, A Wark, L Tan, Y Yang, S Li, K Tang, H Chen, A Powell, Y Itan, D Fuller, J Lohmueller, J Mao, A Schachar, M Paymer, E Hostetter, E Byrne, M Burnett, AP McMahon, MG Thomas, DE Lieberman, L Jin, CJ Tabin, BA Morgan, and PC Sabeti. 2013. Modeling recent human evolution in mice by expression of a selected EDAR variant. *Cell* 152, no. 4: 691-702.

Kaufman, PK, FD Enfield, and RE Comstock. 1977. Stabilizing Selection for Pupa Weight in TRIBOLIUM CASTANEUM. *Genetics* 87, no. 2: 327-341.

Khersonsky, O, C Roodveldt, and DS Tawfik. 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* 10, no. 5: 498-508.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* 217, no. 5129: 624-626.

Kimura, M. 1977. Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* 267, no. 5608: 275-276.

King, MC, and AC Wilson. 1975. Evolution at two levels in humans and chimpanzees. *Science* 188, no. 4184: 107-116.

Koop, BF, DA Tagle, M Goodman, and JL Slightom. 1989. A molecular view of primate phylogeny and important systematic and evolutionary questions. *Mol Biol Evol* 6, no. 6: 580-612.

Kumar, V, and P Chambon. 1988. The estrogen receptor binds tightly to its responsive element as a ligand-induced homodimer. *Cell* 55, no. 1: 145-156.

Landry, CR, PJ Wittkopp, CH Taubes, JM Ranz, AG Clark, and DL Hartl. 2005. Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics* 171, no. 4: 1813-1822.

Laskowski, RA, DS Moss, and JM Thornton. 1993. Main-chain bond lengths and bond angles in protein structures. *J Mol Biol* 231, no. 4: 1049-1067.

Leibniz, GW. 1710. Causa Dei Asserta per Iustitiam Eius: Cum caeteris eius Perfectionibus, Cunctisque Actionibus Conciliatam. *books.google.com*

Lewontin, Richard C. 1974. *The genetic basis of evolutionary change*. Columbia University Press New York.

Li, H, and AD Johnson. 2010. Evolution of transcription networks--lessons from yeasts. *Curr Biol* 20, no. 17: R746-R753.

Li, Y, DD Wu, AR Boyko, GD Wang, SF Wu, DM Irwin, and YP Zhang. 2014. Population variation revealed high-altitude adaptation of Tibetan mastiffs. *Mol Biol Evol* 31, no. 5: 1200-1205.

Lin, Fu, and Renxiao Wang. 2010. Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. *Journal of Chemical Theory and Computation* 6, no. 6: 1852-1870.

Lisewski, AM. 2008. Random amino acid mutations and protein misfolding lead to Shannon limit in sequence-structure communication. *PLoS One* 3, no. 9: e3110.

Liu, S, ED Lorenzen, M Fumagalli, B Li, K Harris, Z Xiong, L Zhou, TS Korneliussen, M Somel, C Babbitt, G Wray, J Li, W He, Z Wang, W Fu, X Xiang, CC Morgan, A Doherty, MJ O'Connell, JO McInerney, EW Born, L Dalen, R Dietz, L Orlando, C Sonne, G Zhang, R Nielsen, E Willerslev, and J Wang. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* 157, no. 4: 785-794.

Luisi, BF, WX Xu, Z Otwinowski, LP Freedman, KR Yamamoto, and PB Sigler. 1991. Crystallographic analysis of the interaction of the glucocorticoid receptor with DNA. *Nature* 352, no. 6335: 497-505.

Lundback, T, C Cairns, JA Gustafsson, J Carlstedt-Duke, and T Hard. 1993. Thermodynamics of the glucocorticoid receptor-DNA interaction: binding of wild-type GR DBD to different response elements. *Biochemistry* 32, no. 19: 5074-5082.

Lynch, M, and V Katju. 2004. The altered evolutionary trajectories of gene duplicates. *Trends Genet* 20, no. 11: 544-549.

Lynch, VJ, G May, and GP Wagner. 2011. Regulatory evolution through divergence of a phosphoswitch in the transcription factor CEBPB. *Nature* 480, no. 7377: 383-386.

MacCallum, C, and E Hill. 2006. Being positive about selection. *PLoS Biol* 4, no. 3: e87.
Markova-Raina, P, and D Petrov. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res* 21, no. 6: 863-874.

Mayr, Ernst. 1983. How to carry out the adaptationist program? *American Naturalist* 324-334.

McDonald, JH, and M Kreitman. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351, no. 6328: 652-654.

McKeown, AN, JT Bridgham, DW Anderson, MN Murphy, EA Ortlund, and JW Thornton. 2014. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* 159, no. 1: 58-68.

Meijsing, SH, MA Pufall, AY So, DL Bates, L Chen, and KR Yamamoto. 2009. DNA binding site sequence directs glucocorticoid receptor structure and activity. *Science* 324, no. 5925: 407-410.

Mekel-Bobrov, N, SL Gilbert, PD Evans, EJ Vallender, JR Anderson, RR Hudson, SA Tishkoff, and BT Lahn. 2005. Ongoing adaptive evolution of ASPM, a brain size determinant in Homo sapiens. *Science* 309, no. 5741: 1720-1722.

Mendel, Gregor. 1866. Versuche über Pflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brunn* 4: 344.

Meyerguz, L, J Kleinberg, and R Elber. 2007. The network of sequence flow between protein structures. *Proc Natl Acad Sci U S A* 104, no. 28: 11627-11632.

Natarajan, C, N Inoguchi, RE Weber, A Fago, H Moriyama, and JF Storz. 2013. Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* 340, no. 6138: 1324-1327.

Nelson, CC, SC Hendy, RJ Shukin, H Cheng, N Bruchovsky, BF Koop, and PS Rennie. 1999. Determinants of DNA sequence specificity of the androgen, progesterone, and glucocorticoid receptors: evidence for differential steroid receptor response elements. *Mol Endocrinol* 13, no. 12: 2090-2107.

Nunes, MD, PO Wengel, M Kreissl, and C Schlotterer. 2010. Multiple hybridization events between Drosophila simulans and Drosophila mauritiana are supported by mtDNA introgression. *Mol Ecol* 19, no. 21: 4695-4707.

Ohno, Susumu. 1970. *Evolution by gene duplication.* Berlin, New York: Springer-Verlag.
Ohta, T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor Popul Biol* 10, no. 3: 254-275.

Otwinowski, Z, and W Minor. 1997. Processing of X-ray crystallographic data in oscillation mode. *Methods Enzymol* 276, 307-326.

Pace, C Nick, and J Martin Scholtz. 1997. Measuring the conformational stability of a protein. *Protein structure: A practical approach* 2, 299-321.

Pavlidis, P, JD Jensen, W Stephan, and A Stamatakis. 2012. A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. *Mol Biol Evol* 29, no. 10: 3237-3248.

Phillips, PC. 2008. Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9, no. 11: 855-867.

Postlethwait, JH, YL Yan, MA Gates, S Horne, A Amores, A Brownlie, A Donovan, ES Egan, A Force, Z Gong, C Goutel, A Fritz, R Kelsh, E Knapik, E Liao, B Paw, D Ransom, A Singer, M Thomson, TS Abduljabbar, P Yelick, D Beier, JS Joly, D Larhammar, F Rosa, M Westerfield, LI Zon, SL Johnson, and WS Talbot. 1998. Vertebrate genome evolution and the zebrafish gene map. *Nat Genet* 18, no. 4: 345-349.

Prakash, S, RC Lewontin, and JL Hubby. 1969. A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of Drosophila pseudoobscura. *Genetics* 61, no. 4: 841-858.

Pritchard, JK, and A Di Rienzo. 2010. Adaptation - not by sweeps alone. *Nat Rev Genet* 11, no. 10: 665-667.

Pryor, KD, and B Leiting. 1997. High-level expression of soluble protein in Escherichia coli using a His6-tag and maltose-binding-protein double-affinity fusion system. *Protein Expr Purif* 10, no. 3: 309-319.

Qian, W, JR Yang, NM Pearson, C Maclean, and J Zhang. 2012. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* 8, no. 3: e1002603.

Reimchen, Thomas E. 1994. Predators and morphological evolution in threespine stickleback. *The evolutionary biology of the threespine stickleback* 240-276.

Rockah-Shmuel, L, and DS Tawfik. 2012. Evolutionary transitions to new DNA methyltransferases through target site expansion and shrinkage. *Nucleic Acids Res* 40, no. 22: 11627-11637.

Roemer, SC, DC Donham, L Sherman, VH Pon, DP Edwards, and ME Churchill. 2006. Structure of the progesterone receptor-deoxyribonucleic acid complex: novel interactions required for binding to half-site response elements. *Mol Endocrinol* 20, no. 12: 3042-3052.

Rohs, R, X Jin, SM West, R Joshi, B Honig, and RS Mann. 2010. Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* 79, 233-269.

Rokas, A, BL Williams, N King, and SB Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425, no. 6960: 798-804.

Sabeti, PC, P Varilly, B Fry, J Lohmueller, E Hostetter, C Cotsapas, X Xie, EH Byrne, SA McCarroll, R Gaudet, SF Schaffner, ES Lander, KA Frazer, DG Ballinger, DR Cox, DA Hinds, LL Stuve, RA Gibbs, JW Belmont, A Boudreau, P Hardenbol, SM Leal, S Pasternak, DA Wheeler, TD Willis, F Yu, H Yang, C Zeng, Y Gao, H Hu, W Hu, C Li, W Lin, S Liu, H Pan, X Tang, J Wang, W Wang, J Yu, B Zhang, Q Zhang, H Zhao, H Zhao, J Zhou, SB Gabriel, R Barry, B Blumenstiel, A Camargo, M Defelice, M Faggart, M Goyette, S Gupta, J Moore, H Nguyen, RC Onofrio, M Parkin, J Roy, E Stahl, E Winchester, L Ziaugra, D Altshuler, Y Shen, Z Yao, W Huang, X Chu, Y He, L Jin, Y Liu, Y Shen, W Sun, H Wang, Y Wang, Y Wang, X Xiong, L Xu, MM Waye, SK Tsui, H Xue, JT Wong, LM Galver, JB Fan, K Gunderson, SS Murray, AR Oliphant, MS Chee, A Montpetit, F Chagnon, V Ferretti, M Leboeuf, JF Olivier, MS Phillips, S Roumy, C Sallee, A Verner, TJ Hudson, PY Kwok, D Cai, DC Koboldt, RD Miller, L Pawlikowska, P Taillon-Miller, M Xiao, LC Tsui, W Mak, YQ Song, PK Tam, Y Nakamura, T Kawaguchi, T Kitamoto, T Morizono, A Nagashima, Y Ohnishi, A Sekine, T Tanaka, T Tsunoda, P Deloukas, CP Bird, M Delgado, ET Dermitzakis, R Gwilliam, S Hunt, J Morrison, D Powell, BE Stranger, P Whittaker, DR Bentley, MJ Daly, PI de Bakker, J Barrett, YR Chretien, J Maller, S McCarroll, N Patterson, I Pe'er, A Price, S Purcell, DJ Richter, P Sabeti, R Saxena, SF Schaffner, PC Sham, P Varilly, D Altshuler, LD Stein, L Krishnan, AV Smith, MK Tello-Ruiz, GA Thorisson, A Chakravarti, PE Chen, DJ Cutler, CS Kashuk, S Lin, GR Abecasis, W Guan, Y Li, HM Munro, ZS Qin, DJ Thomas, G McVean, A Auton, L Bottolo, N Cardin, S Eyheramendy, C Freeman, J Marchini, S Myers, C Spencer, M Stephens, P Donnelly, LR Cardon, G Clarke, DM Evans, AP Morris, BS Weir, T Tsunoda, TA Johnson, JC Mullikin, ST Sherry, M Feolo, A Skol, H Zhang, C Zeng, H Zhao, I Matsuda, Y Fukushima, DR Macer, E Suda, CN Rotimi, CA Adebamowo, I Ajayi, T Aniagwu, PA Marshall, C Nkwodimmah, CD Royal, MF Leppert, M Dixon, A Peiffer, R Qiu, A Kent, K Kato, N Niikawa, IF Adewole, BM Knoppers, MW Foster, EW Clayton, J Watkin, RA Gibbs, JW Belmont, D Muzny, L Nazareth, E Sodergren, GM Weinstock, DA Wheeler, I Yakub, SB Gabriel, RC Onofrio, DJ Richter, L Ziaugra, BW Birren, MJ Daly, D Altshuler, RK Wilson, LL Fulton, J Rogers, J Burton, NP Carter, CM Clee, M Griffiths, MC Jones, K McLay, RW Plumb, MT Ross, SK Sims, DL Willey, Z Chen, H Han, L Kang, M Godbout, JC Wallenburg, P L'Archeveque, G Bellemare, K Saeki, H Wang, D An, H Fu, Q Li, Z Wang, R Wang, AL Holden, LD Brooks, JE McEwen, MS Guyer, VO Wang, JL Peterson, M Shi, J Spiegel, LM Sung, LF Zacharia, FS Collins, K Kennedy, R Jamieson, and J Stewart. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, no. 7164: 913-918.

Sanger, F, and AR Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94, no. 3: 441-448.

Sayou, C, M Monniaux, MH Nanao, E Moyroud, SF Brockington, E Thevenon, H Chahtane, N Warthmann, M Melkonian, Y Zhang, GK Wong, D Weigel, F Parcy, and R Dumas. 2014. A promiscuous intermediate underlies the evolution of LEAFY DNA binding specificity. *Science* 343, no. 6171: 645-648.

Schmidt, Michael W, Kim K Baldridge, Jerry A Boatz, Steven T Elbert, Mark S Gordon, Jan H Jensen, Shiro Koseki, Nikita Matsunaga, Kiet A Nguyen, and Shujun Su. 1993. General atomic and molecular electronic structure system. *Journal of Computational Chemistry* 14, no. 11: 1347-1363.

Schuchardt, KL, BT Didier, T Elsethagen, L Sun, V Gurumoorthi, J Chase, J Li, and TL Windus. 2007. Basis set exchange: a community database for computational sciences. *J Chem Inf Model* 47, no. 3: 1045-1052.

Schwabe, JW, L Chapman, JT Finch, D Rhodes, and D Neuhaus. 1993. DNA recognition by the oestrogen receptor: from solution to the crystal. *Structure* 1, no. 3: 187-204.

Schwabe, JW, and D Rhodes. 1991. Beyond zinc fingers: steroid hormone receptors have a novel structural motif for DNA recognition. *Trends Biochem Sci* 16, no. 8: 291-296.

Shao, H, LC Burrage, DS Sinasac, AE Hill, SR Ernest, W O'Brien, HW Courtland, KJ Jepsen, A Kirby, EJ Kulbokas, MJ Daly, KW Broman, ES Lander, and JH Nadeau. 2008. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc Natl Acad Sci U S A* 105, no. 50: 19910-19914.

Smith, JM. 1970. Natural selection and the concept of a protein space. *Nature* 225, no. 5232: 563-564.

So, AY, C Chaivorapol, EC Bolton, H Li, and KR Yamamoto. 2007. Determinants of cell- and gene-specific transcriptional regulation by the glucocorticoid receptor. *PLoS Genet* 3, no. 6: e94.

Stadler, BM, PF Stadler, GP Wagner, and W Fontana. 2001. The topology of the possible: formal spaces underlying patterns of evolutionary change. *J Theor Biol* 213, no. 2: 241-274.

Stapley, J, J Reger, PG Feulner, C Smadja, J Galindo, R Ekblom, C Bennison, AD Ball, AP Beckerman, and J Slate. 2010. Adaptation genomics: the next generation. *Trends Ecol Evol* 25, no. 12: 705-712.

Steane, DA, BM Potts, E McLean, SM Prober, WD Stock, RE Vaillancourt, and M Byrne. 2014. Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Mol Ecol* 23, no. 10: 2500-2513.

Steiner, CC, JN Weber, and HE Hoekstra. 2007. Adaptive variation in beach mice produced by two interacting pigmentation genes. *PLoS Biol* 5, no. 9: e219.

Stergachis, AB, E Haugen, A Shafer, W Fu, B Vernot, A Reynolds, A Raubitschek, S Ziegler, EM LeProust, JM Akey, and JA Stamatoyannopoulos. 2013. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* 342, no. 6164: 1367-1372.

Stern, DL, and V Orgogozo. 2009. Is genetic evolution predictable? *Science* 323, no. 5915: 746-751.

Storz, JF, AM Runck, SJ Sabatino, JK Kelly, N Ferrand, H Moriyama, RE Weber, and A Fago. 2009. Evolutionary and functional insights into the mechanism underlying high-altitude adaptation of deer mouse hemoglobin. *Proc Natl Acad Sci U S A* 106, no. 34: 14450-14455.

Storz, JF, SJ Sabatino, FG Hoffmann, EJ Gering, H Moriyama, N Ferrand, B Monteiro, and MW Nachman. 2007. The molecular basis of high-altitude adaptation in deer mice. *PLoS Genet* 3, no. 3: e45.

Tan, J, Y Yang, K Tang, PC Sabeti, L Jin, and S Wang. 2013. The adaptive variant EDARV370A is associated with straight hair in East Asians. *Hum Genet* 132, no. 10: 1187-1191.

Teichmann, M, G Dieci, C Pascali, and G Boldina. 2010. General transcription factors and subunits of RNA polymerase III: Paralogs for promoter- and cell type-specific transcription in multicellular eukaryotes. *Transcription* 1, no. 3: 130-135.

Teichmann, SA, and MM Babu. 2004. Gene regulatory network growth by duplication. *Nat Genet* 36, no. 5: 492-496.

Thornton, JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet* 5, no. 5: 366-375.

Timpson, N, J Heron, GD Smith, and W Enard. 2007. Comment on papers by Evans et al. and Mekel-Bobrov et al. on Evidence for Positive Selection of MCPH1 and ASPM. *Science* 317, no. 5841: 1036; author reply 1036.

True, JR, and ES Haag. 2001. Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev* 3, no. 2: 109-119.

Tuch, BB, DJ Galgoczy, AD Hernday, H Li, and AD Johnson. 2008. The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 6, no. 2: e38.

Tyulmenkov, VV, SC Jernigan, and CM Klinge. 2000. Comparison of transcriptional synergy of estrogen receptors alpha and beta from multiple tandem estrogen response elements. *Mol Cell Endocrinol* 165, no. 1-2: 151-161.

Udpa, N, R Ronen, D Zhou, J Liang, T Stobdan, O Appenzeller, Y Yin, Y Du, L Guo, R Cao, Y Wang, X Jin, C Huang, W Jia, D Cao, G Guo, VE Claydon, R Hainsworth, JL Gamboa, M Zibenigus, G Zenebe, J Xue, S Liu, KA Frazer, Y Li, V Bafna, and GG Haddad. 2014. Whole genome sequencing of Ethiopian highlanders reveals conserved hypoxia tolerance genes. *Genome Biol* 15, no. 2: R36.

Umesono, K, and RM Evans. 1989. Determinants of target gene specificity for steroid/thyroid hormone receptors. *Cell* 57, no. 7: 1139-1146.

Vandepoele, K, W De Vos, JS Taylor, A Meyer, and Y Van de Peer. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci U S A* 101, no. 6: 1638-1643.

Vignieri, SN, JG Larson, and HE Hoekstra. 2010. The selective advantage of crypsis in mice. *Evolution* 64, no. 7: 2153-2158.

Voltaire, François. 1759. *Candide: ou, L'optimiste*. La Sirène.

von Hippel, PH, and OG Berg. 1986. On the specificity of DNA-protein interactions. *Proc Natl Acad Sci U S A* 83, no. 6: 1608-1612.

Wagner, A. 2008. Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* 9, no. 12: 965-974.

Walker, Jeffrey A, and Michael A Bell. 2000. Net evolutionary trajectories of body shape evolution within a microgeographic radiation of threespine sticklebacks (Gasterosteus aculeatus). *Journal of Zoology* 252, no. 3: 293-302.

Watson, LC, KM Kuchenbecker, BJ Schiller, JD Gross, MA Pufall, and KR Yamamoto. 2013. The glucocorticoid receptor dimer interface allosterically transmits sequence-specific DNA signals. *Nat Struct Mol Biol* 20, no. 7: 876-883.

Welboren, WJ, MA van Driel, EM Janssen-Megens, SJ van Heeringen, FC Sweep, PN Span, and HG Stunnenberg. 2009. ChIP-Seq of ERalpha and RNA polymerase II defines genes differentially responding to ligands. *EMBO J* 28, no. 10: 1418-1428.

Welch, AJ, OC Bedoya-Reina, L Carretero-Paulet, W Miller, KD Rode, and C Lindqvist. 2014. Polar bears exhibit genome-wide signatures of bioenergetic adaptation to life in the arctic environment. *Genome Biol Evol* 6, no. 2: 433-450.

Wikstrom, A, H Berglund, C Hambraeus, S van den Berg, and T Hard. 1999. Conformational dynamics and molecular recognition: backbone dynamics of the estrogen receptor DNA-binding domain. *J Mol Biol* 289, no. 4: 963-979.

Wittkopp, PJ, BK Haerum, and AG Clark. 2008. Independent effects of cis- and trans-regulatory variation on gene expression in Drosophila melanogaster. *Genetics* 178, no. 3: 1831-1835.

Wray, GA. 2007. The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8, no. 3: 206-216.

Wright, S. 1932. The Roles of Mutation, Inbreeding, Crossbreeding, and Selection in Evolution. *Proceedings of the Sixth International Congress of Genetics* 1, 356-366.

Wuren, T, TS Simonson, G Qin, J Xing, CD Huff, DJ Witherspoon, LB Jorde, and RL Ge. 2014. Shared and unique signals of high-altitude adaptation in geographically distinct Tibetan populations. *PLoS One* 9, no. 3: e88252.

Yang, Z, S Kumar, and M Nei. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, no. 4: 1641-1650.

Yokoyama, S, T Tada, H Zhang, and L Britt. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proc Natl Acad Sci U S A* 105, no. 36: 13480-13485.

Zhuang, H, MS Chien, and H Matsunami. 2009. Dynamic functional evolution of an odorant receptor for sex-steroid-derived odors in primates. *Proc Natl Acad Sci U S A* 106, no. 50: 21247-21251.

Zilliacus, J, K Dahlman-Wright, A Wright, JA Gustafsson, and J Carlstedt-Duke. 1991. DNA binding specificity of mutant glucocorticoid receptor DNA-binding domains. *J Biol Chem* 266, no. 5: 3101-3106.

Zilliacus, J, AP Wright, U Norinder, JA Gustafsson, and J Carlstedt-Duke. 1992. Determinants for DNA-binding site recognition by the glucocorticoid receptor. *J Biol Chem* 267, no. 35: 24941-24947.