MOLECULAR MECHANISMS FOR THE EVOLUTION OF DNA SPECIFICITY IN A TRANSCRIPTION FACTOR FAMILY

by

ALESIA NICOLE MCKEOWN

A DISSERTATION

Presented to the Department of Chemistry and Biochemistry
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: Alesia Nicole McKeown

Title: Molecular Mechanisms for the Evolution of DNA Specificity in a Transcription Factor Family

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

Kenneth Prehoda	Chairperson
Joseph Thornton	Advisor
Bradley Nolen	Core Member
Patrick Phillips	Core Member
Raghuveer Parthasarathy	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

Alesia Nicole McKeown

Doctor of Philosophy

Department of Chemistry and Biochemistry

December 2014

Title: Molecular Mechanisms for the Evolution of DNA Specificity in a Transcription Factor Family

Transcription factors (TFs) bind to specific DNA sequences near target genes to precisely coordinate their regulation. Despite the central role of transcription factors in development and homeostasis, the mechanisms by which TFs have evolved to bind and regulate distinct DNA sequences are poorly understood.

This dissertation details the highly collaborative work to determine the genetic, biochemical and biophysical mechanisms by which distinct DNA-binding specificities evolved in the steroid receptor (SR) family of transcription factors. Using ancestral protein reconstruction, we resurrected and functionally characterized the historical transition in DNA-binding specificity between ancient SR proteins. We found that DNA-binding specificity evolved by changes in the energetic components of binding; interactions at the protein-DNA interface were weakened while inter-protein cooperativity was greatly improved.

We identified a group of fourteen historical substitutions that were sufficient to recapitulate the derived protein's binding function. Three of these substitutions, which we defined as function-switching, were sufficient to change DNA specificity; however, their introduction greatly decreased binding affinity and was deleterious for protein function. A group of eleven permissive substitutions, which had no effect on DNA specificity, allowed for the protein to tolerate the deleterious effects of the function-switching substitutions. They non-specifically increased binding affinity by improving interactions at the protein-DNA interface and increasing inter-protein cooperativity.

We then dissected the functional role of individual substitutions in both the function-switching and permissive groups. We first determined the binding affinity of all possible combinations of function-switching substitutions for a library of DNA sequences. This allowed for us to functionally characterize the sequence space that separated the ancestral and derived DNA-binding specificities as well as identify the genetic determinants for DNA specificity. Lastly, we dissected the effects of the permissive substitutions on the energetics of DNA binding to determine the mechanisms by which they exerted their permissive effect. Together, this work provides insight into the molecular determinants of DNA specificity and identifies the molecular mechanisms by which these interactions changed during the evolution of novel specificity in an important transcription factor family.

This dissertation includes previously published and unpublished co-authored material.

CURRICULUM VITAE

NAME OF AUTHOR: Alesia McKeown

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
University of North Carolina at Wilmington

DEGREES AWARDED:

Doctor of Philosophy, Chemistry, 2014, University of Oregon
Bachelor of Science, Chemistry, 2009, University of North Carolina at Wilmington

AREAS OF SPECIAL INTEREST:

Molecular Evolution
Biochemistry
Structural Biology

PROFESSIONAL EXPERIENCE:

Graduate Teaching Assistant, Department of Chemistry, University of Oregon, Eugene, 2009-2010

GRANTS, AWARDS, AND HONORS:

NIH training grant in Molecular Biology and Biophysics, 2010-2013

B.S. awarded *summa cum laude* and with honors, University of North Carolina at Wilmington, 2009

PUBLICATIONS:

McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW. 2014. Evolution of DNA specificity in a transcription factor family produced a new gene regulatory module. *Cell* 159: 58-68.

McKeown AN, Naro JL, Huskins L, Almeida PF. 2011. A thermodynamic approach to the mechanism of cell-penetrating peptides in model membranes. *Biochemistry* 50: 654-662.

Clark KS, Svetlovics J, **McKeown AN**, Huskins, L, Almeida PF. 2011. What determines the activity of antimicrobial and cytolytic peptides in model membranes? *Biochemistry*. 50: 7919-7932.

Settles EI, Loftus AF, **McKeown AN**, Parthasarathy R. 2010. The vesicle trafficking protein Sar1 lowers lipid membrane rigidity. *Biophysical Journal*. 99: 1539-1545.

ACKNOWLEDGMENTS

This dissertation is a product of the collaborative ideas and efforts of many brilliant minds, to all of whom I am eternally grateful.

I first want to thank my advisor, Joe Thornton, for his mentorship during my graduate career. His guidance has helped me to become a better scholar, writer and scientist. I also want to thank my committee members, Ken Prehoda, Brad Nolen, Raghu Parthasarathy and Patrick Phillips, for their diverse ideas and constructive feedback on my project throughout the years.

Additionally, I want to thank Mike Harms for always being available and teaching me nearly everything I know about protein biochemistry and biophysics. I also want to thank all of the members of the Thornton lab –both past and present—for fostering a supportive and intellectual environment and making our lab such a great place to call home for the past five years.

I especially want to recognize my friend and fellow lab mate, Dave Anderson. Our collaborative project was both the most rewarding and energizing part of my graduate work. I could not have envisioned a better person to have in the trenches with me, fighting the good fight. Long live the soul and spirit of the unstoppable Da'Lesia.

For the published work in Chapter II, I specifically thank Geeta Eick for phylogenetic analysis and Mike Harms for extensive advice. I also thank Vincent Lynch, Pete von Hippel, members of the Thornton Lab, and Will Hudson for comments on the manuscript. The University of Oregon ACISS cluster provided computing resources.

I am very fortunate to have had such a supportive group of friends during these past few years. A special thanks to Luke Helgeson, Javier Fierro Jr. and Andrew Loftus for being my backpacking companions and fellow adventurers. These weekend distractions with them were, hands down, my favorite part about graduate school.

I also want to acknowledge my undergraduate advisor, Paulo Almeida, for being the catalyst of my journey into the exciting and infinite world of scientific research. His continual investment into my development as a scientist has made all the difference.

This dissertation is dedicated to my mother, who has
always been my biggest fan.
Her support and encouragement has made all of this possible.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

Chapter II

Chapter III

Chapter IV

Figure Page

LIST OF TABLES

Table Page

Chapter III

CHAPTER I
INTRODUCTION

Studies of molecular evolution allow resolution of key questions that exist at the interface of biochemistry and evolutionary biology

The fields of evolutionary biology and biochemistry have long been treated as separate entities (Dietrich, 1998). While biochemists ask questions relating to the molecular mechanisms by which systems function, evolutionary biologists investigate the processes by which these systems came to be. However, the independent goals of these two fields are not mutually exclusive. Rather, there exist many questions at their interface that, upon investigation using approaches from both fields, would greatly impact the understanding of molecular systems and the evolutionary processes that created them (Dean and Thornton, 2007; Harms and Thornton, 2013).

The field of evolutionary biology can greatly benefit from the application of molecular biology and biochemical approaches. Currently, numerous studies of evolution depend on analysis of sequence variants within and between populations to understand the genetic mechanisms by which organisms evolved (Fay and Wu, 2003; Ghedin et al., 2005; Kasahara et al., 2007; Lindblad-Toh et al., 2011; Liu et al., 2014). The application of molecular and biochemical techniques to investigate the effects of these historical sequence changes is a critical component in determining how evolutionary changes in an organism's genetic sequence translated to changes in phenotypes (Alberch, 1991; Barrett and Hoekstra, 2011; Wilke, 2012). A molecular evolutionary approach will therefore result in the resolution of several key questions in evolutionary biology. How does functional novelty arise? Do novel functions mainly evolve by few substitutions of large effect or by a large group of substitutions, each of small effect (Orr, 2005)? What are the roles of promiscuous intermediates in the evolution of novelty? Does novelty evolve by exploitation of latent ancestral functions (Tawfik, 2010) or by establishing new interactions completely *de novo*? How does epistasis shape the evolutionary processes that give rise to functional novelty (Kondrashov et al., 2002; Phillips, 2008; Breen et al., 2012; McCandlish et al., 2013)? Does it constrain some mutational pathways while permitting others? What is the role of permissive substitutions in evolution of novel function? How did neutral and non-neutral evolutionary processes, such as selection and drift, give rise to the observed functional diversity (Wagner, 2008; Barrett and Hoekstra, 2011)? Resolution of these questions would be greatly facilitated by the application of tools and approaches from the fields of molecular biology and biochemistry.

Studying molecular systems in an evolutionary framework will also allow a better understanding of the molecular determinants of protein function. By understanding the mechanisms by which evolution "tinkered" with ancestral proteins to give rise to their derived forms, we can begin to understand the biophysical and biochemical constraints that govern the protein's sequence-structure-function relationships. This results in the resolution of several questions regarding the molecular determinants of protein function and how they change to give rise to functional novelty. How do changes in the protein give rise to functional novelty (Soskine and Tawfik, 2010)? Do novel functions evolve by mutations in residues that only occur at important binding interfaces or do mutations throughout the protein coordinate to give rise to functional novelty? Is a novel function the product of the exploitation of a promiscuous or latent activity (Aharoni et al., 2005)? If so, what are the roles of positive and negative interactions in eliminating the ancestral function while establishing the derived function? What is the source of intra-protein epistasis and what are the mechanisms for this genetic epistasis even in the absence of a physical interaction? Further, how have a protein's evolutionary history shaped its biophysical architecture (DePristo et al., 2005; Worth et al., 2009; Harms and Thornton, 2013)?

A synthesis of the fields of biochemistry and evolutionary biology has led to a greater understanding of the molecular determinants of protein function across diverse macromolecular systems and offered mechanistic insights into the evolutionary processes by which these systems changed to give rise to functional novelty (Ortlund et al., 2007; Yokoyama et al., 2008; Bloom et al., 2010; Lynch et al., 2011; Finnigan et al., 2012; Natarajan et al., 2013). In this dissertation, we take a similar approach to investigate the precise molecular mechanisms by which a family of transcription factors evolved to specifically recognize distinct DNA sequences.

What are the molecular mechanisms by which regulatory networks evolve

In 1969, Britten and Davidson proposed that changes in gene regulation were the dominant mechanisms for the evolution of novel traits (Britten and Davidson, 1969). Since then, multiple studies have found that changes in gene regulatory networks have led to the evolution of many diverse traits across species (Quattrocchio et al., 1999; Mann and Morata, 2000; Babu et al., 2004; Shapiro et al., 2004; Olson, 2006; Prud'homme et al., 2007; Lynch et al., 2008; Peter and Davidson, 2011). However, despite the central role of regulatory network diversification in evolution, the mechanisms by which these networks evolve are poorly understood.

At their simplest, gene regulatory networks are built from interconnected modules that orchestrate a cascade of interactions between cellular stimuli, transcription factors (TFs) and target DNA response elements. A trademark of these networks is specificity; a TF responds to a specific cellular signal and then binds to a specific *cis*-acting DNA response element (RE) to regulate a specific target gene. Over time, changes in the specific interactions of these interacting components can have drastic effects on regulatory network architecture (Babu et al., 2004; Teichmann and Babu, 2004; Erwin and Davidson, 2009; Peter and Davidson, 2011) thereby leading to differential regulation of many cellular processes.

Attempts to understand how regulatory networks evolve have traditionally focused on how changes in *cis*-acting elements, such as target gene REs, evolved to allow regulation by a novel, pre-existing transcription factor (Wray, 2007; Carroll, 2008; Peter and Davidson, 2011). Investigations into the role of transcription factor diversification in the evolution of regulatory networks is much more rare. Of the studies addressing TF diversification, the focus has primarily been on changes in protein-protein (Baker et al., 2011; Brayer et al., 2011; Lynch et al., 2011; Baker et al., 2012) or protein-ligand interactions (Bridgham et al., 2006; Bridgham et al., 2009; Eick et al., 2012). Despite the diverse specificities of modern day TFs for DNA (Babu et al., 2004; Badis et al., 2009; Baker et al., 2011; Jolma et al., 2013; Nakagawa et al., 2013) there exist only one study (Sayou et al., 2014) that has investigated the genetic mechanisms by which naturally occurring TFs have evolved to specifically recognize distinct RE sequences. As such, little is known about the evolutionary processes by which modern-day TFs evolved to give rise to such diverse DNA-binding specificities. Does novel specificity evolve by a discrete switch or by subfunctionalization of a promiscuous intermediate TF? How many substitutions are required to cause a switch in specificity? Do function-switching substitutions solely occur at the protein-DNA interface or are they scattered throughout the structure of the protein? What are the roles of permissive substitutions in the evolution of novel specificity and how do these residues interact to give rise to a novel function? Lastly, how does the biophysical architecture of protein-DNA interactions shape TF evolution?

The uncertainty in the evolutionary mechanisms that contribute to TF diversity mirrors uncertainty in our understanding of the biochemical and biophysical mechanisms that give rise to specific protein-DNA interactions. Common approaches to understand the precise molecular mechanisms of protein-DNA recognition have largely relied on structural and biochemical analysis of specific protein-DNA complexes (Luisi et al., 1991; Schwabe et al., 1993; Keller et al., 1995; Wuttke et al., 1997; Grazulis et al., 2002; Campagne et al., 2010). In these studies, many have identified the importance of hydrogen-bonding and van der Waals interactions in forming a high-affinity interaction between a DNA-binding protein and its preferred DNA sequence (von Hippel, 1994; Garvie and Wolberger, 2001; Coulocheri et al., 2007; Rohs et al., 2010). Although these positive interactions contribute to high-affinity binding, a protein's binding specificity is not solely determined by high-affinity interactions with its preferred sequence. Rather, its specificity is determined by its distribution of affinities for target and off-target, non-preferred sequences (von Hippel, 1994; Pan et al., 2010; Stormo and Zhao, 2010). Given that very few studies have investigated the molecular interactions between proteins and both their target and non-target sequences (Winkler et al., 1993; Sapienza et al., 2014), many questions remain unanswered regarding the molecular mechanisms that govern specific protein-DNA interactions. What are the roles of positive and negative interactions in determining specificity? Is specificity largely due to differences in positive interactions—like hydrogen bonding and van der Waals forces—or do negative interactions—such as unpaired polar atoms and steric clashes (von Hippel and Berg, 1986)—also contribute? Is a protein's affinity for DNA determined by residues that participate in direct polar interactions with the DNA or do residues outside of the protein-DNA interface also play a role in establishing specific, high affinity interactions?

Given these unresolved questions, a molecular evolutionary approach to investigate the mechanisms by which transcription factors evolve can lend valuable insight into both the molecular determinants of DNA-binding specificity as well as the evolutionary processes by which regulatory networks evolve.

A molecular evolutionary approach to investigate the mechanisms of transcription factor evolution

The steroid hormone receptors (SRs) are a great model system to study the evolutionary and biochemical mechanisms for the evolution of DNA specificity. SRs are a class of ligand-activated transcription factors that regulate the classic response to sex and adrenal steroid hormones in vertebrate development, reproduction and physiology (Bentley, 1998). These proteins contain a highly conserved DNA-binding domain (DBD) that binds directly to specific DNA sequences upstream of target genes (Bain et al., 2007). All SRs bind cooperatively as dimers to an inverted palindromic DNA repeat consisting of two six-nucleotide half sites separated by a variable three-nucleotide spacer (Beato et al., 1989; Umesono and Evans, 1989; Hard et al., 1990; Lundback et al., 1993; So et al., 2007; Welboren et al., 2009). SRs group into two well-defined phylogenetic clades, each characterized by a distinct DNA-binding specificity. Estrogen receptors specifically bind to the estrogen response element (ERE), a palindrome of AGGTCA (Welboren et al., 2009); androgen, progestagen, glucocorticoid and mineralocorticoid receptors specifically bind to steroid response elements (SREs), palindromes of AGAACA and AGGACA (Chusacultanachai et al., 1999; So et al., 2007). Given the functional diversity of SRs, they represent a great model system to investigate the mechanisms by which a family of biologically important transcription factors evolved to recognize novel DNA sequences.

This thesis details my collaborative work to determine the precise molecular mechanisms for the evolution of novel DNA-binding specificity in the SR family of transcription factors. It is divided into three parts.

Chapter II details the genetic, biochemical and biophysical characterization of the ancestral proteins between which novel specificity evolved. It also identifies and functionally characterizes a set of historical substitutions sufficient to recapitulate the functional transition in DNA specificity. We divide this set of substitutions into subgroups based on their functional role in the evolution of novel specificity; we define them as function-switching substitutions, which are the main determinants of novel specificity, and permissive substitutions, which, by themselves have no effect on specificity but were required for the protein to tolerate the function-switching substitutions. This chapter includes published co-authored work with Jamie T. Bridgham, David W. Anderson, Michael N. Murphy, Eric A. Ortlund and Joseph W. Thornton.

The remaining chapters are directed at understanding the mechanisms of the function-switching and permissive groups of substitutions individually.

Chapter III details the biochemical and biophysical characterization of all possible combinations of the function-switching substitutions. This chapter serves to characterize the independent and epistatic effects of the individual function-switching mutations on protein affinity and specificity. Characterizing all combinations of these substitutions results in a better understanding of the sequence space that separates proteins with the ancestral and derived binding functions and allows us to speculate on the most likely mutational pathways that were taken by the evolving ancestral protein. This chapter includes unpublished co-authored work with David W. Anderson and Joseph W. Thornton.

Chapter IV addresses the role of permissive substitutions in the evolution of novel specificity. It details the biophysical and biochemical characterization of the permissive substitutions and the epistatic interactions between them. Determining the effects of these substitutions helps to elucidate the interdependence of distinct protein residues in determining protein function and results in a better understanding of the biophysical and biochemical mechanisms by which they exerted their permissive effects. This chapter includes unpublished co-authored work with David W. Anderson and Joseph W. Thornton.

Together, this work elucidates the genetic, biochemical and biophysical mechanisms for the evolution of novel DNA specificity in an important family of transcription factors. It results in a better understanding of the evolutionary mechanisms that contributed to a molecular innovation and informs our knowledge of how the biophysical architecture of a molecular system shapes its evolution and evolvability.

CHAPTER II

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

Reproduced from Alesia N. McKeown, Jamie T. Bridgham, Dave W. Anderson, Michael N. Murphy,Eric A. Ortlund, and Joseph W. Thornton; 2014. *Cell* **159**:58-68. Copyright 2014, Cell.

ANM, JTB and JWT conceived the project. All authors designed the experiments and analyzed data. JTB performed the functional characterization of ancestral proteins and their variants and identified key historical substitutions; ANM performed the biochemical and biophysical characterization of ancestral proteins and their variants; DWA performed the molecular dynamics simulations; MNM and EAO performed X-ray crystallography and preliminary biophysical characterizations. ANM and JWT wrote the paper, with contributions from all authors.

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 and Sanchez-Pacheco, 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 1A, (Beato et al., 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 1B). 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 et al., 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.

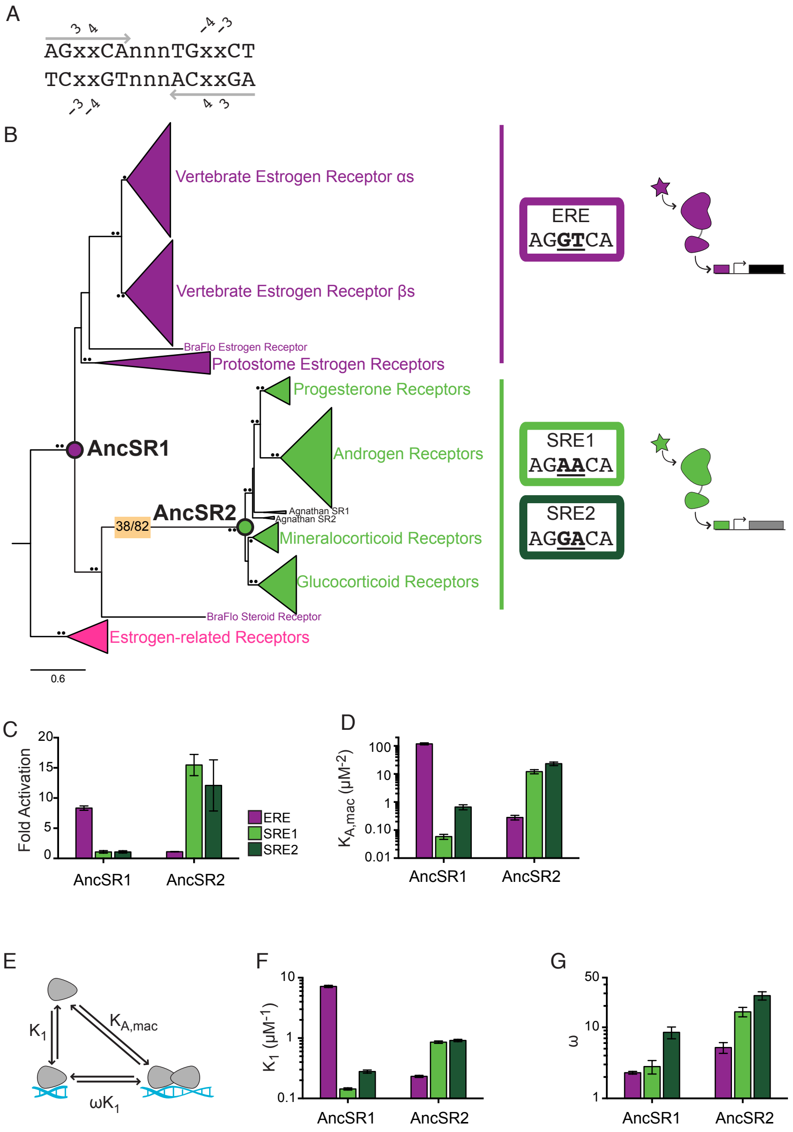
RESULTS

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 1B). 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 (Figure S1). 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 (Figure 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 1C), 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 1B) 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.

Figure 1 (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.



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 (Figure S2A). 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 1B). 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 1D, 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).

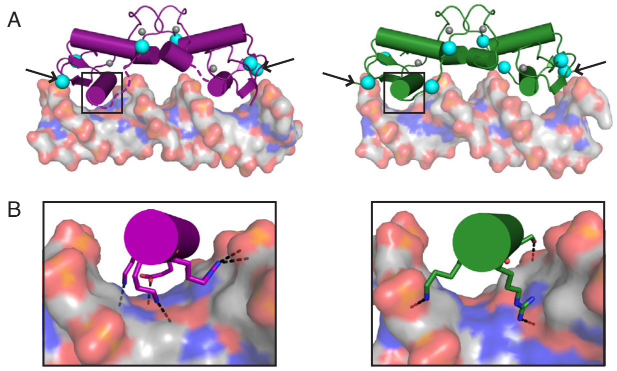
The macroscopic affinity of an SR dimer for a palindromic DNA sequence is determined by two components: the half-site binding affinity (K_1) of each monomer for its half-site and the binding cooperativity (ω) between half-sites, defined as the fold excess of the macroscopic affinity beyond that expected if each monomer binds independently (Figure 1E, (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 1F-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 2, 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 2B). 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 2C). These differences are consistent with AncSR1's greater affinity for DNA half-sites and AncSR2's greater cooperativity of dimeric binding.



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 2. 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.

____

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 3A, 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).

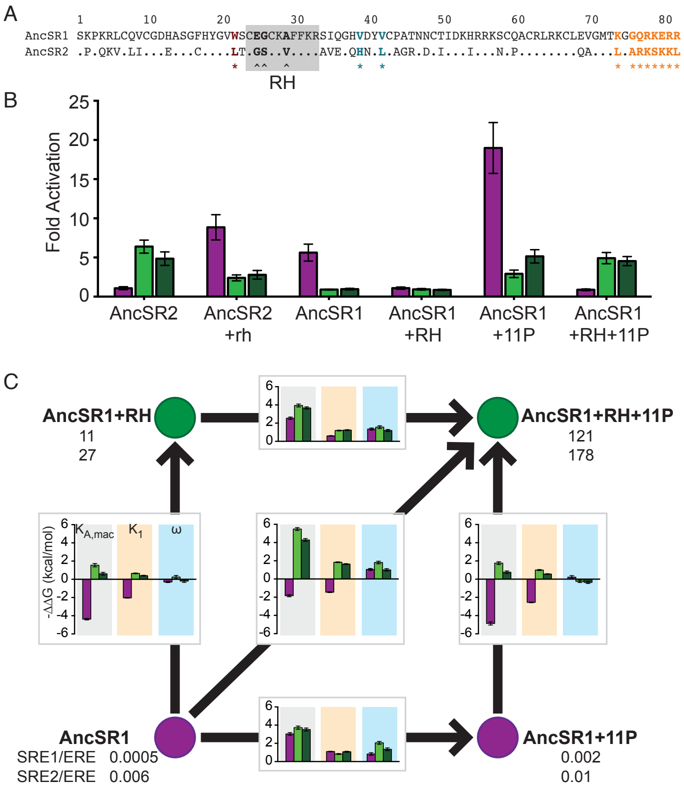


Figure 3. Genetic basis for evolution of new DNA specificity. (A) AncSR1 and AncSR2 sequences. Substitutions between AncSR1 and AncSR2 are shown. Dots, conserved sites. ^, recognition helix (RH) and *, permissive substitutions. Grey box, RH. (B) Effect of RH and 11 permissive (11P) substitutions in luciferase 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 Figs. S2-S4; Tables S3 and S5.

———————————————

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 3B). 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 3B). Surprisingly, activation was entirely abolished on all REs tested (Figure 3B). 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 3 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 3B). 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 ERa 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 (Figure S2E). Adding the 11P into the nonfunctional ERa+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 (Figure 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 3C, Table S3). These effects are entirely attributable to changes in half-site binding affinity, as the RH substitutions do not affect cooperativity (Figure 3C).

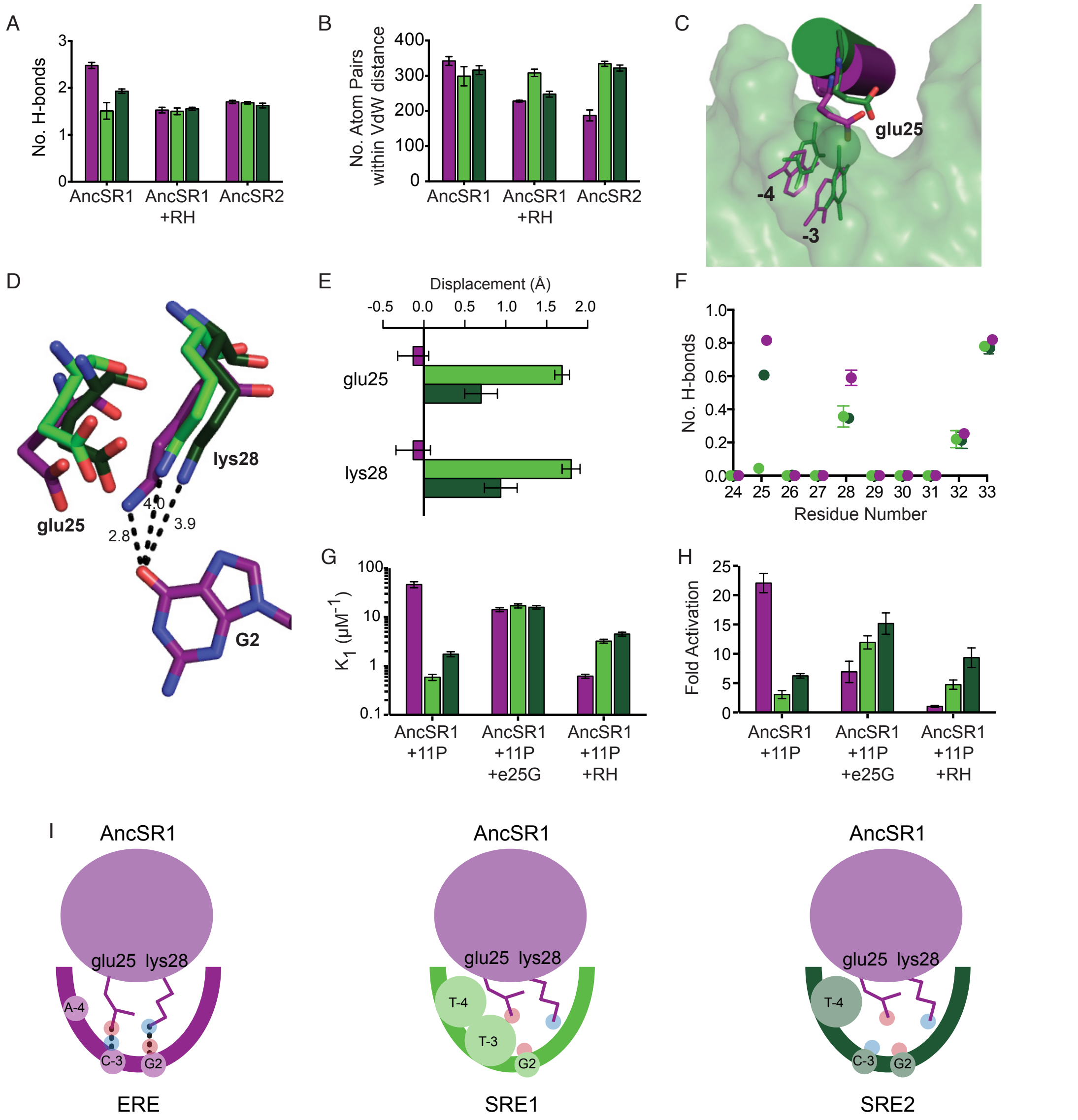
To understand the atom-level mechanisms for the effects of the RH mutations, we compared crystal structures of the ancestral DBDs containing the ancestral or derived RH amino acids in complex with both ERE and SRE1; we also performed molecular dynamics (MD) simulations of AncSR1, AncSR1+RH, and AncSR2, 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 4A, 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 4A). 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 4B). 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 4C, 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 4D-E). As a result, Lys28 forms fewer hydrogen bonds on SREs compared to ERE (Figure 4F). 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.

Figure 4 (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 Figs. 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 Figs. S5-S6.



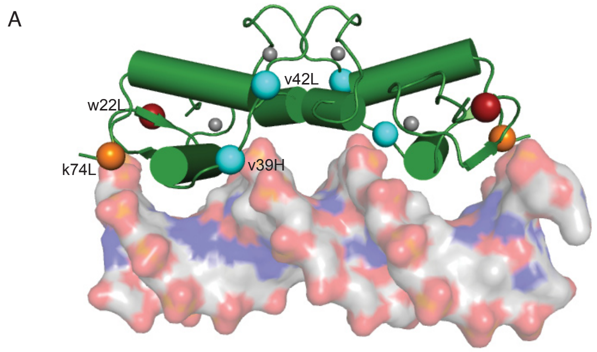
To test the hypothesis that removing glu25 improves SRE recognition by relieving negative interactions, we used site-directed mutagenesis to introduce e25G alone into AncSR1 containing the permissive mutations. We found, as predicted, that SRE affinity and activation were enhanced, despite the fact that Gly25 makes no apparent favorable interactions with SREs (Figure 4G-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 4G-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 4F), 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 4I). 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

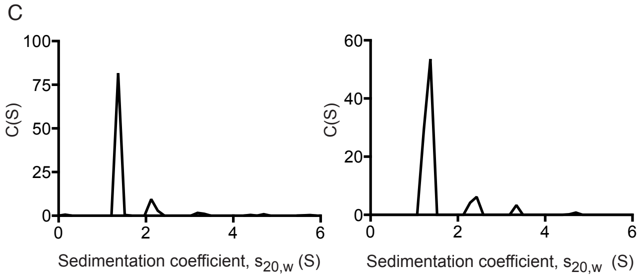
Permissive substitutions are often thought to act by increasing thermodynamic stability, allowing the protein to tolerate mutations that confer new functions but compromise stability (Bershtein et al., 2006; Gong et al., 2013). Using reversible chemical denaturation, however, we found that the 11P substitutions do not increase stability, and the RH substitutions do not decrease stability (Figure 5A-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 3C), 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



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



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 5. 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.

are located on the loop that links the RH to the dimerization surface (Figure 5A). In AncSR1, as in human ERa, 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 (Figure 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 (Berglund et al., 1997; Wikstrom et al., 1999; Meijsing et al., 2009; 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 5A), 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 6A). 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 6B). A similar effect would have ensued if the DBD and/or LBD became promiscuous (Figure 6C-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 6E).

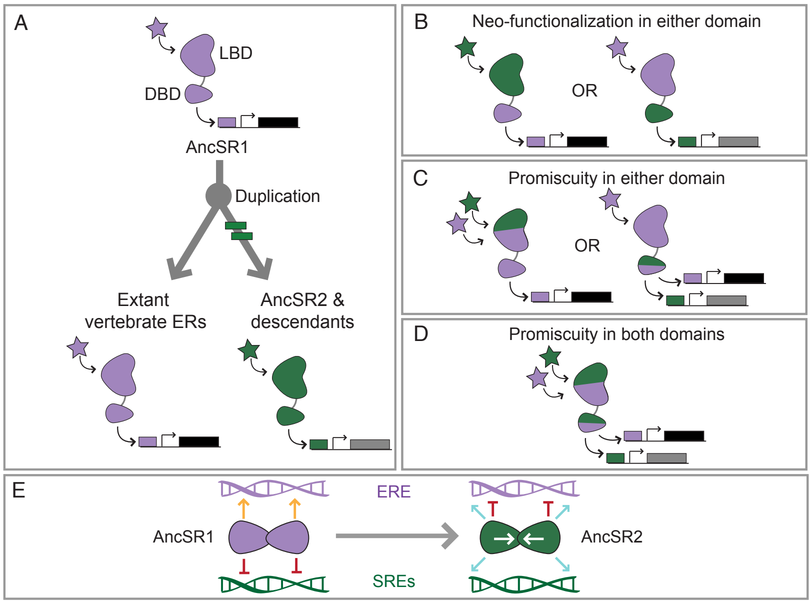

A
LBD
DBD
AncSR1
Duplication
Extant
vertebrate ERs
AncSR2 &
descendants
B
Neo-functionalization in either domain
OR
C
Promiscuity in either domain
OR
D
Promiscuity in both domains
E
ERE
AncSR1
AncSR2
SREs

Figure 6. 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.

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).

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.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found in Appendix A. It includes 6 figures, 6 tables and the Extended Experimental Procedures.

BRIDGE TO CHAPTER III

In Chapter II, we identified a minimal set of substitutions that were sufficient to recapitulate the historical change in DNA specificity. We divided this set of substitutions into two groups: the function-switching substitutions and the permissive substitutions. In Chapter III, we dissect the function-switching mutations and determine the genetic and biochemical mechanisms by which they caused a change in DNA specificity.

CHAPTER III

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

This chapter contains unpublished co-authored material. David W. Anderson and I contributed equally to the design and development of this project. I performed the biochemical binding assays for each protein genotype bound to all 16 REs. DWA performed the molecular dynamics simulations and developed and applied the linear modeling approach for statistical analysis of the data set. DWA and I contributed equally to the writing of this manuscript; the author line of the paper will explicitly indicate this equal contribution.

"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) (Welboren et al., 2009) (Beato et al., 1989; Umesono and Evans, 1989; Lundback et al., 1993). 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

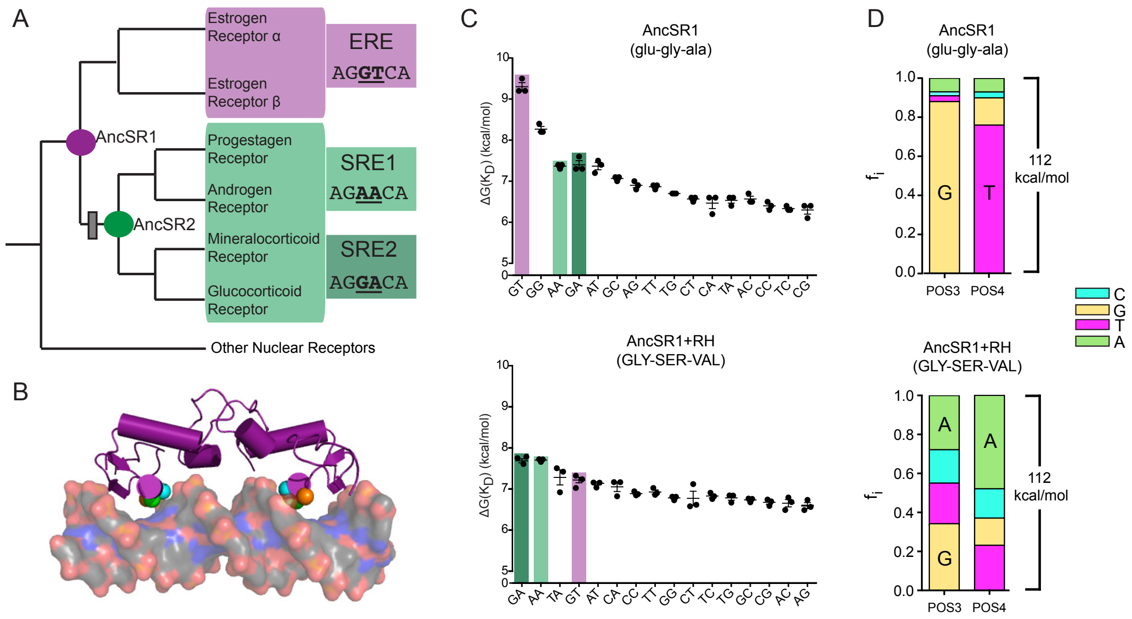


Figure 1. The derived RH causes a switch in DNA-binding preference and specificity. (A) SR receptors group into two well-defined clades based on their DNA-binding specificity. Phylogenetic relationships of extant receptors are shown with the DNA-binding specificity of each receptor indicated by color; 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.

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.

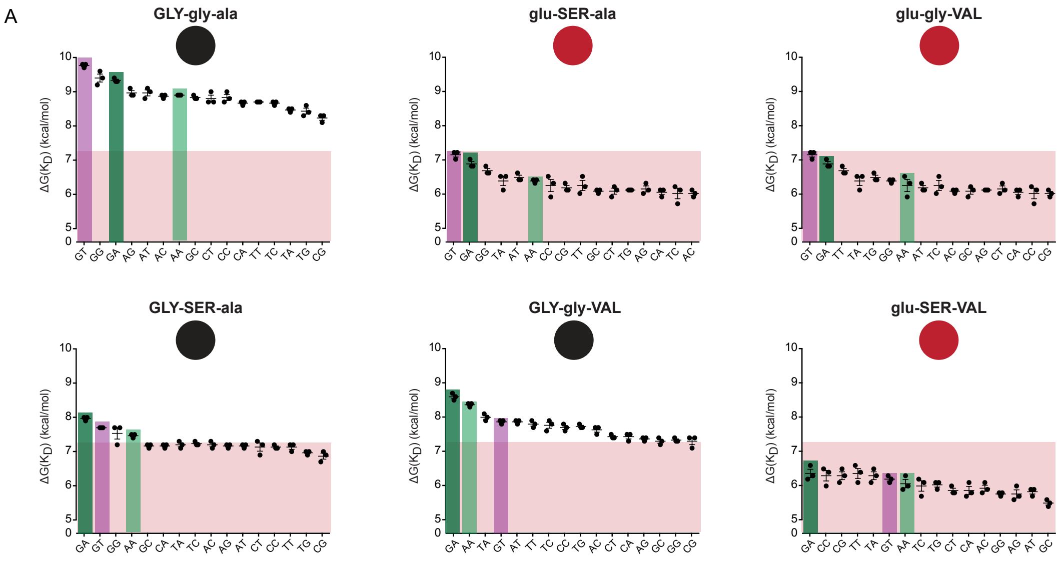
Intermediate protein sequences were either promiscuous or low affinity

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 overfitting (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 S1), and the effects that were significant in both models will be the focus of our discussion.

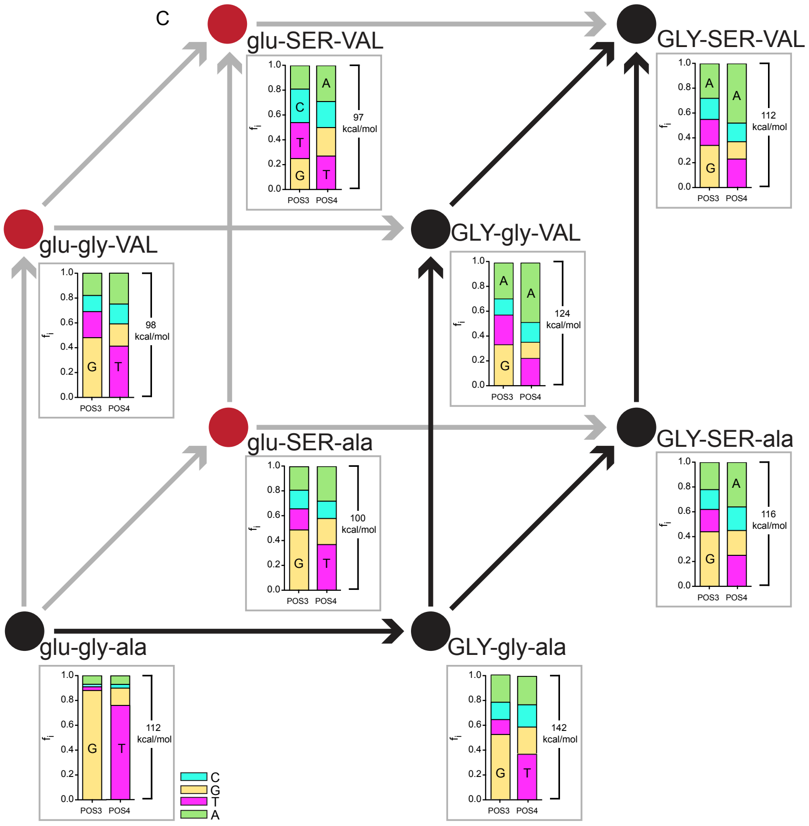
Considering the effects of the substitutions in the RH, we uncovered three first-order terms and two second-order epistatic terms (Figure 2B). The first-order terms represent the general effect of each substitution on binding affinity averaged across all 16 REs and all protein genotype backgrounds. We observed that glu25GLY increased

Figure 2 (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.



B

First-Order Effect	Fold Effect on K_D
glu25GLY	**5.08**
gly26SER	**0.29**
ala29VAL	**0.14**
Second-Order Effect	**Fold Effect on K_D**
glu25, gly26	**0.22**
SER26, ala29	**0.23**



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

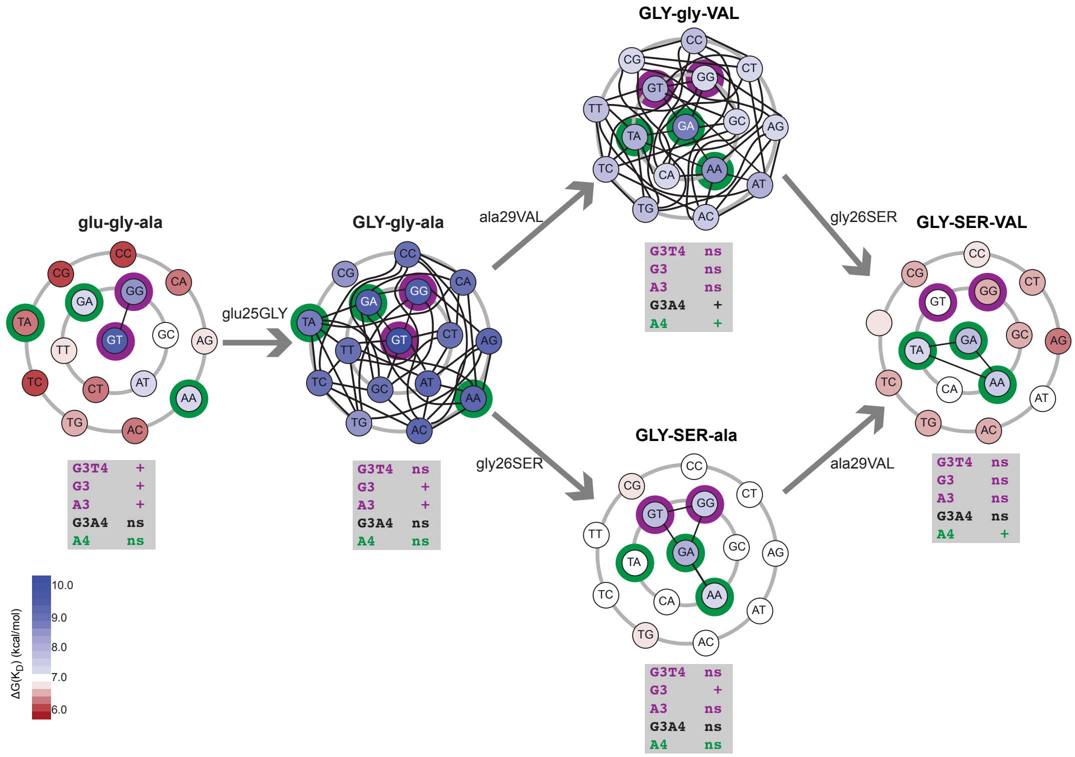


Figure 3. Protein promiscuity increases the size of the high-affinity RE sequence space. Maps of the RE sequence space for each high-affinity protein sequence. RE sequences are colored based on their binding affinity: blue, binding affinity greater than the mean binding affinity; white, mean binding 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.

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).

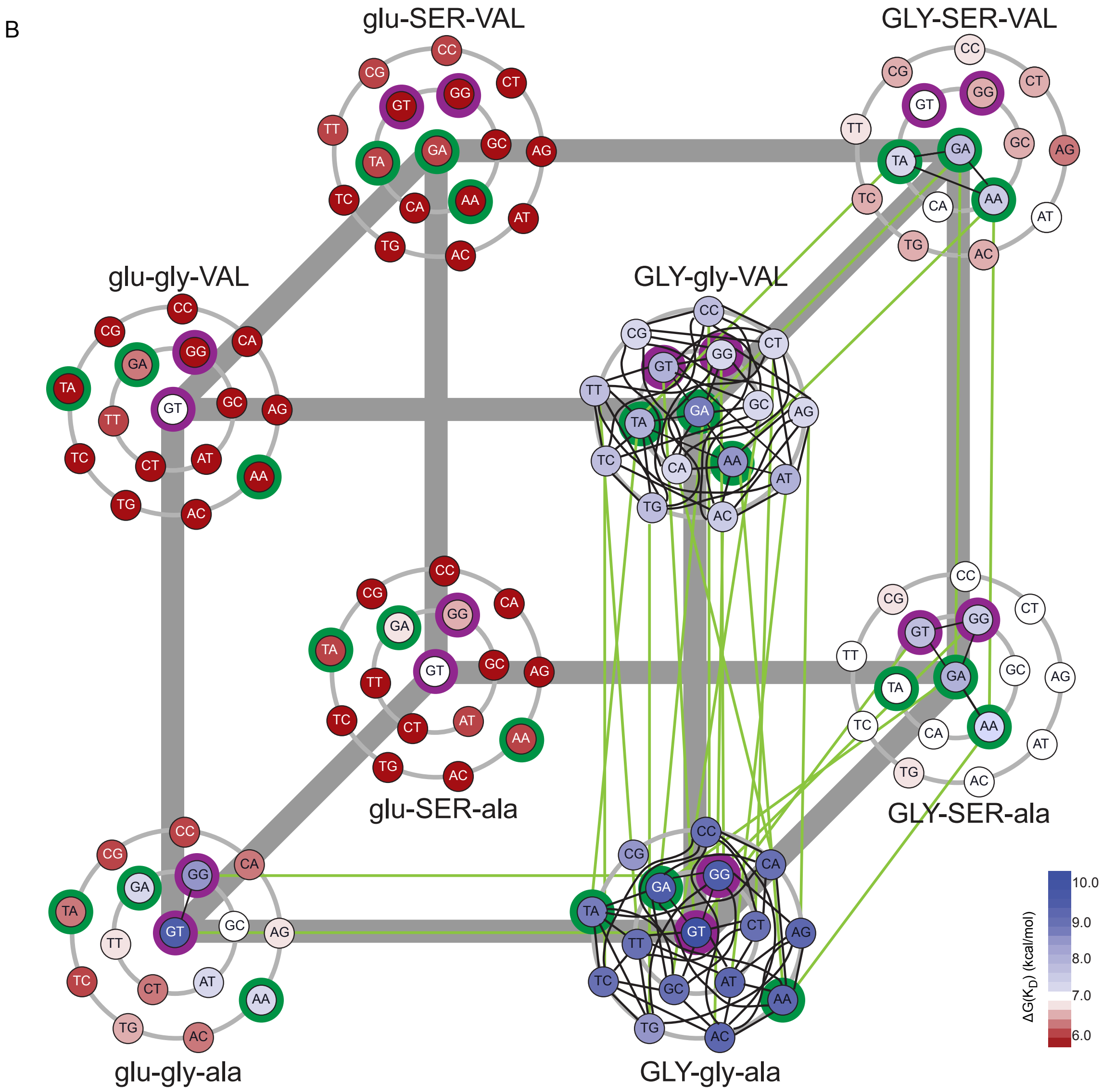
We also identified six epistatic terms between the protein and the RE. These terms indicate the effects of specific individual amino acid states on binding to REs with specific nucleotide states that were preferred by either the ancestral or derived proteins. In particular, we identified 4 epistatic 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

Figure 4 (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 single 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

Genetic Effect	Effect on ΔG(K_D)
A4	+
glu25, A4	-
ala29, A4	-
gly26, G3	+
ala29, G3	+
gly26, C3	-
ala29, T3	-

B



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 4B).

We observed two distinct mutational pathways in the TF by which high-affinity interactions with a novel RE could evolve (Figure 4B). 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 4B). 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

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

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 the 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 mutational pathways could have enabled the evolution of novel binding 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.

EXPERIMENTAL PROCEDURES

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 strands, with differences underlined, were as follows: CCAGGCCA, CCAGGGCA, CCAGCTCA, CCAGCACA, CCAGCCCA, CCAGCGCA, CCAGTTCA, CCAGTACA, CCAGTCCA, CCAGTGCA, CCAGACCA, CCAGAGCA, CCAGGTCA, CCAGAACA, CCAGGACA, CCAGATCA. Complementary reverse strands were also ordered.

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, we measured binding affinity to the half-site REs in triplicate and fit the data to a single-site binding model.

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 C_0 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 (Carroll, 2008). 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

The crystal structure of AncSR1 bound to ERE (PDB: 4OLN) was 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 (Hoops et al., 1991; Lin and Wang, 2010) as previously described (McKeown et al., 2014). Zinc finger partial charges were derived using the RED III.4 pipeline (Dupradeau et al., 2010) as previously described (McKeown et al., 2014). 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.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found in Appendix B. It includes 2 figures and 2 tables.

BRIDGE TO CHAPTER IV

In Chapter III, we dissected the individual and epistatic effects of the function-switching mutations on DNA specificity. In Chapter IV, we perform a similar study in which we dissect the permissive substitutions into sub-groups and assay their individual and epistatic effects on DNA binding affinity and inter-protein cooperativity. We then apply molecular dynamics simulations to determine the molecular mechanisms by which the permissive sub-groups allowed for the protein to tolerate the deleterious effects of the function-switching substitutions.

CHAPTER IV

MOLECULAR MECHANISMS FOR THE FUNCTIONAL ROLE OF PERMISSIVE SUBSTITUTIONS IN THE EVOLUTION OF NOVEL DNA SPECIFICITY IN STEROID RECEPTORS

This chapter contains unpublished co-authored material. I performed the biochemical binding assays for each protein genotype. David W. Anderson performed the molecular dynamics simulations. I analyzed and interpreted all of the data and wrote the manuscript.

INTRODUCTION

What are the roles of historical substitutions in the evolution of protein function?

A long-standing goal in molecular evolution is to identify historical substitutions that contributed to the evolution of novelty. By identifying the number and effects of substitutions necessary for a derived function across diverse systems, we can begin to answer key questions in molecular evolution. What is the distribution of effect sizes of protein substitutions that cause a novel function to evolve (Orr, 2005; Soskine and Tawfik, 2010)? Are the effect sizes of these substitutions independent of one another or do these substitutions interact epistatically to modulate each other's effect on protein function (Phillips, 2008; Breen et al., 2012; McCandlish et al., 2013)? What are the biochemical and biophysical mechanisms that mediate the effects of these functionally important substitutions (Worth et al., 2009; Harms and Thornton, 2013)? How does a protein's biophysical architecture affect the its evolution?

Many studies have aimed to answer these questions across an array of diverse systems. In numerous cases, the evolution of a novel protein function requires two sets of substitutions: function-switching substitutions and permissive substitutions (Ortlund et al., 2007; Bloom et al., 2010; Thomas et al., 2010; Lynch et al., 2011; Gong et al., 2013; Harms and Thornton, 2014; McKeown et al., 2014). Function-switching substitutions are large-effect substitutions and are the main determinants of a derived function. However, these function-switching mutations are often deleterious to protein function and are therefore not tolerated in the starting genetic background (Smith, 1970). Permissive substitutions cause biochemical effects that “buffer” the protein to allow for introduction of the function-switching substitutions. By themselves, the permissive substitutions do not result in a novel function, but are required for the protein to tolerate the deleterious effects of the function-switching substitutions.

Biophysical mechanisms of permissive substitutions vary across molecular systems

The molecular mechanisms by which permissive substitutions exert their effects vary. In some cases, permissive substitutions operate on global protein stability (Bershtein et al., 2006; Tokuriki et al., 2008; Thomas et al., 2010; Gong et al., 2013). In these examples, permissive substitutions non-specifically improve protein stability so that introduction of structurally destabilizing function-switching mutations can be introduced without falling below a critical functional threshold. Others have found that permissive substitutions function on a more local scale, operating under precise constraints so as to increase stability of a specific region of the protein and in a way that is compatible with the ancestral background (Davis et al., 2009; Harms and Thornton, 2014). Alternatively, some permissive substitutions are known to have no effect on stability, but instead are required to directly modulate the effects of the function-switching mutations (Aharoni et al., 2005; Yokoyama et al., 2008; Martin et al., 2009; Field and Matz, 2010). In one such case (Field and Matz, 2010), the permissive substitutions acted to prime the ligand-binding site by introducing residues important for establishing a chemical environment for a novel autocatalytic mechanism. Although, by themselves, these permissive substitutions had no effect on protein function, the chemical environment that they established allowed the function-switching mutations to change the autocatalytic properties of the protein and allow for a novel function. These findings suggest that the mechanisms by which the permissive substitutions act is dependent on the biophysical architecture and functional properties of the system under investigation.

To date, the biophysical mechanisms for the functional effects of permissive substitutions have been primarily investigated in enzymes (Aharoni et al., 2005; Bershtein et al., 2006; Tokuriki et al., 2008; Thomas et al., 2010) and proteins that recognize small ligands (Martin et al., 2009; Harms and Thornton, 2014). One study

(Gong et al., 2013) has investigated the mechanisms of permissive substitutions for a protein that is part of a larger macromolecular complex, but did so in the absence of the protein's interacting partners. Therefore, the role of permissive substitutions in the evolution of multi-component complexes that interact across an extended binding interface is still unknown. What are the mechanisms by which permissive substitutions affect the interactions within the protein, and between its interacting partner, to allow the protein to tolerate the function-switching substitutions? Do permissive substitutions function by improving the stability of free protein or do they operate on a more global scale, functioning instead to improve the stability of the complex? Are the effects of the permissive substitutions localized to the binding interface or are they spread throughout the protein's structure? Is the architecture of the protein such that the permissive substitutions interact epistatically or are their effects largely independent of one another? Answering these questions will allow us to gain a better understanding of the determinants of protein function as well as elucidate the evolutionary processes from which they arose.

Novel DNA specificity in the steroid receptor family of transcription factors evolved by the coordinated effects of function-switching and permissive substitutions

Steroid receptors (SRs) are a good model system to study the mechanisms of permissive substitutions in the evolution of novel function in a multi-component interacting system. SRs are a class of ligand-activated transcription factors that regulate the vertebrate response to adrenal and sex hormones (Bentley, 1998). All SRs contain a highly conserved DNA-binding domain that binds as homodimers to a specific DNA response element that consists of a six nucleotide inverted palindromic repeat separated by a variable three-nucleotide spacer (Figure 1A)(Bain et al., 2007). The architecture of these REs is such that SRs bind in a head-to-head manner, with direct protein-DNA recognition occurring by insertion of a 10-residue recognition helix (RH) into the DNA major groove(Luisi et al., 1991; Schwabe et al., 1993). Cooperative protein-protein interactions occur at the dimerization interface of the two protein monomers.

There are two phylogenetic classes of SRs, each characterized by a distinct DNA-binding specificity (Figure 1B); estrogen receptors specifically bind to ERE, a palindrome of AGGTCA (Beato et al., 1989; Welboren et al., 2009), while androgen, progestagen, mineralocorticoid and glucocorticoid receptors specifically bind to SRES, palindromes of AGAACA or AGGACA (Beato et al., 1989; Zilliacus et al., 1992; Chusacultanachai et al., 1999; So et al., 2007). We have previously reported on the mechanisms by which SRs evolved to recognize divergent DNA response element (RE) sequences (McKeown et al., 2014). By ancestrally reconstructing the ancestor of all SRs (AncSR1) and the ancestor of all androgen, progestagen, mineralocorticoid and glucocorticoid receptors, we determined that the preference for SREs evolved through neofunctionalization of an ancestor that preferentially bound ERE (Figure 1B). This change in specificity was caused by three function-switching substitutions of large effect that occurred in the protein's recognition helix (RH) (Figure 1A, 1C). Although these RH substitutions were the main determinants of the derived DNA-binding specificity, they were deleterious to overall protein function, resulting in a protein that was unable to activate transcription of any RE in a cell-based functional assay. Introduction of a group of 11 permissive substitutions was sufficient to allow for the protein to tolerate the deleterious effects of the function-switching RH substitutions. These substitutions did not alter protein specificity. Together, the function-switching RH substitutions and the permissive substitutions completed a change in specificity and recapitulated the derived protein function.

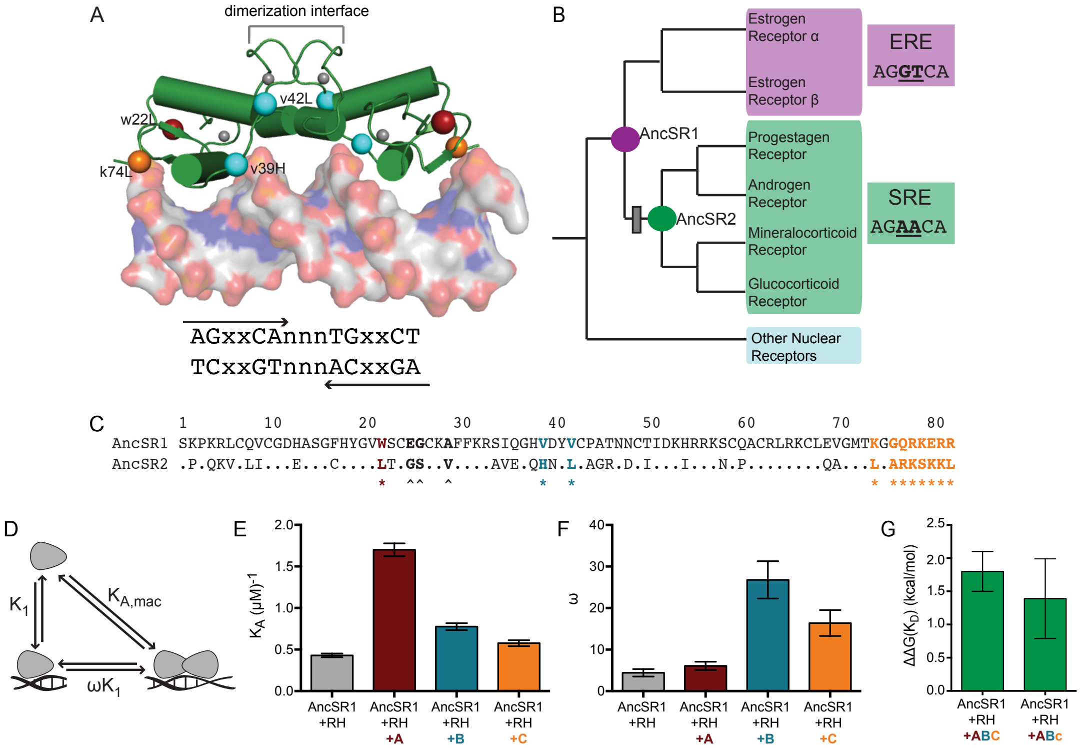
The permissive substitutions allowed for the protein to tolerate the RH substitutions by non-specifically improving DNA affinity

Biophysical characterization of the effects of each group of substitutions suggests that the permissive substitutions functioned to improve the stability of the complex by increasing macroscopic binding affinity (McKeown et al., 2014). We found that the derived RH caused a significant decrease in DNA binding affinity by removing positive interactions at the protein-DNA interface while also establishing new negative ones. These changes resulted in a low-affinity interaction and thus destabilized the protein-DNA complex. Conversely, the permissive substitutions stabilized the complex by non-specifically increasing macroscopic binding affinity. Together, these results suggest that the permissive substitutions functioned to improve the stability of the protein-DNA complex, greatly increasing DNA-binding affinity so as the deleterious RH substitutions could be introduced without ever causing the complex to exist below the thermodynamic threshold for high-affinity interactions.

Given that SRs bind cooperatively as homodimers to their specific RE, their macroscopic binding affinity contains contributions from both single-site affinity and inter-protein cooperativity (Figure 1D) (Hard et al., 1990). We previously determined that the increase in macroscopic binding affinity caused by the permissive substitutions was due to increases in both single-site affinity and cooperativity (McKeown et al., 2014). The goal of this work is to uncover the precise molecular mechanisms by which individual sub-groups of the permissive substitutions contributed to these change in the

———————————————

Figure 1 (next page). The three groups of permissive substitutions occur throughout the protein and have unique effects on single-site affinity and cooperativity. (A) Steroid receptors group into two well-defined clades based on their DNA-binding specificity. Estrogen receptors bind the estrogen response element (ERE); progestagen, androgen, mineralocorticoid and glucocorticoid receptors bind the steroid response element (SRE). Response element half-sites are shown to the right of the phylogeny with differences indicated in bold and underlined. Receptors are colored based on their DNA-binding specificities; purple, ERE and green, SRE. SRE-specificity evolved on the branch between AncSR1 and AncSR2, indicated with a gray box. (B) Sequence alignment of AncSR1 and AncSR2 shows linear position of permissive and function-switching mutations. The three recognition helix (RH) substitutions capable of switching specificity are shown with ^. Permissive substitutions are shown with * and colored by group membership. Red, group A; cyan, group B; orange, group C. (C) Permissive substitutions occur at the protein-DNA and protein-protein interfaces. X-ray crystal structure of AncSR2 bound to SRE. Permissive substitutions are indicated with Cα as spheres and colored by group membership. Colors are same as in A. Protein is shown in cartoon; DNA is shown as surface and colored by atom type. (D) The macroscopic binding affinity ($K_{A,mac}$) for binding to a palindromic RE sequence has contributions from both single-site affinity (K_1) and cooperativity (ω). Permissive substitutions could affect either of these parameters. (E-F) The permissive substitutions have unique effects on single-site affinity (D) and cooperativity (E). Bars are colored by group; colors are same as in A. Values are mean ± SEM for three replicate experiments. (G)SRs bind with greater affinity to DNA sequences containing extended flanking sequences on the SRE half-site regardless of the identity of group C. Unlabeled short (2 flanking nucleotides) and long (6 flanking nucleotides) single-site SRE oligos competed against labeled short single-site SRE oligo for binding to purified DBD. The difference in the competitive binding affinity for the long versus short oligo is reported. Lower-case and upper-case letters indicate the ancestral and derived group states, respectively. Values are mean ± SEM of three replicate experiments.



energetic components of binding. To this end, we divided the group of 11 permissive substitutions into three subgroups, A-C, based on their location in the protein sequence and structure (Figure 1A, 1C). We assayed the function of all possible combinations of permissive groups and determined their individual and epistatic effects on single-site affinity and cooperativity. We then used molecular dynamics simulations to probe the molecular mechanisms for these observed effects. This approach elucidated the functional effects and biochemical mechanisms by which the permissive substitutions permitted the evolution of novel DNA specificity and allowed us to speculate on the evolutionary processes by which they arose.

RESULTS

Individual groups of permissive substitutions have unique effects on single-site binding affinity and cooperativity

To determine how each sub-group of permissive substitutions allowed the protein to tolerate the derived RH, we introduced the individual groups into the AncSR1+RH background and assayed their effects on single-site affinity and cooperativity when binding SRE (Figure 1E-F). We found that each sub-group affected macroscopic binding affinity by causing differential effects on single-site affinity and cooperativity.

Group A consists of a single substitution, w22L, which does not directly contact DNA. However, its flanking residues make both polar and non-polar contacts to DNA (Figure 1A). We hypothesized that group A would influence single-site affinity but would have minimal to no effect on protein cooperativity. Concordantly, introducing group A caused a dramatic increase in single-site affinity and had no effect on cooperativity (Figure 1E-F). These results suggest that group A, affects single-site affinity by changing interactions with nearby residues and DNA.

Group B consists of two substitutions, v39H and v42L. Residue 39 occurs in the flexible region leading to the dimerization interface and residue 42 occurs in the dimerization interface (Figure 1A, 1C). The derived His39 is in close proximity to the DNA backbone and forms a hydrogen bond with the backbone phosphate in crystal structures of the human GR (Luisi et al., 1991). We hypothesized that these substitutions would cause an increase in both single-site affinity and cooperativity. When we introduced group B into AncSR1+RH, we observed a significant increase in both single-site affinity and cooperativity, with a greater effect on cooperativity (Figure 1E-F). These results indicate that the permissive effect of group B is primarily to increase inter-protein cooperativity, but also plays a role in increasing single-site affinity.

Group C is a group of 8 substitutions that occur in the positively charged and unstructured C-terminal tail (Figure 1A, 1C). This region of SRs has not been completely resolved in any x-ray crystal structures of extant or ancestral proteins. However, partial resolution of the C-terminal tail in structures of extant steroid receptors (Roemer et al., 2006; Meijsing et al., 2009; Helsen et al., 2012) show that residues within this region can hydrogen bond to the DNA backbone and minor groove. Consistent with these structures, DNA-binding proteins have been proposed to use a positively charged C-terminal extension for non-specific interactions with the DNA backbone, thereby improving non-specific binding and facilitating in DNA scanning and site recognition (von Hippel, 2007). We therefore hypothesized that group C contributed to the permissive effect by increasing single-site affinity through polar contacts with the DNA backbone.

Contrary to our prediction, introduction of group C causes only a slight increase in single-site affinity, but a dramatic increase in cooperativity (Figure 1E-F). It is possible that the minimal effect that group C had on single-site affinity was due to a limitation of our assay. If the C-terminal tail functions to interact with nucleotides flanking the consensus half-site, this contribution would be immeasurable in our current FP assays because the DNA fragments lack extensions outside of the consensus site. To determine if group C interacts non-specifically with nucleotides flanking the RE, we performed competition experiments in which unlabeled "long" half-site oligos with six flanking nucleotides compete with pre-bound labeled "short" half-site oligos that contain only two flanking nucleotides for binding to proteins.

To determine if the protein interacts with flanking nucleotides, we first measured the competition binding affinity of AncSR1+RH+ABC for single-site short and long oligos. We found that AncSR1+RH+ABC bound with greater affinity to longer DNA oligos (Figure 1G). To determine if the derived group C was responsible for this improvement in affinity, we reverted the group back to its ancestral state, resulting in a protein genotype of AncSR1+RH+ABc, and assayed its competition affinity for long and short oligos. We found that AncSR1+RH+ABc also bound longer DNA oligos with higher affinity (Figure 1G). Further, the improvement in affinity for the long DNA oligos was not significantly different in the presence of the ancestral or the derived group C substitutions. These results imply two things. First, they suggest that the protein may interact with flanking nucleotide sequences to increase single-site affinity. Second, they imply that the permissive group C does not cause an increase in single-site affinity by improving interactions with flanking sequences relative to the ancestral states.

The strong effect that group C had on cooperativity was remarkable as group C occurs very distant from the dimerization interface. Given that group C does not directly contact the dimerization interface, these results imply that the derived residues may function through an allosteric mechanism, causing a change in the conformation of either the DNA or protein upon binding that facilitates inter-protein cooperativity. A potential allosteric mechanism in the DNA is consistent with circular permutation studies showing that human GR, which contains all of the group C substitutions, causes a characteristic bend in the DNA upon binding (Petz et al., 1997), however, additional studies are required to determine if this bending is due to interaction between group C and DNA.

It is also possible that group C allosterically modulates the dimerization interface through inter-protein interactions. Crystal structures of a closely-related protein, the human estrogen-related receptor, indicates that the polar C-terminal extension has the potential to make inter-protein hydrogen bonds on the surface of the protein (Gearhart et al., 2003). Although this mechanism has never been observed in other SR proteins, it is possible that the C-terminal tail could modulate the protein's structure through interactions with adjacent residues and indirectly improve interactions at the dimerization interface. Additional experiments are required to test these hypotheses.

The individual groups of permissive substitutions interact epistatically to alter inter-protein cooperativity

We next wanted to determine if the permissive substitutions functioned independently or interacted epistatically to cause increases in single-site affinity and cooperativity. Epistasis is the phenomenon by which the effect of a given substitution, or group of substitutions, is modulated by the presence or absence of a different substitution at a separate site in the protein. Epistasitically interacting substitutions are common in protein evolution (Bridgham et al., 2009; Lunzer et al., 2010; Breen et al., 2012; McCandlish et al., 2013; Parera and Martinez, 2014) and result in a rugged genotype-phenotype landscape that can impede the capacity of directional selection to drive the evolution of a novel function (Weinreich et al., 2005; Phillips, 2008; Kvitek and Sherlock, 2011). Therefore, determining the role of epistatic interactions in modulating the permissive effect of these individual groups of permissive substitutions is an important goal in understanding the role of these substitutions in permitting or constraining the evolution of the SR DBD.

To determine if the effects of the permissive groups functioned independently or epistatically, we measured the thermodynamic effects of each group upon introduction to all possible genetic backgrounds. We identified significant epistasis in the effect that the three groups of substitutions had on cooperativity. The effect that group A had on cooperativity was dependent on the presence of groups B and C. When group A is introduced alone or in the presence of only group B or group C, it had no significant effect on cooperativity (Figure 2). However, if introduced in the presence of both groups B and C, it caused a large increase in protein cooperativity. This indicates that the combined groups B and C interact epistatically with group A to modulate its effect on protein cooperativity.

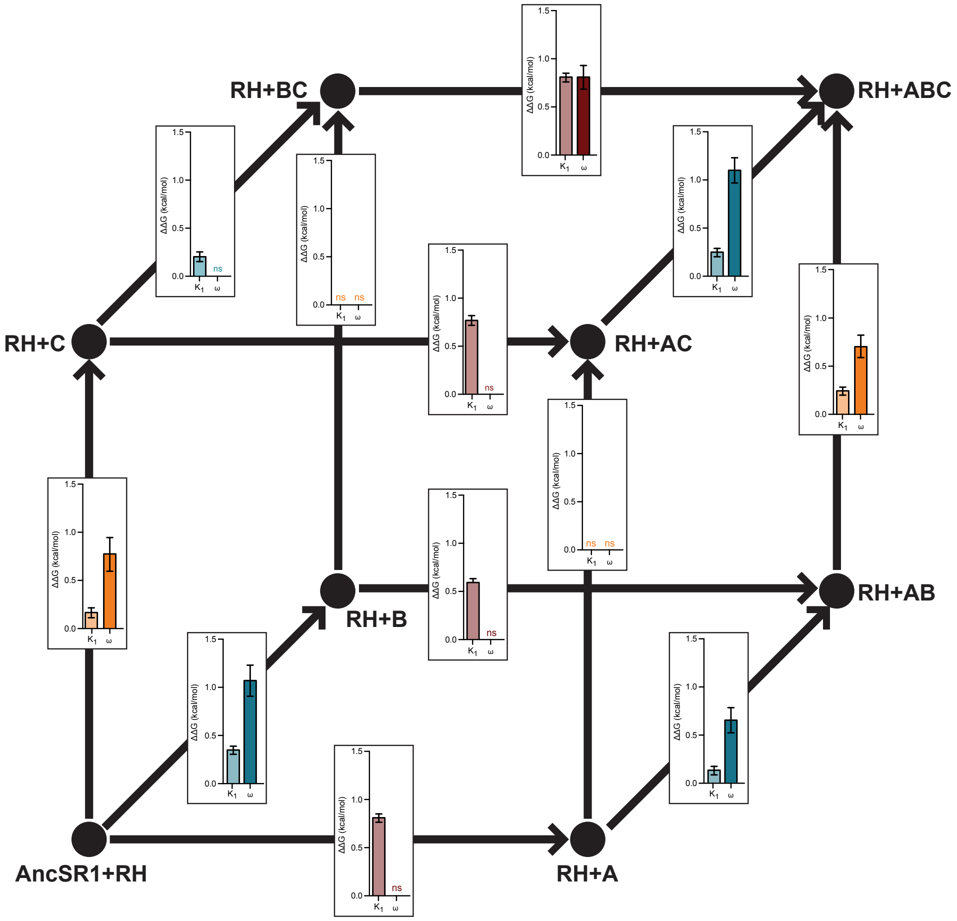


Figure 2. The permissive substitutions act independently to alter single-site affinity, but interact epistatically to alter cooperativity.
Genetic cube identifying the effect of each group of permissive substitutions on single-site affinity (left, lighter bar) and cooperativity (right, darker bar) in all combinations of permissive backgrounds. Vertices indicate specific protein sequence; lower-case and upper-case letters indicate ancestral and derived states, respectively. Edges connecting vertices indicate introduction of individual group of substitutions. Boxes on each edge are the effect of each group of substitutions on the Gibbs free energy of single-site affinity and cooperativity. Values are mean ± SEM of three replicate experiments. ns-no significant difference, as determined by Fisher's LSD, $p<0.005$.

The effect that group B had on cooperativity was also dependent on the presence of groups A and C (Figure 2). In most cases, group B caused a strong positive effect on protein cooperativity. However, when introduced into AncSR1+RH+C, it had no effect. As such, group C has a negative epistatic interaction with group B.

We also observed that the effect of group C changed depending on the presence of groups A and B (Figure 2). Group C had strongly positive effects when introduced alone or when introduced in a background that contains both groups A and B. However, when introduced in the presence of either group A or group B, group C had no significant effect on cooperativity. This implies that groups A and B interact epistatically with group C.

Together, these results indicate that all three groups of substitutions interact epistatically to elicit their permissive effect. This is remarkable given their placement throughout the protein structure. It implies that even residues that occur a great distance away from each other can interact to modulate the other's effect on protein function. These observations lead to an outstanding question in protein biochemistry: what are the mechanisms by which residues that do not interact physically give rise to genetic epistasis? We hypothesize that these effects are potentially due to small-scale conformational or dynamic rearrangements that may not be detected in crystal structures. This possibility is illustrated by the epistatic interactions that the individual groups A and B have on the cooperative effects of C. Structurally, groups A and B occur on either side of the protein's recognition helix (Figure 1A). We hypothesize that these two groups may interact by differentially affecting the conformation of the recognition helix, altering its position in the DNA major groove. These small changes in the position of the recognition helix may alter the symmetry of the complex, indirectly influencing the complementary surfaces of the dimerization interface and leading to changes in cooperativity afforded by group C. However, many additional structural and dynamic experiments are required to test this hypothesis and to resolve the role of conformational rearrangements to explain epistatic interactions.

The individual groups of permissive substitutions function independently to improve single-site affinity

We next wanted to determine if the individual groups of permissive substitutions interacted epistatically to affect single-site affinity. Although the magnitude of the effect varies slightly across genetic backgrounds, the individual effects of the permissive groups on single-site affinity remain largely independent of genetic background (Figure 2). The biggest variation in effect occurs with group C, but even these differences are very small, alternating between causing a very small increase of about 0.2 kcal/mol to having an insignificant effect. These results indicate that the effect of each permissive group on single-site affinity is independent of genetic background.

Together with the observed epistatic effects on cooperativity, these results indicate that the epistatic nature of individual groups of substitutions need not apply to all of the physical properties of the protein. In the case of even a small protein like the SR DBD, small groups of substitutions can differentially affect the thermodynamic components of complex formation and differentially interact with distant groups to achieve these effects. Given that two of these three groups contain more than a single substitution, it is possible that the extent of epistasis between these historically important substitutions could be even greater, as residues within groups could also interact epistatically. In order to determine if this is the case, additional experiments that functionally characterize all possible combinations of the 11 permissive substitutions are required, but are beyond the scope of this work.

The permissive increase in single-site affinity is not due to an increase in hydrogen bonding or packing at the protein-DNA interface

We next sought to determine the molecular mechanisms by which the permissive substitutions improved single-site affinity by using molecular dynamics (MD) simulations. We first investigated whether the permissive substitutions acted by a non-allosteric mechanism, improving single-site affinity through changes in hydrogen bonding and packing across the protein-DNA interface. To measure these biophysical interactions, we modeled in the recognition helix substitutions as well as groups A and B onto a previously solved crystal structure of AncSR1 (McKeown et al., 2014). We were unable to model in group C as this region is not resolved in any crystal structures and would potentially introduce a high degree of error into the simulations.

We found that the increases in single-site affinity afforded by the permissive groups A and B were not due to changes in hydrogen bonding or packing interactions at the protein-DNA interface. We first measured the number of hydrogen bonds formed across the protein-DNA interface for both AncSR1+RH and AncSR1+RH+AB and did not observe a change in the number of hydrogen bonds formed upon introduction of these permissive groups (Figure 3A). We next measured the degree of packing across the protein-DNA interface for AncSR1+RH and AncSR1+RH+AB. We did not observe an increase in packing efficiency upon introduction of groups A and B (Figure 3B). These results imply that the improvement in single-site affinity was not due to the evolution of novel positive interactions at the protein-DNA interface.

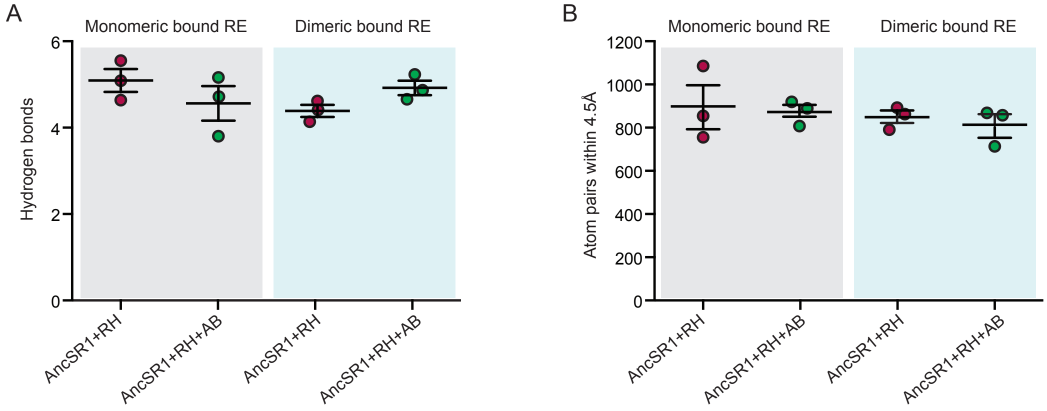


Figure 3. The permissive substitutions do not cause an increase in hydrogen bonding or packing at the protein-DNA interface. (A) Groups A and B do not improve hydrogen bonding between protein and DNA. The number of hydrogen bonds formed across the entire protein-DNA interface was calculated for AncSR1+RH (pink) and AncSR1+RH+AB (green). Values, calculated for monomeric (gray box) and dimeric (blue box) protein bound DNA complexes, are for three replicate MD simulations; lines indicate mean and SEM. (B) Groups A and B do not improve packing efficiency at the protein-DNA interface. Packing efficiency across the entire protein-DNA interface was calculated for AncSR1+RH (pink) and AncSR1+RH+AB (green). Values, calculated for monomeric (gray box) and dimeric (blue box) protein bound DNA complexes, are for three replicate MD simulations; lines indicate mean and SEM.

Permissive substitutions increase cooperativity in part by improving packing interactions but not by increasing hydrogen bonding at the dimerization interface

We next wanted to determine if the permissive substitutions caused an increase in cooperativity by evolving novel positive interactions at the protein dimerization interface. We measured the number of hydrogen bonds formed between protein monomers at the dimerization interface (residues 39-57) and found that the permissive groups A and B did not cause a significant increase in the number of hydrogen bonds formed at the interface (Figure 4A). We next measured the packing efficiency at the dimerization interface and observed that introduction of groups A and B caused a significant increase in packing efficiency (Figure 4B). To identify which residues contributed to this increase in packing, we calculated the packing of each individual residue in the dimerization interface and compared these values between AncSR1+RH and AncSR1+RH+AB (Figure 4C). This led to the identification of a single residue, residue 42, that differed in the its packing between protein constructs. This result was of interest because residue 42 is the site of one of the permissive substitutions in group B, v42L. These results imply that substitution of the smaller valine with the larger leucine at position 42 contributed to cooperativity by improving packing at the dimerization interface, thereby driving dimeric complex formation via the hydrophobic effect.

Measurements from MD simulations do not provide evidence of an entropic mechanism for the increase in single-site affinity caused by permissive groups A and B

In the absence of any novel biophysical interactions at the protein-DNA interface, we hypothesized that groups A and B may have affected single-site affinity by altering the conformational entropy of the system so as to decrease the entropic cost of complex formation. Similar mechanisms, by which permissive substitutions alter the conformational flexibility of ground state structures, have been observed in molecular evolution studies of lab-derived and naturally occurring enyzmes (Jackson et al., 2009; Tokuriki and Tawfik, 2009; Thomas et al., 2010). To investigate this possibility, we wanted to determine if introduction of the permissive substitutions altered the conformation of the free protein to more closely resemble the bound form. By

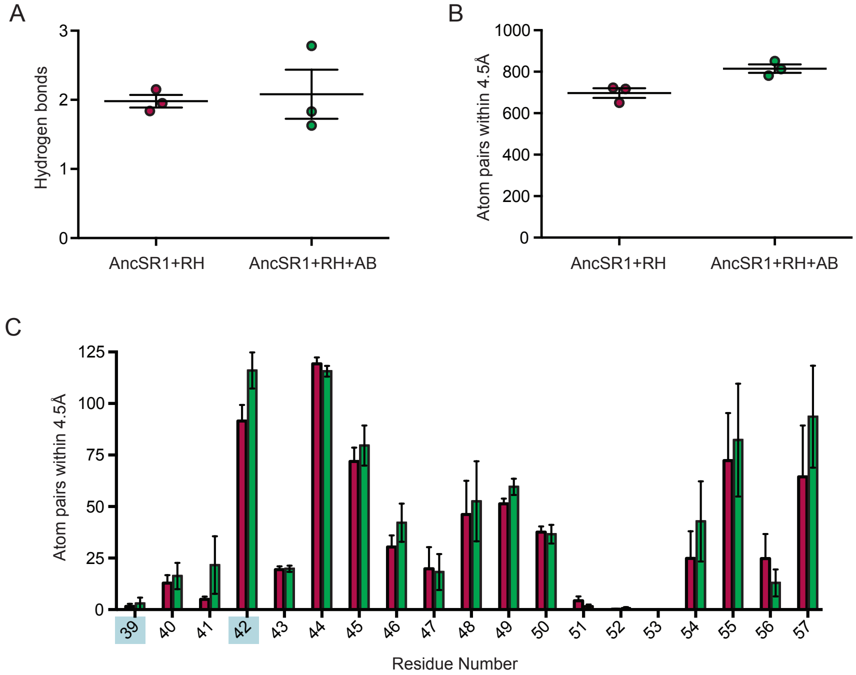


Figure 4. The effect of the permissive substitutions on cooperativity is not due to an increase in hydrogen bonding, but can be partially explained by an increase in packing at the dimerization interface. (A) Groups A and B do not increase the number of hydrogen bonds at the protein dimerization interface. The number of hydrogen bonds formed at the protein dimerization interface (residues 39-57) are indicated for AncSR1+RH (pink) and AncSR1+RH+AB (green). Values are mean ± SEM of three replicate MD simulations. (B) Groups A and B increase packing efficiency at the protein dimerization interface. Packing efficiency was determined by calculating the number of atom pairs between protein monomers (residues 39-57) within 4.5Å. Values are mean ± SEM of three replicate MD simulations. (C) The permissive substitution v42L improves packing efficiency at the protein dimerization interface. Packing efficiency was calculated for each individual residue in the dimerization interface (residues 39-57) for AncSR1+RH (pink) and AncSR1+RH+AB (green). Values are mean ± SEM of three replicate MD simulations.

———————————————

functioning to alter the conformational differences between the bound and free protein, the permissive substitutions would decrease the entropic cost of binding and would therefore improve single-site affinity and/or cooperativity.

To determine if the permissive substitutions changed the conformation of the protein to more closely resemble the bound form, we determined the mean position structures for all proteins across the MD trajectory. We then aligned the structures of free and bound proteins and reported the RMSD of the alignment (Figure 5A-B). If the permissive substitutions altered the conformation of the protein to more closely resemble the bound form, we would expect a lower RMSD for the alignment of free and bound proteins in the presence of the permissive groups A and B. Comparison of the mean position structures for the free and bound proteins showed that the variation of RMSDs from the alignments of the free and bound forms was not significantly different from the variation within the three replicates of either form (Figure 5B). This implies that the structural variation between replicates of the free protein is of the same magnitude as the structural variation between the free and bound structures. These data therefore do not provide sufficient evidence to imply that the permissive substitutions function by structuring the protein to more closely resemble the bound conformation.

To determine if the permissive substitutions improved single-site affinity and/or cooperativity by decreasing the entropic dynamics of the free protein, we measured the flexibility of the protein backbone. We determined the backbone flexibility by calculating the RMSD of all Cα proteins across each 50ns trajectory (Figure 5C). We then compared the RMSDs of the protein in the absence and presence of the groups A and B. If the permissive substitutions decreased the entropy of the protein, we would expect for a decrease in backbone flexibility, manifest by a smaller RMSD of the backbone atoms in the presence of these groups.

Upon comparison of the Cα RMSD for AncSR1+RH and AncSR1+RH+AB we found that the permissive groups do not cause a significant change in the flexibility of the protein in the MD simulations. We observed that the highest RMSD values were for residues that occurred within the dimerization domain (residues 39-57), but introduction of the permissive groups A and B did not significantly change the protein's flexibility across this region (Figure 5C). These data do no support a role of the permissive substitutions in improving single-site affinity and/or cooperativity by decreasing the dynamics of the free or bound protein.

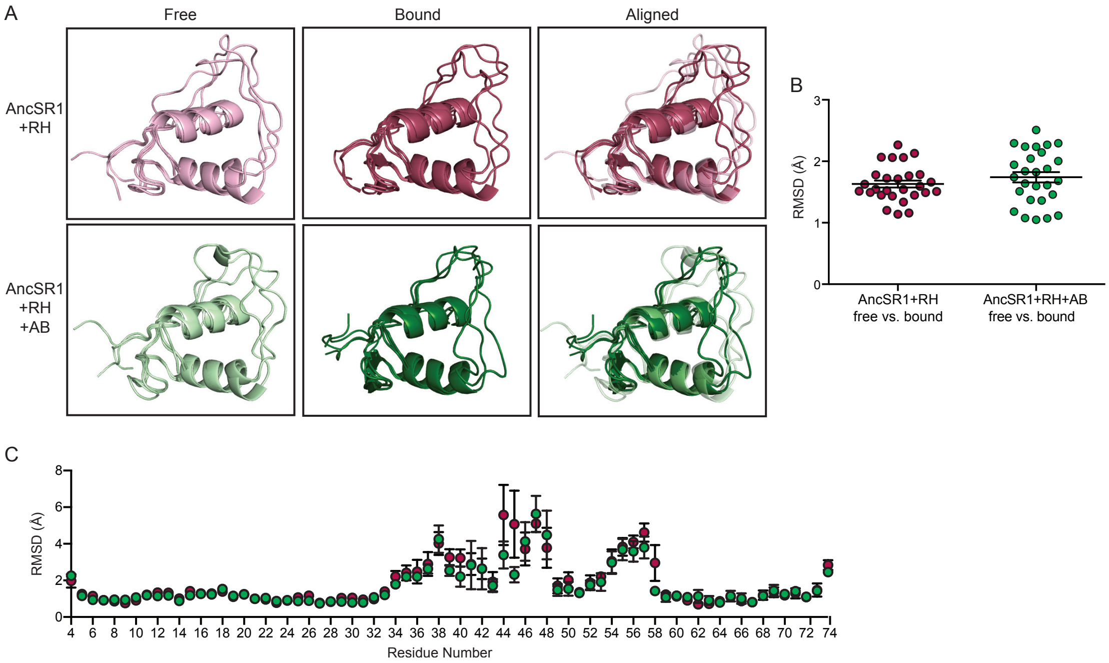


Figure 5. The permissive substitutions do not significantly alter protein conformation or backbone flexibility.
(A) Permissive groups A and B do not alter the conformation of the protein to more closely resemble the bound form. Mean position structures of three independent MD simulations of AncSR1+RH (top) and AncSR1+RH+AB (bottom) when free in solution (left) and when bound as a monomer (middle). Aligned free and bound structures are on the right. (B) Mean position structures of free proteins were compared with structures of monomeric and dimeric bound proteins. RMSD of the alignment was reported. Circles represent the RMSD of the alignments of free and bound proteins for all possible comparisons between replicates. Lines are mean ± SEM. Introduction of the permissive substitutions does not decrease the RMSD of the alignment between free and bound protein conformations. (C) Permissive groups A and B do not stabilize specific regions of the protein. RMSD of Cα atom for each residue in the MD simulations. Values are mean ± the SEM of three replicate MD simulations. For all panels: pink, AncSR1+RH; green, AncSR1+RH+AB.

———————————————————

Despite these findings, it is still possible that the permissive substitutions contribute to single-site affinity by decreasing the number of possible conformations that could be sampled by the free protein. Therefore, instead of forcing the protein into a single conformation that resembled the bound form, the permissive substitutions could decrease the entropic cost of binding by decreasing the number of single conformations that could be sampled by the free protein. If this were the case, our previous measurements would not be able to detect these differences. In order to determine if this is the mechanism by which these substitutions function, we would have to identify and enumerate all of the conformations that were sampled by the free protein in the MD simulation. If the permissive groups functioned to decrease the entropy of the free protein, we would expect for introduction of the permissive groups to greatly decrease the number of conformational states sampled by the protein.

The permissive substitutions may contribute to single-site affinity and cooperativity through an entropic mechanism that is not measureable by molecular dynamics

From this molecular dynamics approach, we were only able to identify one potential mechanism by which the permissive substitutions contributed to an increase in the macroscopic binding affinity for SRE and allowed for the protein to permit the derived RH. Although we tried to utilize MD to examine the entropic contribution to binding, we were unsuccessful in our attempts. However, several lines of evidence point to the potential of a mechanism driven by entropic forces to explain the improvement in the thermodynamics of complex formation.

First, NMR relaxation studies on human ER and human GR show differences in dynamics of both the free and bound proteins wherein human ER is much more dynamic than human GR (Wikstrom et al., 1999). These dynamic experiments suggest that novel specificity may have evolved concurrently with changes in protein dynamics.

Second, we find that an increase in enthalpically significant, positive interactions, specifically hydrogen bonding and packing, are insufficient to completely account for an increase in either single-site affinity or cooperativity (Figures 4-5). Since an improvement in the ΔG of single-site affinity and cooperativity must be due to changes in enthalpic or entropic components, the absence of improvement in enthalpic components therefore implies that the observed differences in ΔG of single-site affinity and cooperativity may solely be due to differences in entropy.

Third, biophysical characterization of the thermodynamic components of binding for extant receptors by isothermal titration calorimetry demonstrate that the binding of human GR, which contains all of the permissive substitutions, has greater entropic contributions to binding relative to the enthalpically-driven binding of human ER (Lundback and Hard, 1996; Deegan et al., 2010). Although these differences could be due to variation in the release of interfacial solvent given differences in surface complementation (Ha et al., 1989), it could also be due to significant changes in conformational dynamics (Spolar and Record, 1994). These observations suggest that the functional divergence in SRs may have also occurred by changes in the entropic contributions to binding.

The lack of evidence in the MD simulations to support this hypothesis may be due to the inability of this method to accurately describe the chemical system at such fine detail. To this end, we are currently employing NMR to seek resolution of these potential mechanisms. Specifically, we are using NMR to characterize the backbone dynamics of protein residues in AncSR1+RH and AncSR1+RH+AB when free in solution and when bound to a single-site SRE DNA fragment. Comparison of these dynamics will allow for us to identify changes in the dynamics of the free and bound proteins and to determine how permissive groups A and B affect these changes. This will allow for us to directly address the role of permissive substitutions in altering the entropic cost of binding to facilitate the evolution of novel DNA specificity.

DISCUSSION

The molecular roles of permissive substitutions are context-specific; they function to specifically counter-act the destabilizing effects of the function-switching mutations

Our studies suggest that, in this case, the permissive substitutions did not operate on global or local protein stability, but instead functioned to improve the stability of the complex. By acting at the protein-DNA and protein-protein interfaces, these permissive substitutions exploited the thermodynamic components of cooperative, dimeric binding to realize a high-affinity interaction even in the presence of the destabilizing RH substitutions. This work implies that the function of the permissive substitutions varies depending on the specific effect of the function-switching substitutions. Specifically, the role of permissive substitutions is to alter the protein in order to maintain the physical properties of the system that were negatively affected by the function-switching substitutions. In the case of DNA-binding proteins, this critical property was high-affinity binding.

Since function-switching substitutions cause varying effects across different molecular systems, it is not surprising that there exist no general mechanism for the functional role of permissive substitutions. However, given the shared biophysical architecture across protein-DNA interactions, we propose that non-specific increases in macroscopic DNA-binding affinity is a common mechanism by which permissive substitutions function to allow for the evolution of novel DNA-binding specificity. Additional studies in other DNA-binding model systems are required to test this hypothesis.

The effect of permissive substitutions is not constrained by their proximity to the destabilizing function-switching substitutions

Our study also shows that functionally important substitutions are not confined to the region of the protein that is destabilized by the function-switching substitutions. Instead, we find that the permissive substitutions occur throughout the protein and even in regions that are very distant from the binding interface. These results imply that the occurrence and effect of a permissive substitution is not constrained by its proximity to the destabilizing mutation(s).

Further, we find that the permissive substitutions do not function in an "equal and opposite" way as the destabilizing functions (Davis et al., 2009). As indicated by previous work (McKeown et al., 2014), we find that the RH substitutions function to solely alter changes in single-site binding affinity. If the permissives functioned in a completely equal and opposite way, we would have expected for their effect to only be due to changes in single-site binding affinity. Although we do observe a noticeable effect of the permissive substitutions on single-site affinity, the increase is not sufficient to off-set the affinity-reducing effects of the derived RH. Further, we find that the role of the permissive substitutions is to greatly improve inter-protein cooperativity. This indicates that the combined functional effects of the permissive substitutions on macroscopic binding affinity resulted from the exploitation of existing energetic components defined by the biophysical architecture of the ancestral SR complex.

The permissive substitutions may have been selected for their improvement in DNA affinity

It is thought that the "nearly neutral" nature of permissive substitutions implies that these substitutions cannot be agents of selection as their effects are invisible to evolution. In this regime, the accumulation of these mutations is contingent on the chance wanderings of the protein through its neutral sequence space. However, defining a permissive substitution as neutral depends on the function being investigated. During the evolution of SRs, the permissive substitutions are neutral for specificity and therefore could not have been fixed by selection for a novel specificity. However, the permissives are not neutral for SRE affinity. Instead they have a very large positive effect, increasing SRE affinity by improving protein-DNA and protein-protein interactions. Therefore, if selection was operating to improve affinity with SRE, then the effects of the permissive substitutions could have been the agents of positive selection. Dissecting these two possibilities is impossible given what we know about the evolutionary history of the SR family. However, these observations imply that permissive substitutions may not just serve as a neutral "buffer" for function-switching mutations, but, depending on their immediate biochemical effects, may themselves be the primary determinants of a selected function.

EXPERIMENTAL PROCEDURES

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: SRE-half – CCAGAACAGAG, CTCTGTTCTGG; SRE-full – CCAGAACAGAGTGTTCTGA, TCAGAACACTCTGTTCTGG. 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.

FP competition binding assays

DNA constructs were ordered as HPLC-purified single stranded oligos from Eurofins Operon. Fluorescently labeled oligos were ordered with a covalent 6-FAM modification on the 5’ end of the forward strand. Sequences of forward and reverse strands, respectively, for short and long oligo sequences were as follows: SRE-half-short – CCAGAACAGAG, CTCTGTTCTGG; SRE-half-long – ATTCAGCCAGAACAGAG, CTCTGTTCTGGCTGAAT. Forward and reverse strands were re-suspended in duplex buffer (20mM Tris [pH 7.7], 130 mM NaCl) to a concentration of 200 μM. Equimolar concentrations of complementary forwards 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 100 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. Protein was incubated for 20 minutes with 10nM of labeled short oligos at a concentration that was 10 times the K_D (as determined previously by direct FP binding). 30 μL of pre-bound complex was pipette onto a black, NBS-coated 384 well plate (Corning 3575). 60 μL of a range of unlabeled short or long oligo was then titrated in triplicate into wells containing the pre-bound labeled complex. Plates were incubated for 15 minutes and sample fluorescence polarization was read using a Perkin Elmer Victor X5, exciting at 495nm and measuring emission polarization at 520nm. Data was fit to a model of single-site competition with ligand depletion as previously described by Wang and colleagues (Wang 1993).

Molecular dynamics simulations

The crystal structure of AncSR1 bound to ERE was 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) (Hoops et al., 1991; 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.

BRIDGE TO CHAPTER V

In the preceding chapters, we dissected the functional effects of groups of historical substitutions and determined their contribution to the evolution of novel DNA-binding specificity. In Chapter V, we synthesize how this work has contributed to a greater understanding of the molecular determinants of DNA-binding specificity as well as the evolutionary processes by which they evolved.

CHAPTER V
CONCLUSION

This in depth study of the molecular mechanisms by which a family of transcription factors evolved their diverse DNA-binding functions has allowed for resolution of key questions existing at the interface of biochemistry and evolutionary biology. The results of this work offer insights into the determinants of DNA-binding specificity as well as general mechanisms for the evolutionary processes by which they evolved. Given the similar biophysical architecture observed across DNA-binding proteins, we believe that this work elucidates general principles by which novel DNA specificity can evolve while also offering general principles for the evolution of novel function in other molecular systems.

Diverse functions do not have to evolve by partitioning the function of a promiscuous ancestor, but may evolve through exploitation of a latent function

A common question in evolutionary biology is whether novel functions and phenotypes evolve through refinement of a promiscuous protein that was capable of performing multiple functions. In the SR transcription factor family, the functional differences that we observe in extant receptors could have been due to the independent evolution of both ERE and SRE-specificity on the post-duplication lineages leading from a promiscuous common ancestor that bound both ERE and SRE with high affinity. Alternatively, we found that the functions of modern-day SRs evolved by neo-functionalization on one lineage following a duplication event; the ancestral DNA-binding function was conserved on the lineage leading to modern-day estrogen receptors while AncSR2 and its descendants realized a completely novel DNA specificity.

Although the evolution of modern-day SR specificity did not occur through refinement of a promiscuous ancestor that bound both ERE and SRE with high affinity, it did occur through exploitation of a latent binding function. These results imply that the evolution of a novel phenotype may be facilitated by improving existing interactions instead of establishing novel interactions completely *de novo*. Similar mechanisms have been observed in many other systems (Bridgham et al., 2006; Khersonsky et al., 2006;

Coyle et al., 2013), potentially making this a general mechanism for the evolution of a novel function.

Novel function can be realized through substitution of a few key residues of large effect

This work also allowed for us to determine the number and effect size of substitutions that contributed to the evolution of a novel function. Careful biochemical characterization of the historical substitutions sufficient for the derived function showed that novel DNA-binding specificity was realized through three, large-effect substitutions. Functional dissection of these three substitutions showed that each had a drastic effect on protein function, having a dual role in actively eliminating the ancestral preference while also helping to established the derived preferences. These data indicate that the protein evolved a novel function by sampling a very minimal region of its sequence space.

Functional characterization of the combinatorial complete set of genotypes possible for these three substitutions also allowed for us to identify potential mutational pathways by which substitutions in both the protein and the RE could occur while still maintaining a high-affinity interaction. The size of this system's high-affinity network suggests that it was highly evolvable and had many mutational pathways that would allow for the evolution of novel interactions between the protein and RE. Further, we were able to identify many mutational pathways by which compensatory mutations in the RE could permit diversification of protein binding and still maintain an ancestral connection. The size of the high-affinity network that exists within this sequence space separating these two functions indicates that there were multiple ways in which the protein and/or the RE could wander its way across this space and ultimately connect functional spaces that might otherwise appear completely discrete.

Evolution of substitutions necessary for a novel function can affect multiple protein properties, but must be compatible with the system's biophysical architecture

This work also allowed for us to identify the functional and structural effects of specificity-switching mutations and to determine why permissive substitutions were required for the protein to tolerate them. Only three substitutions were required to realize a novel specificity, but these specificity-switching mutations were negative for protein function and resulted in a low-affinity protein that was not capable of driving expression from either RE. Their deleterious effect on function was due to the biophysical mechanisms by which they caused a change in DNA-binding specificity. Instead of evolving novel positive interactions with the newly preferred DNA sequence, these substitutions mainly operated to affect negative interactions at the interface. These new negative interactions greatly weakened the strength of interactions at the protein-DNA interface, forcing the protein below a thermodynamic threshold and resulting in a very low affinity intermediate.

Although changes in negative interactions may have been the most accessible mechanism by which evolution could realize a novel specificity, this mechanism was not compatible with the biophysical architecture of protein-DNA interactions. In order to compensate for this incompatibility, a suite of additional permissive substitutions were required to maintain a high affinity complex. These permissive substitutions were not just localized to the protein-DNA interface and did not function solely to establish novel interactions at the protein-DNA interactions. Instead, they were spread throughout the entire domain and functioned to increase affinity by improving interactions at the protein-DNA interface while also greatly increasing inter-protein cooperativity. Together, these data indicate that the complexity of this potential mutational pathway leading to novel specificity was a product of the molecular constraints imposed by the biophysical architecture of the system to maintain a high-affinity interaction between the protein and DNA.

Evolution of novel specificity was greatly shaped by epistatic interactions within and between the interacting molecules of the evolving system

Characterizing the individual and combined effects of the historical substitutions on DNA-binding function also allowed for us to identify the presence and effects of epistasis on the evolution of the system. We observed extensive epistasis both within and between the interacting molecules of this system. Although most of these epistatic interactions were between structurally adjacent residues and nucleotides, we also observed significant epistasis between residues that occurred in very distant regions of the protein. Further, the epistatic effects between residues varied across the protein's thermodynamic properties, differentially affecting the energetic components of association. These results imply that the function of a given protein is not solely due to the sum of its parts, but is a product of the individual and epistatic effects of residues structurally adjacent and distant from interfaces that directly contact a specific ligand. This distribution of functionally significant residues, and the existence of epistatic interactions between them, implies that substitutions that occur on all functional interfaces of the protein can and do contribute to the evolution of a novel function.

The epistatic interactions that we observed between the protein and the DNA also suggest a level of complexity that has not yet been reported for evolving systems. Considering the evolution of the SR transcriptional module as a whole, it is evident that the function of the system arises from the interconnected sequence spaces of the protein and of the RE. The epistasis that arises from this interconnectedness implies that movement through the system's sequence space by mutation of one macromolecule directly affects the potential mutational pathways available to the other macromolecular that would maintain a high-affinity interaction. As such, the inter-molecular epistasis between the protein and DNA directly shaped the evolvability of the system, likely permitting some mutational pathways while potentially constraining others. Together, these results imply that the evolution and evolvability of a multi-component system is a product of the interactions within and between its interacting parts and cannot be determined solely by studying its macromolecules in isolation. Instead, we must aim to understand how each of its interacting parts evolved together to maintain ancestral connections and/or give rise to functional novelty. Approaching molecular studies of interacting systems in this way will lend insights into the molecular determinants of multi-component interactions as well as the evolutionary processes by which they evolved.

APPENDIX A

CHAPTER II SUPPLEMENTAL INFORMATION

SUPPLEMENTAL FIGURES

Figure S1 (next page): Inference of the ML steroid receptor phylogeny and reconstruction of AncSR1 and AncSR2 with high confidence; related to Figure 1. 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.

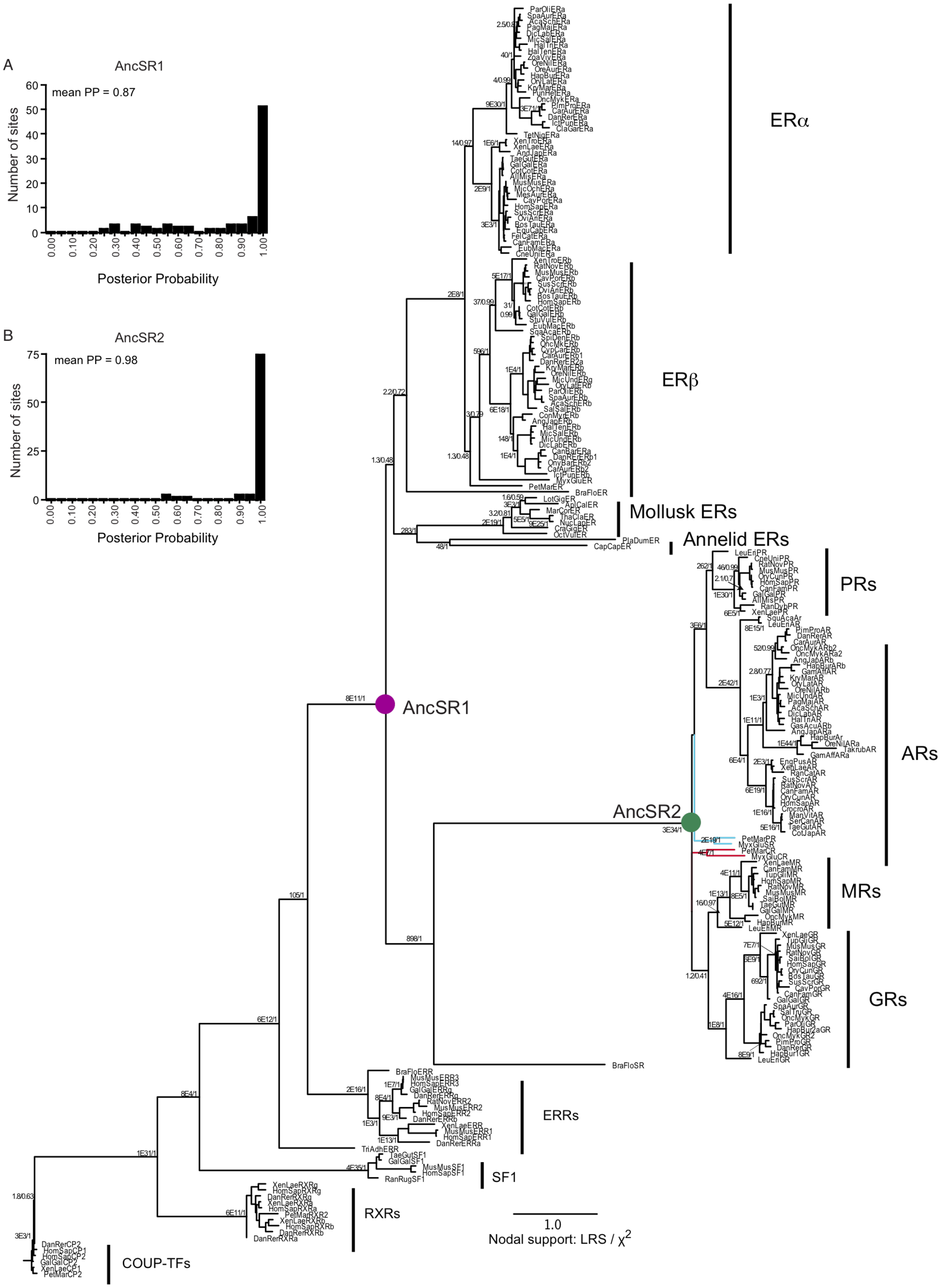
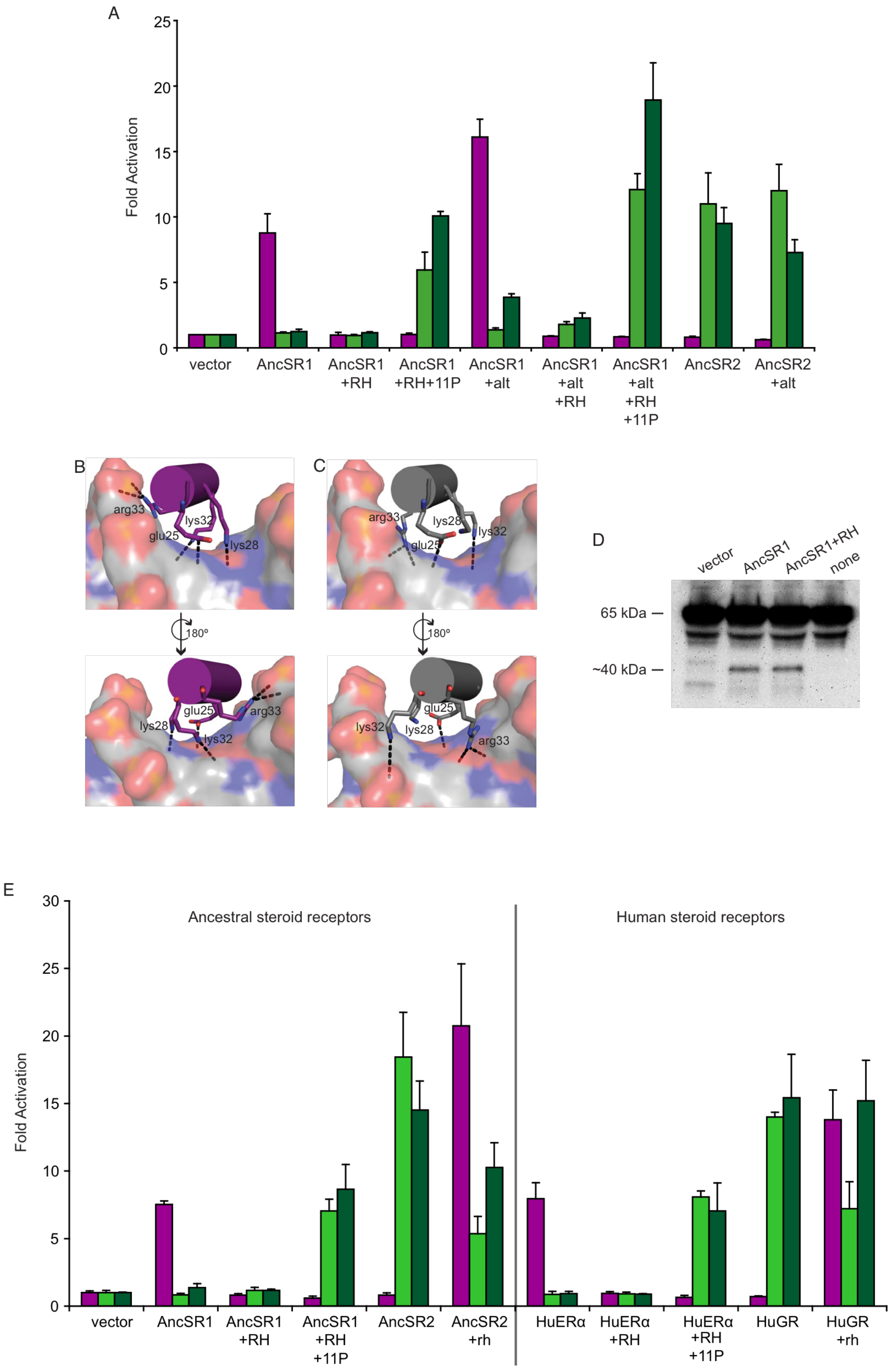

A
AncSR1
mean PP = 0.87
Number of sites
0
10
20
30
40
50
60
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Posterior Probability
B
AncSR2
mean PP = 0.98
Number of sites
0
25
50
75
0.00
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
1.00
Posterior Probability
ERα
ERβ
Mollusk ERs
Annelid ERs
PRs
ARs
MRs
GRs
AncSR1
AncSR2
ERRs
SF1
RXRs
COUP-TFs
BraFloSR
BraFloER
PetMarER
MyxGluER
PlaDumER
CapCapER
LotGigER
AplCalER
MarCorER
ThaClaER
NucLapER
CraGigER
OctVulER
PetMarPR
MyxGluSR
PetMarCR
MyxGluCR
BraFloERR
TriAdhERR
TaeGutSF1
GalGalSF1
MusMusSF1
HomSapSF1
RanRugSF1
XenLaeRXRg
HomSapRXRg
DanRerRXRg
XenLaeRXRa
HomSapRXRa
PetMarRXR2
XenLaeRXRb
HomSapRXRb
DanRerRXRb
DanRerRXRa
DanRerCP2
HomSapCP1
HomSapCP2
GalGalCP2
XenLaeCP1
PetMarCP2
8E11/1
3E34/1
898/1
105/1
6E12/1
8E4/1
1E31/1
1.8/0.63
3E3/1
2E16/1
6E11/1
4E35/1
2E8/1
2.2/0.72
1.3/0.48
283/1
48/1
1.0
Nodal support: LRS / χ^2

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 3. (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
0
5
10
15
20
25
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
Fold Activation
0
5
10
15
20
25
30
Ancestral steroid receptors
Human steroid receptors
vector
AncSR1
AncSR1 +RH
AncSR1 +RH +11P
AncSR2
AncSR2 +rh
HuERα
HuERα +RH
HuERα +RH +11P
HuGR
HuGR +rh

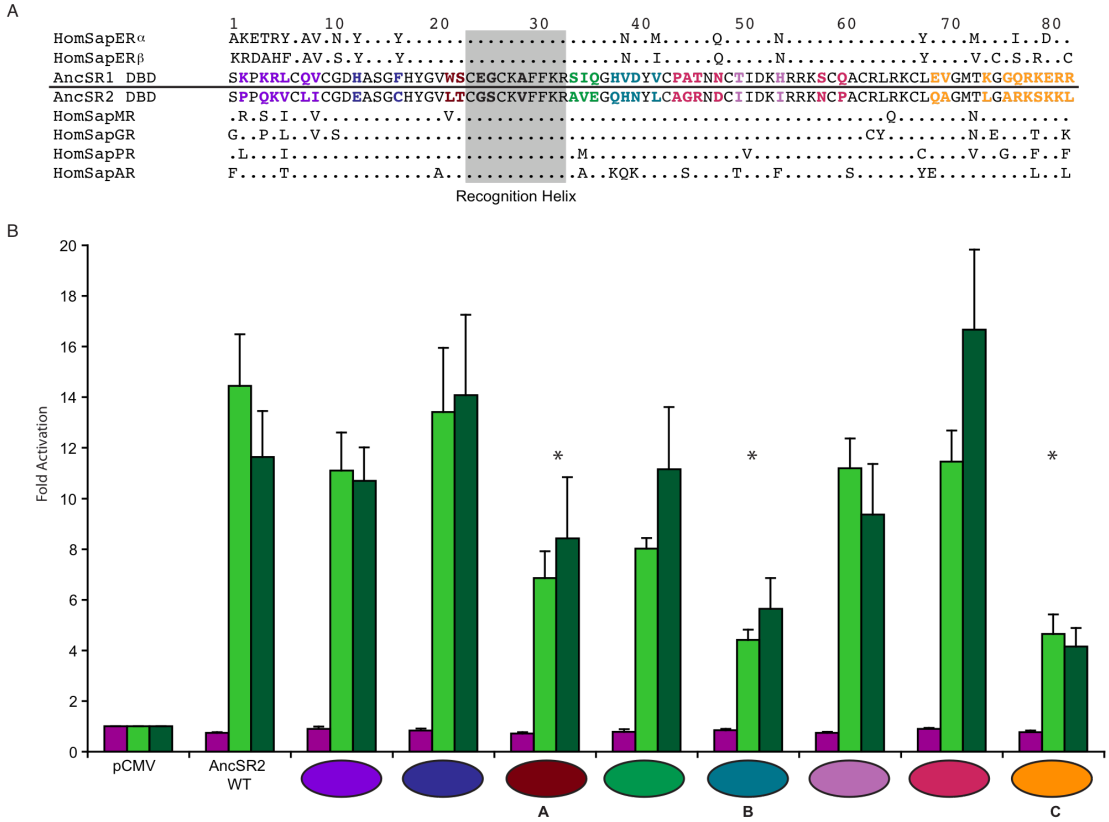


Figure S3: Three groups of permissive substitutions are required to support the derived specificity; related to Figure 3. (A) Alignment of ancestral and human DBDs shows amino acid differences; residues that are conserved between human DBDs and their closest ancestral sequence are indicated by '.' In 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 3. (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 Figure S3), 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 Table S5). (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.

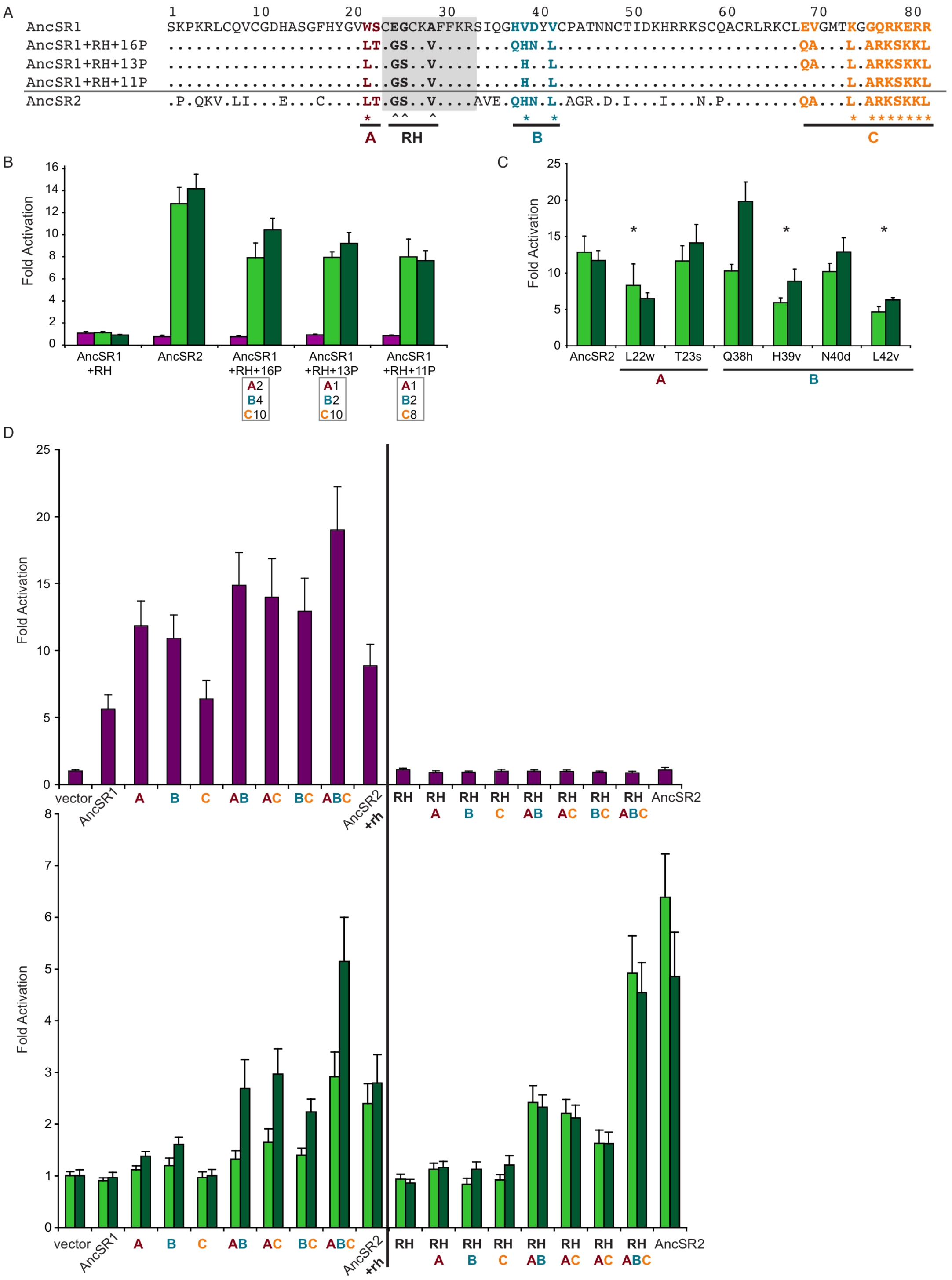

A
1 10 20 30 40 50 60 70 80
AncSR1 SKPKRLCQVCGDHASGFHYGVWSCEGCKAFFKRSIQGHVDYVCPATNNCTIDKHRRKSCQACRLRKCLEVGMTKGGQRKERR
AncSR1+RH+16P .....................LT.GS..V........QHN.L..........................QA...L.ARKSKKL
AncSR1+RH+13P .....................L..GS..V.........H..L..........................QA...L.ARKSKKL
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
0 2 4 6 8 10 12 14 16
AncSR1 +RH
AncSR2
AncSR1 +RH+16P
A2 B4 C10
AncSR1 +RH+13P
A1 B2 C10
AncSR1 +RH+11P
A1 B2 C8
C
Fold Activation
0 5 10 15 20 25
AncSR2
L22w
T23s
Q38h
H39v
N40d
L42v
*
*
*
A
B
D
Fold Activation
0 5 10 15 20 25
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
Fold Activation
0 1 2 3 4 5 6 7 8
vector
AncSR1
A
B
C
AB
AC
BC
ABC
AncSR2 +rh
RH
RH A
RH B
RH C
RH AB
RH AC
RH AC
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 4.** (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.

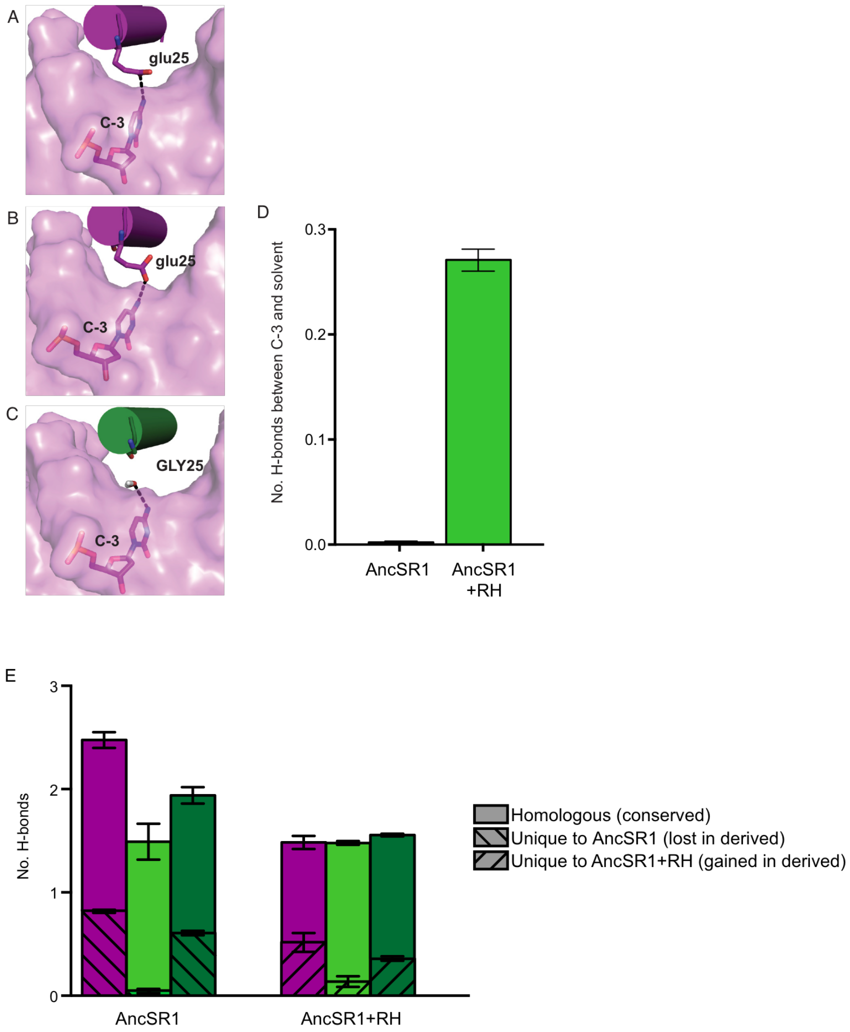
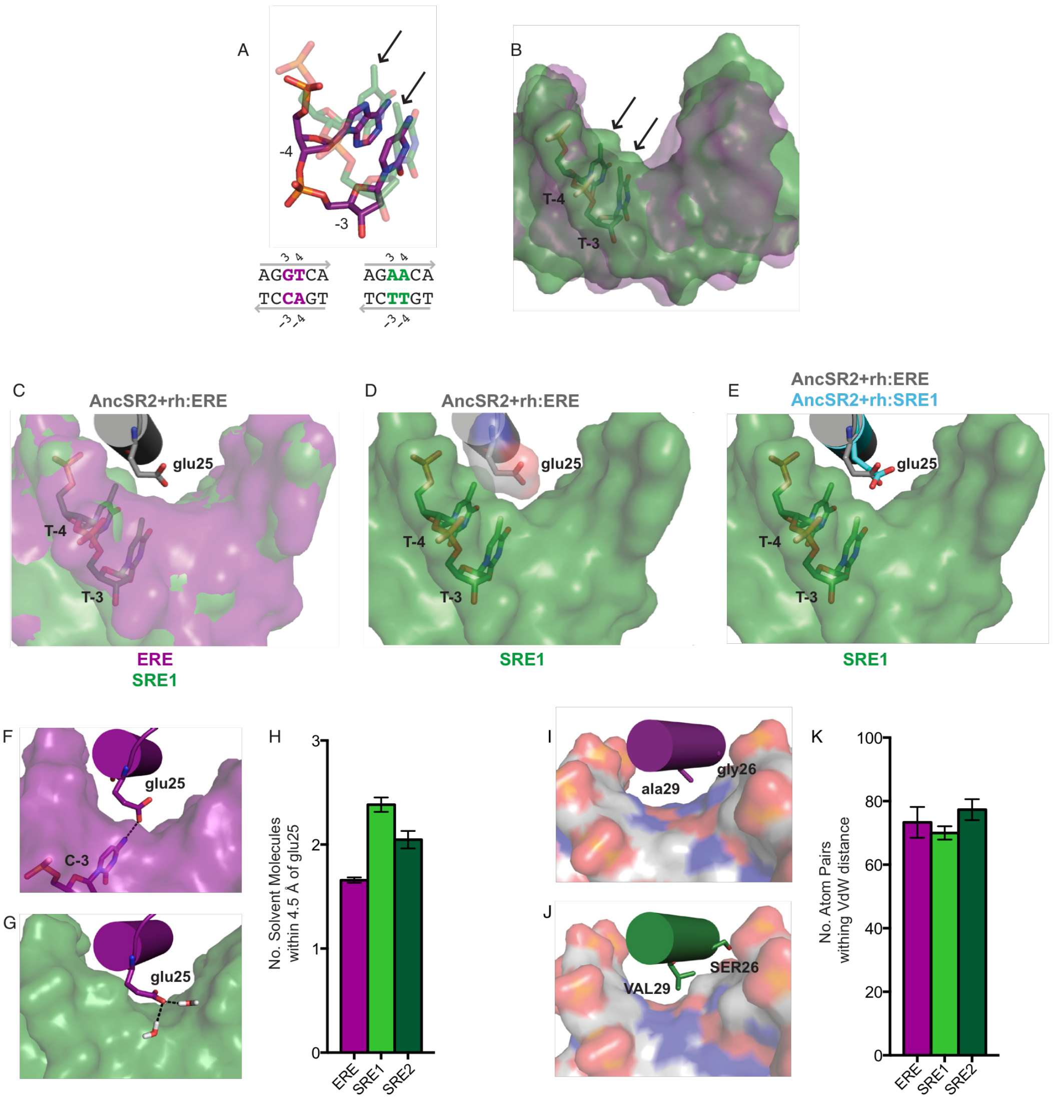

A
glu25
C-3
B
glu25
C-3
C
GLY25
C-3
D
No. H-bonds between C-3 and solvent
0.3
0.2
0.1
0.0
AncSR1
AncSR1 +RH
E
No. H-bonds
3
2
1
0
AncSR1
AncSR1+RH
Homologous (conserved)
Unique to AncSR1 (lost in derived)
Unique to AncSR1+RH (gained in derived)

Figure S6 (next page): The ancestral and derived RH exclude binding to non-target REs through negative interactions; related to Figure 4. (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
3 4
AGGTCA
TCCAGT
-3 -4
3 4
AGAACA
TCTTGT
-3 -4
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
0
1
2
3
ERE
SRE1
SRE2
I
ala29
gly26
J
VAL29
SER26
K
No. Atom Pairs withing VdW distance
0
20
40
60
80
100
ERE
SRE1
SRE2

SUPPLEMENTAL TABLES

Table S1 (next page): Posterior probabilities for each amino acid residue of AncSR1 and AncSR2 DBDs; related to Figure 1. 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	Cephalocordate ER	ERRs	Cephalocordate 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	Cephalocordate ER	ERRs	Cephalocordate 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 1, and Figures 3-4. 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 2 and Experimental Procedures.

	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	30.60-2.70	37.60-2.25	29.12-2.35
	(1.53-1.50)*	(2.80-2.70)*	(2.33-2.25)*	(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 3 and Experimental Procedures. 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; related to Experimental Procedures. (A) Zn-Cys interactions terms. (B) Partial charges for Cys and Zn atoms within each zinc finger.

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_I 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_I 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 5th 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) (Hoops et al., 1991; 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 B

CHAPTER III SUPPLEMENTAL INFORMATION

SUPPLEMENTAL FIGURES

Figure S1 (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.

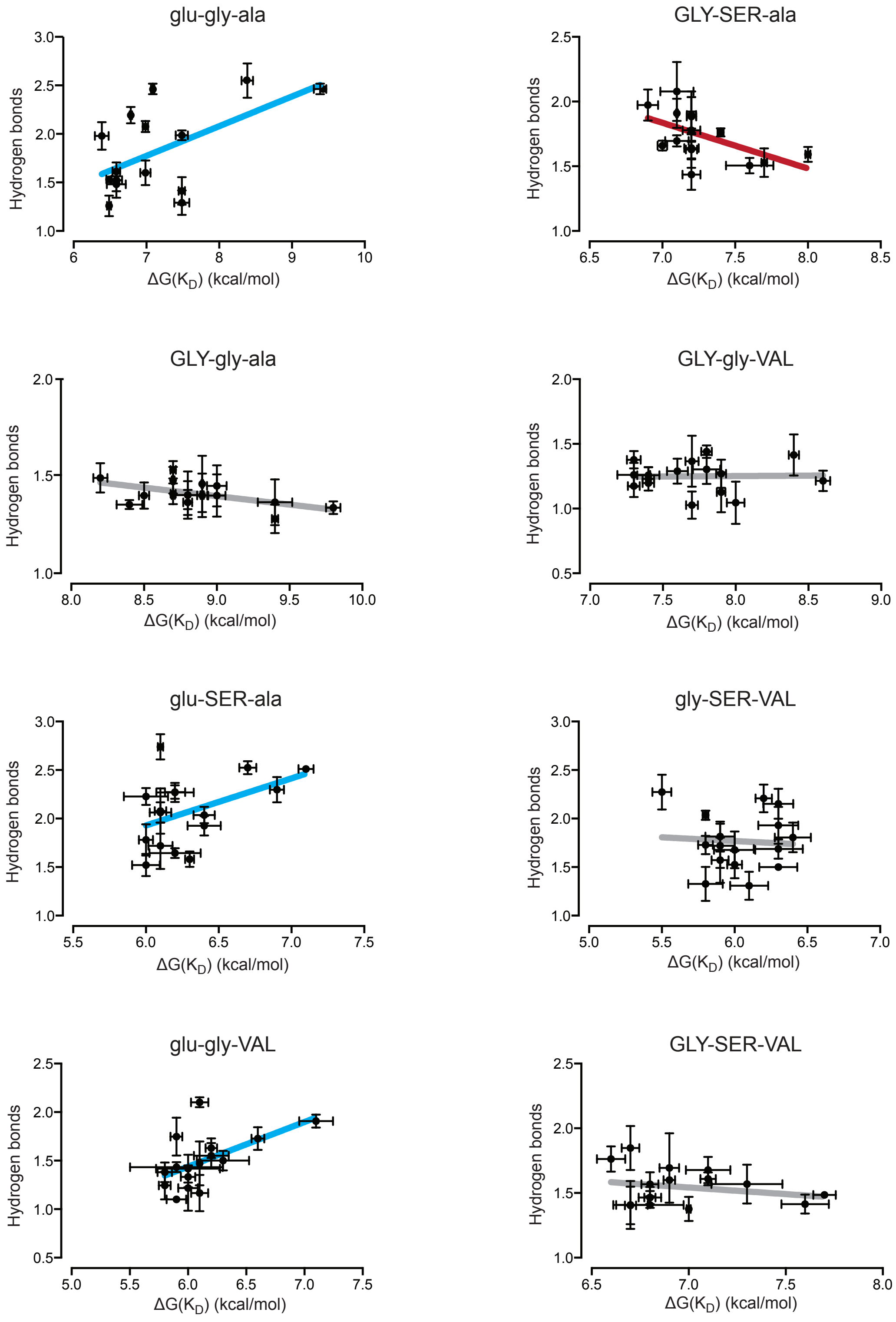
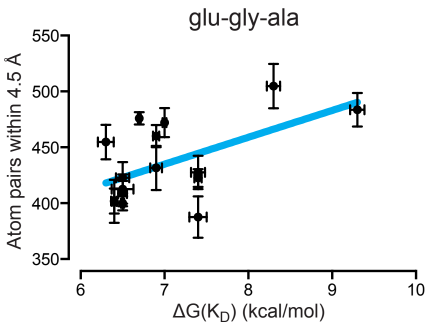
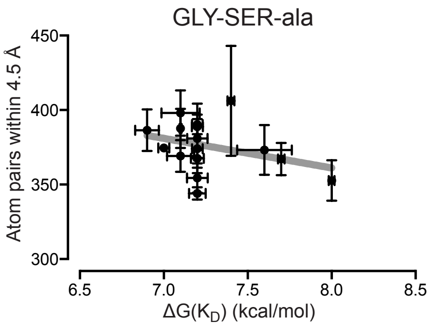
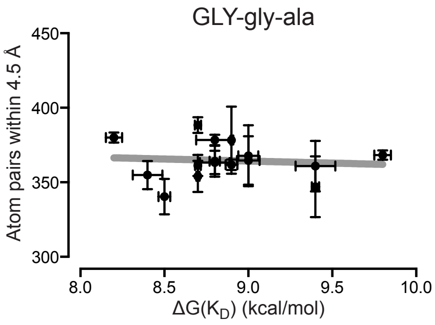
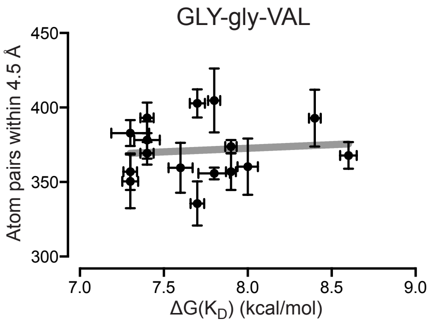

glu-gly-ala
Hydrogen bonds
3.0
2.5
2.0
1.5
1.0
6
7
8
9
10
ΔG(K_D) (kcal/mol)
GLY-SER-ala
Hydrogen bonds
2.5
2.0
1.5
1.0
6.5
7.0
7.5
8.0
8.5
ΔG(K_D) (kcal/mol)
GLY-gly-ala
Hydrogen bonds
2.0
1.5
1.0
8.0
8.5
9.0
9.5
10.0
ΔG(K_D) (kcal/mol)
GLY-gly-VAL
Hydrogen bonds
2.0
1.5
1.0
0.5
7.0
7.5
8.0
8.5
9.0
ΔG(K_D) (kcal/mol)
glu-SER-ala
Hydrogen bonds
3.0
2.5
2.0
1.5
1.0
5.5
6.0
6.5
7.0
7.5
ΔG(K_D) (kcal/mol)
gly-SER-VAL
Hydrogen bonds
3.0
2.5
2.0
1.5
1.0
5.0
5.5
6.0
6.5
7.0
ΔG(K_D) (kcal/mol)
glu-gly-VAL
Hydrogen bonds
2.5
2.0
1.5
1.0
0.5
5.0
5.5
6.0
6.5
7.0
7.5
ΔG(K_D) (kcal/mol)
GLY-SER-VAL
Hydrogen bonds
2.5
2.0
1.5
1.0
6.5
7.0
7.5
8.0
ΔG(K_D) (kcal/mol)

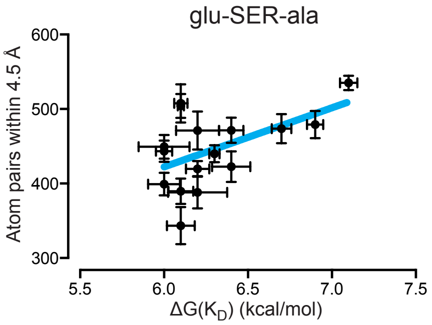
Figure S2 (next page). Packing efficiency is insufficient to account for variation in binding affinity across the transition from AncSR1 to AncSR1+RH.
Linear modeling of packing efficiency data versus binding affinity. Packing efficiency has a positive correlation (blue line) with binding affinity for 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.

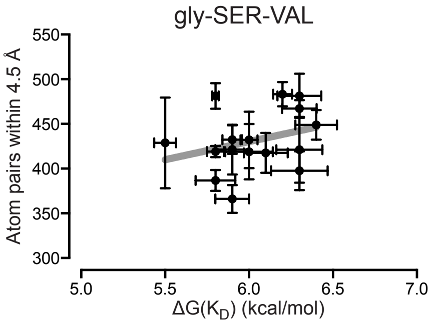

glu-gly-ala
Atom pairs within 4.5 Å
550
500
450
400
350
6
7
8
9
10
ΔG(K_D) (kcal/mol)

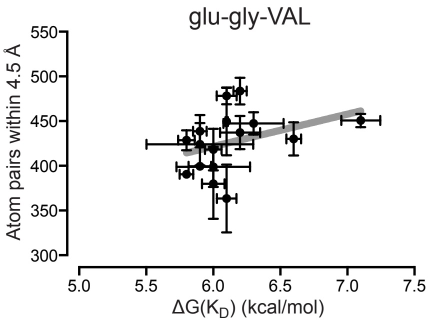

GLY-SER-ala
Atom pairs within 4.5 Å
450
400
350
300
6.5
7.0
7.5
8.0
8.5
ΔG(K_D) (kcal/mol)

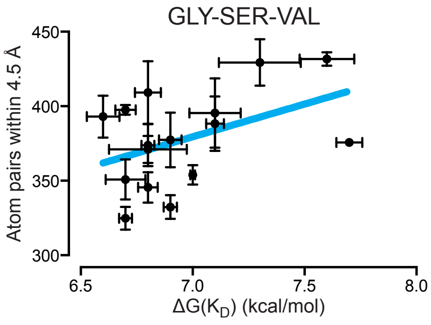

GLY-gly-ala
Atom pairs within 4.5 Å
450
400
350
300
8.0
8.5
9.0
9.5
10.0
ΔG(K_D) (kcal/mol)


GLY-gly-VAL
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)


glu-SER-ala
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)


gly-SER-VAL
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)


glu-gly-VAL
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)


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

SUPPLEMENTAL TABLES

Table S1 (next page): 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 ($^\Phi$) 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

<u>glu-gly-ala</u>

	AIC-optimized Model		Global Model	
Genetic Term	**Effect (Fold Affinity)**	**p**	**Effect (Fold Affinity)**	**p**
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

<u>GLY-gly-ala</u>

	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

<u>GLY-gly-VAL</u>

	AIC-optimized Model		Global Model	
Genetic Term	**Effect (Fold Affinity)**	**p**	**Effect (Fold Affinity)**	**p**
C3	0.560	6.3e-5	0.540	1.0e-4
G3 [Φ]	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 [Φ]	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 [Φ]	N/A	N/A	0.347	3.6e-5

GLY-SER-ala

Genetic Term	AIC-optimized Model		Global Model	
	Effect (Fold Affinity)	p	Effect (Fold Affinity)	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		Global Model	
	Effect (Fold Affinity)	p	Effect (Fold Affinity)	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		Global Model	
	Effect (Fold Affinity)	p	Effect (Fold Affinity)	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 Φ	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 S2: All first and second order terms from global linear models. Data was fit to a global model as described in the experimental procedures. 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

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.

Aharoni, A, L Gaidukov, O Khersonsky, S McQ Gould, C Roodveldt, and DS Tawfik. 2005. The ‘evolvability’ of promiscuous protein functions. *Nat Genet* 37, no. 1: 73-76.

Alberch, P. 1991. From genes to phenotype: dynamical systems and evolvability. *Genetica* 84, no. 1: 5-11.

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.

Babu, MM, NM Luscombe, L Aravind, M Gerstein, and SA Teichmann. 2004. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14, no. 3: 283-291.

Badis, G, MF Berger, AA Philippakis, S Talukder, AR Gehrke, SA Jaeger, ET Chan, G Metzler, A Vedenko, X Chen, H Kuznetsov, CF Wang, D Coburn, DE Newburger, Q Morris, TR Hughes, and ML Bulyk. 2009. Diversity and complexity in DNA recognition by transcription factors. *Science* 324, no. 5935: 1720-1723.

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.

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, G Chalepakis, M Schauer, and EP Slater. 1989. DNA regulatory elements for steroid hormones. *J Steroid Biochem* 32, no. 5: 737-747.

Beato, M, and A Sanchez-Pacheco. 1996. Interaction of steroid hormone receptors with the transcription initiation complex. *Endocr Rev* 17, no. 6: 587-609.

Bentley, P. J. 1998. *Comparative vertebrate endocrinology*. Cambridge, UK ; New York: Cambridge University Press.

Berglund, H, M Wolf-Watz, T Lundback, S van den Berg, and T Hard. 1997. Structure and dynamics of the glucocorticoid receptor DNA-binding domain: comparison of wild type and a mutant with altered specificity. *Biochemistry* 36, no. 37: 11188-11197.

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, LI Gong, and D Baltimore. 2010. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* 328, no. 5983: 1272-1275.

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.

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.

Britten, RJ, and EH Davidson. 1969. Gene regulation for higher cells: a theory. *Science* 165, no. 3891: 349-357.

Campagne, S, O Saurel, V Gervais, and A Milon. 2010. Structural determinants of specific DNA-recognition by the THAP zinc finger. *Nucleic Acids Res* 38, no. 10: 3466-3476.

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.

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.

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.

Darden, T, D York, and L 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.

Davis, BH, AF Poon, and MC Whitlock. 2009. Compensatory mutations are repeatable and clustered within proteins. *Proc Biol Sci* 276, no. 1663: 1823-1827.

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

Deegan, BJ, KL Seldeen, CB McDonald, V Bhat, and A Farooq. 2010. Binding of the ERalpha nuclear receptor to DNA is coupled to proton uptake. *Biochemistry* 49, no. 29: 5978-5988.

DePristo, MA, DM Weinreich, and DL Hartl. 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6, no. 9: 678-687.

Dietrich, MR. 1998. Paradox and persuasion: negotiating the place of molecular evolution within evolutionary biology. *J Hist Biol* 31, no. 1: 85-111.

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 RED 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.

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.

Erwin, DH, and EH Davidson. 2009. The evolution of hierarchical gene regulatory networks. *Nat Rev Genet* 10, no. 2: 141-148.

Fay, JC, and CI Wu. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu Rev Genomics Hum Genet* 4, 213-235.

Field, SF, and MV Matz. 2010. Retracing evolution of red fluorescence in GFP-like proteins from Faviina corals. *Mol Biol Evol* 27, no. 2: 225-233.

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 correlations between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399-433.

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

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

Gearhart, MD, SM Holmbeck, RM Evans, HJ Dyson, and PE Wright. 2003. Monomeric complex of human orphan estrogen related receptor-2 with DNA: a pseudo-dimer interface mediates extended half-site recognition. *J Mol Biol* 327, no. 4: 819-832.
Ghedin, E, NA Sengamalay, M Shumway, J Zaborsky, T Feldblyum, V Subbu, DJ Spiro, J Sitz, H Koo, P Bolotov, D Dernovoy, T Tatusova, Y Bao, K St George, J Taylor, DJ Lipman, CM Fraser, JK Taubenberger, and SL Salzberg. 2005. Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution. *Nature* 437, no. 7062: 1162-1166.

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.

Granovsky, AA, and PC Gamess. "Firefly version 7.1. G." http://classic.chem.msu.su/gran/firefly/index.html.

Grazulis, S, M Deibert, R Rimseliene, R Skirgaila, G Sasnauskas, A Lagunavicius, V Repin, C Urbanke, R Huber, and V Siksnys. 2002. Crystal structure of the Bse634I restriction endonuclease: comparison of two enzymes recognizing the same DNA sequence. *Nucleic Acids Res* 30, no. 4: 876-885.

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.

Ha, JH, RS Spolar, and MT Jr Record. 1989. Role of the hydrophobic effect in stability of site-specific protein-DNA complexes. *J Mol Biol* 209, no. 4: 801-816.

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.

Harms, MJ, and JW Thornton. 2014. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature* 512, no. 7513: 203-207.

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.

Hess, B, H Bekker, HJC Berendsen, and JGEM Fraaije. 1997. LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry* 18, no. 12: 1463-1472.

Hoops, SC, KW Anderson, and KM Merz. 1991. Force field design for metalloproteins. *J. Am. Chem. Soc.* 113, no. 22: 8262-8270.

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.

Humphrey, W, A Dalke, and K Schulten. 1996. VMD: visual molecular dynamics. *J Mol Graph* 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.

Jackson, CJ, JL Foo, N Tokuriki, L Afriat, PD Carr, HK Kim, G Schenk, DS Tawfik, and DL Ollis. 2009. Conformational sampling, catalysis, and evolution of the bacterial phosphotriesterase. *Proc Natl Acad Sci U S A* 106, no. 51: 21631-21636.

Jolma, A, J Yan, T Whitington, J Toivonen, KR Nitta, P Rastas, E Morgunova, M Enge, M Taipale, G Wei, K Palin, JM Vaquerizas, R Vincentelli, NM Luscombe, TR Hughes, P Lemaire, E Ukkonen, T Kivioja, and J Taipale. 2013. DNA-binding specificities of human transcription factors. *Cell* 152, no. 1-2: 327-339.

Kasahara, M, K Naruse, S Sasaki, Y Nakatani, W Qu, B Ahsan, T Yamada, Y Nagayasu, K Doi, Y Kasai, T Jindo, D Kobayashi, A Shimada, A Toyoda, Y Kuroki, A Fujiyama, T Sasaki, A Shimizu, S Asakawa, N Shimizu, S Hashimoto, J Yang, Y Lee, K Matsushima, S Sugano, M Sakaizumi, T Narita, K Ohishi, S Haga, F Ohta, H Nomoto, K Nogata, T Morishita, T Endo, T Shin-I, H Takeda, S Morishita, and Y Kohara. 2007. The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, no. 7145: 714-719.

Keller, W, P Konig, and TJ Richmond. 1995. Crystal structure of a bZIP/DNA complex at 2.2 A: determinants of DNA specific recognition. *J Mol Biol* 254, no. 4: 657-667.
Khersonsky, O, C Roodveldt, and DS Tawfik. 2006. Enzyme promiscuity: evolutionary and mechanistic aspects. *Curr Opin Chem Biol* 10, no. 5: 498-508.

Kondrashov, AS, S Sunyaev, and FA Kondrashov. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* 99, no. 23: 14878-14883.
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.

Kvitek, DJ, and G Sherlock. 2011. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape. *PLoS Genet* 7, no. 4: e1002056.

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, MW MacArthur, DS Moss, and JM Thornton. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography* 26, no. 2: 283-291.

Lin, F, and R Wang. 2010. Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. *J. Chem. Theory Comput.* 6, no. 6: 1852-1870.

Lindblad-Toh, K, M Garber, O Zuk, MF Lin, BJ Parker, S Washietl, P Kheradpour, J Ernst, G Jordan, E Mauceli, LD Ward, CB Lowe, AK Holloway, M Clamp, S Gnerre, J Alfoldi, K Beal, J Chang, H Clawson, J Cuff, F Di Palma, S Fitzgerald, P Flicek, M Guttman, MJ Hubisz, DB Jaffe, I Jungreis, WJ Kent, D Kostka, M Lara, AL Martins, T Massingham, I Moltke, BJ Raney, MD Rasmussen, J Robinson, A Stark, AJ Vilella, J Wen, X Xie, MC Zody, J Baldwin, T Bloom, CW Chin, D Heiman, R Nicol, C Nusbaum, S Young, J Wilkinson, KC Worley, CL Kovar, DM Muzny, RA Gibbs, A Cree, HH Dihn, G Fowler, S Jhangiani, V Joshi, S Lee, LR Lewis, LV Nazareth, G Okwuonu, J Santibanez, WC Warren, ER Mardis, GM Weinstock, RK Wilson, K Delehaunty, D Dooling, C Fronik, L Fulton, B Fulton, T Graves, P Minx, E Sodergren, E Birney, EH Margulies, J Herrero, ED Green, D Haussler, A Siepel, N Goldman, KS Pollard, JS Pedersen, ES Lander, and M Kellis. 2011. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* 478, no. 7370: 476-482.

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.

Lundback, T, and T Hard. 1996. Sequence-specific DNA-binding dominated by dehydration. *Proc Natl Acad Sci U S A* 93, no. 10: 4754-4759.

Lunzer, M, GB Golding, and AM Dean. 2010. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* 6, no. 10: e1001162.

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.

Lynch, VJ, A Tanzer, Y Wang, FC Leung, B Gellersen, D Emera, and GP Wagner. 2008. Adaptive changes in the transcription factor HoxA-11 are essential for the evolution of pregnancy in mammals. *Proc Natl Acad Sci U S A* 105, no. 39: 14928-14933.

Mann, RS, and G Morata. 2000. The developmental and molecular biology of genes that subdivide the body of Drosophila. *Annu Rev Cell Dev Biol* 16, 243-271.

Martin, RE, RV Marchetti, AI Cowan, SM Howitt, S Broer, and K Kirk. 2009. Chloroquine transport via the malaria parasite's chloroquine resistance transporter. *Science* 325, no. 5948: 1680-1682.

McCandlish, DM, E Rajon, P Shah, Y Ding, and JB Plotkin. 2013. The role of epistasis in protein evolution. *Nature* 497, no. 7451: E1-2; discussion E2.

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.

Nakagawa, S, SS Gisselbrecht, JM Rogers, DL Hartl, and ML Bulyk. 2013. DNA-binding specificity changes in the evolution of forkhead transcription factors. *Proc Natl Acad Sci U S A* 110, no. 30: 12349-12354.

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.

Ohno, Susumu. 1970. *Evolution by gene duplication*. Berlin, New York: Springer-Verlag.
Olson, EN. 2006. Gene regulatory networks in the evolution and development of the heart. *Science* 313, no. 5795: 1922-1927.

Orr, HA. 2005. The genetic theory of adaptation: a brief history. *Nat Rev Genet* 6, no. 2: 119-127.

Ortlund, EA, JT Bridgham, MR Redinbo, and JW Thornton. 2007. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* 317, no. 5844: 1544-1548.

Otwinowski, Zbyszek, and Wladek Minor. 1997. Processing of X-ray diffraction data collected in oscillation mode. *Methods in Enzymology* 276, 307-326.

Pace, CN, and JM Scholtz. 1997. *Measuring the conformational stability of a protein*. Protein structure: A practical approach. New York: Oxford University Press.

Pan, Y, CJ Tsai, B Ma, and R Nussinov. 2010. Mechanisms of transcription factor selectivity. *Trends Genet* 26, no. 2: 75-83.

Parera, M, and MA Martinez. 2014. Strong epistatic interactions within a single protein. *Mol Biol Evol* 31, no. 6: 1546-1553.

Peter, IS, and EH Davidson. 2011. Evolution of gene regulatory networks controlling body plan development. *Cell* 144, no. 6: 970-985.

Petz, LN, AM Nardulli, J Kim, KB Horwitz, LP Freedman, and DJ Shapiro. 1997. DNA bending is induced by binding of the glucocorticoid receptor DNA binding domain and progesterone receptors to their response element. *J Steroid Biochem Mol Biol* 60, no. 1-2: 31-41.

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.

Prud'homme, B, N Gompel, and SB Carroll. 2007. Emerging principles of regulatory evolution. *Proc Natl Acad Sci U S A* 104 Suppl 1, 8605-8612.

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.

Quattrocchio, F, J Wing, K van der Woude, E Souer, N de Vetten, J Mol, and R Koes. 1999. Molecular analysis of the anthocyanin2 gene of petunia and its role in the evolution of flower color. *Plant Cell* 11, no. 8: 1433-1444.

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.

Sapienza, PJ, T Niu, MR Kurpiewski, A Grigorescu, and L Jen-Jacobson. 2014. Thermodynamic and structural basis for relaxation of specificity in protein-DNA recognition. *J Mol Biol* 426, no. 1: 84-104.

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, MW, KK Baldridge, JA Boatz, ST Elbert, MS Gordon, JH Jensen, and S Koseki. 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, and D Rhodes. 1993. The crystal structure of the estrogen receptor DNA-binding domain bound to DNA: how receptors discriminate between their response elements. *Cell* 75, no. 3: 567-578.

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.

Shapiro, MD, ME Marks, CL Peichel, BK Blackman, KS Nereng, B Jonsson, D Schluter, and DM Kingsley. 2004. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature* 428, no. 6984: 717-723.

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.

Soskine, M, and DS Tawfik. 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11, no. 8: 572-582.

Spolar, RS, and MT Jr Record. 1994. Coupling of local folding to site-specific binding of proteins to DNA. *Science* 263, no. 5148: 777-784.

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.

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

Stormo, GD, and Y Zhao. 2010. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* 11, no. 11: 751-760.

Tawfik, DS. 2010. Messy biology and the origins of evolutionary innovations. *Nat Chem Biol* 6, no. 10: 692-696.

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.

Thomas, VL, AC McReynolds, and BK Shoichet. 2010. Structural bases for stability-function tradeoffs in antibiotic resistance. *J Mol Biol* 396, no. 1: 47-59.

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

Tokuriki, N, F Stricher, L Serrano, and DS Tawfik. 2008. How protein stability and new functions trade off. *PLoS Comput Biol* 4, no. 2: e1000002.

Tokuriki, N, and DS Tawfik. 2009. Protein dynamism and evolvability. *Science* 324, no. 5924: 203-207.

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

Tuch, BB, H Li, and AD Johnson. 2008. Evolution of eukaryotic transcription circuits. *Science* 319, no. 5871: 1797-1799.

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.

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

von Hippel, PH. 1994. Protein-DNA recognition: new perspectives and underlying themes. *Science* 263, no. 5148: 769-770.

von Hippel, PH. 2007. From “simple” DNA-protein interactions to the macromolecular machines of gene expression. *Annu Rev Biophys Biomol Struct* 36, 79-105.

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.

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.

Weinreich, DM, RA Watson, and L Chao. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59, no. 6: 1165-1174.

Welboren, WJ, FC Sweep, PN Span, and HG Stunnenberg. 2009. Genomic actions of estrogen receptor alpha: what are the targets and how are they regulated? *Endocr Relat Cancer* 16, no. 4: 1073-1089.

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.

Wilke, CO. 2012. Bringing molecules back into molecular evolution. *PLoS Comput Biol* 8, no. 6: e1002572.

Winkler, FK, DW Banner, C Oefner, D Tsernoglou, RS Brown, SP Heathman, RK Bryan, PD Martin, K Petratos, and KS Wilson. 1993. The crystal structure of EcoRV endonuclease and of its complexes with cognate and non-cognate DNA fragments. *EMBO J* 12, no. 5: 1781-1795.

Wittkopp, PJ, BK Haerum, and AG Clark. 2008. Regulatory changes underlying expression differences within and between Drosophila species. *Nat Genet* 40, no. 3: 346-350.

Worth, CL, S Gong, and TL Blundell. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol* 10, no. 10: 709-720.

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

Wuttke, DS, MP Foster, DA Case, JM Gottesfeld, and PE Wright. 1997. Solution structure of the first three zinc fingers of TFIIIA bound to the cognate DNA sequence: determinants of affinity and sequence specificity. *J Mol Biol* 273, no. 1: 183-206.

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, H Yang, and WT Starmer. 2008. Molecular basis of spectral tuning in the red- and green-sensitive (M/LWS) pigments in vertebrates. *Genetics* 179, no. 4: 2037-2043.

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.