THE UNIVERSITY OF CHICAGO

COMPUTATIONAL AND EXPERIMENTAL APPROACHES TO PROTEIN EVOLUTION: WHY A CLASSIC SELECTION TEST CAN MISLEAD, AND HOW TRANSCRIPTION FACTORS EVOLVE DNA SPECIFICITIES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF HUMAN GENETICS

BY
AARTI VENKAT

CHICAGO, ILLINOIS
MARCH 2018

To my dear friends and loved ones who made this possible

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my advisor, Dr. Joseph Thornton for his wonderful support and guidance during my PhD. Being trained as a computational biologist, I joined Joe's lab to learn how to do experiments. Joe taught me how to articulate and identify high impact questions in science — two of the most important things to learn as a scientist. Working with Joe has been incredibly rewarding, and also very challenging. During my PhD, I completed two projects in protein evolution: one computational, and the other experimental. While pursuing two disparate research areas in my graduate career was difficult, I learnt an enormous amount during this process through my communications with Joe. This included developments on both scientific and personal fronts, and I feel elated I got this opportunity. Joe is an exceptional writer, and I was able to learn some of it during my graduate school. I am very thankful to him for so many things, including introducing me to the best beer on this planet, la fin du monde, and his dog Violet, who remains blissfully unaware of the fact that she made me shed my fear of dogs.

My committee members, Erin Adams, Allan Drummond and Vincent Lynch have provided exceptional support, and I feel so lucky to have had them. Specifically, Erin and Allan. I could not have asked for anything better. I have gone into their office and talked for hours and hours, and they have generously lent me their time, advice, valuable support, and recommendations for postdocs. I would also like to acknowledge Anna Di Rienzo and Yoav Gilad for their timely help and invaluable guidance. Even though she was not on my committee, Anna provided me exceptional help, and I am very grateful to her for that.

The Thornton lab has been an incredibly strong and dynamic environment, and I have learnt so much from my colleagues. Thank you so much Arvind Pillai, Georg Hochberg, Tyler Starr, Mo Siddiq, Yeonwoo Park and Brian Metzger. I want to specifically thank Georg and Tyler for having read and extensively commented on anything I ever wrote in graduate school, including papers, grants and thesis. I learnt so much from them. Tyler has

been a phenomenal colleague who has always given me the best advice. Arvind has been a wonderful friend, and I have lost count of how many times in a day I went for coffee walks and food trips with him. He has listened to me for hours and hours replaying the tape of my life, and has never complained. I also want to thank Carrie and Qinwen, two of Joe's former postdocs for still keeping in touch with me, and providing me great advice when I needed them the most.

Matt Hahn, our collaborator on the MNM project was wonderful to work with. He always responded so fast to the gazillion drafts sent back and forth, that I remain amazed how he does that.

I am very thankful to Elena Solomaha at the biophysics core to be there with me, during the most horrible phases of my graduate life with ITC experiments. She has seen my blood, sweat and tears, and somehow emerge alive from the process. To whoever reading this, please do not let this get you down with ITC! Its not for everyone, but perhaps could be the one for you. I am also thankful to 'NMR Joe' for being a great source of joy during my meaningless water-to-water ITC titrations.

Graduate administrators Candice and Sue have been fabulous. We have become great friends after extensive interactions, and I cannot thank them enough for being patient and listening to my rants. Also, thank you very much Bonnie and Jeff, for helping me enormously with travel reimbursements and awards.

My friends in HG and GGSB have been a huge pillar of strength and support. I specifically want to thank Lauren, Diedre and Iuri for being there with me the whole time, and never getting tired of me. I have bugged them to no end, and they still love me. I feel very lucky that I have met them during graduate school, and I will always be in touch with them. Also Katie, Bryan, Andrei and Alex. Thank you so much for all your support. Special thanks to my cardio kickboxing instructor, Sidra, at the Ratner gym — Those exercises got me through some difficult times.

I am very happy I re-connected with my family after years of missing communication. They came from Bombay for my defense, and were elated to see me speak. I could not have imagined this a few years ago. They have been an incredible source of strength throughout this process.

Lastly, I want to thank Brendan Finucane, and his parents, Matt and Sandy Finucane for being absolutely wonderful. I completed this PhD under several challenging circumstances and Brendan has been a wonderful companion. I am very lucky to have met him.

ABSTRACT

An important goal of molecular evolution is to reveal the historical processes and mechanisms by which diverse molecular systems have evolved their present day forms and functions. This research program requires an integration of computational and mechanistic approaches to learn about the evolutionary processes that fix mutations in genes, and the genetic and biophysical mechanisms by which the historical mutations that fixed in evolution cause molecular functions to diverge. This dissertation describes one such functional synthesis.

I first focused on a classic computational phylogenetic method that infers lineage-specific adaptation on protein-coding genes: the branch-site test. I show that a newly discovered phenomenon in molecularstudies of mutation, multinucleotide mutations, causes a very strong bias towards false inferences of positive selection by the branch-site test; this bias is so pervasive that it can potentially explain many or even most inferences of positive selection in human and fruitfly genomes that have been made based on the test. These results call into question thousands of previously published inferences of positive selection, and suggest that the importance of adaptation in shaping genes and genomes could be vastly distorted. A version of the branch-site test that incorporates MNMs to partially reduce this bias is developed.

Next, to provide a mechanistic explanation of how historical mutations changed protein functions, I investigated the evolution of novel DNA-specificity in steroid and related receptors, a gene family composed of two functionally diverged clades of transcription factors: : 1) Steroid Receptors (SRs), which bind as a cooperative dimer to an inverted palindrome of a 6-bp half-site; and 2) Estrogen Related Receptors, which bind as monomers to an extended 9-bp half-site, containing a 5'-flanking extension of the 6-bp SR half-site. Using a combination of ancestral sequence reconstruction with biochemical and cell-based characterizations of protein function, I show that present-day SRs evolved from an ancestral ERR-like receptor with a preference for a specific 3-bp flanking sequence in addition to the 6-bp corehalf-site.

Six mutations in two structural groups occurred on the lineage leading to modern SRs and were sufficient for the loss of preference for the 3-bp flank. This change was offset by increasing affinity for the 6-bp half-site on the SR lineage, with no changes to cooperative binding. These findings show that new protein functions can evolve through changes in their thermodynamic properties without the radical evolution of novel interfaces. Together, the two projects aim to provide a complete understanding of molecular evolution — from generating hypotheses about functional change to establishing causality.

This dissertation includes unpublished material, with the MNM chapter on bioarxiv.

CHAPTER 1

INTRODUCTION

1.1 Evolutionary biology is an integrative science

A central goal of evolutionary biology is to explain the origin of diversity in species, their molecular systems and organization. This diversity of living systems can be attributed in part to the diverse complement of genes carried by them. Genes tend to be organized into gene families that have functionally diversified over evolution to produce a complex repertoire of proteins with different chemical and physical properties and organization. To explain the origin of diverse molecular systems, one must therefore understand the mechanisms and dynamics by which genes, and the proteins they encode for, evolve new functions. In this dissertation, I provide an example that shows how this broad research program is effectively pursued using a functional synthesis of computational and experimental techniques two fields generally considered very disparate. The complementary strengths and weaknesses of the two fields enable a deeper and richer understanding of molecular evolution from generating hypotheses about functional change in proteins to demonstrating causal effects [39].

This thesis aims to bring together the ideas, concepts and techniques from the two fields, to understand the mechanisms and dynamics by which proteins evolve new functions. I am interested in how novel protein functions evolve, both at the level of the underlying evolutionary forces that fix mutations, and the genetic and biophysical mechanisms by which specific historical mutations that fixed during evolution change protein functions. By pursuing both computational and experimental projects, the hope has been to demonstrate that a combination of the two approaches can provide a more complete picture of molecular evolution than either can in isolation.

1.2 Computational advances in phylogenetics

How do genes evolve new functions? One way is gene duplication followed by functional diversification, a mechanism that underlies many evolutionary innovations [45, 50, 62]. Inferring the timing of this duplication and diversification process is the first step toward understanding how novel functions evolve. Computational phylogenetics methods provide a powerful means to reconstruct the evolutionary history of genes and gene functions. Thanks to statistical developments in phylogenetics, estimation of a gene phylogeny from an alignment of nucleotide or amino-acid sequences is now possible with, parametric and non-parameter methods based on maximum-likelihood (ML), maximum parsimony and Bayesian approaches [60, 137, 81, 152]. Parsimony methods are simple and intuitive, but the lack of an explicit model makes it impossible to incorporate molecular understanding of sequence evolution [160].

Modern inferences of phylogenies are based almost entirely on ML or Bayesian approaches. Both methods share many statistical properties, although ML methods provide a powerful framework to define model parameters as fixed unknown constants, as opposed to Bayesian methods where model parameters are random variable parameters that depend on prior distributions [160]. Recent developments in ML based gene phylogeny inference include optimizations of likelihoods to enable a fast and efficient search across the space of tree topologies while ensuring an accurate reconstruction of the gene phylogeny [60, 61, 87, 101, 68, 89, 67, 84, 148, 42]. Standard ML softwares can easily estimate a gene phylogeny from 100-150 gene sequences with up to 10,000 character states on a desktop computer, and have formed the basis for thousands of phylogenetic inferences [60, 137, 61].

The development of computational methods for estimating gene phylogenies is a significant contribution to the field of evolutionary biology because it is the first step to infer the genetic basis of evolutionary change. The gene phylogeny provides an opportunity to map extant gene functions and construct hypotheses about the historical interval and mutations

underlying the genes functional change. Although the change in genes sequences and functions need not be adaptive, evolutionary biologists in the Panglossian and Post-Panglossian paradigm seek to provide adaptationst explanations for the form and function of molecules [12]. A central goal in evolutionary biology is to identify mutations that were adaptive for the switch in gene function [39, 63, 64].

1.3 Identifying adaptive mutations — Dangers of model violation

Recent decades have witnessed the development of a variety of sequence-based population genetic and phylogenetic approaches to identify adaptive mutations. Approaches based on codon-based sequence analysis, fixation indices, departures from demographic models, LD, extended haplotype heterozygosity, and associations between allele frequencies and environmental variables have been useful in identifying putative mutations associated with adaptation [103, 102, 104, 111, 157, 158, 35, 99, 162]. Computational models are at best an approximation of our understanding of the evolutionary process; so these inferences should be integrated with studies of functional or phenotypic differences between populations or species to establish causality. But this synthesis is indicated as future work and rarely done [52, 91, 122]. I argue here that such computational inferences are not decisive in isolation because these patterns can be forged by chance or the result of unincorporated model complexities. I provide several examples of unreasonable model assumptions routinely used in popular tests of adaptation that have the potential to cause misleading inferences.

Often simplifying assumptions are made about the evolutionary process in the models of adaptation, either for mathematical simplicity, or because of a lack of detailed molecular understanding of mechanisms of evolutional change. Violations of model assumptions by unincorporated forms of complexity remain alternate explanations consistent with adaptation. In codon-based phylogenetic tests of adaptation, a common approach to identifying sites under positive selection is to look for adaptations on certain branches of a phylogeny (branch

test), or certain sites across lineages (sites test), or on a small fraction of sites on one lineage such as the branch-site test [160, 111, 157, 158, 159]. Positive selection is modeled with the parameter , which represents the ratio of the instantaneous rates of non-synonymous (dN) and synonymous substitutions (dS) ($\omega = dN/dS$). A non-synonymous substitution changes the underlying amino-acid while a synonymous substitution does not. A dN significantly greater from dS is therefore considered evidence for lineage-specific protein adaptation. In the null model, ω is constrained to values ≤ 1. In the positive selection model $\omega > 1$, so this model will always have a likelihood that is greater or equal to that of the null model.

One unrealistic assumption that underlies all codon-based tests surrounds synonymous changes that they have neutral effects on function and fitness. No heterogeneity in dS is allowed [103, 102, 104, 3, 55, 57, 83, 134] Several studies have shown the effect of synonymous substitutions on the function of molecules: for translation accuracy, folding, or transcription factor binding, but little efforts have been dedicated to incorporation of these effects in sequence models [46, 114, 143]. Selective constraints on synonymous substitutions could therefore vary across codons. Even variation in recombination or mutation rates could cause dS to vary, and the overall rate of dN/dS will vary for reasons other than adaptation. These forms of heterogeneity are real in biological datasets, and not accounting for them can cause misleading inferences of adaptation [41, 110, 129, 156].

It is also assumed that dN does not depend on the amino acids being exchanged, and that the equilibrium frequency of codons is identical among sites and does not depend on the amino acid the codon produces. If these assumptions are violated, neither the null or positive selection models will be an accurate description of the evolutionary process. The positive selection model could fit the data better than the null model to incorporate the unmodeled complexity, leading to an inference of adaptation.

An additional important form of model violation comes from unincorporated complexity in the mutational process. Mutations lead to genetic variation, providing the raw material

for evolution. Modeling mutations is the first step in describing evolutionary change, but incorporation of mechanistic knowledge of mutations into phylogenetic tests of adaptation is still at its infancy. It has been routinely assumed that most mutations arise independently at individual sites due to errors in the DNA replication process, and therefore substitutions are independently and identically distributed as a Poisson process [125]. But mechanistic work on mutations shows that multi-nucleotide mutations (MNMs), clusters of closely-spaced mutations, can arise from neutral processes such as the action of error-prone polymerase slippage. Ample molecular studies on mutation demonstrate the unreasonable nature of the independence assumption [110, 66, 93, 107, 147, 166]. I discuss below how this assumption can cause misleading inferences of adaptation in phylogenetic tests, with additional implications for population genetic tests.

An excess of clustered mutations, such as those within a single codon, will be very rarely expected under the assumption of independence. This is very problematic for the branch-site test for example, which seeks to infer adaptation on a few codons on a single lineage. The signature of multiple mutations in MNM codons can provide strong evidence for adaptation because such data will be readily explained by selectionist interpretations such as rapid emergence of separate mutations, compensatory evolution, mutational hotspots, or repeated fixation of adaptive alleles. But if MNM arise from neutral processes, inferring positive selection is an incorrect inference of adaptation due to model violation from unincorporated mutation complexity.

Population-genetic tests of adaptation, such as the McDonald-Kreitman, could also be similarly affected [97]. The test infers adaptation by comparing the relative number of non-synonymous and synonymous substitutions in species divergence data to the relative number of non-synonymous and synonymous polymorphisms in one or a few species. An excess of non-synonymous divergence compared to the expectation from the two types of intra-population polymorphisms is considered evidence for adaptation, which MNMs could

readily provide. Closely spaced mutations between two haplotypes can also cause the age of alleles to be overestimated leading to false interpretations of ancient balancing selection [66].

Population and phylogenetic tests of adaptation have been applied to many thousands of genes and gene families and have provided the foundation for a great number of inferences of positive selection. Simulation studies have shown that the tests are powerful and accurate when the model used for analyzing sequences was the same one used to simulate the data [158, 164, 167]. Limited simulations have been performed under conditions of model violations, calling for thorough investigations of the performance of a wide-variety of tests, under a variety of evolutionary conditions [166, 165] . In Chapter 1 of this thesis, I investigate the effect of one such violation in detail the effect of unincorporated mutation complexity on inferences of adaptation made by the branch-site test. Careful computational studies are required to assess robustness to model violation and their inferences should be followed upon by experimental studies to test hypotheses about sequence causes.

1.4 Integration of experimental and computational approaches

To fully address how historical mutations change protein functions, computational inferences should be further experimentally tested. When combined with molecular biology and protein biochemistry, hypotheses that emerge from computational studies about function-changing mutations in genes can be explicitly tested, making for a deeper and richer inference of the evolutionary process. Molecular biology is a reductive science with high priority placed on establishing causality. Therefore, the effect of a mutation on gene or protein function can be explicitly tested and the inferences of adaptation or function-change from statistical analyses further corroborated or refuted [39, 63, 64]. Using this synthesis, long-standing questions in evolutionary biology pertaining to the genetic and physical basis of gene evolution can be answered such as: i) How many mutations underlie a switch in gene function? 2) Does

evolution proceed by few mutations of large effect, or many mutations of small effect? 3) How does complexity arise in molecular systems? 4) Do new functions evolve through radical de novo changes in gene and protein organization, or through tinkering of ancestral properties? Additional questions pertaining to the diversity in present-day protein properties and the genetics that enabled functional diversity in proteins can be addressed with the combination of evolutionary biology, molecular biology and protein biochemistry.

Evolutionary biochemistry fuses evolutionary biology, phylogenetics, molecular biology and protein biochemistry to understand how protein architecture constrains evolution, and how evolution constrains protein architecture and accessible functions [64]. Biochemistry seeks to understand the physical and chemical properties and functions of proteins in isolation. Evolutionary biology, in turn, seeks to traverse through functions over evolutionary time scales. Many open questions in protein evolution can therefore be formulated at the interface of evolution and biochemistry, such as those concerning the evolution of allostery, DNA specificity, cooperativity, or protein folding [64]. With a purely biochemical approach, these questions can be difficult to answer because there could be a vast space of protein sequences that would need to be functionally characterized to learn about the map that connects protein sequence to its function.

The knowledge gap in understanding how the proteins physical properties determine their functions persists because most biochemical studies of proteins have ignored the evolutionary history of the protein. However, by focusing on the extant diversity in protein functions, and considering an evolutionary approach, we can characterize how specific amino acid changes that occurred during evolution changed protein functions. Since evolution has already conducted this experiment over billions of years, the problem of understanding how the map of protein sequence and function are connected, and how they evolve is easier, as focusing on historical substitutions that already occurred makes the goal tractable. This fusion, of the techniques from phylogenetics, molecular biology, and protein biochemistry

form the core of ancestral sequence resurrection (ASR), an approach that traces in detail evolutionary changes in proteins sequences, structures, and biochemical, characteristics to shed light on the molecular mechanisms that underlie the evolution of new protein functions [63, 64, 69, 98, 139].

ASR reconstructs the historical trajectory of a family of gene or protein sequences using phylogenetic techniques [63, 64]. Starting from an alignment of gene sequences for recent divergences, or protein sequences for more ancient divergences, a phylogenetic tree is inferred using ML or Bayesian approaches. Statistical techniques based on ML are then used to infer ancestral sequences at the interior nodes of the tree. These computationally inferred sequences are then physically synthesized, and experimentally characterized using in vitro or in vivo assays. This approach allows an explicit functional characterization of ancestral and derived genotypes on a phylogeny, using which the historical interval underlying a function change can be inferred. The effect of sequence substitutions that occurred on the interval can also be experimentally tested, either singly or in combination to infer the causal effects of substitutions on the function of the protein, including epistatic interactions. Ambiguity in reconstructions can also be tested by introducing ambiguous states onto the most-likely reconstructed sequence, and their effects on functions tested for robustness of inference. Overall, ASR has proved extremely useful to understand why many living systems are organized how they are [39, 63, 64, 69, 32, 59, 65, 74, 123, 130].

Despite its benefits, ASR cannot be readily applied to all proteins. ASR requires a confident alignment with good phylogenetic signal, which some rapidly evolving proteins such as intrinsically disordered families, may not provide [69]. Even if a good alignment can be obtained, there are dangers of reconstructing incompatible amino-acid states due to model violations resulting from assumption of independence among sites. Unincorporation of epistasis in the substitution models could be a bigger issue for reconstructions done using Bayesian approaches that tend to incorporate low probability ancestral states onto the

most plausible sequence as opposed to maximum-likelihood reconstructions. Reconstruction of incompatible states for pairs of amino acids that covary or epistatsically interact could have detrimental effects on function; nevetheless an alignment with good phylogenetic signal should enable correct reconstructions of these couplings, as this information should be encoded in the patterns of conservation of individual states. Further, mutation complexity or substitution rate heterogeneity is not incorporated in the ASR substitution models; For example, MNMs from gene alignments could be incorrectly constructed. These considerations suggest that there is considerable room for improvement in the development of ASR, including the development of algorithms to improve multiple sequence alignments [27, 155]. Despite the limitations, the benefits of ASR the integration of phylogenetic reconstructions with functional characterization of proteins, still has a lot of potential to illuminate the causes of structure and function in underexplored protein families that can be studied successfully using ASR.

Recent developments in ASR have shed light on the molecular basis of specificity between proteins and their interacting partners in a range of biological systems [98, 65, 1, 9, 40, 71, 72, 115, 154]. These studies have shown that specificities of related proteins can not only evolve by partitioning or enhancing ancestral functions, but also de novo from ancestral proteins that lacked those functions [71, 72, 115, 154, 19, 108, 112, 141]. However, limited studies have revealed the genetic and biochemical mechanisms by which a switch in specificity evolves in proteins [98, 65, 1, 2, 18, 48]. Chapter 2 of this thesis describes some insights into the genetic and thermodynamic mechanisms by which transcription factors evolve new DNA preferences. Proper functioning of gene regulatory networks depends on the distinct DNA specificities of transcription factors, but the molecular mechanisms by which DNA specificities evolve in transcription factors have been poorly understood.

Transcription factor-DNA specificity depends on different physical modes of DNA recognition, often involving core DNA motifs and flanking sequences. At the mechanistic level,

base-specific affinity is largely derived from direct physical interactions between protein side-chains and base-specific functional groups in core DNA motif, often in the major groove [109, 118, 142]. Flanking sequences outside the core motif can also contribute to DNA affinity; it is now being increasingly recognized that paralogous transcription factors can recognize the same core DNA motif, yet regulate distinct genes depending on the composition of flanking sequences [28, 58, 131, 78, 7]. These preferences have been shown to be especially important in vivo for development [118, 28, 94, 22, 105, 121]. Although the diversity in the preferences of paralogous proteins for core motifs and flanking sequences has been recognized, virtually nothing is known about how the affinities for core motifs and flanking sequences are gained or lost during protein evolution, including its genetic basis. What is the energetic contribution of these interactions, and how do they evolve? In particular, how do the enthalpic and entropic components of binding to the core half-site and flanking sequences evolve to produce proteins with new thermodynamic properties? Additionally how does cooperative binding of proteins modulate this affinity? [133, 136, 140, 79, 163, 54, 82]. Very little is known about the individual contribution of the single-site affinity and cooperativity to the overall DNA affinity, or the interaction and evolution of these energetic parameters of binding to generate new transcription factor-DNA specificities [98]. Can novel DNA specificity evolve through a change in either affinity or cooperativity alone, or do both evolve concomitantly? And finally, does novel DNA specificity only evolve through drastic shifts in protein-DNA thermodynamics, or do subtle changes in affinity and cooperativity result in large changes in response element preferences? Chapter 2 of this thesis unravels some answers to these questions, with a specific focus on the evolution of flanking sequence preferences in transcription factors.

1.5 Summary of thesis

In Chapter one, I describe a computational project that explores the consequences of unrealistic model assumptions made by widely-used tests of adaptation for inferences of lineage-specific positive selection. With a focus on multi-nucleotide mutations and molecular mechanism, the project investigates how a classic phylogenetic test of adaptation, the branch-site test, makes frequent false inferences of lineage-specific positive selection due to the nature of its oversimplified codon substitution model. The test has been the basis for thousands of inferences of positive selection in literature. I showed that the majority of genes claimed to be under positive selection on the human and fly lineages are artifacts of model violation resulting from unincorporated neutral mutational processes, and suggest that many published inferences of positive selection on protein-coding genes in other species would similarly be artifacts. This chapter includes unpublished co-authored work with my advisor Joseph W. Thornton, and collaborator Matthew Hahn, now posted on bioarxiv - https://www.biorxiv.org/content/early/2017/07/20/165969.1

In Chapter two, I describe an experimental project to investigate the genetic and thermodynamic mechanisms by which mutations that fixed during protein evolution changed protein functions. I focused on transcription factor-DNA specificity, with an emphasis on the evolution of flanking sequence preferences in proteins, an understudied research topic. I used the steroid and related receptor family of transcription factors (SRRs) as a model system. SRRs comprise of two functionally diverged clades: 1) Steroid Receptors (SRs), which bind as a cooperative dimer to an inverted palindrome of a 6-bp half-site AGGTCA; and 2) Estrogen Related Receptors (ERRs), which bind as monomers to an extended 9-bp half-site (TCAAGGCTA), containing a 5'-flanking extension of the 6-bp SR half-site. The great diversity in flanking sequence specificities together with cooperative binding makes the SRR family a great model system to explore questions pertaining to the evolution of novel DNA-specificities, cooperativity, and the relationship between these thermodynamic param-

eters of binding. Using integrative approaches, I showed that the preference for a specific flanking sequence was lost on the SR lineage from an ERR-like ancestor, a transition that was accomplished in six mutations through a change in half-site affinity alone, with little to no changes to cooperative binding. This study provides a mechanistic explanation for how flanking sequence preferences are lost in proteins, and illustrates that new DNA specificity can evolve by subtly shifting the thermodynamic parameters of binding. This chapter describes unpublished work.

CHAPTER 2

MULTINUCLEOTIDE MUTATIONS CAUSE FALSE INFERENCES OF LINEAGE-SPECIFIC POSITIVE SELECTION

2.1 Abstract[1]

Phylogenetic tests of adaptive evolution, which infer positive selection from an excess of nonsynonymous changes, assume that nucleotide substitutions occur singly and independently. But recent research has shown that multiple errors at adjacent sites often occur in single events during DNA replication. These multinucleotide mutations (MNMs) are overwhelmingly likely to be nonsynonymous. We therefore evaluated whether phylogenetic tests of adaptive evolution, such as the widely used branch-site test, might misinterpret sequence patterns produced by MNMs as false support for positive selection. We explored two genome-wide datasets comprising thousands of coding alignments — one from mammals and one from flies — and found that codons with multiple differences (CMDs) account for virtually all the support for positive selection inferred by the branch-site test. Simulations under genome-wide, empirically derived conditions without positive selection show that realistic rates of MNMs cause a strong and systematic bias in the branch-site and related tests; the bias is sufficient to produce false positive inferences approximately as often as the branch-site test infers positive selection from the empirical data. Our analysis indicates that genes may often be inferred to be under positive selection simply because they stochastically accumulated one or a few MNMs. Because these tests provide no reliable means to distinguish sequence patterns produced by authentic positive selection from those caused by neutral fixation of MNMs, many published inferences of adaptive evolution using these techniques may therefore be artifacts of model violation caused by unincorporated neutral mutational

1. Citation for chapter: https://www.biorxiv.org/content/early/2017/07/20/165969.1

processes. We develop an alternative model that incorporates MNMs and may be helpfully in reducing this bias.

2.2 Introduction

Identifying genes that evolved under the influence of positive natural selection is a central goal in molecular evolutionary biology. During recent decades, likelihood-based phylogenetic methods have been developed to identify gene sequences that retain putative signatures of past positive selection [103, 102, 104, 111, 159, 57, 134, 167, 85, 151]. Perhaps the most widely used of these is the branch-site test (BST) of episodic selection, which allows positive selection to affect only some codons on one or a few specified branches of a phylogeny, and therefore has relatively high power compared to methods that detect selection across an entire sequence or an entire phylogenetic tree [158, 159, 167]. The BST has been the basis for published claims of lineage-specific adaptive evolution in many thousands of individual genes [45, 50, 62, 86, 120].

The BST and related methods use a likelihood ratio test to compare how well two mixture models of sequence evolution on a phylogeny fit an alignment of coding sequence data. The null model constrains all codons to evolve with rates of nonsynonymous substitution (dN) less than or equal to the rate of synonymous substitution (dS), as expected under purifying selection and drift. In the positive selection model, some sites are allowed to have dN>dS on a branch or branches of interest. If the increase in likelihood of this model given the data is greater than expected due to chance alone, the null model is rejected and adaptive evolution is inferred. The BST has been shown to be conservative, with a low rate of false positive inferences, when sequences are generated under an evolutionary process corresponding to the null model [158, 167] . It is widely appreciated that likelihood ratio tests can become biased if the underlying probabilistic model is incorrect [165]. The effect on the BST of a few forms of model violation — such as an unequal distribution of selective effects among sites,

positive selection on non-foreground lineages, high sequence divergence, and non-allelic gene conversion — have been previously studied [3, 83, 107, 166, 24] and the test has been found to be reasonably robust to most, but not all forms of violation examined [55, 164, 167].

Recent research in molecular genetics and genomics suggests a potentially important phenomenon that has not been incorporated into models used in tests of positive selection: the propensity of DNA polymerases to produce mutations at neighboring sites. All implementations of the BST and other likelihood-based tests of adaptive evolution use models in which mutations occur only at individual nucleotide sites and are fixed singly and independently. Codons with multiple differences between them can be interconverted only by serial single-nucleotide substitutions, the probability of which is the product of the probabilities of each independent event. Recent molecular studies have shown, however, that mutations affecting adjacent nucleotide sites often occur during replication, apparently because certain DNA microstructures recruit error-prone polymerases that lack proofreading activity and therefore make multiple errors close together [125, 124, 92, 96, 128, 144, 4, 16, 30]. Consistent with these mechanisms, genetic studies of human trios and mutation-accumulation experiments in laboratory organisms indicate that de novo mutations occur in tandem or at nearby sites more frequently than expected if each occurred independently [125, 66, 16, 30, 29, 148], and these multinucleotide mutations (MNMs) are enriched in transversions [66, 51, 168]. The precise frequency at which MNMs occur is difficult to estimate, but a recent compilation of genetic studies in humans concluded that about 0.4% of mutations, polymorphisms, and substitutions are at directly adjacent sites (counting each tandem pair as one event) [29]. In Drosophila melanogaster genomes, analysis of rare polymorphisms and mutation-accumulation experiments estimated that 1.3% of all mutations are at adjacent sites [5]. Although the methods and data sources in these studies differ, these findings suggest that tandem MNMs probably account for on the order of 1% of mutational events.

We hypothesized that these mutational processes might lead to false signatures of positive

selection in the BST. Because of the structure of the genetic code, virtually all MNMs in coding sequences are nonsynonymous, and most would comprise multiple nonsynonymous nucleotide changes if they were to occur by single nucleotide steps (Supplementary Table 2.1). The enrichment of transversions in MNMs further increases the propensity for MNMs to produce nonsynonymous changes, because transversions are more likely than transitions to be nonsynonymous. MNMs are therefore likely to produce codons with multiple differences (CMDs) that contain an apparent excess of nonsynonymous substitutions. When these CMDs are assessed using a method that treats all substitutions as independent events, a model that allows dN to exceed dS at some sites may have a higher likelihood than one that restricts dN/dS to values ≤ 1. Further, the assumption that all mutations have the same transversion-transition rate might exacerbate the tendency to misinterpret MNM-produced nonsynonymous changes as evidence for positive selection. Of course, CMDs can also be driven to fixation by positive selection [158, 6, 13, 117] — and the same is true of transversion-rich substitutions — but these considerations suggest that failing to incorporate MNMs in likelihood models might make tests of adaptive evolution susceptible to false positive inferences. The BST and related tests might be particularly sensitive to this problem because they seek signatures of positive selection acting on small numbers of codons on one or a few specified branches of the tree [147]. Simulation studies suggest that MNMs may elevate false positive rates in some selection tests [38], but there has been no comprehensive analysis of the effect of MNMs, particularly on the branch-site test or under realistic, genome-scale conditions.

2.3 Results

To understand the effect of MNMs on the accuracy of the branch-site and related tests of adaptive evolution, we analyzed in detail two previously published genome-wide datasets, which represent classic examples of the application of these tests [45, 86, 88]. The mammalian

dataset consists of coding sequences of 16,541 genes from six eutherian mammals; we retained for analysis only the 6,868 genes with complete species coverage. The fly dataset consists of 8,564 genes from six species in the melanogaster subgroup clade, all of which had complete coverage (Supplementary Fig. 2.10). The fly genes have higher sequence divergence than those in the mammalian dataset, allowing us to examine the performance of the BST under different evolutionary conditions.

We used the classic BST to identify genes putatively under positive selection ($P < 0.05$) on the human lineage in the mammalian dataset and on all six terminal lineages in flies. 82 genes in humans and 3,938 in flies yielded significant tests (Supplementary Table 2.2). To facilitate further analysis of CMDs, we filtered CMDs in gaps, and imposed a quality control filter that kept only those genes in which all CMDs on the branch of interest were reconstructed identically between null and positive selection models; we also applied a multiple testing correction (FDR < 0.20). In flies, 443 genes were retained after these steps. Thirty human genes passed the reconstruction filter, but none met the FDR threshold, consistent with previous analyses of these data [86] ; nevertheless, we included the 30 initially significant human genes because this lineage is the object of intense interest and because its short length contrasts with the fly branches. These two groups constitute the "BST-significant" sets of genes in flies and humans.

2.3.1 CMDs provide virtually all support for positive selection

We sought to determine how much of the evidence for positive selection comes from CMDs. We first observed that CMDs were dramatically enriched in BST-significant genes compared to non-BST-significant genes (Fig. 2.1a). In humans, BST-significant genes contain one CMD on average, while BST-nonsignificant genes contain none (Supplementary Fig. 2.11). The pattern is similar but less extreme in flies, with the average number of CMDs per BST-significant gene greater than that in non-significant genes (Supplementary Fig. 2.11). When

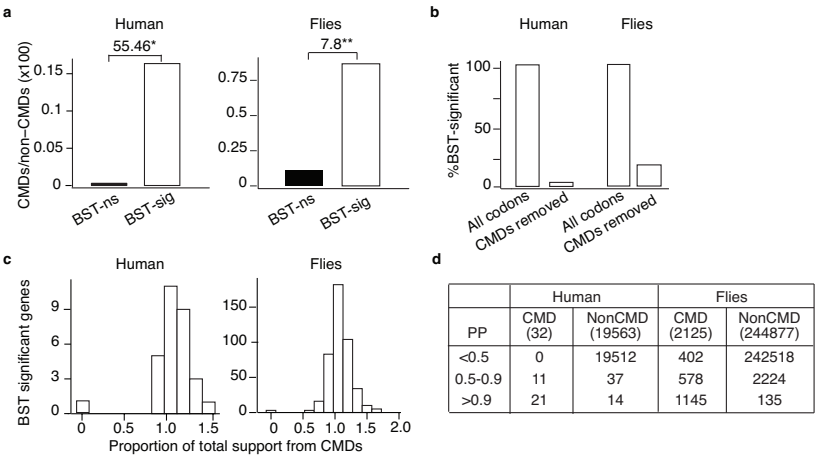
CMD-containing codons are excluded from the alignments, the vast majority of genes that were BST-significant lose their signature of selection in both datasets (Fig. 2.1b).

We next calculated the fraction of statistical support for positive selection that comes from CMDs. The total support for positive selection in an alignment is defined as the difference between the log-likelihood of the positive selection model and that of the null model, summed across all codons in the alignment. The fraction of support from CMDs is the support from CMD-containing codons divided by the total support across the entire alignment. CMDs account for $> 95\%$ of the support for positive selection in virtually all BST-significant genes in both datasets; in about 70% of genes, CMDs provide all the support (Fig. 2.1c).

Finally, we examined the BST's a posteriori identification of sites under positive selection. We found that CMDs were far more likely to be classified as positively selected than non-CMDs. Among genes that were BST-significant on the human lineage, every CMD was inferred to be under positive selection using a Bayes Empirical Bayes posterior probability (PP) cutoff > 0.5. Using a more stringent cutoff of PP> 0.9, 66 percent of CMDs were classified as positively selected, compared to 0.07% of non-CMDs. In the fly dataset, CMDs accounted for 90% of codons with BEB > 0.9, although they represent less than 1% of all codons (Fig. 2.1d). CMDs are therefore the primary drivers of the signature of selection identified in the BST. A single CMD provides sufficient statistical support to yield a signature of positive selection on the human lineage, and only a few CMDs in a gene are enough to do the same in flies.

2.3.2 Incorporating MNMs eliminates the signature of positive selection in many genes

CMDs might be enriched in BST-positive genes because of an MNM-induced bias or because they were fixed by positive selection. To incorporate MNMs into a BST framework, we



	Human		Flies	
PP	CMD (32)	NonCMD (19563)	CMD (2125)	NonCMD (244877)
<0.5	0	19512	402	242518
0.5-0.9	11	37	578	2224
>0.9	21	14	1145	135

Figure 2.1: **Codons with multiple nucleotide differences (CMDs) drive branch-site signatures of selection**. (a) CMDs are enriched in genes with a signature of positive selection. Codons were classified by the number of nucleotide differences between the ancestral and terminal states on branches tested for positive selection. CMDs have ≥ 2 differences; non-CMDs have ≤ 1 difference. The CMD/non-CMD ratio is shown for genes with a significant signature of selection in the classic BST (BST-sig) and those without (BST-ns). Fold-enrichment is shown as the odds ratio. *, $P = 4e-4$ by χ^2 test; **, $P = 1e-41$ by 'Fisher's' exact test. (b) Percentage of genes that retain a signature of positive selection when CMDs are excluded from the branch-sites test analysis. (c) Distribution across BST-significant genes of the proportion of total support for the positive selection model that is provided by CMDs. Total support is the difference in log-likelihood between the positive selection and null models, summed over all codons in the alignment. Support from CMDs is summed over codons with multiple differences. The proportion of support from CMDs can be greater than 1 if the log-likelihood difference between models is negative at non-CMDs. (d) Most codons classified as positively selected are CMDs. The number of CMDs and non-CMDS in BST-significant genes are shown according to their Bayes Empirical Bayes posterior probability (PP) of being in the positively selected class.

developed a codon model in which double-nucleotide changes are allowed, with the parameter δ serving as a multiplier that modifies the rate of each double-nucleotide substitution relative to single-nucleotide substitutions. We implemented a version of the BST (BS+MNM) that is identical to the classic version, except that both the null and positive selection models allow MNMs. Simulations under conditions derived from a sample of genes in the mammalian dataset show that the method estimates the parameters used to generate the sequences with reasonable accuracy (Supplementary Fig. 2.12).

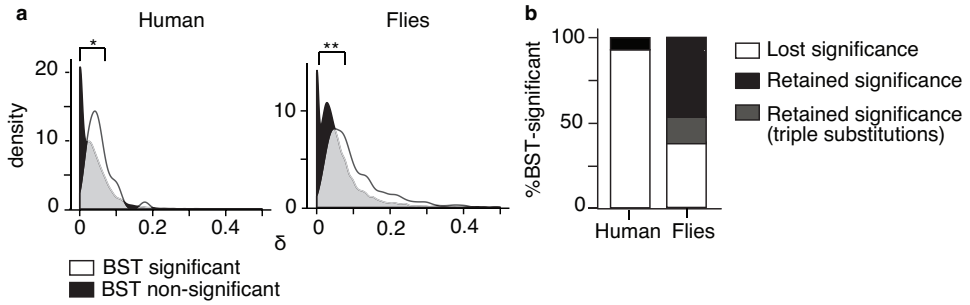


Figure 2.2: **Incorporating MNMs into the branch-sites model eliminates the signature of positive selection in many genes**. The mammalian and fly datasets were reanalyzed using a version of the BST that allows MNMs (BS+MNM) by including a parameter δ, the rate of double substitutions relative to single substitutions. (a) The distribution of ML estimates of δ across genes with (white) and without (black) a significant result in the classic BST is shown for empirical alignments. Median estimates of δ for BST-significant and BST-nonsignificant genes are 0.047 and 0.026 in humans, respectively, and 0.107 and 0.062 in flies. *, P=6.7e-4; **, P=1e-8 by Mann-Whitney U Test. (b) Proportion of genes with a significant result in the BST that lose or retain that signature using the BS+MNM test. Genes that remain significant but contain CMDs with three differences, which are not incorporated into BS+MNM, are also shown.

We first fit the BS+MNM null model to all alignments in the mammalian and fly datasets. The average estimate of δ across all genes was 0.026 in mammals and 0.062 in flies, with δ in both cases about twice as high in the subset of BST-significant genes as in BST-nonsignificant genes (Fig. 2.2a). Using a likelihood-ratio test, we found significant support for the BS+MNM null model (compared to the classic BST null model) in 22% of human genes and $> 50\%$ of fly genes (Supplementary Table 2.3); simulations without MNMs showed that this comparison has a very low false-positive rate (Supplementary Table 2.4).

We then used this BS+MNM test to evaluate the empirical sequences for positive selection. We found that 96% of the BST-significant genes on the human lineage lost significance in the BS+MNM test (Figs. 2.2b, Supplementary Table 2.5). In flies, 38% of the BST-significant genes lost significance; a substantial fraction of those that retained significance were enriched in triple substitutions, a process not accounted for in our model (Figs. 2.2b, Supplementary Table 2.5).

2.3.3 MNMs cause false positive inferences on a genome-wide scale

That the BS+MNM test eliminates the signature of positive selection from many genes could arise from several causes, including: 1) 2) the more complex BS+MNM model may have reduced power to identify authentic positive selection compared to the BST, 2) incorporating MNMs may ameliorate a bias towards false positive inference in the classic BST that is caused by MNMs, and 3) the additional δ parameter in the BS+MNM test may allow it to incorporate other forms of sequence complexity, potentially ameliorating a bias caused by other forms of model violation.

We addressed these possibilities in two ways. First, we performed power analyses of the BS+MNM test using simulations in which positive selection is present in the generating model. We simulated sequence data on the mammalian and fly phylogenies using genome-wide averages for all parameters of the BST positive selection model, but we varied the strength of positive selection (ω_2) and the proportion of sites under positive selection. We then applied the BS+MNM test to these data and found that it can reliably detect strong positive selection ($\omega_2 > 20$) when it affects more than 10% of sites in a typical gene, or moderate positive selection ($10 < \omega_2 < 20$) that affects a larger fraction of sites (Supplementary Fig. 2.13a). The test's power is similar to that of the classic BST, with slightly reductions under only a few conditions on the fly lineage (Supplementary Figs. 2.13a—c). Thus, although some genes may have lost their signature of selection because of reduced power in the BS+MNM test, it appears unlikely that a difference in power is the primary cause of the dramatic reduction in the number of positive results when the test is used.

Second, we used simulations under null conditions to directly evaluate the frequency of false positive inferences by the classic BST when sequences are generated with realistic rates of multinucleotide mutation. For every gene in the mammalian and fly datasets, we simulated sequence evolution under the null BS+MNM model without positive selection using parameters derived from the alignments, including δ. These parameters generate

sequences with an observed frequency of tandem substitutions of 1.6% in humans and 3.2% in the *D. melanogaster* lineage in flies, approximately in the same range as observed in other studies and slightly higher than the observed frequencies in the empirical sequences (1.3% and 1.6%, respectively), presumably because the BS+MNM model captures some but not all aspects of real sequence evolution (Supplementary Table 2.6) [29, 5].

We then analyzed these positive-selection-free simulated data using the classic BST. In both humans and flies, the number of genes with significant results — all of which are false positive inferences — was greater than the number of genes that the BST had concluded were under positive selection using the empirical data (Fig. 2.3a). In flies, almost 9 percent of tests were false positives $P < 0.05$, despite the conservative approach the method uses to calculate P-values [158, 167], compared to just 1 percent under control simulations without MNMs. Further, more than 1,700 of these false positive tests survived FDR adjustment, compared to just four in the control simulations (Supplementary Table 2.2). In humans, the fraction of false positive inferences is lower, consistent with the test's reduced power in this dataset, but still about three times greater than in the control simulations.

These false inferences are caused primarily by MNM-induced bias, because simulating data under identical control conditions without MNMs ($\delta = 0$) produced few positive tests. All other parameters were identical between the generating model and analysis models, so other forms of model violation do not contribute to the bias observed in the simulation experiments. Taken together, these findings indicate that MNMs under realistic evolutionary conditions produce a strong and widespread bias in the BST toward false inferences of positive selection. This bias is strong enough to cause the BST to make false inferences of positive selection at about the same rate as it infers selection in the real genomes of humans and flies. In the simulations, every positive result is false; in the tests of real sequences, the fraction is unknown.

2.3.4 Systematic bias caused by chance MNMs in longer genes

We next sought to identify the causal factors that determine whether a gene yields a false positive result in the BST because of MNM-induced bias. Most genes are only several hundred codons long, and only a few percent of mutations are MNMs, so on phylogenetic branches of short to moderate length many genes will contain no CMDs caused by multinucleotide mutations. We therefore hypothesized that the propensity for a gene to produce a BST-significant result will depend on factors that increase the probability it will contain one or more fixed MNMs by chance, including its length and the gene-specific rate at which MNMs occur within it.

We first tested for an effect of gene length on the results of the branch-site test. As predicted, we observed that BST-significant empirical genes were on average 100 and 16 codons longer than non-significant genes in the human and fly empirical datasets, respectively (Fig. 2.3b). The relationship between length and propensity to yield a BST positive result could arise because genes that present a larger "target" are more likely to undergo MNMs than shorter genes; alternatively, longer genes, by including more sites for analysis, might increase the power of the BST to detect authentic positive selection. However, in genome-wide simulations under the null model with no positive selection (but with $\delta > 0$), genes with false positive BSTs are longer than the non-significant genes by an average of 26 and 31 codons using the human and fly parameters, respectively (Supplementary Fig. 2.14). This finding cannot be attributed to increased power to detect true positive selection and supports the conclusion that mutational target size is a determinant of a gene's propensity to manifest MNM-induced bias by chance alone.

To directly test the causal relationship between sequence length and false-positive bias in the BST, we simulated sequence evolution at increasing sequence lengths, using evolutionary parameters derived from each of the BST-significant genes in the mammalian and fly datasets. For each gene's parameters, we simulated 50 replicate alignments under the

BS+MNM null model and then analyzed them using the classic BST (Supplementary Fig. 2.15a). The false positive rate for any gene's simulations is defined as the fraction of replicates with a significant LRT in the classic BST, using a P-value cutoff of 0.05. When sequences 5,000 codons long were simulated, 96% of BST-significant genes in the mammalian dataset yielded an FPR greater than 0.05, with a median FPR across genes of 0.39; simulating sequences 10,000 codons long increased this fraction to 100% and the median FPR to 0.56 (Fig. 2.3c). In flies, 99% of genes had FPR¿0.05 (median FPR 0.74) when genes 5,000 codons long were analyzed, which increased to 100% of genes (median FPR 0.90) at sequence length of 10,000 codons (Fig. 2.3c). Control simulations under identical conditions but with $\delta = 0$ led to very low FPRs (median 0.02 to 0.03 for both datasets), even with very long sequences (grey dots in Fig. 2.3c). A similar systematic and length-dependent bias also resulted when sequences were simulated under gene-specific conditions, but with δ fixed to its average across the thousands of BST-nonsignificant genes in each dataset (Supplementary Fig. 2.15b). Although the sequence lengths tested are longer than most real genes, these experiments directly establish that a gene's probability of returning a significant BST result in the absence of positive selection is directly related to the target size it presents for chance fixation of MNMs.

We next evaluated whether the gene-specific rate of multinucleotide mutation affects a gene's propensity to yield a positive result in the BST. As predicted, we observed that BST-significant genes in the empirical datasets had higher estimated δ than nonsignificant genes (Fig. 2.2a). Genes producing false positive results in the genome-wide null simulations under empirical conditions also tended to have higher δ (Fig 2.3d); this result that cannot be attributed to the possibility that a higher δ might be the result of the model fitting an excess of CMDs caused by positive selection, because positive selection was absent from the generating model.

To directly test the effect of the neutral MNM substitution rate on the BST, we simulated

sequences 5,000 codons long under the null BS+MNM model, with a variable δ and all other parameters fixed to their averages across all genes. We found that increasing δ led to a monotonic increase in the frequency of false positive inferences. The FPR was > 0.05 when δ was only 0.001 and 0.013 on the human and fly lineages, respectively. When δ was equal to its genome-wide average (0.026 and 0.062 in mammals and flies), false positive inferences occurred at rates of 22 and 17 percent, respectively (Fig. 2.3e). As δ increased, so too did the inferred value of the parameter ω_2, which represents the inferred intensity of positive selection in the model (Fig. 2.3f).

Typical evolutionary conditions are therefore sufficient to cause a strong and systemic bias in the BST. MNMs are rare, however, so longer genes and those with higher rates of multinucleotide mutation are more likely to undergo this process and manifest the bias. This view is further supported by the fact that fewer genes are BST-positive on the human branch — which is so short that substitutions of any type are rare, and MNMs even more so — than on the fly phylogeny, where branches are longer, more CMDs are apparent, and hundreds of genes have BST signatures of selection. Taken together, these findings suggest that many genes with BST-significant results in empirical datasets may simply be those that happened to fix multinucleotide substitutions by chance alone.

2.3.5 Transversion-enrichment in CMDs exacerbates bias in the branch-site test

MNMs tend to produce more transversions than classical single-site mutational processes, so if CMDs are produced by MNMs, they should be transversion-rich [66, 51, 168]. As predicted, the transversion:transition ratio is elevated in CMDs relative to that in non-CMDs by factors of three and two in mammals and flies, respectively (Fig. 2.4a). In the subset of BST-significant genes, CMDs have an even more elevated transversion:transition ratio, as expected if transversion-rich MNMs bias the test (Fig. 2.4a). These data are consistent

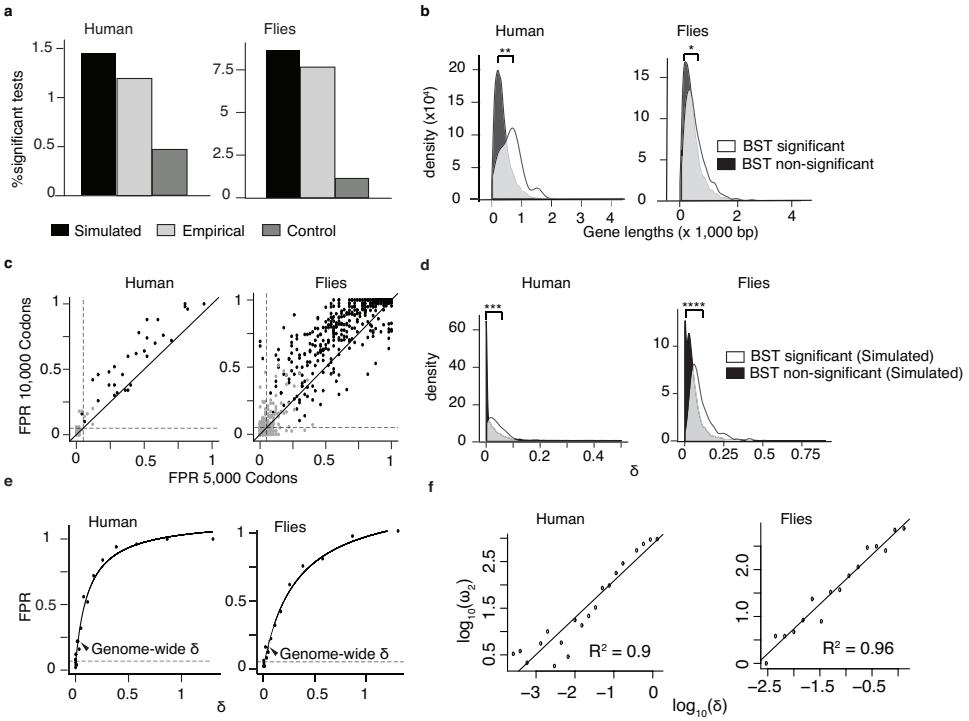


Figure 2.3: **MNMs cause a strong bias in the branch-site test under realistic conditions** For each gene in the mammalian and fly datasets, the parameters of the BS+MNM null model were estimated by maximum likelihood. We then simulated sequence evolution under each gene's inferred null parameters and used the classic BS test on the simulated alignments to test for positive selection on the human and terminal fly lineages. (a) The fraction of all tests that are BST-significant ($P < 0.05$) is shown for the data simulated under the BS+MNM null model, the original empirical sequence alignments, and a control dataset simulated with $\delta = 0$. Each gene's length in the simulation was identical to its empirical length. (b) BST-significant genes are longer than BS non-significant genes. The probability density of gene lengths in the two categories is shown for the empirical mammalian and fly datasets. Median lengths in BST-significant and non-significant genes, respectively, were 642 and 343 bp in humans; in flies, 448 and 399 bp. The difference between the two distributions was evaluated using a Mann-Whitney U test. *, P=8e-4; **, P=8e-5. (c) Systematic bias in the BST. For each gene with a significant result in the branch-site test using the empirical data, we simulated 50 replicates using the BS+MNM null model and the ML parameter estimates for that gene at lengths of 5,000 and 10,000 codons; these data were then analyzed using the BST. The false positive rate (FPR) for any gene's simulation (black points) is the proportion of replicates with $P < 0.05$. Gray points show FPR for control simulations with $\delta = 0$. Dashed lines, FPR of 0.05. The solid diagonal line has a slope of 1. (d) The distribution of ML estimates of δ across genes with (white) and without (black) a signature of positive selection in the classic BST is shown for data simulated under the BS+MNM null model. Median δ in BST-significant and BST-nonsignificant genes = 0.03 and 0.0009 in humans, 0.04 and 0.08 in flies. Difference between the distributions was evaluated using a Mann-Whitney U Test: ***, P=1e-12; ****, P=1e-199.

Figure 2.3: (continued) (e) Increasing the MNM rate increases bias in the BST. Sequences 5000 codons long were simulated using the BS+MNM model and the median value of each model parameter and branch length across all genes in each dataset, but δ was allowed to vary. The rate of false positives ($P < 0.05$) in 50 replicates at each value of δ is shown. Solid line, hyperbolic fit to the data; dotted line, FPR level of 5%. Arrowhead, median δ across all genes. (f) Monotonic relationship between δ and inferred ω_2. Sequences simulated in (e) were used to infer the ω_2 estimated by BST, and the relationship plotted. The best-fit regression line is shown along with the R^2.

with the hypothesis that a transversion-rich MNM process produced many of the CMDs in BST-significant genes, but it is also possible that positive selection could have enriched for transversions.

To test whether transversion-enrichment in MNMs exacerbates the BST's bias, we developed an elaboration of the BS+MNM model in which an additional parameter allows MNMs to have a different transversion:transition ratio (κ_2) than single-site substitutions do (κ_1). We estimated the maximum likelihood estimates of the model's parameters for every gene in the mammalian and fly datasets and simulated sequences using genome-wide median values for all model parameters and branch lengths, except for κ_2, which we varied. Sequences 10,000 codons long were used, because simulating shorter sequences resulted in a high variance in the realized transversion:transition ratio. We analyzed these data using the classic BST and calculated the fraction of replicates in which positive selection was inferred. We found that increasing κ_2 caused a rapid and monotonic increase in the false positive rate, indicating that transversion enrichment in MNMs does exacerbate the test's bias. The bias is strong: when κ_2/κ_1 is increased from 1 to 2, the FPR approximately doubles (Fig. 2.4b). Thus, realistic rates of MNM generation and transversion enrichment together cause an even stronger bias in the BST. This result cannot be accounted for by positive selection driving fixation of transversions, because no positive selection was present in the simulations.

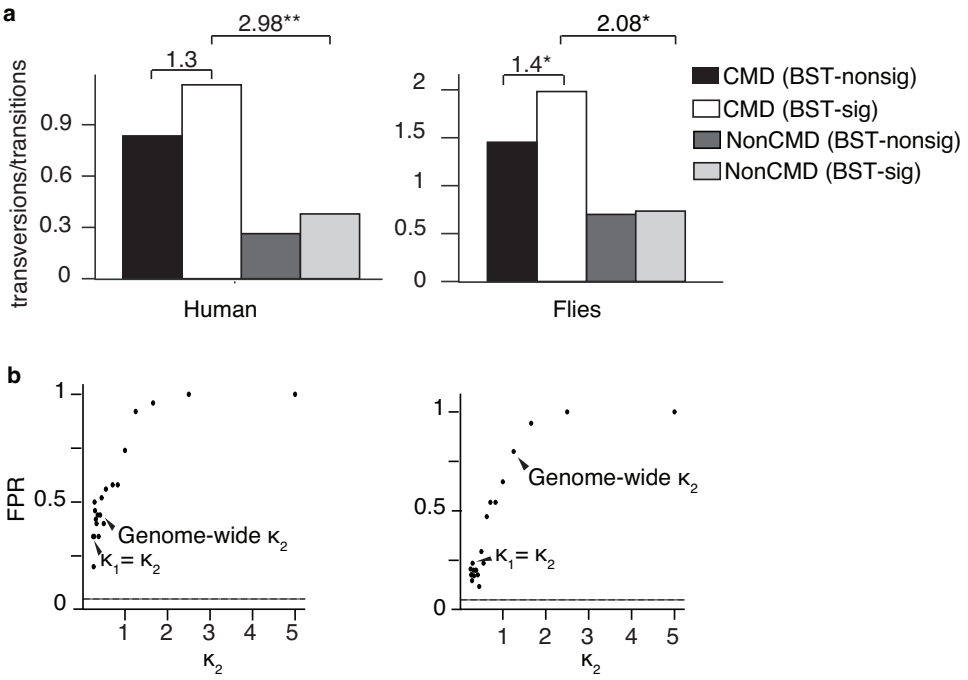


Figure 2.4: **Transversion-enrichment in CMDs biases the BST** (a) The ratio of transversions to transitions observed in CMDs and in non-CMDs is shown for BST-significant and BST-nonsignificant genes. Fold-enrichment is shown as the odds ratio. *, P=5e-4; **, P=3e-25 by Fisher's exact test. (b) Increasing the transversion rate in MNMs increases bias of the BST. Sequences 10,000 codons long were simulated using an elaboration of the BS+MNM model that allows MNMs to have a transversion:transition rate (κ_2) different from that in single-nucleotide substitutions (κ_1). 50 replicate alignments were simulated under the null model using the average value of every model parameter and branch length across all genes in each dataset, except κ_2 was allowed to vary. The rate of false positives ($P < 0.05$) at each value of κ_2 is shown. Dotted line, FPR of 5%.

2.3.6 MNMs affect a newer test of positive selection

In recent years, newer likelihood-based methods have been introduced to test for episodic site-specific positive selection [103, 102, 111]. All these methods are based on models of sequence evolution that, like the BST, do not allow MNMs but instead model CMDs as the result of serial site-specific substitutions. We therefore hypothesized that these methods might also be biased by MNMs. We chose a recent branch-site test, BUSTED [102], which was developed primarily to test for episodic selection across an entire tree. We tested its performance on alignments 5,000 codons long that were simulated using the BS+MNM null

model and parameters estimated from the BST-significant gene alignments in humans and flies. To test for MNM-induced bias, we compared results when δ was assigned to three different values: zero, its average across all alignments in the mammalian or fly datasets, or its gene-specific value in each of the BST-significant genes (Supplementary Fig. 2.6a). We found that BUSTED was sensitive to an MNM-induced bias. When $\delta = 0$, virtually no genes' parameters led to frequent false positive inferences, with a median $FPR < 0.03$ across genes (Fig. 2.5). But when δ was assigned to its empirically estimated gene-specific value, the parameters from every gene in humans and the majority in flies yielded false positive rates > 0.05, with median FPRs of 0.29 and 0.5, respectively (Fig. 2.5). Frequent false positive inferences were evident when sequences were simulated using genome-wide average estimates of δ, as well.

2.3.7 CMDs that invoke multiple nonsynonymous steps drive the signature of positive selection

Finally, we sought further insight into the reasons why CMDs yield a false signature of positive selection in the BST and related tests. In standard models of codon evolution, CMDs are interpreted as the result of two or more serial independent substitutions, even though they can be produced by MNMs in a single mutational event. We hypothesized that CMDs that imply multiple nonsynonymous nucleotide substitutions under these models would provide the strongest support for the positive selection model. We therefore classified CMDs in the empirical datasets by the minimum number of nonsynonymous single-nucleotide substitutions required from the ancestral to derived codon state under standard codon models. As predicted, we found that CMDs that imply more than one nonsynonymous step are dramatically enriched in BST-significant genes (Fig. 2.6a).

We also examined the statistical support provided by different kinds of CMDs. As the number of nonsynonymous steps increased, the statistical support provided for the positive

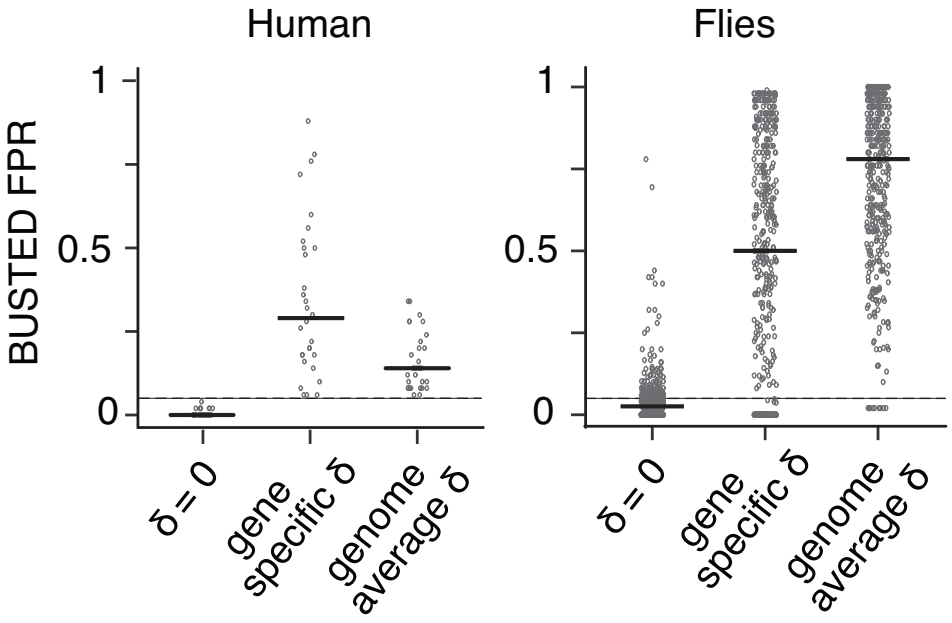


Figure 2.5: **MNMs bias newer tests of positive selection.** False positive inferences under realistic conditions using BUSTED. For every BST-significant gene in each dataset, 50 replicate alignments 5,000 codons long were simulated using the BS+MNM null model and parameter values estimated from the empirical sequences. These alignments were then analyzed for a signature of positive selection ($P < 0.05$) using BUSTED. δ was assigned to its gene-specific estimate, to its average across all genes in each dataset, or to zero. FPR is the proportion of replicate alignments for each gene with $P < 0.05$. Each dot represents the FPR for one gene; black bar, median across genes.

selection model also increased (Fig. 2.6b). CMDs that imply one nonsynonymous and one synonymous step typically provide weak to moderate support for the positive selection model, but CMDs that imply two nonsynonymous steps provide very strong support. In many cases, a single CMD in this latter category is sufficient to yield a statistically significant signature of positive selection.

2.4 Discussion

Our results demonstrate that the branch-site test suffers from a strong and systematic bias toward false positive inferences. This bias is caused by a mismatch between the method's underlying codon model of evolution — which assumes that a codon with multiple differ-

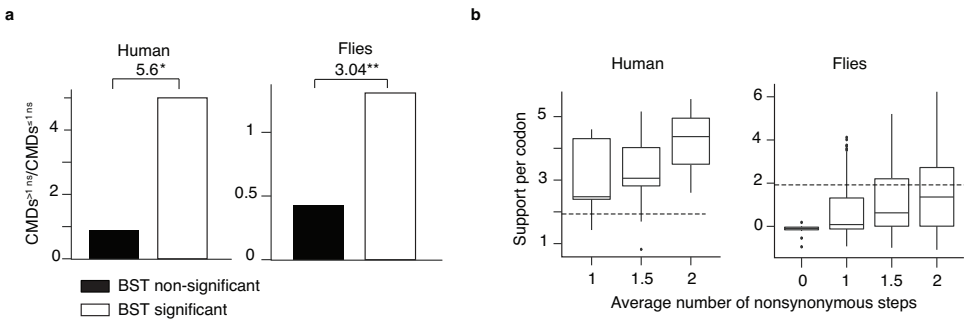


Figure 2.6: **CMDs implying multiple nonsynonymous steps drive the branch-site test** (a) For every CMD, the mean of the number of nonsynonymous single-nucleotide steps on the two direct paths between the ancestral and derived states was calculated. In BST-significant and BST-nonsignificant genes, the ratio of CMDs invoking more than one nonsynonmous step to those invoking one or fewer such steps is shown. Fold-enrichment is shown as the odds ratio. *, P=9e-04; **P= 1.6e-67 by Fisher's exact test. (b) Support for the positive selection model provided by CMDs depends on the number of implied non-synonymous single-nucleotide steps. Support is the log-likelihood difference between the positive selection and null models of the BS test given the data at a single codon site. Box plots show the distribution of support by CMDs in BST significant genes categorized according to the mean number of implied nonsynonymous steps. Dotted line, support of 1.92, at which the BST yields a significant result for an entire gene ($P < 0.05$). In human BST-significant genes, no CMDs imply zero non-synonymous changes.

ences can be produced only by two or more independent substitution events — and the recently discovered phenomenon of multinucleotide mutation, which produces such codons in a single event. Because of the structure of the genetic code and the high transversion rates that characterize MNMs, most codons produced by this mechanism cause more than one nonsynonymous single-nucleotide change. Confronted with this kind of codon data, the likelihood calculated by the BST is determined by the product of the probabilities of the individual mutations. Under the null model, the probability of such compound events is extremely small, but it can increase dramatically when dN/dS exceeds one, as the positive selection model allows. This increase in likelihood afforded by the positive selection model is much greater than it would be if the substitution were interpreted as the result of a single multinucleotide event. Indeed, our results show that a single codon comprising two non-synonymous substitutions is often sufficient to yield a statistically significant signature of

positive selection in the BST for an entire gene.

As a result, CMDs are the primary drivers of positive results by the BST. Virtually all statistical support for positive selection in real alignments comes from CMD-containing sites; removing them from the alignment or incorporating MNMs into the BST's model eliminates the signature of selection from the majority of genes. CMDs can be produced by either positive selection or by neutral evolution under multinucleotide mutation. In the former case, the BST will be correct; however, the test cannot reliably distinguish CMDs that represent authentic evidence of positive selection from those caused by MNM-induced bias. The bias is strong and pervasive under realistic conditions. Indeed, when sequences were simulated under the null model using parameters estimated from the fly and mammalian datasets, the number of genes with false positive BSTs was approximately the same as the number of positive BST results when the empirical data were analyzed. There is therefore no excess of BST-positive results in these genomes beyond that potentially attributable to MNM-induced bias. Worse, these null simulations did not include the elevated transversion rate that characterizes MNMs, which exacerbates the test's bias. Taken together, these results suggest the possibility that MNM-induced bias could explain many of the BST's inferences of positive selection in these datasets.

Are our findings from these datasets generalizable? MNMs appear to be a property of all eukaryotic replication processes, and the MNM rates that we observed in mammals and flies are in the same range as those previously identified in genetic and molecular studies in a variety of eukaryotic species [125, 29, 5]. Both datasets comprise a small number of taxa, but the BST seeks evidence of selection on individual branches, so it seems unlikely that larger trees will somehow inoculate the test against MNM-induced bias. We observed strong bias on lineages with divergence levels ranging from very low (on the human terminal branch) to moderate (the fly branches), so this problem does not appear to be unique to highly diverged sequences or phylogenies with long branches. We must therefore consider the

possibility that many of the thousands of previously published reports of positive selection based on the BST could simply be the ones that happened by chance to neutrally fix one or more multinucleotide mutations.

We do not contend that the BST is always wrong or that molecular adaptive evolution does not occur. Indeed, some of the CMDs in BST-significant genes may have evolved because of authentic positive selection, either by repeated substitution of single nucleotides in a codon or selection on MNMs. But because the BST test cannot distinguish the kinds of sequence data produced by positive selection from those produced by neutral evolution of MNMs, it provides no reliable evidence of a gene's adaptive history — not even prima facie evidence. There are numerous cases of strongly supported adaptive evolution, such as those involving host-parasite and intracellular genetic conflicts, that have produced sequence signatures of positive selection that are likely to be authentic [132]. The persuasive evidence in these cases, however, comes from sources other than the BST.

If the BST and other tests based on the single-step codon model are unreliable in the face of multinucleotide mutation, what should researchers do? The BS+MNM test could be used to accommodate multinucleotide mutation; our results suggest this may be a promising approach. But there are many forms of evolutionary complexity that are not incorporated in this model, including MNMs that affect three consecutive nucleotides in a codon, elevated transversion probability within MNMs, and many other kinds of heterogeneity that might bias the BS+MNM test [110, 93, 17]. Other models are also available to incorporate MNMs [151], but their accuracy and robustness are not well characterized, either. More work is therefore required before the BS+MNM or similar models can be used with confidence in the branch-site or similar tests.

A complementary approach is to use functional experiments to explicitly test hypotheses that specific historical changes in molecular sequence caused changes in function or phenotype thought to have mediated adaptation [10, 25]. Indeed, the bias we observed may help

to explain why some molecular experiments have shown that codons with a high posterior probability of positive selection in the BST do not contribute to putative adaptive functions, whereas the codon changes that do confer those functions have low or moderate PPs [49]. Experimental tests provide the most convincing evidence of a gene's putative adaptive history, but they require time-consuming laboratory and fieldwork [12, 130] , so it is not clear how to implement them on a genome-wide scale. Future research may develop and validate more robust models to detect positive selection, and these may help to identify candidate genes for which specific, testable hypotheses of past molecular adaptation on specific phylogenetic lineages can be formulated. The test primarily used for this purpose till now, however, is unreliable.

2.5 Methods

2.5.1 Datasets, quality control, and inference of BST-significant genes

We analyzed two previously published comprehensive datasets of protein-coding alignments on a genomic scale, one in six mammals, the other in six Drosophila species [45, 86, 88]. We aimed to apply the branch-site test on every terminal lineage in the Drosophila dataset, and on the human lineage in the mammal dataset. We only retained gene alignments without gross misalignments, possessing complete coverage in all fly species, and minimally all primate species. We then applied the branch-site test as implemented in CODEML 4.7 to each alignment, assuming the phylogenetic relationships reported in the published studies (Supplementary Fig. 2.10) [45, 86]. Branch lengths and model parameters were estimated for each alignment by maximum likelihood (ML), and the F3x4 model was used for codon frequencies. We tested each gene in mammals for selection on the terminal branch leading to humans; in flies, each gene was tested separately for selection on each of the six terminal branches, and we express the fraction of positive inferences across genes as the proportion of

all tests conducted [167]. As is standard practice, we calculated P-values using a likelihood ratio test with 1 df (χ_1^2) which makes the test conservative under the null hypothesis [167]. Genes were initially identified as having a putative BST signature of selection at $P < 0.05$. We then applied a correction for multiple testing to a false discovery rate $FDR < 0.20$ using the q-value package in R (avaitable at http://github.com/jdstorey/qvalue). To facilitate unambiguous analysis of CMDs, we removed genes containing CMDs falling in gaps. We also removed genes for which the ML ancestral reconstructions reported by CODEML at the base of the tested branch differed between the null and positive selection models, yielding a set of genes with CMDs that do not depend upon which model is chosen. In flies, 443 gene-tests ("genes") were retained after these filters and constitute the BST-significant set of genes from this dataset. No genes on the human lineage were significant after FDR correction, so we retained as the BST-significant set from this dataset those genes that passed the ancestral reconstruction filter and had $P < 0.05$ (Supplementary Table 2.2). The BST-nonsignificant set of genes comprises all genes that pass the alignment and ancestral reconstruction filter that are not in the BST-significant set (n=6757, humans; n=6883, flies). We also repeated our analysis of CMD enrichment (see below) using a gene set that had not been filtered for reconstruction consistency and found that our conclusions were unchanged (Supplementary Table 2.7)

We only considered genes where the ancestral codons (both CMD and non-CMD codons) have the same reconstruction under the BST null and BST alternate models. In doing so, we have also excluded CMDs in codons with gaps in the alignment. For example, in the human dataset, of the 82 genes that initially provided support for positive selection, 30 genes consist of unambiguously reconstructed codons under the null and alternate model (the BST-significant gene set). In 49 genes, CMDs fall in gaps. We did not consider the ancestral codon reconstructions at these sites, and excluded these from our analyses due to alignment ambiguities. The remaining 3 genes have CMDs that do not fall in gaps, for which

the ancestral codons were reconstructed differently under the null and alternate models. If we re-consider these 3 "positively selected" genes that were excluded, we find 3 additional CMDs, one in each of the genes. Including these genes made little to no difference to our CMD enrichment results.

2.5.2 Support for positive selection

CMDs were identified in BST-significant and BST-nonsignificant genes as codons with 2 or 3 observed nucleotide differences between the ML states at the ancestral and extant nodes for the branch being tested; non-CMDs are codons with 0 or 1 differences on the branch tested. CMDs were not assessed on branches not tested.

To determine the role of CMDs in significant results from the BST, we excluded codon positions in BST-significant genes containing CMDs, reanalyzed the data using the BST, and calculated the fraction of tests that retained a significant result $P < 0.05$.

We quantified the proportion of statistical support for positive selection in BST-significant genes that comes from CMDs as follows. The site-specific support provided by one codon site in an alignment is the difference between the log-likelihoods of the positive selection model and the null model given the data at that site. Support for positive selection provided by all CMDs in a gene (supportCMD) is the support summed over all CMD sites in the alignment. The proportion of support provided by CMDs is supportCMD / (support CMD + support nonCMD). This proportion can be greater than 1 if support by non-CMDs is negative, as occurs if the likelihood of the null model at non-CMD sites is higher than that of the positive selection model, given the parameters of each model estimated by ML over all sites.

Sites were classified a posteriori as under positive selection if their Bayes Empirical Bayes posterior probability of being in class 2 ($\omega_2 > 1$) under the positive selection model in CODEML was > 0.5 (moderate support) or > 0.9 (strong support).

We categorized observed CMDs by the minimum number of nonsynonymous single-

nucleotide steps implied under the Goldman-Yang model between the ancestral and derived states. For each CMD comprising two nucleotide differences, there are two paths by which they can be interconverted by two single nucleotide steps. We determined whether the steps on these paths would be nonsynonymous or synonymous using the standard genetic code and then calculated the mean number of nonsynonymous steps averaged over the two paths. Paths involving stop-codons were not included. We conducted a similar analysis for all possible CMDs in the universal genetic code table.

2.5.3 BS+MNM codon substitution model and test

The codon substitution model of the classic BST is based on the Goldman-Yang (GY) model [159]. Sequence evolution is modeled as a Markov process, where the matrix element q_{ij}, the instantaneous rate of change from ancestral codon i to derived codon j, is defined for four types of changes: synonymous transitions and transversions, and nonsynonymous transitions and transversions (see q_{ij}, Fig. 2.7).

$$q_{ij} = \begin{cases} \kappa\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \kappa\omega\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ 0 & \text{two or more differences} \end{cases} \quad \ldots\ldots\ldots \quad (1)$$

Figure 2.7: **Goldman and Yang codon substitution model**.

Three parameters are estimated from the data by maximum-likelihood: ω, the ratio of nonsynonymous substitution rate to the synonymous substitution rate (dN/dS); π_j, the equilibrium frequency of codon j; and κ, the transversion:transition rate ratio. Element q_{ij} is zero for substitutions involving more than one difference, so codons with multiple differences can only evolve through intermediate codons that are a single change away. A scaling factor

applied to the matrix ensures that branch lengths are interpreted as the expected number of substitutions per codon.

We developed a modification of the GY model that incorporates MNMs using the parameter, δ, which represents the relative instantaneous rate of double substitutions to that of single substitutions (see q_{ij} equation 2, Fig. 2.8). When $\delta = 0$, the BS+MNM model reduces to the classic BST model that does not incorporate MNMs (q_{ij} equation 1,Fig. 2.7). Triple substitutions have an instantaneous rate of zero.

$$q_{ij} = \begin{cases} \kappa\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \kappa\omega\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ \omega\delta\kappa^2\pi_j & \text{non-synonymous, 2 transversions} \\ \omega\delta\pi_j & \text{non-synonymous, 2 transitions} \quad \ldots\ldots\ldots \\ \omega\delta\kappa\pi_j & \text{non-synonymous, 1 transversion, 1 transition} \\ \delta\pi_j & \text{synonymous, 2 transitions} \\ \delta\kappa^2\pi_j & \text{synonymous, 2 transversions} \\ \delta\kappa\pi_j & \text{synonymous, 1 transversion, 1 transition} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Figure 2.8: **BS+MNM codon substitution model**.

The BS+MNM test of positive selection is identical to the BST, except it utilizes this MNM codon model. We implemented this test by modifying the branch-site test batch file (YangNielsenBranchSite2005.bf) in Hyphy 2.2.6 software by declaring δ a global variable, incorporating it into the codon table, and allowing it to be optimized by ML as it other model parameters are.

We validated the BS+MNM implementation by simulating 50 replicate alignments using the BS+MNM null model in Hyphy under genome-median parameters (see below). We then used the BS+MNM procedure to find the ML estimate of each parameter, including branch lengths, given each alignment and the topology of the phylogeny used to generate the sequences. We compared the distribution of estimates over replicates to the "true" values used to generate the sequences (Supplementary Fig. 2.12).

To test if there is statistical support in the data for the BS+MNM null model relative to the standard BST null model, we performed an LRT with 1 df, comparing the fit of the BS+MNM null model and the BST null model on our empirical genes. Briefly, for each of the 6868 human genes, we tested if the BS+MNM null model fit the data better than the BST null model at $P < 0.05$ and also applied and adjustment for multiple testing ($FDR < 0.2$). We performed similar LRTs for each of the six terminal lineages in flies. To determine whether this test might be prone to falsely infer support for the BS+MNM model, we simulated control sequences under the null BST model with parameters derived from the empirical sequences and performed the LRT as described above. Only 2 percent of genes in humans and 2.6 percent in flies yielded significant support for BS+MNM at $P < 0.05$. Zero human genes and 0.006 percent of fly genes retained significance after multiple testing adjustment ($FDR < 0.2$). (Supplementary Table 2.4).

2.5.4 Simulations and analysis of false-positive bias

To characterize bias in the BST and other tests of selection, we conducted sequence simulations in the absence of positive selection under empirically derived conditions. We used the BS+MNM method we implemented in Hyphy to estimate by maximum likelihood (ML) the gene-specific branch lengths and parameters of the null BS+MNM model for every gene in the mammalian and fly datasets. We also calculated the genome-wide median of each parameter over all genes in each dataset (the "genome-average" parameter value). Probability density characterizations for parameters and gene length were performed using the density function in R.

We simulated sequence evolution under the BS+MNM null model using either gene-specific or genome-median parameters. First, we simulated a "pseudo-genome" without positive selection by simulating one replicate of each of the 6868 and 8564 mammalian and fly alignment, each at its empirical length, using the BS+MNM null model and the ML

parameter estimates inferred for that gene from the empirical data. We then ran the BST on these sequences, testing for signatures of positive selection on the human lineage and each terminal fly lineage (Supplementary Table 2.2). Control simulations were conducted under identical conditions but with $\delta = 0$.

To test the effect of gene length on bias in the BST, we focused on genes in the BST-significant set. For each gene's gene-specific parameters, we simulated 50 replicates alignments of length 5,000 or 10,000 codons. We analyzed these alignments using the BST, assigning the human branch as foreground for mammalian genes or, for flies, the same branch that produced a significant result when the empirical data were analyzed. The false positive rate (FPR) for any gene's parameters is the fraction of replicates yielding a positive test ($P < 0.05$). We also repeated these simulations and analyses using the genome-median value of δ. For control experiments without MNMs, we set $\delta = 0$ in the simulations.

To test the effect of the rate at which MNM substitutions are produced on false positive inference rates, we simulated evolution of alignments 5,000 codons long under the BS+MNM null model, using genome-median estimates for all parameters except δ, which we varied. At each value of δ, we simulated 50 replicates. We analyzed each replicate using the BST for selection on the human or D. simulans lineages and calculated the proportion of replicates for each value of δ that yielded a false positive inference ($P < 0.05$).

We computed the observed proportion of tandem substitutions as a fraction of all substitutions on the human and D. melanogaster lineages in both empirical and simulated datasets. For each of the 6868 genes in the curated mammalian dataset, we aligned the human gene to the inferred sequence of the human-chimp ancestor, identified all substitutions as differences between these sequences, and calculated the proportion of tandem substitutions, T, as the number of substitutions at adjacent sites divided by the sum of substitutions at adjacent sites and those at non-adjacent sites across all sites in the dataset. Differences at adjacent sites were counted as a single tandem substitution. For each of the 8564 genes in the fly

dataset, we aligned the D. melanogaster sequence to the D. melanogaster/D. simulans ancestor and followed the procedure described above. For simulated sequences, we repeated this procedure using the sequences simulated under the BS+MNM null model and parameters estimated from each gene in the empirical datasets.

2.5.5 BUSTED

To examine the accuracy of BUSTED, we used Hyphy software 2.2.6 (batch files BUSTED.bf and QuickSelectionDetection.bf). We analyzed the 5,000 codon-long alignments simulated under the BS+MNM null model, using parameters estimated by ML for each BST-significant gene, with δ assigned either to its gene-specific estimate, its genome-average, or to zero. We applied BUSTED to the replicate alignments to test for selection ($P < 0.05$) on the human lineage or the same fly lineage that was significant for that gene in the BST of the empirical data.

2.5.6 Power analyses

To characterize the statistical power of the BST and BS+MNM tests, we simulated sequence evolution with positive selection of variable intensity and pervasiveness (Supplementary Fig. 2.13). Specifically, we used the BS positive model in Hyphy to simulate sequence evolution with the human and D. simulans terminal branches as the foreground branches. We used genome-average estimates of all parameters, including gene length (418 and 510 codons for mammals and flies, respectively), but we varied ω_2 and p_2. 20 replicate alignments were simulated under each set of conditions and then analyzed using the BST, the BS+MNM test, or BUSTED. For each set of conditions, the true positive rate was calculated as the fraction of replicates yielding a significant test of positive selection ($P < 0.05$ for BST and BS+MNM, $FDR < 0.20$ for at least one site in the alignment for BUSTED).

2.5.7 BS+MNM+ κ_2 model

We developed the BS+MNM+ κ_2 model, which incorporates into the BS+MNM model (q_{ij} Fig. 2.8) two different transversion:transition rate ratio parameters, κ_1 for single-site substitutions and κ_2 for MNMs (q_{ij} Fig. 2.9). All free parameters of the model are estimated by ML given a sequence alignment. This model was implemented by further modifying our BS+MNM batchfile in Hyphy 2.2.6 software by declaring κ_2 a global variable, incorporating it into the codon table, and allowing it to be optimized by ML as other parameters are in the batch file.

$$q_{ij} = \begin{cases} \kappa_1\pi_j & \text{synonymous transversion} \\ \pi_j & \text{synonymous transition} \\ \omega\kappa_1\pi_j & \text{non-synonymous transversion} \\ \omega\pi_j & \text{non-synonymous transition} \\ \omega\delta\kappa_2^2\pi_j & \text{non-synonymous, 2 transversions} \\ \omega\delta\pi_j & \text{non-synonymous, 2 transitions} \\ \omega\delta\kappa_2\pi_j & \text{non-synonymous, 1 transversion, 1 transition} \\ \delta\pi_j & \text{synonymous, 2 transitions} \\ \delta\kappa_2^2\pi_j & \text{synonymous, 2 transversions} \\ \delta\kappa_2\pi_j & \text{synonymous, 1 transversion, 1 transition} \\ 0 & \text{otherwise} \end{cases} \quad \ldots\ldots\ldots \quad (3)$$

Figure 2.9: **BS+MNM+$\kappa_2$codon substitution model**.

For validation, we estimated the parameters of the BS+MNM+ κ_2 null model by ML for every alignment in each dataset and calculated the genome-average median estimate of each parameter (Supplementary Fig. 2.16). We then simulated 50 replicate alignments of length 418 and 510 codons in the mammalian and fly datasets respectively, under the BS+MNM+ κ_2 null model with all model parameters set to their genome-wide median. We then estimated each parameter by ML under the null model given each alignment and compared the distribution of estimates to the parameters used to generate the alignments. We found that most parameters were estimated accurately, but estimates of κ_2 had very high variance (Supplementary Fig. 2.16) , presumably because the quantity of data in a single gene, in which CMDs are typically rare, is inadequate to support a robust estimate

of this parameter. We therefore limited our use of this model to simulations rather than inference. To determine the effect of the MNM-specific transversion:transition rate on false-positive bias in the BST, we simulated sequences 10,000 codons long under the BS+MNM+κ_2 null model, using genome-median parameters except κ_2, which we varied. For each value of κ_2, we simulated 50 replicates, applied the BST, and calculated the FPR as the fraction of replicates yielding a positive inference ($P < 0.05$).

2.5.8 Data availability

The empirical alignments reanalyzed in this study are available in the supplementary information of the original publications that generated these data [101, 84, 129]

2.5.9 Code availability

The custom HYPHY batch codes for the BS+MNM and BS+MNM+κ_2 tests are avaitable as supplementary files and at https://github.com/JoeThorntonlab/MNM_SelectionTests.

2.6 Acknowledgements

We are grateful to the members of the Thornton lab for discussion and helpful comments. We thank the Beagle2, Midway2, and Tarbell supercomputing clusters at the University of Chicago. We also thank the developers of HyPhy for presenting an open source platform that allows limitless customization of standard analyses. Funding was provided by NIH R01GM104397 and R01GM121931 (JWT), NSF DEB-1601781 (JWT and AV), NSF DBI-1564611 (MWH), and the Precision Health Initiative of Indiana University (MWH).

2.7 Author contributions

Analyses were designed by all authors, performed by AV, and interpreted by all authors. The manuscript was written by AV and JWT with contributions from MWH.

2.8 Competing financial interests

The authors declare no competing financial interests.

2.9 Supplementary Information

2.9.1 Supplementary Figures

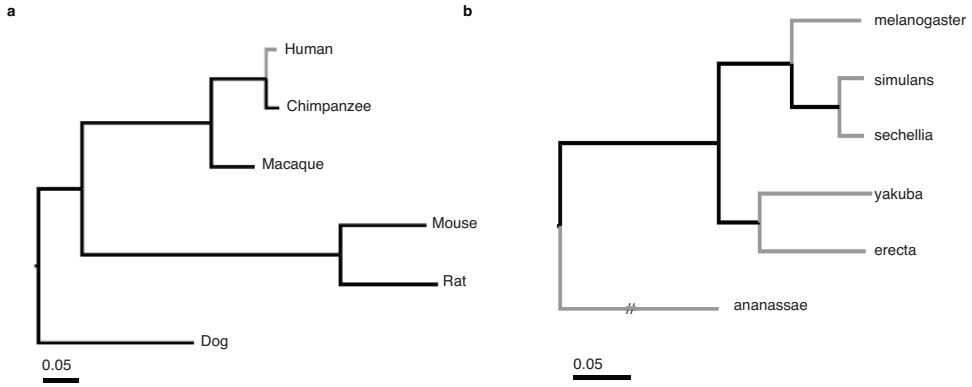


Figure 2.10: **Phylogenies of mammalian and Drosophila species used in this analysis**. The BST was used to identify genes under positive selection on each of the lineages colored in grey. Branch lengths are proportional to the median length across all genes analyzed. Scale bar, expected nucleotide substitutions per codon; the hatched branch, shortened for display, has length 0.62

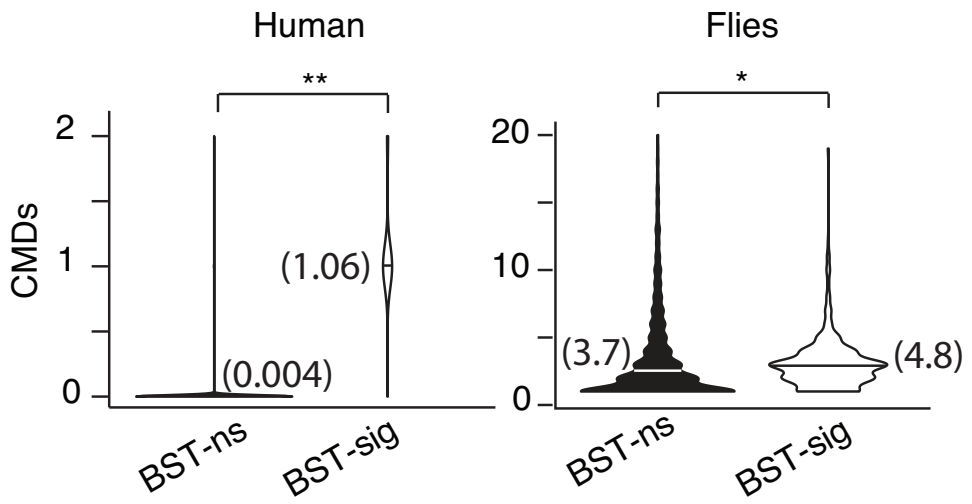


Figure 2.11: **Distribution of the number of CMDs per gene in BST-significant (BST-sig) and BST-nonsignificant (BST-ns) genes**. Horizontal line in each violin plot indicates the median. *, $P = 4e - 10$; **, $P < 2e - 16$, Mann Whitney U test.

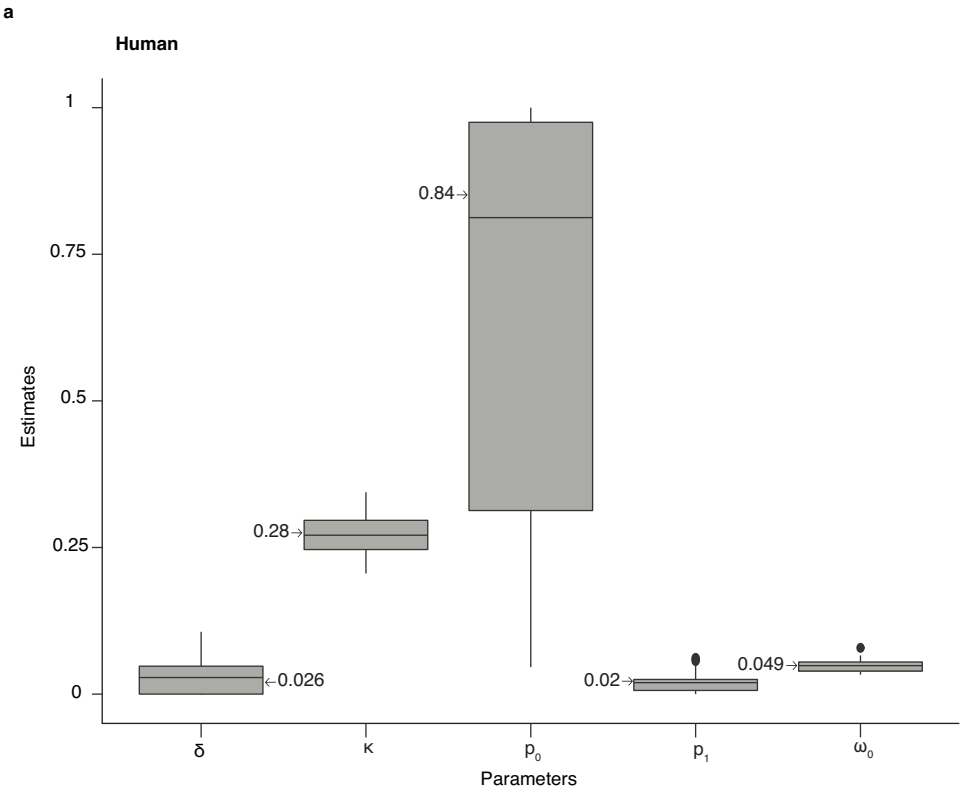


Figure 2.12: **Validation of parameter estimation in the BS+MNM model**

.

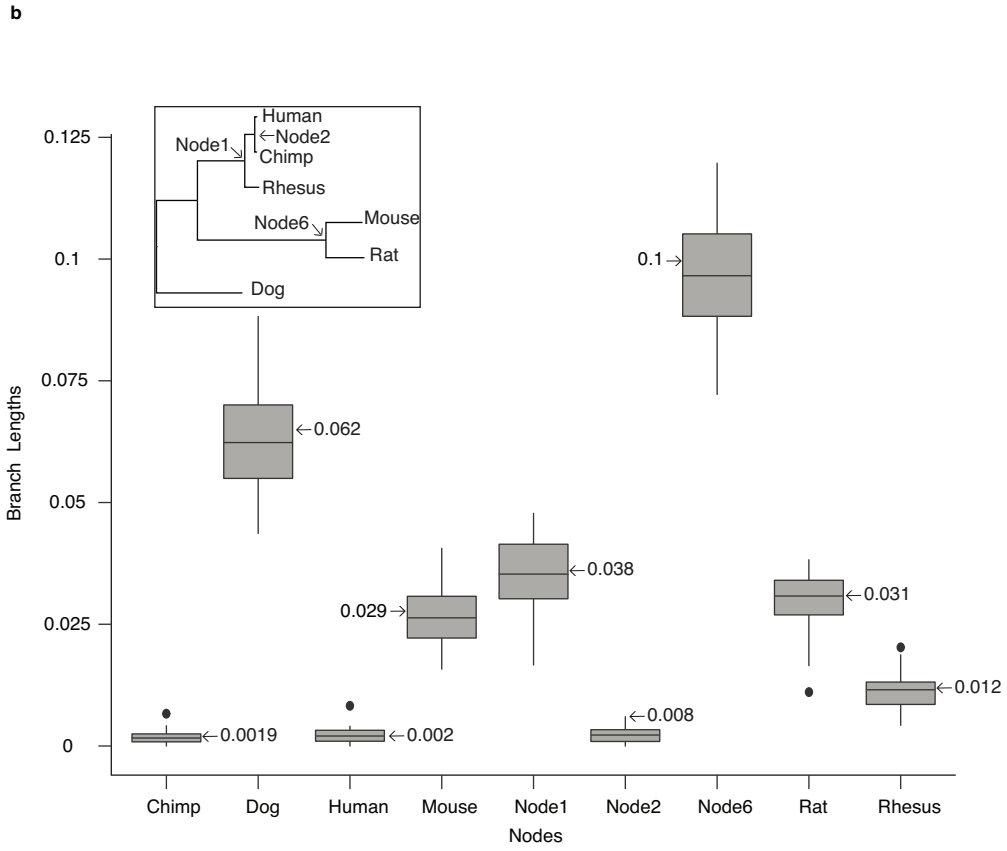


Figure 2.12: (continued)

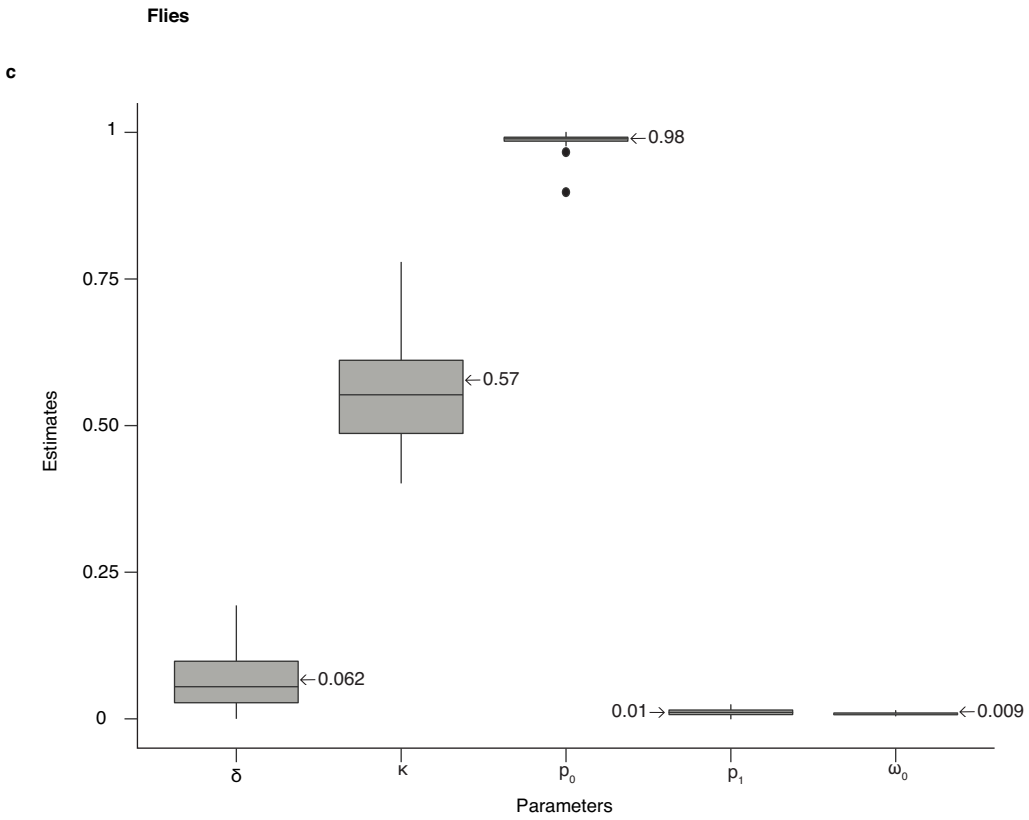


Figure 2.12: (continued)

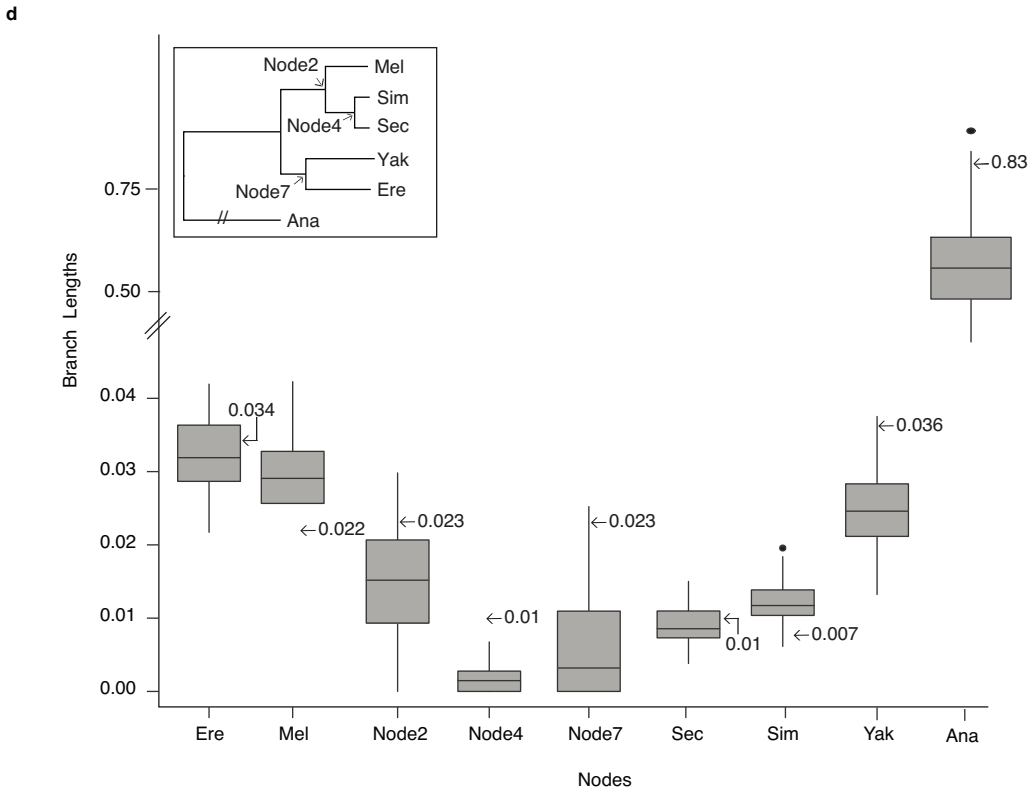


Figure 2.12: (continued). 50 replicate alignments were simulated under the BS+MNM null model with genome-average parameters, including gene length; the BS+MNM null model was then used to estimate the parameters from each replicate. (a) Box plots show the distribution of ML estimates across alignments for model parameters: δ, the MNM parameter; κ, the transversion:transition rate multiplier; p_0, mixture weight for submodel 0; ω_0, nonsynonymous/synonymous rate multiplier for submodel 0; p_1, mixture weight for submodel 1; and branch lengths (b) in the mammalian dataset. The branch length for each node on the x-axis represents the branch length to the labeled node from its immediate ancestral node. Parameter estimates for Drosophila dataset are shown (c) and (d). Arrows indicate the generating parameters used in the simulation. Node names in panels (b) and (d) correspond to those in the phylogenies shown in the inset.

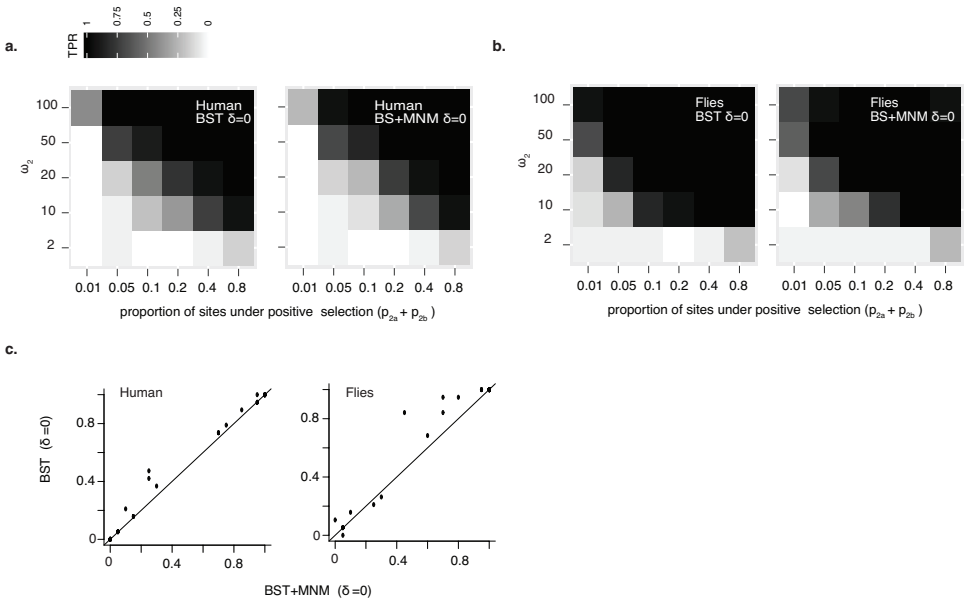


Figure 2.13: **Analysis of power to detect positive selection by BST and BS+MNM, under conditions of no bias ($\delta = 0$), and bias (genome-wide average δ)** (a) Power of the classic BST and BS+MNM test to detect positive selection in the human dataset under conditions of no bias ($\delta = 0$). Sequence alignments of genome-average length were simulated with positive selection using the BS+MNM model and genome-average model parameters including branch lengths. Simulations contained no MNMs ($\delta = 0$). The proportion of sites under positive selection (p_2) and the strength of positive selection (ω_2) were varied, with 20 replicate simulations under each set of conditions. The BST and BS+MNM test were applied and the rate of true positive inferences (TPR) was calculated as the fraction of replicates under each condition with a significant result ($P < 0.05$). Each cell in the grid represents a set of conditions, shaded by TPR according to the heat map key, shown at top. Left panel, heat map showing TPR of BST; Right panel, heat map showing TPR of the BS+MNM test (b) Same as (a), but shown for the fly dataset. (c) Scatter plot showing a power comparison of the BS+MNM test and the classic BST on human and fly sequences simulated with no MNMs. The data are the TPRs shown in panels (a) and (b) respectively.

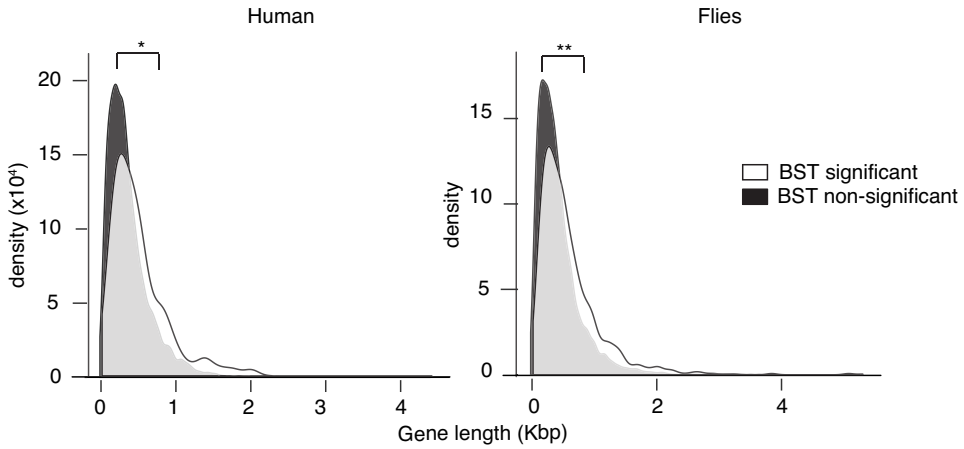


Figure 2.14: **Longer genes are more likely to yield false positive BST results**. For each empirical gene in the mammalian and fly datasets, the parameters of the BS+MNM null model were estimated by maximum likelihood. We then simulated sequence evolution under each gene's inferred null parameters and empirical length and used the classic BST on the simulated alignments to test for positive selection on the human and terminal fly lineages. The distribution of the lengths of genes yielding a BST-significant test ($P < 0.05$) or a BST-nonsignificant test is shown. Median lengths in BST-significant and non-significant genes, respectively, were 422 and 343 bp in humans; in flies, 484 and 391 bp. Differences between distributions were evaluated using Mann-Whitney U test. *, P=5e-3; **, P=2e-23.

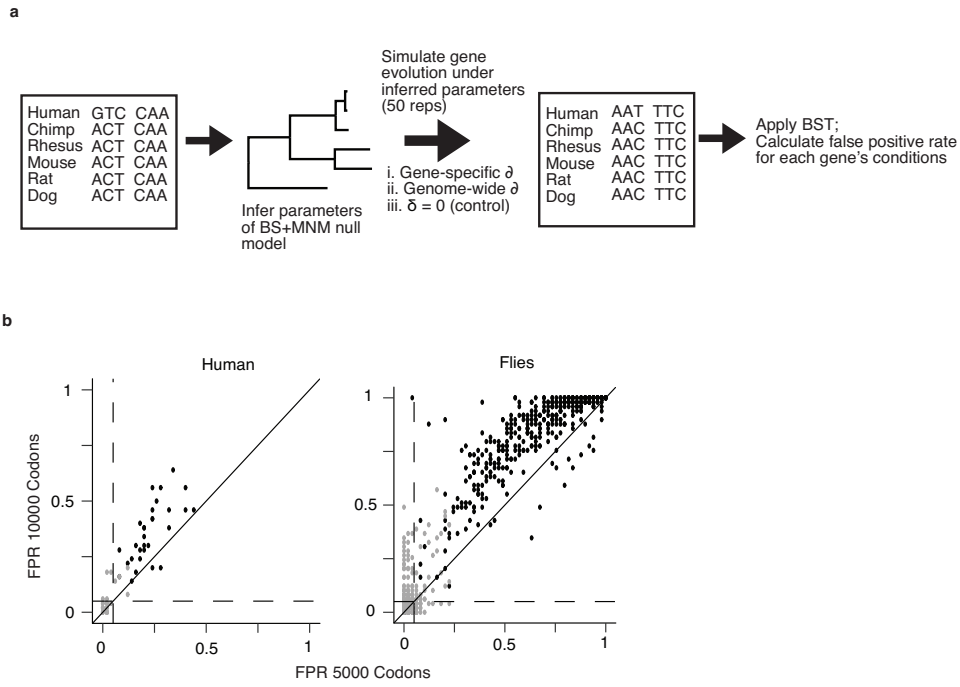


Figure 2.15: **MNMs bias the classic BST** (a) Scheme to simulate genes with MNMs without positive selection. For each BST-significant empirical gene, the parameters of the BS+MNM null model were estimated. Using each set of parameters, 50 replicate alignments were simulated, with δ (the relative rate of MNM substitution) assigned to its gene-specific value, its median across all genes in the dataset (genome-wide average), or to zero. The classic BST was applied to simulated data, and the false positive rate (FPR) for each set of generating parameters was calculated as the fraction of replicates with a positive result ($P < 0.05$). (b) Systematic bias in the BST using the genome-average MNM rate. 50 replicate alignments 5,000 or 10,000 codons long were simulated under the BS+MNM null model using gene-specific parameters inferred as in (a). Each black point represents FPRs for sequences 5,000 and 10,000 codons long simulated under one empirical gene's parameters and the genome-wide average δ. Gray points show FPRs for control simulations with $\delta = 0$. Dashed lines, FPR of 0.05. Diagonal line has a slope of 1.

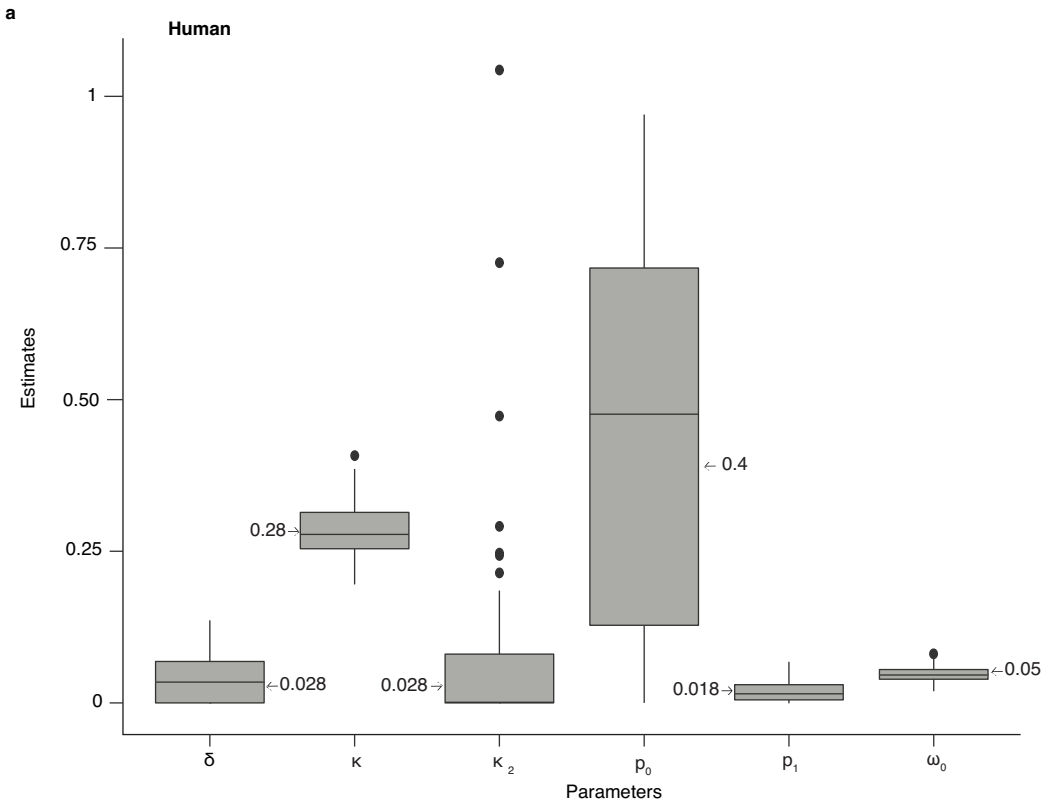


Figure 2.16: **Validation of parameter estimates by BS+MNM+κ_2 model**

.

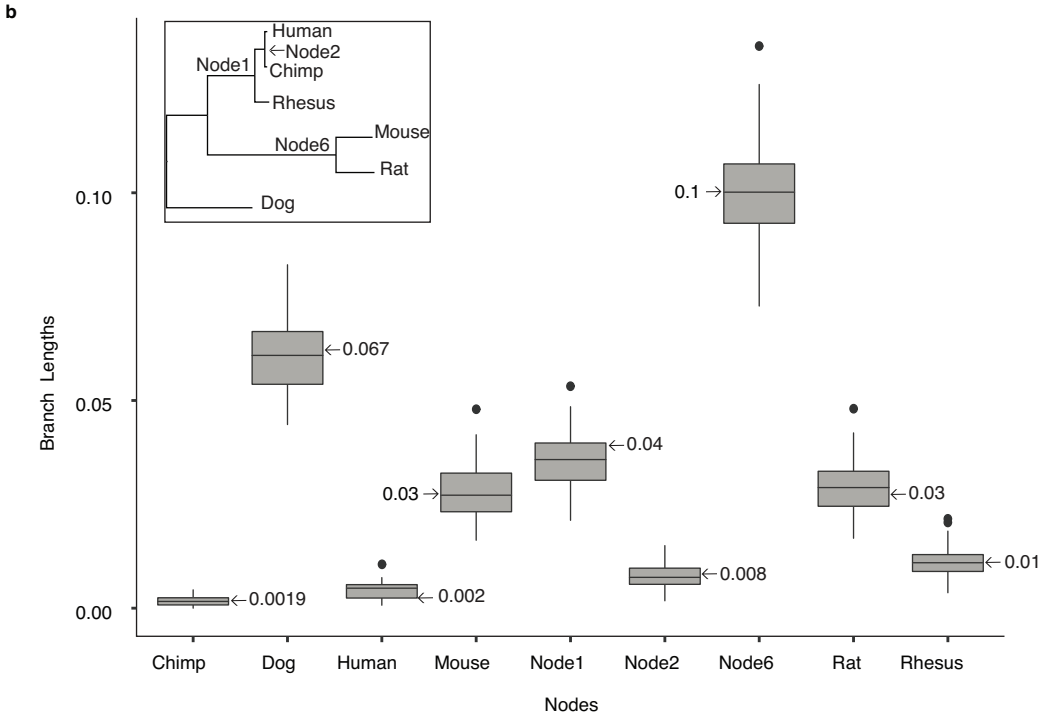


Figure 2.16: (continued)

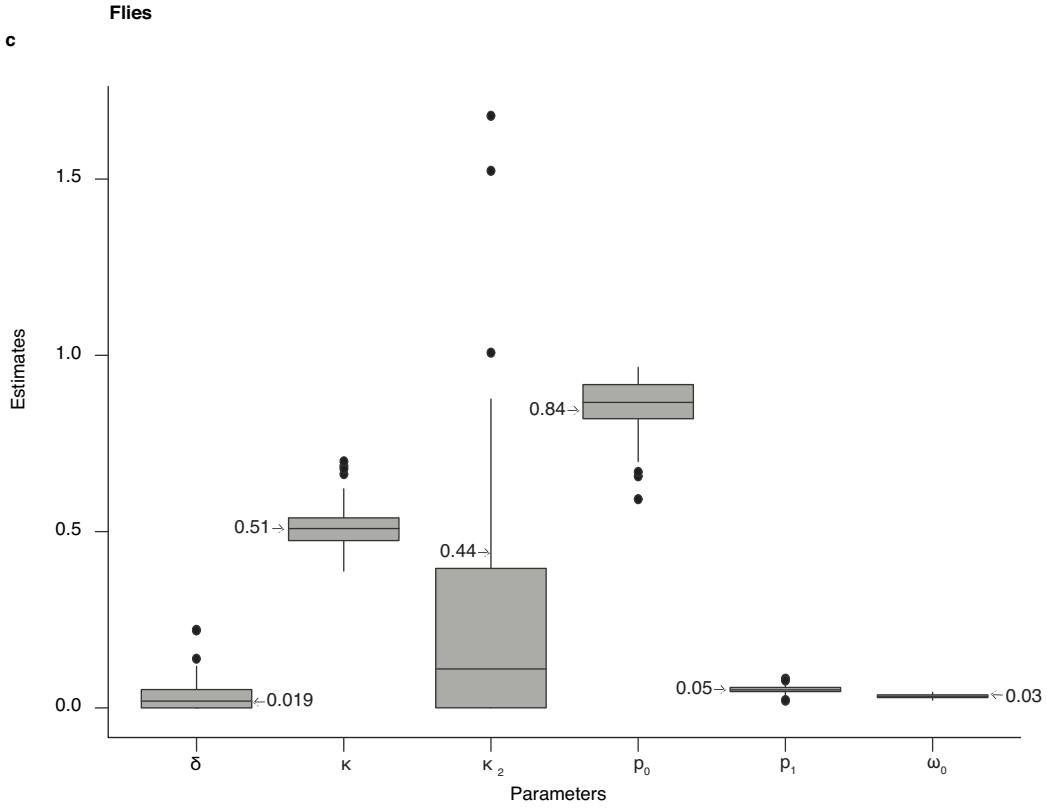


Figure 2.16: (continued)

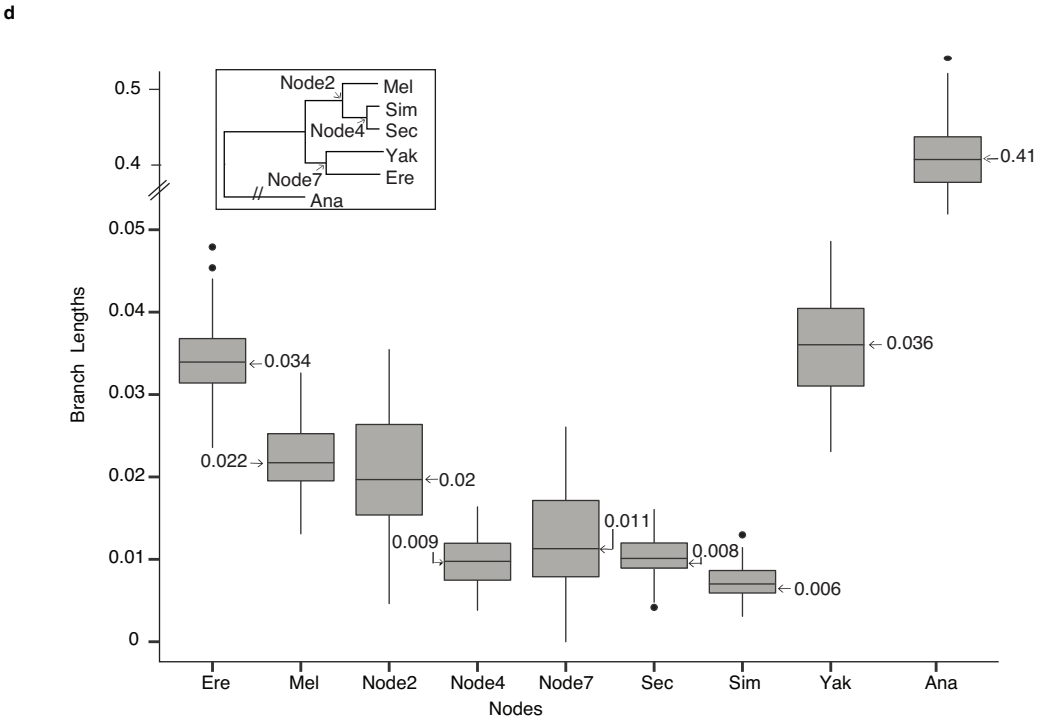


Figure 2.16: (continued). 50 replicate alignments were simulated under the BS+MNM+ κ_2 null model using genome-wide median parameters. Parameters were then re-estimated given each alignment using the same model. Box plots show the distribution of estimates of model parameters (a) and branch length (b) in humans and in flies (c, d). For detailed explanation of model parameters, see Supplementary Fig. 2.12. κ_2 is the transversion:transition rate ratio for CMDs. The branch length for each node on the x-axis represents the branch length from the ancestor to that node in question. Arrows, generating parameters used to simulate the data. Node names correspond to those in the phylogenies shown in the inset. The median value of κ_2 is less than the median κ_1, whereas the mean κ_2 is greater than the mean κ_1, as reported in Fig. 2.4b. This difference is likely caused by the rarity of CMDs with multiple transversions in each codon, which leads to large variance in the estimate of κ_2 in many individual genes.

2.9.2 Supplementary Tables

Average steps on path (nonsyn, sun)	Number of pairs
2	8
0.5,1.5	0
1,1	588
1.5,0.5	308
2,0	548

Table 2.1: **Paths between codon pairs** For each possible codon pair separated by 2 nucleotide differences, the universal genetic code was parsed to tabulate the mean number of nonsynonymous and synonymous steps (nonsyn, syn) on the two direct paths between them. Paths with stop codons were excluded.

	Empirical (H)	Simulated (H)	Control (H)	Empirical (F)	Simulated (F)	Control (F)
All genes	16,541	6,868	6,868	8,564	8,564	8564
Genes with complete species coverage	6,868	6,868	6,868	8,564	8,564	8564
BS tests significant at $P < 0.05$ (% of tests)	82 (1.1%)	99 (1.4%)	32 (0.5%)	3,938 (7.6%)	4,444 (8.6%)	582 (1.1%)
BS tests significant after correction ($FDR < 20\%$)	0	0	0	2,147	1,755	4
BS-significant genes with unambiguous ancestral codon reconstructions	30*			443*		

Table 2.2: **Filtering steps** Filtering steps are described in Methods. Empirical data are all genome-wide coding alignments from previously published studies by Kosiol et al., and Larracuente et al (see Methods). Simulated data are alignments simulated under the BS+MNM null model using parameters derived from the empirical data. Control data are alignments simulated as above but with MNM rate parameter $\delta = 0$. *, in humans, the empirical BS-significant set includes genes that pass the filter for ancestral reconstruction of CMDs but not for FDR adjustment; in flies, both criteria are met. The total number of tests on the 6 fly lineages is 51,384. H = Humans; F=Flies.

Lineage	%genes
human	22 (1545/6868)
mel	60 (5080/8564)
sim	59 (5089/8564)
sec	58 (4994/8564)
yak	59 (5089/8564)
ere	58 (4957/8564)
ana	51 (4297/8564)

Table 2.3: **Proportion of empirical genes fit better by the BS+MNM null model compared to the BS null model** In the empirical dataset, about 22% of the genes on the human lineage, and at least half of the genes on each fly lineage were fit better by the BS+MNM null model compared to the BS null model (LRT, 1 df), indicating that there is statistical support in the data for including MNMs on top of the standard GY framework.

Lineage	#tests significant at $P < 0.05$	#tests significant after $FDR < 0.2$
human	162/6868 (2%)	0
mel	203/8564 (2%)	1
sim	253/8564 (3%)	0
sec	236/8564 (2.7%)	0
yak	251/8564 (3%)	0
ere	210/8564 (2.4%)	2
ana	217/8564 (2.5%)	0

Table 2.4: **Number of genes significant in null simulations (BS+MNM null model compared to the BS null model)** The BS+MNM null model was compared to the BS null model for genes simulated with $\delta = 0$. The BS+MNM null model is significant for $< 5\%$ of null simulations across humans and flies. Only a handful of genes remain significant after FDR correction across all lineages tested.

Species	Lost significance	Retained significance	Retained significance (triple)
Human	28/30	2/30	0/30
Flies	174/458	213/458	71/458

Table 2.5: **Proportion of genes that lost and retained significance after the BS+MNM test was applied to BS significant genes.** There were 30 tests conducted on the human lineage, one for each BS-significant gene. In flies, the 443 genes correspond to 458 tests, conducted on every lineage in which the gene is significant (One gene can have more than one lineage under positive selection). The number of genes that lost and retained significance is as indicated.

	Empirical	Simulated
Humans	295 / 23,335 (1.26%)	326 / 20,040 (1.63%)
D. melanogaster	3,895 / 249,344 (1.56%)	9,069 / 278,978 (3.25%)

Table 2.6: **Observed frequency of tandem substitutions on the human and melanogaster lineages in both empirical and simulated datasets.** The number of tandem substitutions on the human and melanogaster lineages as computed from the empirical and simulated datasets is indicated.

	#CMDs	#non-CMDs	#CMDs / #non-CMDs
BS+ (filtered)	32	19563	0.00163
BS+ (non-filtered)	35	20923	0.00167

Table 2.7: **No effect of filtering based on ancestral state reconstruction on CMD enrichment.** The ratio of CMDs/non-CMDs is shown for the original 30 BS+ (filtered) genes, and the new BS+ (non-filtered) set which consists of 33 BS+ genes. The almost identical ratio suggests that our selection criterion does not enrich CMDs in BS+ genes.

CHAPTER 3

EVOLUTION OF TRANSCRIPTION FACTOR DNA SPECIFICITY IN STEROID AND RELATED RECEPTORS

3.1 Abstract[1]

Many human transcription factors possess distinct preferences for sequences flanking core motifs, but little is known about how such preferences evolve in proteins. We investigated this question in the steroid and related receptor family of transcription factors (SRRs) which comprise of two functionally diverged clades: 1) Steroid Receptors (SRs), which bind as a cooperative dimer to an inverted palindrome of a 6-bp half-site AGGTCA; and 2) Estrogen Related Receptors (ERRs), which bind as monomers to an extended 9-bp half-site (TCAAGGCTA), containing a 5'-flanking extension of the 6-bp SR half-site. Through ancestral sequence reconstructions, mobility shift assays, isothermal titration calorimetry and in vivo reporter gene activation assays, we show that the ancestral SRR DNA binding domain behaves similarly to ERRs; therefore, the ancestral preference for a particular 5'-flanking sequence was lost on the lineage leading to extant SRs. This transition only involved changes to the half-site affinity, with little to no changes to cooperative binding – all ancestors and their modern descendants retain weak DNA-binding cooperativity, including the human ERR which has been described as a monomeric DNA-binder. After duplication of the ancestral SRR, only six mutations were sufficient to recapitulate the preferences of the SR ancestor, suggesting that a change in binding specificity evolved by few mutations of large effect. We show that non-overlapping DNA-binding specificities evolved in paralogous proteins through tinkering with thermodynamic basis of specificity, without a radical reorganization or evolution of novel interfaces. This study provides a mechanistic basis for how flanking sequence preferences are lost in transcription factors.

1. This chapter contains unpublished material

3.2 Introduction

3.2.1 Evolution of transcription-factor DNA binding: affinity, cooperativity and gene activation

Proper functioning of gene regulatory networks depends on the distinct DNA specificities of transcription factors. These proteins possess the ability to recognize and bind to specific response elements in the genome, a task that involves discriminating among thousands of other response elements. Even seemingly subtle differences between response elements can form the basis for discrimination. High-throughput experimental studies have shown that great diversity exists in the intrinsic DNA-binding specificities of transcription factors, a diversity that underlies distinct gene regulatory programs [78, 7, 106, 21, 161, 95, 11, 145, 14]. Further, most transcription factors are the products of gene duplication; and yet many paralogous proteins regulate distinct, non-overlapping genes, suggesting the evolution of mechanisms to minimize competition for the same response elements [98, 8, 20, 33, 76, 135, 23]. While the mechanisms and dynamics by which new binding sites in the DNA are created or destroyed by mutations has been fairly well-established [76, 56, 116, 75, 149, 146, 26], little is known about how the extant diversity in protein functions emerged during evolution. How do the DNA-binding specificities of transcription factors evolve over time to produce new DNA gene regulatory programs?

Extant transcription factors employ different physical modes of DNA recognition to achieve specificity, often involving core DNA motifs and flanking sequences. At the mechanistic level, base-specific affinity is largely derived from direct physical interactions between protein side-chains and base-specific functional groups in core DNA motif, often in the major groove [109, 118, 142]. Flanking sequences outside the core motif can also contribute to DNA affinity; it is now being increasingly recognized that different transcription factors can recognize the same core DNA motif, yet regulate distinct genes depending on the com-

position of flanking sequences [28, 58, 131, 78, 7]. These preferences have been shown to be especially important in vivo for development [118, 28, 94, 22, 105, 121]. It is unclear how the affinities for core motifs and flanking sequences are gained or lost during protein evolution. What is the energetic contribution of these interactions, and how do they evolve? In particular, how do the enthalpic and entropic components of binding to the core half-site and flanking sequences evolve to produce proteins with new thermodynamic properties? An additional key thermodynamic component that modulates affinities is cooperative binding, a major determinant of the specificity of binding that causes a sharp switch like effect on gene expression [133, 136, 140, 79, 163, 54, 82]. The individual contribution of the single-site affinity and cooperativity to the overall DNA affinity is poorly understood. Further, how do affinity and cooperativity interact and evolve to generate new transcription factor-DNA specificities? Can novel DNA specificity evolve through a change in either affinity or cooperativity alone, or do both evolve concomitantly? And finally, does novel DNA specificity only evolve through drastic shifts in protein-DNA thermodynamics, or do subtle changes in affinity and cooperativity result in large changes in response element preferences?

These fundamental questions about the evolution of protein-DNA specificity fall at the interface of protein biochemistry and evolution. The knowledge gap in understanding how proteins' physical proteins determine DNA specificity persists because most biochemical studies of proteins have ignored the evolutionary history of the protein [39, 63, 64]. Using ancestral sequence reconstruction (ASR) and biochemical characterizations, ancient evolutionary changes in proteins' sequences, biochemical, and biophysical characteristics can be traced, and the thermodynamics and genetics by which new specificities emerged dissected [64, 69].

3.2.2 Steroid and related receptors

We used ASR to elucidate the genetic and physical basis of the shift in DNA-specificity by uncovering an ancient evolutionary transition in the steroid and related receptors (SRRs), a family of metazoan transcription factors. SRRs are biologically important transcription factors that bind directly to specific DNA response elements through their DBD and activate transcription of nearby target genes. SRRs include the intracellular receptors for steroid hormones that mediate the effects of gonadal and adrenal steroids on reproduction, development, behavior, stress, immunity, metabolism, and homeostasis throughout the vertebrates, as well as other closely related proteins involved in embryonic development and other diverse biological processes [98, 15, 70, 90, 126, 153]. They consist of two paralogous sub-families that differ in their specificity for DNA - steroid receptors (SRs) and estrogen related receptors (ERRs) (Fig. 3.1A).

SRs are ligand activated transcription factors, comprising of receptors for estrogens, androgens, progestogens, and corticosteroids. The SR DBD binds as a dimer to an inverted palindromic DNA sequence consisting of two short 6-bp half-sites separated by a variable 3-bp spacer [98] (Supplementary Fig. 3.6). The DNA-bound crystal structure of a human SR, the Estrogen receptor α DBD (hER α), shows that like the crystal structure of other SRs, it binds as a dimer to a 15-bp response element (EREpal), with the Recognition Helix (RH) of each monomer establishing base-specific contacts with a 6-bp half-site sequence AGGTCA (EREhalf) (Figs. 3.1 B, Supplementary Fig. 3.6) [126]. The first zinc-finger within each DBD monomer recognizes EREhalf within the major groove, while the second zinc finger is responsible for the homodimerization of the two DBD domains upon DNA binding. It has previously been established that all extant SRs descend from an ER-like ancestor with weak-DNA binding cooperativity [98].

ERRs, on the other hand, are orphan transcription factors that do not require a ligand [70]. The ERR DBD has a monomeric 9-bp half-site preference, and full-length proteins

only activate from multiple copies of the 9-bp response element separated by variable spacers (Fig. 3.1 C) [53, 77, 169]. The crystal structure of extant human ERR β DBD (hERR β) bound to the 9-bp DNA demonstrates a striking similarity to hER α DBD — namely, that the RH of the hERR β DBD is hER α like, recognizing the same 6-bp EREhalf sequence in the major groove, AGGTCA (Fig. 3.1C) [53]. There is an additional mode of DNA binding to a 3-bp flanking extension TCA in the minor groove, such that the response element preferred by the protein is a 9-bp extended EREhalf sequence (Ext_EREhalf). The C-Terminus extension (CTE) of the DBD is buried in the minor groove and recognizes the 3-bp flanking sequence through few polar contacts, established using the two Arg residues that extend on the opposite sides of the minor groove (Fig. 3.1 C). While the interactions of the ERR CTE with flanking DNA appear to be base-specific, the arginines enriched in the DNA minor groove imply that electrostatic interactions could also be involved, as observed in the Drosophila Hox protein, Sex combs reduced (Scr), bHLH family of TFs, and other ERR-related receptors such as SF1 and Ftz-F1 [163, 153, 31, 34, 44, 43, 80, 119] . The observation from Chip-SEQ of a monomeric 9-bp DNA preference raises the possibility that ERRs may be monomeric receptors with either no cooperativity in their DBD, or anti-cooperativity, an open research question.

The diversity and similarity in the modes of DNA recognition observed in the two paralogs, ERRs and SRs, make the SRR gene family an ideal one to explore the historical trajectory and molecular mechanisms underlying the evolution of novel DNA specificities involving flanking sequences. Here we use ancestral sequence reconstruction coupled with *in vitro* and *in vivo* functional characterizations of ancestral proteins to investigate how these paralogous families evolved distinct modes of DNA recognition. Our results describe how preferences for flanking sequences change during protein evolution, including a detailed mechanistic understanding of the evolutionary relationships between affinity and cooperativity, and its impact on gene activation and evolution.

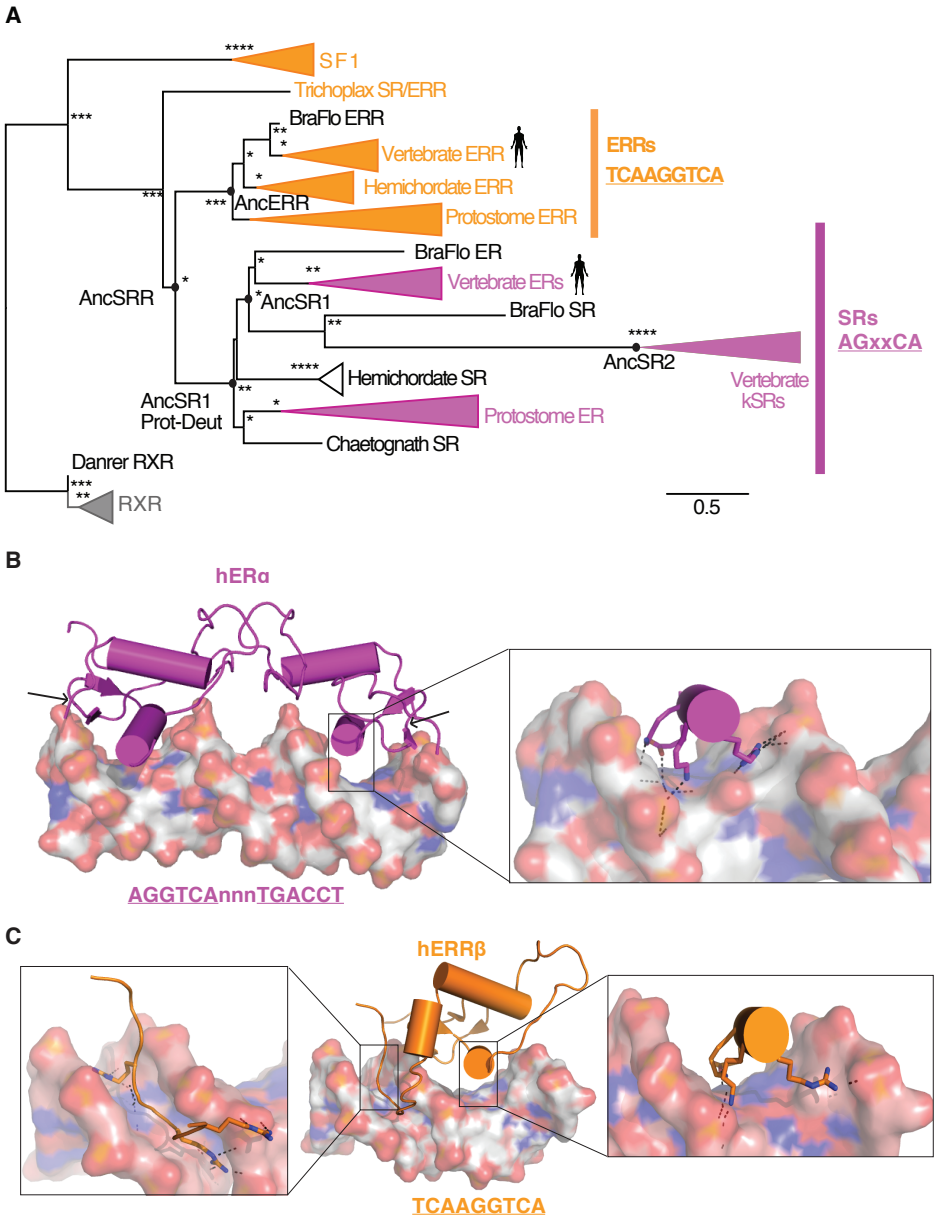


Figure 3.1: **The distinct DNA-binding specificity of ERRs and SRs is retained from the respective ancestors to the present**

.

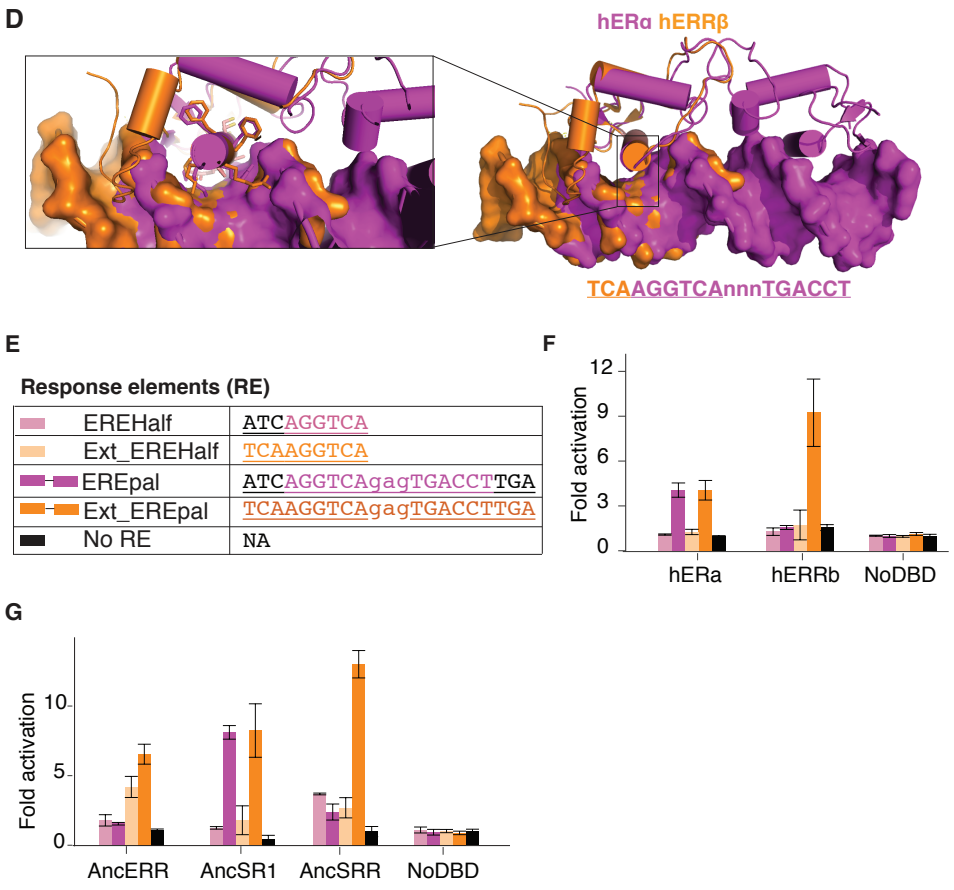


Figure 3.1: (continued) The SRR phylogeny was inferred using maximum likelihood from an alignment of extant SRR sequences across bilateria, and constrained to include syntenic relationships and linked genes. ERRs and SRs were generated by a duplication event at the base of bilateria, and differ in their half-site specificities. Receptors are colored according to the half-site specificities of human proteins in the corresponding clades. See Supplementary Fig. 3.9 for the preference of *Trichoplax adherens* DBD. ERRs bind to a 9-bp half-site response element with TCA extension; SRs bind to a shorter 6-bp response element (shown underlined).

Figure 3.1: (continued) Using maximum likelihood and the AIC-best fit evolutionary model, we reconstructed the DBD+CTE amino acid sequence at three ancestral nodes: 1) the ancestor of all ERRs, AncERR; 2) the ancestor of all chordate SRs (AncSR1); 3) the bilaterian ancestor, (AncSRR). We focused our experiments on the chordate AncSR1 ancestor, as the biochemistry of newly discovered hemichordate receptors and protostome ERs is not very well understood. Also shown are two other nodes discussed in the manuscript (AncSR1 Prot-Deut) and AncSR2. Nodal support is marked by the approximate likelihood ratio statistic, aLRT values (* 1 to 10; ** 10 to 50; *** 50 to 100; **** $>$ 100). Scale bar is in substitutions per site. B. Representative crystal structures of human ER α DNA-binding domain bound to its preferred DNA EREPal. DNA is depicted as surfaces and colored by element. Helices are shown as cylinders. The box shows the 6-bp core half-site in the major groove (EREhalf, shown underlined) recognized by the alpha recognition helix, RH. The zoom-in shows the polar contacts made by the RH with the DNA, depicted as black dashes. The side-chains of residues in the RH is depicted as sticks. The black arrows point to the CTE. C. Representative crystal structure of human ERR β bound to its preferred DNA. Helices are shown as cylinders. ERRs recognize the 9-bp $Ext\_EREhalf$, composed of a similar SR-like 6-bp $EREhalf$. The RH of the ERR DBD is shown boxed, and recognizes EREhalf in the major groove. The zoom-in version of the RH shows the similarity in the mode of binding to ER α DBD. The RH mediates numerous polar contacts with the DNA, depicted as black dashes. The C-terminus extension (CTE) of ERR is boxed and shown alongside the RH. The CTE is buried in the minor groove, contacting the 3-bp flanking extension 'TCA'. A zoom-in view of the ERR β CTE shows the two Arginine residues extend on opposite sides of the minor groove, mediating numerous polar contacts with the flanking extension. The side-chains of residues in the RH and CTE involved in base-specific contacts are depicted as sticks.

Figure 3.1: (continued) D. Overlay of the ERR β DBD and ER α DBD shows that the two structures are very identical, except they diverge in their CTE. The RH of both receptors is zoomed in to show identical contacts in the major groove 6-bp half-site. The side-chains of residues in the RH is depicted as sticks. E. Half-site and palindromic response elements tested in yeast reporter activation assays. REs were integrated into the yeast genome to make stable yeast reporter gene constructs. The ERE and Ext_ERE response elements only differ in their flanking sequences — ERE elements have a random flanking sequence with the same base composition as Ext_ERE elements. F. Fold activation of human ER α and ERR β DBDs on the response elements shown in panel E. Human ERR β DBD only activates from elements with the TCA extension, while the ER α DBD activates from elements with and without the extension. Bars are color coded according to the response elements shown in panel E. Bar heights indicate fold activation relative to the No DBD-No RE control with SEM of three experimental replicates. G. Fold activation of ancestral proteins described in panel A on the response elements shown in panel E. Both AncERR and AncSRR DBD only activate from elements with the TCA extension, while the AncSR1 DBD activates from elements with and without the extension. Bars are color coded according to the response elements shown in panel E. Bar heights indicate fold activation relative to the No DBD-No RE control with error bars showing the SEM of three biological replicates.

3.3 Results

3.3.1 Steroid and related receptors have distinct DNA binding specificities

Although ERRs do not compete with most SR paralogs, there is some reported overlap between the genes regulated by ERs and ERRs [150]. That the RH of ERR β DBD and hER α DBD present nearly identical amino-acids at the DNA interface and bind to the same 6-bp core half-site EREhalf, suggests that they might share transcriptional targets (Fig. 3.1 D). However, the CHIP-SEQ profiles of modern receptors are distinct [78]. We hypothesized from the crystal structures that ERRs, but not ERs, might use their CTE to read flanking sequences, thereby discriminating among the space of response elements bound by the two paralogs (Figs. 3.1B —D).

To test our hypothesis, we looked at the activation of reporter genes by hER α and hERR β DBDs from response elements with different extensions. We transformed human DBDs into yeast reporter strains with genomically integrated response elements fused to the GFP reporter gene, and monitored GFP reporter expression via flow cytometry (Figs. 3.1E, Supplementary Fig. 3.7). The response elements were designed to contain the ERR-preferred extension TCA, and a randomized extension with the same base composition. Consistent with our hypothesis, we found that hER α and hERR β have distinct DNA specificities, distinguished by the presence of the TCA flanks. hERR α binds poorly to and does not activate from response elements lacking the specific TCA flank. hER α does not discriminate between flanks, activating equally well from response elements with the TCA flank and a random flank (Fig. 3.1F).

These data demonstrate that the two human paralogs have evolved distinct DNA specificities based on the flanking sequences. Although extant DBDs can bind to variants of these sequences, the response elements we have tested include physiologically relevant response elements present upstream of genes regulated by the receptors, including those used

to crystallize the DBD-DNA complexes [126, 53, 77, 169] .

3.3.2 Distinct DNA specificity evolved on the SR1 lineage from an ERR-like ancestor

To determine how ERRs and SRs evolved the ability to regulate distinct genes, we first identified the historical interval of evolutionary change using a combination of phylogenetics, ancestral sequence reconstruction and functional characterization of ancestral proteins using our yeast transcriptional reporter assay. The SRR phylogeny was inferred using maximum likelihood from an alignment of extant SRR sequences across bilateria, incorporating new phylogenetically diverse receptors and constraints imposed from syntenic information and linked genes [60, 61, 68] (Figs. 3.1A, Supplementary Fig. 3.8). We reconstructed the DBD+CTE amino acid sequence at three ancestral nodes: the ancestral SRR (AncSRR, the bilaterian ancestor), and its two descendants, Ancestral Estrogen Related Receptor (AncERR), and the ancestor of all chordate SRs (AncSR1). Strong to moderate statistical support was evident in our ancestors, as expressed in terms of the mean posterior probability over sites. (AncSRR, $PP = 0.88$, AncERR, $PP = 0.96$, and AncSR1, $PP = 0.87$) (Supplementary Fig. 3.8). More importantly, these reconstructions contain 4, 12, and 14 sites, respectively, that are ambiguously reconstructed (secondary states with $PP \geq 0.2$). There is no ambiguity in the RH.

The phylogeny of SRRs, together with the extant structures, indicates that SF1 receptors are the closest outgroup clade to the ERRs. SF1, like the ERR, shares the same structural preference for the TCA extension (Fig. 3.1A) [90, 153]. Activation assays with our yeast reporter strains indicated that the *Trichoplax adherens* DBD, the simplest metazoan DBD, also had a preference for the TCA extension (Supplementary Fig. 3.9). The most parsimonious hypothesis that follows from the phylogeny is that ancestral SRR was ERR-like, with a preference for the TCA extension lost on the lineage leading to AncSR1. We conducted

reporter activation assays with our yeast reporter strains, and found that both AncSRR and AncERR activated specifically from palindromic response elements containing the TCA extension (Figs. 3.1E,G). AncSR1 activated from both types of palindromic elements, with and without the extension, suggesting that the preference for the extension was lost on this lineage. Ancestral preferences were conserved to the present, on the lineages leading to hERR β, and hER α, respectively (Figs. 3.1F, G). Across ancestors and extant proteins, little to no activation was observed on half-site elements (Figs. 3.1E,G).

Together, these results are consistent with our parsimonious hypothesis, and indicate that the presence of a specific 3-bp flanking sequence can have a large effect on recognition and activation of target genes. Transcription factors therefore read modifications of flanking sequences very effectively.

3.3.3 Thermodynamic basis of the switch in DNA-specificity — AncSR1 gained macroscopic binding affinity on EREPal

Our flow cytometry results demonstrated distinct DNA preferences for the ancestral proteins on response elements with and without the extension, suggesting that there might be a difference in the binding affinities of ancestors on these target elements. We therefore sought to determine the thermodynamic basis underlying the switch in DNA specificity on the lineage leading to SR1.

We began by estimating the macroscopic binding affinity of ancestors, $K_{A,Mac}$, using electrophoretic mobility shift assays (EMSAs). The $K_{A,Mac}$ describes the protein's macroscopic binding affinity on a full two-site response element, defined as the product of the binding affinity of each monomer on each half-site, which includes any cooperativity between sites.

We found that the $K_{A,Mac}$ of AncERR on Ext_EREPal was about 30-fold higher than EREpal, consistent with its specific activation from Ext_EREPal (Figs. 3.2A, B). The deepest ancestor, AncSRR, exhibited a moderate 4-fold preference for Ext_EREPal compared to

ERE pal; but this ancestor also specifically activated from Ext_EREPal, indicating that even moderate quantitative differences in thermodynamic parameters can result in a difference in activation (Fig. 3.1G). On the lineage leading to AncSR1, this preference for the extension was reduced to 1.7-fold with an increase in $K_{A,Mac}$ on EREpal; Consistent with this, we found that the close affinities of binding on the two response elements led AncSR1 to activate from both types of extensions in Ext_EREPal and EREpal (Figs. 3.1G, 3.2A, B).

A caveat in our EMSA experiments is that they are analyzed assuming equilibrium, though this assumption can be violated under certain conditions. We therefore performed equilibrium isothermal titration calorimetry (ITC) experiments on AncERR and AncSR1 to obtain independent estimates of $K_{A,Mac}$ (Figs. 3.2C — F). Consistent with EMSAs, we found that AncERR has a 30-fold preference for Ext_EREPal compared to EREpal (Figs. 3.2C, D, G, H). The difference in $K_{A,Mac}$ on the two response elements corresponds to a $\Delta\Delta G$ of -2.1 +/- 0.6 kcal/mol, suggesting that about 1 kcal/mol, could be obtained solely by contacting the TCA extension on each half-site of Ext_EREPal. ITC experiments on half-site REs further corroborated this result (Supplementary Fig. 3.10) . In contrast, AncSR1 had no preference for the extension, binding with similar $K_{A,Mac}$ on both elements (Figs. 3.2E — H).

The EMSA and ITC results together suggest the evolution of a new $K_{A,Mac}$ on EREpal following the duplication event. The preferences of post-duplication ancestors from ITC data are consistent with that of EMSAs, indicating that a simple assay like EMSA can still yield reasonable quantitative trends for investigating protein-DNA thermodynamics and its evolution (Supplementary Fig. 3.11). In our ITC experiments, the change in the enthalpy for the deepest ancestor AncSRR was too low to obtain reliable thermodynamic estimates for meaningful comparisons and a polarization of these energetics. However, the EMSA and ITC data together, yield thermodynamics that are consistent with the ancestral activation profiles (Figs. 3.1G, 3.2B, 3.2H).

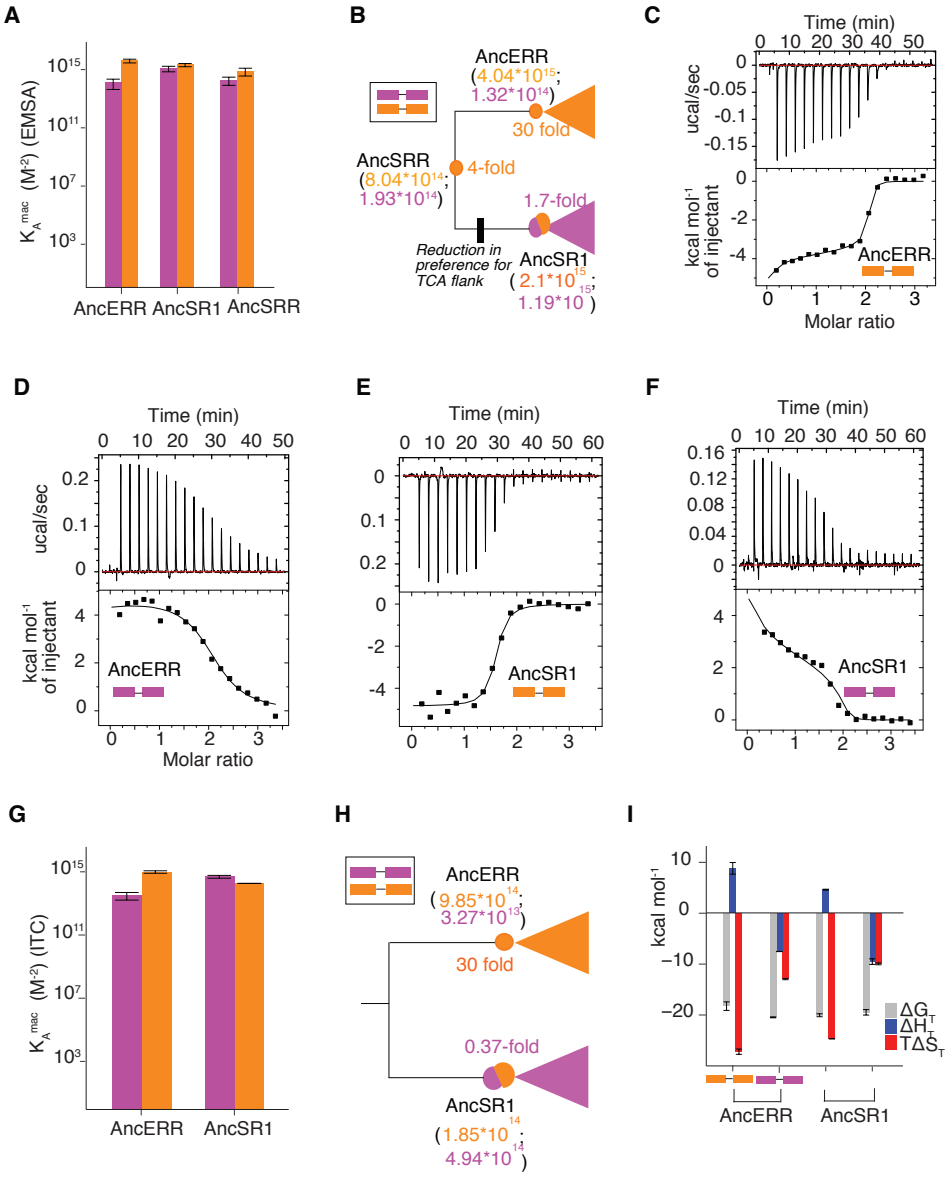


Figure 3.2: **AncSR1 gained macroscopic binding affinity on EREpal**

3.3.4 Evolution of enthalpy and entropy post-duplication

In addition to providing equilibrium ΔG estimates of protein-DNA binding, ITC data also provide valuable information on the mechanism of binding for different response elements by deconvolving the free energy of binding into the enthalpic and entropic components. We asked whether AncERR and AncSR1 differed in their mechanism of binding on EREpal and Ext_EREPal.

Figure 3.2: (continued) A. Estimates of $K_{A,Mac}$ of ancestors on two-site DNA elements Ext_EREpal and EREpal as tested by EMSA. Error bars show mean +/- SEM from triplicate experiments. B. Summary of evolution of flanking sequence preference from the $K_{A,Mac}$ determined from the EMSA fits. The estimated mean macroscopic binding affinities, $K_{A,Mac}$ (units of M^{-2}), of dimers on Ext_EREpal (orange) and EREpal (pink) are mapped on the SRR phylogeny to show the evolutionary trajectory. A cartoon of response elements is shown in the box on the left, color-coded according to Fig. 3.1E. Data show mean $K_{A,Mac}$ from panel A. AncERR and AncSRR prefer extension, while AncSR1 does not. C-F show representative ITC thermograms for post-duplication ancestors, AncERR and AncSR1 tested on Ext_EREpal and EREpal. The heat of dilution of the protein into buffer has been subtracted out by performing control experiments of protein into buffer. G. Estimates of $K_{A,Mac}$ from ITC for AncERR and AncSR1 binding to Ext_EREpal and EREpal. Error bars show mean +/- SEM from two replicates. H. Summary of evolution of flanking sequence preference from the $K_{A,Mac}$ determined from the ITC fits. The estimated mean macroscopic binding affinities, $K_{A,Mac}$ (units of M^{-2}), of dimers on Ext_EREpal (orange) and EREpal (pink) are mapped on the SRR phylogeny to show the evolutionary trajectory. The cartoon of response elements is shown in the box on the left, color-coded according to Fig. 3.1E. Data show mean $K_{A,Mac}$ from panel G. AncERR prefers extension, while AncSR1 does not. I. The decomposition of total free energy, ΔG_T, into the total enthalpic (ΔH_T) and total entropic ($T\Delta S_T$) components of binding for two-site response elements. The x-axis depicts the cartoon of response elements tested, Ext_EREpal and EREpal, for AncERR and AncSR1.

We found that in all cases, binding was entropically favorable. It was enthalpically unfavorable when there is no extension, but enthalpically favorable when there is a TCA extension (Fig. 3.2I). AncSR1's binding to EREpal was less enthalpically ufavorable than AnERR's binding to EREpal, with a $\Delta\Delta H$ of -4.1 kcal/mol ($\Delta\Delta H = \Delta H\ AncSR1_{EREpal}$ - $\Delta H\ AncERR_{EREpal}$) (Fig. 3.2I).

This result suggests two possible scenarios during evolution: i) either AncSR1 evolved substitutions that decrease the enthalpic penalty for binding EREpal, or, ii) substitutions that increase the enthalpic penalty for EREpal-binding evolved on the AncERR lineage. The ITC data do not allow a polarization of the directionality of evolution. However, the activation profiles of ancestors show that AncSRR was ERR-like, making scenario ii) less-likely (Fig. 3.1G). Therefore, AncSR1 likely enhanced its affinity and activation on EREpal through the evolution of substitutions that reduce the unfavorable enthalpy for EREpal

binding (Fig. 3.2I). The evolution of a predominantly major groove binder could have entailed a reduction in the unfavorable enthalpy of binding on this groove post-duplication, even though binding is still primarily entropy-driven, a result consistent with the general observation that favorable enthalpy drives proteins to bind the major groove [113].

3.3.5 Loss of preference for the extension entailed the strengthening of half-site binding affinity

Evolution of a new $K_{A,Mac}$ could involve changes in half-site affinity, cooperativity, or both. To determine if the loss of preference for the extension on the SR1 lineage was enabled through a change in half-site affinity or cooperativity, or both, we conducted half-site binding experiments using EMSA to partition the $K_{A,Mac}$ into the two components — half site binding affinity, and cooperativity.

We first estimated the half-site affinities of all ancestors on half-site response elements, EREhalf and Ext_EREhalf (Fig. 3.1E). Both AncSRR and AncERR preferred the extension by 3-fold and 9-fold, respectively (Figs. 3.3A, B, D, E). In contrast, the ancestral preference for the extension was reduced to 1.5-fold on the lineage leading to AncSR1. The EMSA data also show a 3-fold gain in the affinity for EREhalf on AncSR1 lineage, suggesting additional shifts in function along the SR lineage — Despite apparently reducing preference for the extension, AncSR1 still binds to its target sequences with affinity similar to that of AncSRR on $Ext\_EREhalf$ (Fig. 3.3E).

These results are consistent with the inferences drawn from palindromic response elements, and imply a scenario where AncSR1 enhanced its affinity on EREPal through strengthening its half-site affinity on EREhalf. This way, the AncSR1 DBD compensated for the ancestral loss of affinity from specifically preferring the TCA extension (Figs. 3.2B, 3.3E). Our inferences are robust to uncertainty in the reconstructions, tested by incorporating all ambiguous sites (defined as $PP \geq 0.2$) for each ancestor onto its maximum likelihood

reconstructed sequence (Supplementary Fig. 3.12) . Further, the ancestral preferences for the extension, and loss thereof, are retained to the present on the lineages leading to human ERR β and human ER α, respectively (Supplementary Fig. 3.13).

3.3.6 Loss of preference for the extension was not accompanied by a change in cooperative binding

Loss of recognition of the 5' extension would reduce the specificity of the SR1 transcription factor's half-site motif to 6 bases from 9; But extant SRs bind 12-bp palindromic DNAs as dimers — which should increase the specificity of their binding dramatically. It has been previously established that extant SRs descend from an ancestor with little to no cooperativity [98]. The observation that ERRs have a monomeric preference for Ext_EREhalf, raises the possibility that either ERRs are like SRs, with little to no cooperativity, or anti-cooperative receptors [53, 77].

To test if the loss of preference for the extension was accompanied by the loss of anti-cooperativity on the AncSR1 lineage, we estimated the cooperative binding constants of ancestors using EMSAs . We jointly estimated the half-site binding affinity K_A, and the cooperativity constant, ω, by performing a global fit of the half-site and two-site data using a modification of statistical thermodynamic approaches developed by Senear and Brenowitz [127] (Fig. 3.3F). We found the K_A estimated from the global fit to be comparable to from the half-site data alone, indicating the two-site EMSA data do not generate vastly different half-site affinities (Figs. 3.3D, G).

To verify that our global fit approach can detect strong cooperative binding when it exists, we first tested it on a system known to have strong cooperative behaviour — the binding of ancestral keto-steroid receptor protein (AncSR2) on its consensus motif, the steroid response element [98, 15, 73, 37] (Fig. 3.1A). We recovered a very strong cooperativity constant of ($\omega > 14$) consistent with previous publications, suggesting that our approach can detect

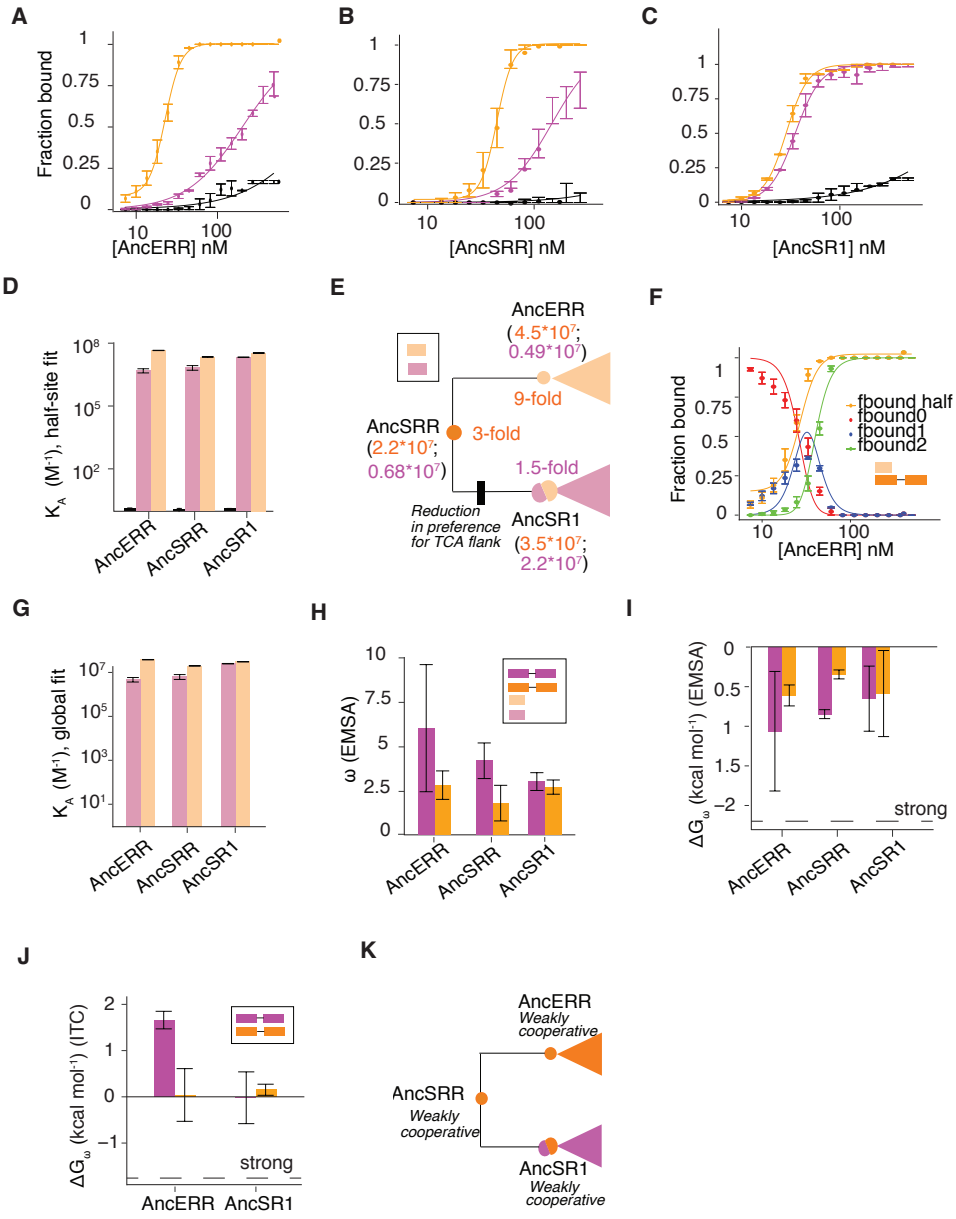


Figure 3.3: **New DNA-specificity evolved through changes in half-site affinity and not cooperativity. AncSR1 gained affinity on the half-site motif EREhalf, with no novel evolution of cooperative binding**. A-C. Model fits show binding of ancestors on half-site DNA elements, EREhalf (magenta) and $Ext\_EREhalf$ (orange), as tested using EMSA for estimating K_A. Black trends are for control oligos, composed of a randomized sequence with the same base composition as the half-site oligos. D. Half-site K_A as estimated from the fits shown in A-C. Error bars show mean +/- SEM of three experimental replicates. Bars are colored according to the color for half-site response elements described in Fig. 3.1E.

Figure 3.3: (continued) E. Summary of the evolution of flanking sequence preference from half-site experiments. The estimated half-site affinities, K_A (units of M^{-1}), of monomers on $Ext\_EREhalf$ (orange) and EREhalf (pink) are mapped on the SRR phylogeny to show the evolutionary trajectory. The cartoon of response elements is shown in the box on the left, color-coded according to Fig. 3.1E. Data show mean half-site affinities from panel D. AncSRR and AncERR prefer extension, with a reduction in this preference on the lineage leading to AncSR1. Bars are colored according to the color for half-site response elements described in Fig. 3.1E. F. Representative global fit of half-site and two-site DNA binding data of AncERR bound to its target elements using EMSA. *fboundi* represents the fraction of DNA bound to exactly 0,1 or 2 ligands. $fbound\_half$ is half-site data from panel A. G. Half-site binding affinity K_A, plotted on the log10 scale, as calculated from the global fit of two-site and half-site data. Global fit K_A estimates agree with those estimated from half-site alone, shown in panel D. Bars are colored according to the color for half-site response elements described in Fig. 3.1E. H. Cooperativity constant, ω, estimated from the global fit of half-site and two-site data using statistical thermodynamic models. The boxed inset shows the response elements tested for estimating ω. I. ΔG_ω estimated from EMSA data shown in H. Dotted line shows the cooperativity (ΔG_ω) of AncSR2, a strongly cooperative protein, as estimated using EMSA. J. ΔG_ω estimated from ITC experiments on AncERR and AncSR1. Dotted line shows the cooperativity (ΔG_ω) of AncSR2, a strongly cooperative protein, as estimated using ITC. The boxed inset shows the response elements tested for estimating ω. K. Proposed trajectory for cooperative binding, as inferred from panels H and J. All ancestors are weakly cooperative.

strong cooperative binding when it exists [98] (Supplementary Fig. 3.14A). We found AncSR2's strong cooperative constant to stand above experimental uncertainty and imprecision in the estimates of K_1 and K_2.

Using the global fit approach on EMSA data, we found that all ancestral proteins — including AncSR1 were weakly cooperative on palindromic DNA elements (Fig. 3.3H). These DBDs did not inherently possess a kSR level of cooperativity, with average ω values ranging from 2 to 6 corresponding to $\Delta G's$ of small effect (Fig. 3.3H, I). The weak DNA-binding cooperativity of ancestral proteins was also found to be conserved to the present, on the modern lineages leading to ERs and ERRs (Supplementary Figs. 3.14B—D). That the human ERR β DBD is weakly cooperative just like the ER α DBD is contrary to the non-cooperative expectation that follows from previous studies [53, 77]. To rule out any artifact of model fitting on the estimates of cooperative constant, we analyzed the AncERR data using an alternate approach that can quantify the cooperative constant from monomer bands alone, capitalizing on the unique strength of EMSAs over other assays (Supplementary Fig. 3.16). Using this approach, we recovered an ω value of 3.16 for AncERR, further corroborating our inferences from the global fitting approach for this protein.

We also analyzed the ITC data of AncERR and AncSR1 binding to Ext_EREPal and EREpal to deduce orthogonal estimates of cooperative constants. We found that ITC can detect strong DNA-binding cooperativity in AncSR2 directly from the estimates of ΔG_1 and ΔG_2 obtained from a two-site ITC experiment, with an inference of cooperativity constant comparable to that of EMSAs and known kSR literature (Supplementary Fig. 3.14 E). Using this simple approach, we found that AncSR1 is weakly cooperative on both response elements, Ext_EREPal and EREpal, a result consistent with EMSA (Figs. 3.3I, J). AncERR also has weak DNA-binding cooperativity on Ext_EREPal, consistent with inference from EMSA. On EREpal, the AncERR DBD has some anti-cooperativity, not observed in our EMSA experiments (Fig. 3.3J). We reasoned that the two-site ITC data by itself could

sometimes contain insufficient information to uniquely disentangle two very similar half-site affinities and enthalpies; the estimate of cooperative constant could be unreliable in these cases. To test if the anti-cooperativity of AncERR is real or the potential inability of the model to uniquely disentangle parameter estimates, we combined with $K_{A,Mac}$ the ΔG_1 estimated separately from ITC experiments of AncERR binding on EREhalf to estimate the cooperative constant (Supplementary Fig. 3.9). This analysis led to a very weak positive DNA-binding cooperativity that is consistent with EMSA, suggesting that the negative cooperativity was associated with non-unique parameter values (Supplementary Fig. 3.14F).

Taken together, these results imply that the evolution of derived preference in AncSR1 was not accompanied by the loss of anticooperativity in AncSR1, All ancestors are weakly cooperative. Loss of extension preference was therefore only accomplished by a gain in half-site affinity in the SRs (Fig. 3.3K).

3.3.7 Genetic basis of the switch in DNA specificity

Next, we sought to identify the evolutionary genetic changes that were sufficient for the shift in function on the AncSR1 lineage. To this end, we investigated the genetic mechanisms underlying the evolution of loss of preference for the extension, and gain of activation from EREpal.

The crystal structure of hERR β DBD:DNA binding suggests that a modular specificity could exist — with the specificity for the half-site EREhalf encoded in the DBD, and the specificity for the 3bp extension TCA encoded in the CTE [53] (Fig. 3.1C).

There are 16 amino acid differences between AncSRR and AncSR1, 3 in the CTE, and 13 in the DBD (Fig. 3.4A). We first investigated the genetic changes underlying the loss of preference for the extension. There are only 3 changes in the CTE on the lineage leading to AncSR1, e75G, q87T, y89R (lowercase denotes ancestral amino acids; upper case denotes derived states) (Fig. 3.4A). All 3 residues are strictly conserved in the ERRs, but not in the

ERs. When we introduced all the derived CTE changes into AncSRR (AncSRR+3CTE), there was still preference for the extension, indicating the involvement of other substitutions in the DBD, or the non-modularity of this function (Fig. 3.4B). We also observed a small increase in activation on EREpal, without a concomitant increase in activation on Ext_EREPal.

We then sought to determine the genetic changes that were sufficient for the gain of activation from EREpal. There are 13 changes between the derived and ancestral DBD proper (Fig. 3.4A). Based on patterns of conservation and sturctural positioning, we introduced 3 of these 13 derived DBD changes onto the AncSRR background, together with the 3 previously tested derived CTE changes (AncSRR 3DBD+CTE). These three residues in the DBD, l22W, t34S and k50T are fully conserved in modern ERs and we hypothesized that they might be important for binding and activation on EREpal. We found that this AncSRR construct, with 3DBD+CTE changes, activated from EREpal, with a fold activation comparable to that of AncSR1 (Fig. 3.4B). This demonstrates that the 3 DBD substitutions that evolved on the lineage to AncSR1 together with the CTE changes, were sufficient for the shift in activation on EREpal. We did not observe a concomitant increase in activation on Ext_EREPal, suggesting either epistasis between the DBD and CTE substitutions, or that our activation assay reached a maximum threshold.

The three derived DBD substitutions are located at critical positions in the protein. I22W is adjacent to several backbone-contacting residues, binding just outside the core RE (Fig. 3.4C). It presumably works by increasing the affinity on EREhalf, by positioning the RH on the 6 bp half-site. The other two substitutions, t34S and k50T, are located in the RH, and on the loop just past the dimer interface, respectively. These substitutions could further enhance half-site affinity and dimer stablization on EREpal (Fig. 3.4C).

One might have expected from the crystal stucture of human ERR β:DNA binding that the CTE should be very important for specifying preference for the extension. Contrary

to this expectation, our results show that on the historical interval between AncSRR and AncSR1, the derived CTE changes by themselves do not encode a modular preference for the extension. If the CTE substitutions by themselves were not important for the loss of preference for the extension, what is the role of the CTE for DNA-binding in ERRs and SRs? To investigate this more broadly, we created CTE truncation mutants of AncERR and AncSR1, and tested their ability to bind half-site and palindromic response elements using EMSA. We found that the CTE is required for DNA binding in AncERR and AncSR1, on every DNA element tested (Supplementary Fig. 3.15). It has been previously shown that the CTE is required for DNA binding by both human ERR β and hER α DBDs, suggesting that the ancestral dependence on the CTE is conserved on the lineages leading to the present [53, 100]. AncSR1 CTE truncation binds with a very weak $K_{A,Mac}$ to EREpal and Ext_EREPal, suggesting that the CTE is still required for high affinity binding. The dependence on the CTE for high affinity binding is therefore a conserved feature in SRR nuclear receptor family.

Our genetic experiments demonstrate that binding and activation from the extension is a non-modular feature of SRRs. As few as six changes in the DBD and CTE were sufficient to recapitulate the gain in activation from EREpal. We hypothesize that these substitutions work by enhancing the 6-bp half-site affinity to producing predominantly major groove binders from an ancestral mode of binding to both the major and minor groove flanking extension.

3.4 Discussion

Our findings indicate that DNA specificities evolved in ERRs and SRs from an ancestral ERR-like receptor. We can now begin to map the evolutionary trajectory of SRRs on the phylogeny, depicting information on structural preferences of the nuclear receptors, their mode of DNA binding and gene activation (Fig. 3.5A).

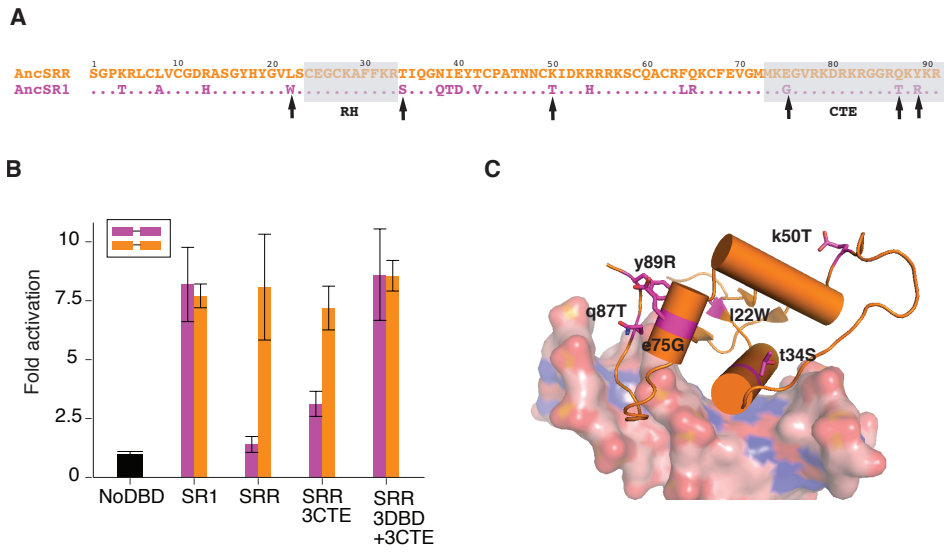
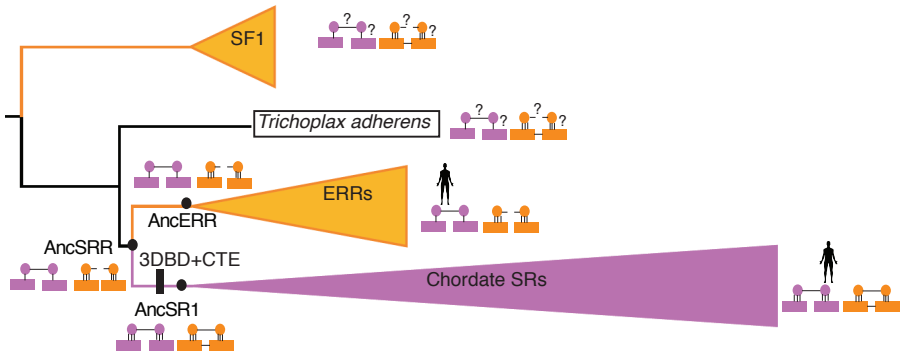


Figure 3.4: **Three mutations in the DBD along with the derived CTE are sufficient to explain the derived specificity**. A. Alignment of AncSRR and AncSR1 sequences. Only residues that are different between the two sequences are shown. Arrows show the substitutions tested in the genetic experiments, in the core DBD and the CTE. The RH and CTE are colored grey. Resides numbering is indicated on top. B. Fold activation of ancestral WT DBDs and select mutants on the palindromic response elements depicted in the box on the top left. Bars are color coded according to the response elements tested, as shown in Fig. 3.1E. SRR with 3 mutations in the DBD and derived CTE recapitulates fold-activation profile of the derived protein AncSR1. Bar heights indicate fold activation relative to the No DBD-No RE control with SEM of three experimental replicates. C. The position of derived substitutions shown in panel A, is mapped on the human ERR β structure.

A



B

	Preference for the TCA extension			
	ITC (K_A^{mac})	EMSA (K_A^{mac})	EMSA (K_1)	Activation in yeast
AncERR	+	+	+	+
AncSRR	na	+	+	+
AncSR1	-	-	-	-

Figure 3.5: **Summary of the evolutionary trajectory of SRRs**. A. AncSRR and AncERR bind with strong affinity and weak cooperativity to elements with the TCA extension. The affinity and cooperativity are depicted by vertical lines connecting the monomers to the half-sites, and protein-protein interface, respectively. The dotted lines connecting the protein-protein dimers on Ext_EREpal elements represent the rare nature of these response elements; very few genomic sites should have this structure; the most likely binding event might be to direct repeats of Ext_EREhalf. The preference of AncSRR is retained to the present on the lineage leading to human ERR β. AncSR1 evolved to bind to elements without the TCA extension, a transition that entailed six mutations, three in the core DBD, and three in the CTE. The weak ancestral DNA-binding cooperativity of AncSRR was however retained in AncSR1, so this transition was only achieved through changes in the half-site affinity. There was no overall change in the affinity of AncSR1 on its target response elements, with the derived protein binding equally well as AncSRR on Ext_EREPal elements. The preference of AncSR1 is retained to the present, on the lineage leading to human ER α. The outgroup receptor clade SF1, and *Trichoplax adherens* DBD are ERR-like, with unknown single-site affinity and cooperativity. Binding and activation from EREpal is a derived feature that evolved on the SR lineage. Palindromic response elements are depicted as a cartoon are colored according to Fig. 3.1E. B. Summary of results pooled from four different experiments support the trajectory shown in panel A. Preference for the extension is indicated as a '+'; '-' indicates no preference; 'na' indicates lack of estimate for the deepest ancestor AncSRR, from ITC experiments.

The SF1 receptor clade, the clade closely related to ERRs, has a similar mechanism of DNA binding to that of ERRs, with a preference for the TCA extension [90, 153, 53]. Consistent with its structural preference, it has been shown to activate from elements with the specific extension. The DNA-binding cooperativity and affinities of SF1 have yet to be characterized (Fig. 3.5A).

The function of the *Trichoplax adherens* SR/ERR gene was unknown until this study. The extant diversity of SRs and ERRs can be traced to this single SR/ERR gene at the base of the bilateria, and subsequent duplications have diversified the functional repertoire of SRR genes. The Trichoplax SR/ERR gene co-orthologous to both ERRs and SRs; so it could be either SR or ERR-like. We distinguished between these possibilities using yeast activation assays and found the DBD was ERR-like, only activating from response elements with the TCA extension (Fig. 3.5A). This result, together with our previous results demonstrating that AncSRR was ERR-like, activating from response elements with the TCA extension, suggests that the gain of activation from EREpal is also a parsimoniously supported derived feature on the SRR phylogeny that evolved on the lineage leading to AncSR1.

Using a combination of different in vitro and in vivo approaches such as mobility shift assays, ITC and yeast activation assays, we propose that the most-likely evolutionary trajectory involved the duplication of an ERR-like ancestor that preferred binding to the TCA extension (Figs. 3.5A, B). This preference was retained on the lineage leading to extant ERRs, but lost on the lineage leading to extant SRs. Human proteins retain the ancestral preferences. The genetic experiments suggest that as few as three derived substitutions in the derived DBD together with the CTE enhanced the affinity on EREpal, likely through an increase in the favorable enthalpy of binding on the 6 bp-half-site (Figs. 3.2B, 3.3E, 3.5B). We found that these derived substitutions are also present in the AncSR1 protostome-deuterostome ancestor, suggesting that the loss of dependence on the flanking sequences evolved before protostome and deuterostome SRs diverged from one another (Fig. 3.1A). Immediately after

duplication, there could have been competition for palindromic response elements with the TCA extension by the AncERR and AncSR1 (Fig. 3.2B). After duplication, the affinity of binding on response elements with the TCA extension and without the extension were strengthened on the lineages leading to AncERR and AncSR1 respectively, such that post-duplication ancestors now retain distinct specificity on their preferred response elements (Figs. 3.2B, 3.3 E).

This shift only entailed an increase in affinity on EREhalf, with no changes in cooperative binding. All ancestral DBDs, including AncERR, and human ERR were found to be weakly cooperative on palindromic response elements. The ERR DBD has a 9-bp monomeric preference, and in this study we ruled out the possibility that ERRs might be non-cooperative. However, the observation that the ERR DBD is weakly cooperative, just like the ER DBD raises a conundrum: why do the CHIP-SEQ experiments and protein binding microarrays not recover a dimeric palindromic invert repeat element of the structure of Ext_EREPal for the cooperative ERR DBD? We suggest that the most likely explanation is the availability of binding sites of the structure of Ext_EREPal. Such 21-bp sites will not be common in the genome, as opposed to the availability of 9-bp Ext_EREhalf response elements. ERRs appear to have the ability to bind and activate from independent Ext_EREhalf elements as monomers, or as weakly cooperatively as dimers when the right response elements and spacings are presented. Our choice of Ext_EREPal has been motivated by the question of cooperative binding of the DBD on a palindromic inverted repeat response element.

Extant transcription factors bind distinct DNA-sequences, but little is known how the key determinants of DNA specificity — namely the single-site affinity and cooperativity of binding, interact and evolve to generate proteins with new thermodynamic properties and response element activation. In this study, we mechanistically dissected one such transition involving ERRs and SRs, by pursuing an integrative approach combining phylogenetics, ancestral protein reconstructions, thermodynamics and gene activation to show how distinct

DNA specificities evolve, with a focus on the loss of flanking sequence preference during protein evolution. Whether the transition in protein function preceded, followed, or accompanied the evolution of EREpal binding sites is unknown Our results have implications for computational prediction of such preferences, and understanding how new gene regulatory systems emerge from ancestral systems.

3.5 Methods

3.5.1 Phylogenetics and ancestral sequence reconstructions

All annotated steroid and related receptor gene sequences were downloaded from Uniprot/Swissprot, Genbank Joint genome institute and flybase databases. 223 total receptors were chosen to minimize redundancy and maximize coverage across the metazoan phylogeny. Eight new receptors were also included (gift from Paul Cziko), which comprised of newly discovered hemichordate SRs, the basal deuterostome SR (*Chaetognath*), protostome ERs and basal deuterostome ERRs. We removed the hinge region and aligned the 82 amino-acid DBD, the 20 amino-acid C-terminus extension and the LBD using the Multiple Sequence Alignment by Log-Expectation (MUSCLE) program [47]. The alignment was manually checked to remove sequences that had $< 80\%$ coverage of the DBD, lineage specific indels, or highly divergent sequences with propensity for long-branch attraction. The ML phylogeny was inferred from the alignment using PHYML v3 [61] 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 [36]. Statistical support for each node was evaluated by the approximate likelihood ratio criterion. The ML Tree was constrained according to the chordate duplication hypothesis to incorporate information from genome-wide synteny and linked genes (https://scholarsbank.uoregon.edu/xmlui/handle/1794/18733). This ensured

the correct relationships between the cyclostome SRs and newly discovered hemichordate SRs. Ancestors were reconstructed using the maximum likelihood method implemented in the PAML software 4.7 in the codeml program [157]. The distribution of posterior probabilities is from all sites in the DBD and the CTE. To determine the robustness of inferences made from ML ancestors to statistical uncertainty in the reconstruction, we identified all plausible alternate reconstructions — sites with alternate posterior probability $PP > 0.2$, and introduced all of them together on the ML reconstruction.

3.5.2 Protein purifications for EMSAs

Ancestral and human ERR β cDNA sequences were synthesized from GeneScript and subcloned into the pETMALc-H10T vector C-terminal to a cassette containing a 6xHis tag, maltose binding protein (MBP) and a TEV protease cleavage site. Human ER α and *Trichoplax adherens* DBDs were a generous gift from Jamie Bridgham. For EMSAs, ancestral and extant proteins were purified according to protocols in the previously published paper with minimal modifications [140]. DBDs were expressed in BL21(DE3) pLysS Rosetta cells. Protein expression was induced by first beginning a starter culture and then adding 1 mM IPTG at A600 of 0.8 —1.2 to a litre of large growth culture the following day. After induction, cells were grown overnight at 15C. Cells were harvested via centrifugation and frozen at -20C overnight in polyprene bottles. The next day, cells were lysed for about 40 mins on ice using B-PER Protein Extraction Reagent Kit without the need for sonication (ThermoScientific). The clear Lysate was loaded onto a pre-equilibrated 5ml HisTrap HP column (GE) and eluted with a linear imidazole gradient (25mMto 1M) in 25mMsodium phosphate and 100mM 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 bME 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 (150mMto 1M) in 25

mM sodium phosphate buffer [pH 8.0]. DBDs were further purified on a Superdex200 10/300 GL size exclusion column (GE) with 10 mM Tris [pH 7.6], 100 mM NaCl, 2 mM bME, 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). Purified protein aliquots were flashed frozen in liquid nitrogen and stored at -80C. Protein folding and aggregation states were measured using circular dichroism (Jasco J-1500 CD) to make sure that they were in good biophysical conditions. For EMSAs, protein aliquots were thawed, spinned and the supernatant was used for Quick StartTM Bradford Assay (Bio-Rad) to quantify protein concentrations.

3.5.3 Protein purifications for ITC

For ITCs, we needed high concentration of proteins (500 uM to 1mM), so proteins were purified using a protocol very similar to that used for EMSA, but optimized to increase the concentration. We incorporated the following modifications: In the Ni-NTA purification using the HisTrap column, we increased the salt concentration to remove non-specific binding and added some glycerol and TCEP reducing agent to favor the protein. The following buffers were used: His Buffer A 25 mM Phosphate buffer, 0.5 M NaCl, 25 mM Imidazole, pH 7.6, 10% Glycerol, 1 mM TCEP; His Buffer B 25 mM Phosphate buffer, 0.5 M NaCl, 25 mM Imidazole, pH 7.6, 10% Glycerol,1 mM TCEP. In the TEV cleavage step, a simple dilution step was used to replace the dialysis step, in order to reduce dramatic protein precipitation during dialysis. The FPLC was run at 2ml/min with linear elution. We diluted the Ni-NTA fractions 4x using Heparin Buffer A, and added 5M NaCl to adjust the salt concentration to a final 250 mM. TEV protease was added and the protein left at 4 deg overnight. We replaced the cation exchange column with heparin purification, an affinity purification to remove his-MBP tag and other impurities. Different salt concentrations for binding to the column needed to be optimized for ancestral DBDs. The following buffer compositions were

used: Heparin Buffer A 25 mM Phosphate buffer, pH 7.6, 1 mM TCEP; Heparin Buffer B 25 mM Phosphate buffer, 1M NaCl, pH 7.6, 1 mM TCEP, and the FPLC run at 2 ml/min with linear elution. We finally concentrate fractions from Heparin elution and load using 1 or 2 ml loop onto a gel filtration column by S75 to remove any aggregates and trace impurities. Gel filtration Buffer: 20 mM Tris, 130 mM NaCl, pH 7.6, 1 mM TCEP. FPLC run at 1 ml/min.

3.5.4 Response element preparation for EMSA and ITC

Half-site and palindromic DNA constructs were ordered from Eurofins Operon as salt-purified single stranded oligos with the forward strand labeled at the 5-end with 6-FAM. Sequences of the oligos are presented in Fig. 3.1E, and we ordered their forward and reverse strands. A 20-bp flanking sequence (AATTGCAACATTACACATCT) was added to prevent hairpin formation and clean annealing of oligos. This sequence was the same for all oligos, including that of control experiments, and no binding was detected to the flanking sequence in our control experiments. 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 mM. Equimolar quantities of complementary forward and reverse strands were combined and placed in a 95C heat block for 10 min then slowly cooled to room temperature. The double stranded product was diluted to 5 uM in water, and stored in 40 uL aliquots at -20 degrees in UV resistant microfuge tubes. On the day of the EMSA experiment, the oligos were dissolved to 5nM concentration in the EMSA buffer.

For ITC, 0.2 umol HPLC purified ssDNA corresponding to all half-site and palindromic response elements were ordered from Eurofins. Response elements were prepared fresh on the day of the ITC experiment after setting up their annealing the previous night. Each ssDNA oligo was dissolved in duplex buffer to achieve a final concentration of 400uM. Annealed oligos were dissolved in the same dialysate as that of the protein after the size exclusion

column run on the day of the experiment.

3.5.5 EMSA experiments

Quantitative F-EMSAs were used to measure the binding affinities of ancestral proteins to fluorescein-labeled dsDNA oligonucleotides (FAM-DNA). FAM-DNAs were produced by annealing equimolar complementary ssDNA strands (HPLC purified; Eurofins) as described above. Typical reactions consisted of 5 nM FAM-DNA equilibrated with varying concentrations of protein in assay buffer for 30min at room temperature (20 mm Tris, pH 7.5, 150 mm NaCl and 1mM DTT). Immediately prior to loading, one-twentieth volume of 30% (v/v) glycerol, 0.01% (w/v) bromcresol green was added to each reaction as a dye marker. A 10 ul sample of each reaction was loaded onto a pre-run 8% native polyacrylamide gel in pre-chilled 0.5TBE buffer. The gels were run for 40 min at 100 volts then immediately scanned using a fluor-imager (Bio-Rad GelDoc) with a blue laser at 491 nm. Pixel intensities corresponding to fraction bound and unbound were estimated using the freely available ImageJ software.

3.5.6 Modeling affinities and cooperativity from EMSA data

To determine K_A and ω with high confidence, we performed two experiments for each protein-DNA pair, one involving half-site binding and one involving palindromic DNA binding. The Standard Langmuir isotherm was used to model the half-site binding data, although some non-idealities resulting from the inability of the imager to image low concentrations of fraction bound necessitated some modifications. We used the following equation for half-site binding: fraction bound $\quad b + (m - b)/(1 + ((K_d/conc)^n))$

where b is the baseline signal, m is the maximal signal, K_d is the dissociation constant, conc is the concentration of the protein [nM], and n is the hill slope.

Estimating K_A and ω from palindromic data alone was only meaningful when there is clear separation of ligated states from the EMSA gels. K_1 and K_2 could not be uniquely

resolved with the palindromic data alone when there is insufficient data for the monomer bound fraction. We added half-site data to that from the palindromic binding experiment to uniquely estimate K_A and ω from five different data series — fraction unbound and fraction bound by monomers (fbound_half) from half-site binding; fraction unbound (fbound0), monomer bound (fbound1) and dimer bound (fbound2) from two-site binding. We applied a global fit, based on the Senear and Brenowitz models (1991) [127] to estimate K_A and ω simultaneously. Once K_1 and K_2 were estimated, the $K_{A,Mac}$ was calculated as their product. The half-site and two-site data were fit to the following equations:

$fbound0 = 1/Z$

$fbound1 = ((k1 + k2) * conc)^n/Z$

$fbound2 = (k1 * k2 * \omega * conc^2)^n/Z$

$fboundhalf = (b + (m - b)/(1 + (1/(2 * k1 * conc)^n)))$

where, ki are the microscopic association constants, and Z is the partition function.

All procedures associated with image process and model fitting were carried out using custom R scripts.

3.5.7 ITC experiments

Isothermal titration calorimetry (ITC) experiments were performed on the MicroCal ITC 200 instrument and data were processed using the Microcal ORIGIN software. Measurements were repeated two times in the ITC buffer composed of 100 mM Tris with 1 mM TCEP. Additionally, 110 mM NaCl was added to adjust the ionic strength of the buffer. The protein was dialyzed into the ITC buffer using the sizing column. The response elements were prepared fresh on the day of the experiment (see section on Response element preparation for EMSA and ITC) by dissolving in the ITC buffer and de-gassed using the ThermoVac accessory for 5min. The experiments were conducted with the protein in the syringe and the DNA in the sample cell. We injected 2ul aliquots of the protein from the syringe into the

calorimetric cell containing 200 uL of duplex DNA at 25C. The change in thermal power as a function of each injection was automatically recorded using Microcal ORIGIN software and the raw data were further processed to yield binding isotherms of heat release per injection as a function of molar ratio of protein to DNA duplex. The heats of mixing and dilution were subtracted from the heat of binding per injection by carrying out control experiments in which the same buffer in the syringe was titrated against the protein or DNA in an identical manner. For half-site binding experiments we used a final 100 uM protein concentration and 10 uM DNA. For two-site binding experiments we used 200 uM protein titrated against 12.5uM DNA.

3.5.8 Modeling affinities and cooperativity from ITC data

To extract macroscopic affinities $K_{A,Mac}$, enthalpy ΔH, entropy ΔS, or half-site affinities K_A ,the binding isotherms were fit to in built one-site, two-site and independent binding models implemented by non-linear least squares regression analysis in the ORIGIN software. The $K_{A,Mac}$ was directly obtained from the two-site fit in Origin, as the product of K_1 and K_2. The cooperativity constant ω was estimated by simply subtracting ΔG_2 from ΔG_1. When non-unique parameter values were obtained, we estimated cooperativity by subtracting the total ΔG of binding on the two-site from $2 * \Delta G_1$ obtained separately from half-site binding experiments.

3.5.9 Flow cytometry DBD activation assays

These activation assays have been published previously [138], and we adapted the same protocol. Briefly, we performed all activation experiments in the Saccharomyces cerevisiae strain K20. Oligonucleotide sequences used for integrating into the yeast genome are shown in Fig. 3.1E, and were ordered from IDT. They were annealed using the annealing procedure described above in the hepes buffer. We constructed yeast reporter strains containing

yeast enhanced GFP (yeGFP) under the control of a minimal CYC1 promoter with two upstream half-site or two-site annealed oligo response elements, integrated into the ADE2 locus. Colony PCR and Sanger sequencing confirmed correct integration of the response elements at the ADE2 locus. The yeast DBD expression plasmid contained the ancestral and extant DNA-binding domains with an N-terminal SV40 nuclear localization sequence and Gal4 activation domain (AD) connected by a flexible linker. Expression of the AD-DBD fusion protein is controlled by the galactose-inducible GAL1 promoter, in the background of the pRS413 plasmid containing a HIS selection marker. For flow cytometry, we conducted all experiments in triplicate using a plate assay. Individual colonies were inoculated in 500 uL SC-His with 2% raffinose, and incubated for 16h at 30C and 225 r.p.m. in an orbital shaker incubator. Cells were back-diluted to 0.25 A600 nm in SC-His with 2% galactose (+ G) to induce DBD expression and grown for an additional 24h. Cells were pelleted and suspended to 1 A600 nm in 1 TBS. We analysed 10,000 cells of each genotype by flow cytometry on a BD LSR-Fortessa 4-15, with 488 nm excitation and 530 nm emission.

3.6 Acknowledgements

I performed all the experiments, analyzed and interpreted the results, and wrote the chapter. I got extensive comments from my advisor, Joe Thornton at each step of this process. I would like to specifically thank my labmate, friend and excellent colleague Tyler Starr for sharing protocols for the yeast flow cytometry DBD activation assay. Additionally, Qing Chen for training me on the ITC machine and sharing experimental protocols and Georg Hochberg for helpful discussions.

3.7 Supplementary Information

3.7.1 Supplementary Figures

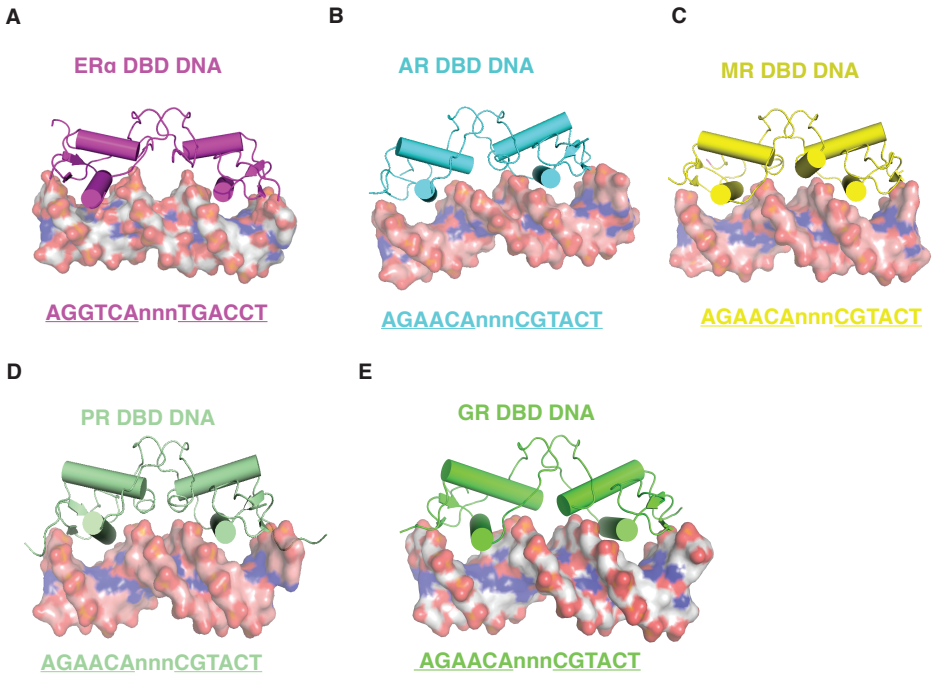


Figure 3.6: **All SRs DBDs bind as dimers to a palindromic invert repeats of two short 6-bp half-sites** Depicted are the crystal structures of human paralogs in the SR clade bound to their preferred DNA elements. Each monomer of the DBD binds to a short 6-bp half-site in the major groove. A 3-bp variable spacer separates the two half-sites.
A. ER α Estrogen receptor α B. AR Androgen receptor C. MR Mineralocorticoid receptor D. PR Progesterone receptor E. GR Glucocorticoid receptor

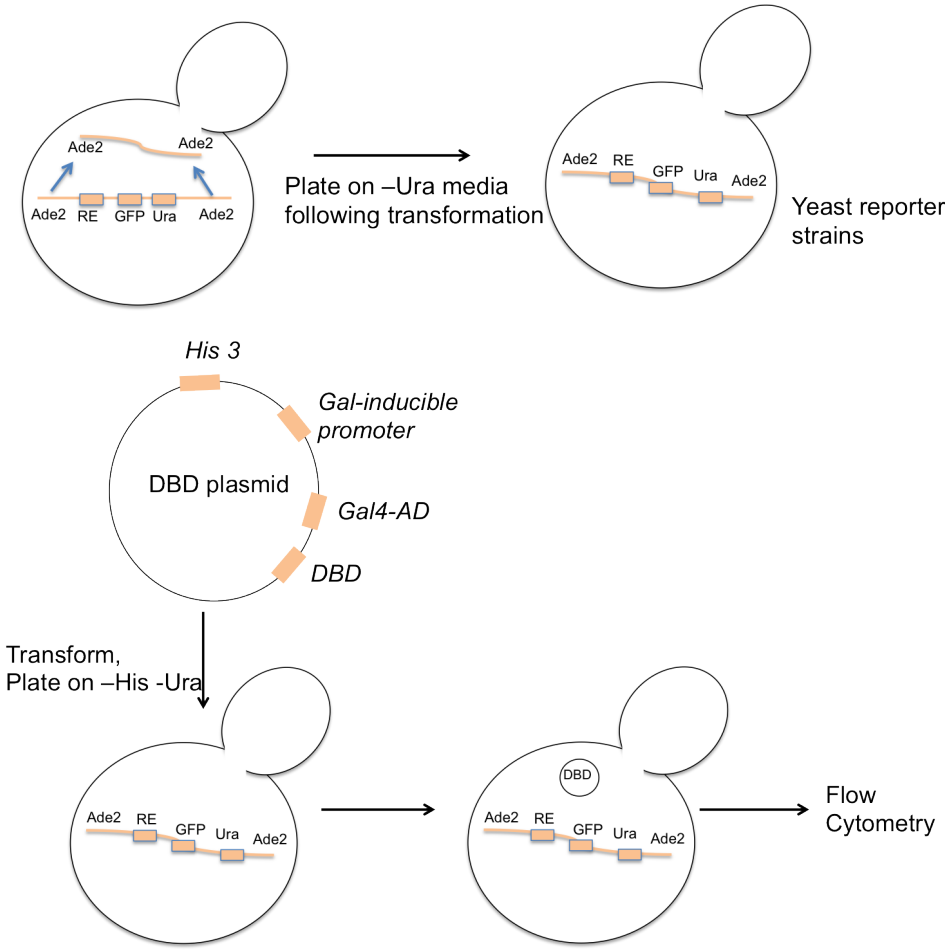


Figure 3.7: **Experimental design of DBD activation assays** Response elements shown in Fig. 3.1E were fused to the GFP reporter gene and selectable Ura marker and then integrated at the Ade2 locus in yeast to yield stable yeast reporter strains. Ancestral and extant DBD plasmids were transformed into the reporter strains for flow cytometry, selected using His markers. DBDs were fused to the Gal4AD activation domain at their N-terminus. DBD expression was induced by galactose a day before flow cytometry was performed.

Figure 3.8: **ML phylogeny of SRRs** Tree is based on alignment from 223 steroid and related receptors, rooted with retinoic acid receptor (RXR) as the outgroup. Scale is in substitutions per site. Nodal support is indicated by approximate likelihood ratio statistics. A few nodes were constrained to incorporate information from genome-wide synteny and linked genes which ensured the correct relationships between clades, cyclostome SRs and newly discovered hemichordate SRs. The constrained nodes have 0 support and are indicated in red. Nodes with aLRT values ≤ 2 are unlabeled. AncSR1, the ancestor of chordate SRs; AncSRR, the duplication node at the base of bilateria; AncERR, the ancestor of ERRs. SF1, Steroidogenic factor receptors; RXR, Retinoic acid receptors. Insets show the distribution of posterior probabilities for the three reconstructed ancestors, A. AncERR; B. AncSR1; C. AncSRR. Ancestors are reconstructed with good statistical confidence, as indicated by the mean posterior probability across sites.

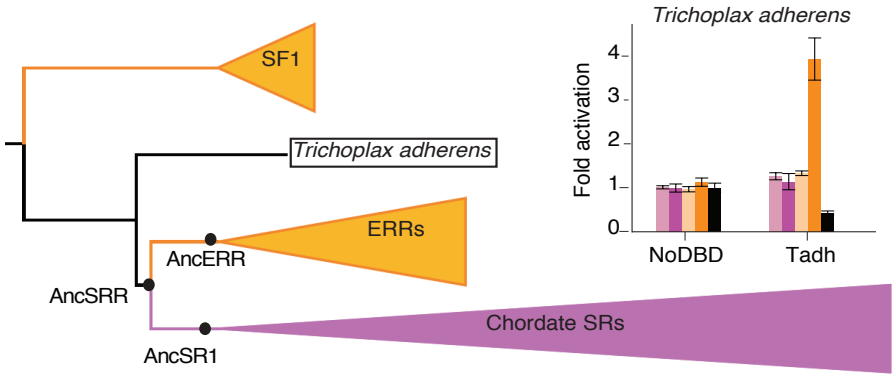


Figure 3.9: ***Trichoplax adherens* DBD is ERR-like** Fold activation of *Trichoplax adherens* DBD on the response elements shown in Fig. 3.1E. The DBD only activates from elements with the TCA extension. Bars are color coded according to the response elements shown in Fig. 3.1E. Bar heights indicate fold activation relative to the No DBD-No RE control with error bars showing the SEM of three experimental replicates.

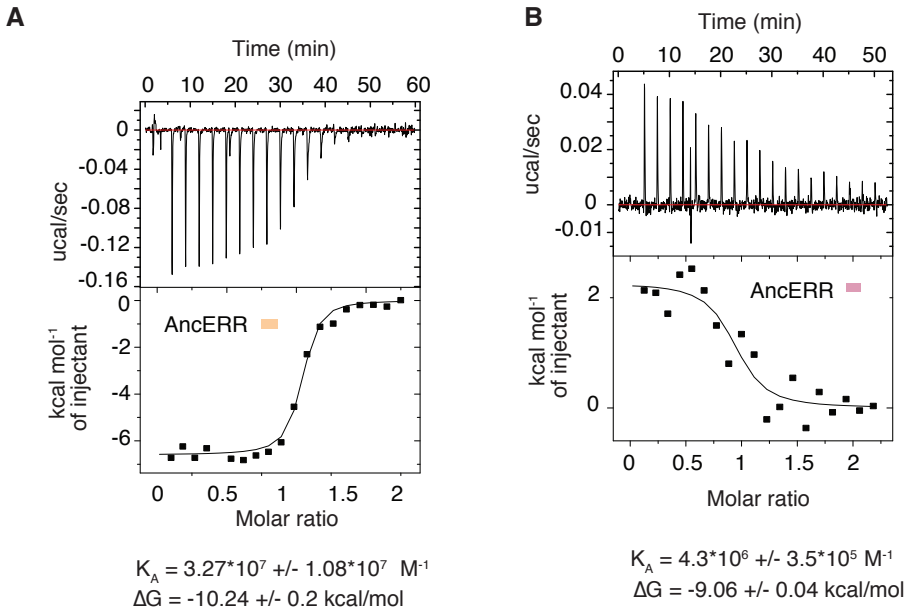


Figure 3.10: **AncERR prefers the extension by 1 kcal/mol** A shows a representative ITC thermogram for AncERR binding on Ext_EREhalf. The response element is indicated in the inset. The heat of dilution of the protein into buffer has been subtracted out by performing control experiments of protein into buffer. Shown below the plot are mean K_A and ΔG estimated from two independent experiments. Errors are SEM from two experiments. B similar to A, except shows a representative ITC thermogram for AncERR binding on EREhalf. The mean K_A and ΔG estimated from two independent experiments are indicated below the plot. Errors are SEM from two experiments.

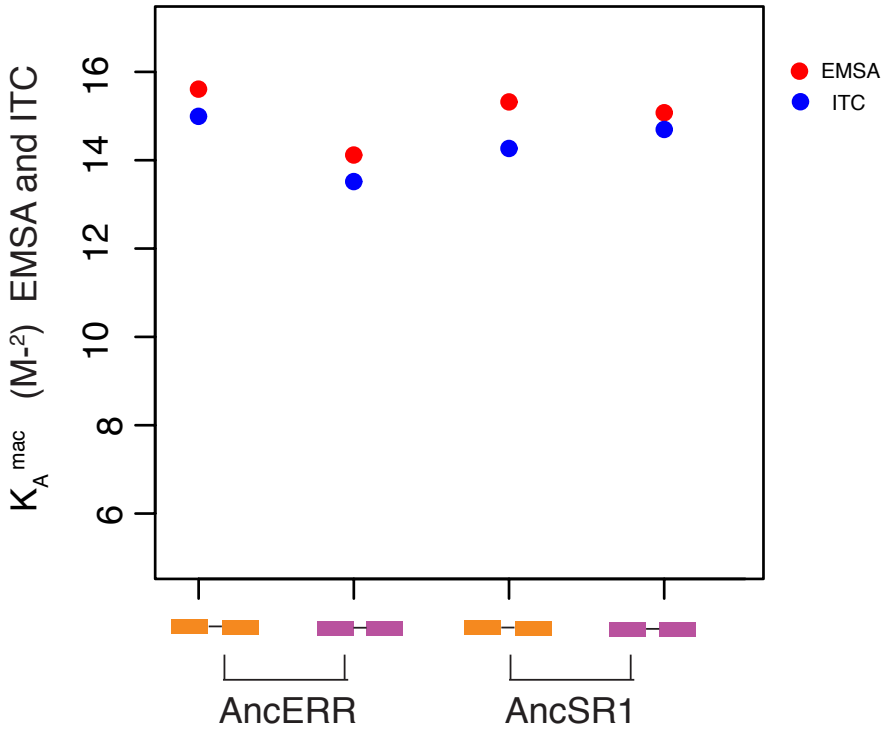


Figure 3.11: $K_{A,Mac}$ **from EMSA and ITC are in good agreement** The macroscopic binding affinities of AncERR and AncSR1 on *Ext_EREpal* (orange), and EREpal (magenta) as estimated by EMSA and ITC are shown. RE cartoons are depicted according to Fig. 3.1E. The affinities estimates from the two different methods are very consistent. AncERR binds better to *Ext_EREPal*. AncSR1 binds equally well, or better to EREpal.

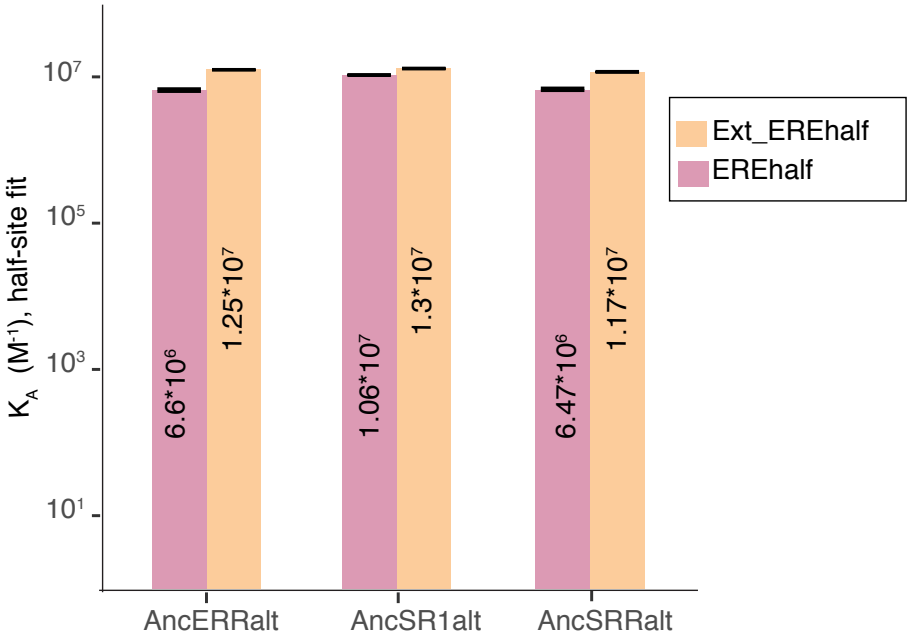


Figure 3.12: **Ancestral DBDs with reconstruction uncertainty show the same evolutionary trajectory as ML DBDs** The binding affinities, K_A, of DBDs with reconstruction uncertainty incorporated (altall). Each DBD contains all alternately reconstructed states and is tested on Ext_EREhalf (orange), and EREhalf (magenta) using EMSA. The actual K_A values from the model fit are indicated on the bars. Errors correspond to errors from the model fit. RE cartoons and colors are according to Fig. 3.1E. AncERRalt and AncSRRalt prefer extension. AncSR1alt binds equally well on both.

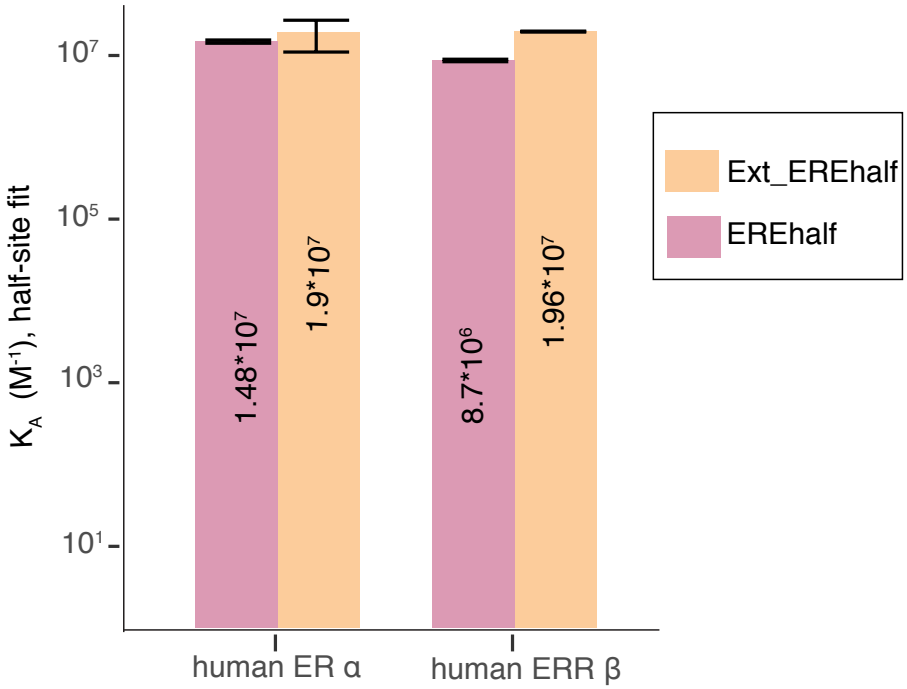


Figure 3.13: **Human ERR β DBD binds better on extended REs than human ER α DBD** The binding affinities, K_A, of human ERR and ER DBDs on Ext_EREhalf (orange), and EREhalf (magenta) as estimated by EMSA. The actual K_A values from the model fit are indicated on the bars. Errors correspond to errors from the model fit. RE cartoons and colors are according to Fig. 3.1E. The human ERR DBD binds better to Ext_EREhalf. Human ER DBD binds equally well on both.

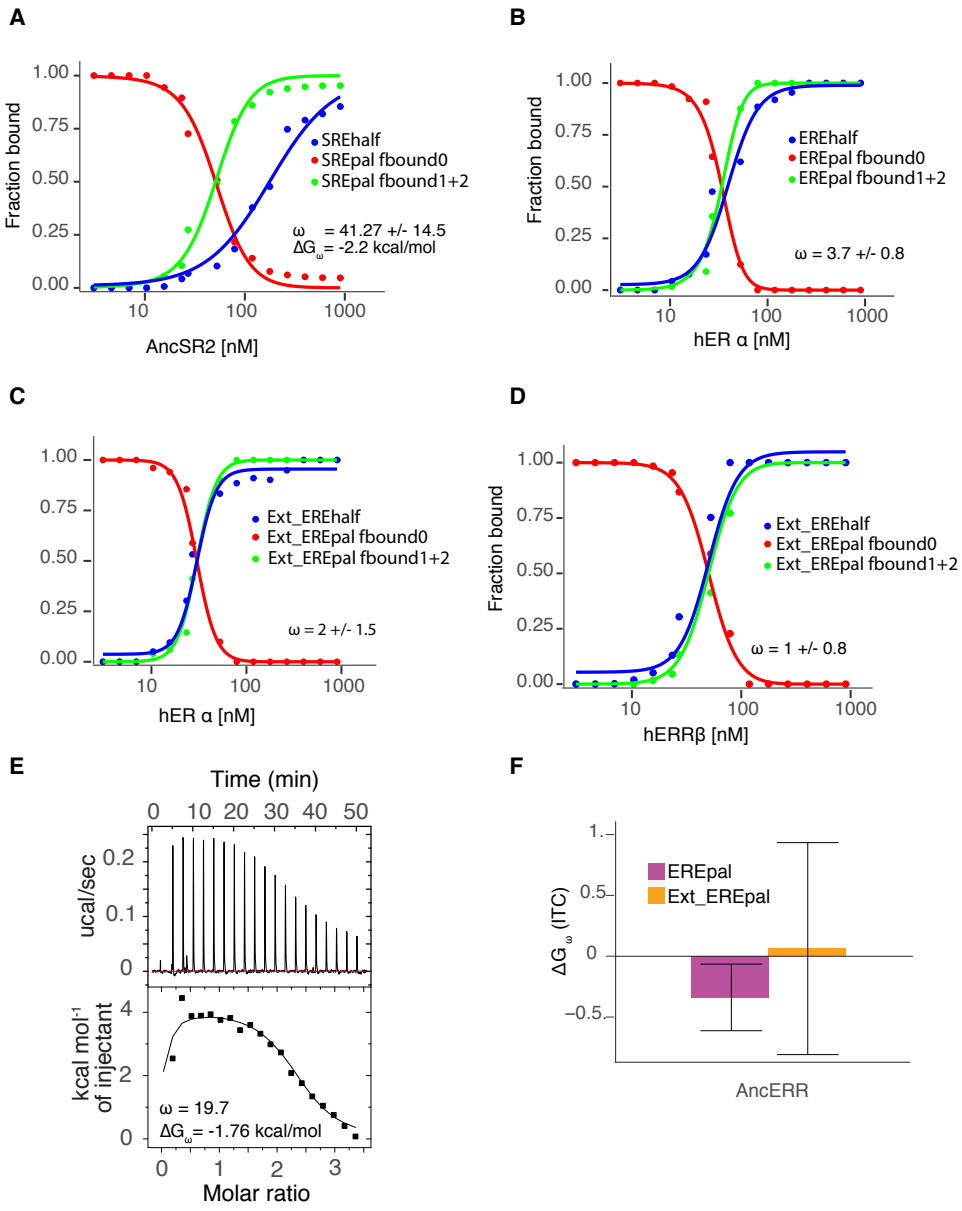


Figure 3.14: **Estimation of cooperativity constants using EMSA and ITC**

Figure 3.14: (continued) A. AncSR2 cooperativity estimated on its preferred steroid response element (SRE) by EMSA. Plotted are the fraction of DNA bound by AncSR2 on a two-site response element (SREpal). From this gel, two data series are extracted: filled red circles show the fraction of DNA bound to 0 ligands, SREpal fbound0, which decreases as the protein concentration increases. Filled green circles show the sum of monomer and dimer bound fractions, SREpal fbound1+2, which increase as the protein concentration increases. Also overlaid is the fraction of DNA bound to AncSR2 from a separate half-site experiment (SREhalf): filled blue circles show the monomer bound fraction on SREhalf. The cooperativity constant ω is estimated by a joint fit of the half-site and two-site data, and is shown on the plot. Error bars on ω indicate errors from model fit. AncSR2 is strongly cooperative.ER α is weakly cooperative, ω is as indicated. B. Cooperativity of human ER α DBD estimated on EREhalf/pal by EMSA. Plotted are the fraction of DNA bound by ER α on the two-site response element (EREpal). From this gel, two data series are extracted: filled red circles show the fraction of DNA bound to 0 ligands, EREpal fbound0, which decreases as the protein concentration increases. Filled green circles show the sum of monomer and dimer bound fractions, EREpal fbound1+2, which increase as the protein concentration increases. Also overlaid is the fraction of DNA bound to ER α from a separate half-site experiment (EREhalf): filled blue circles show the monomer bound fraction on EREhalf. C. Similar to B, except cooperativity of human ER α DBD is estimated on $Ext\_EREhalf$/pal by EMSA. Error bars on ω indicate errors from model fit. ER α is weakly cooperative, ω is as indicated. D. Similar to C-D. Cooperativity of human ERR β DBD is estimated on $Ext\_EREhalf$/pal by EMSA. Error bars on ω indicate errors from model fit. ERR β is weakly cooperative, ω is as indicated. E. Strong cooperativity of AncSR2 is also estimated using ITC. Thermogram shows the binding of AncSR2 on a single palindromic response element, SREpal. Cooperativity constant is estimated from ΔG_1 and ΔG_2, and is indicated on the plot. F. AncERR's cooperativity re-estimated by combining the ΔG_1 from the EREhalf ITC experiment with $K_{A,Mac}$ from two-site ITC experiment on EREpal. Plotted are the mean ΔG_ω from two independent experiments, with the error bars showing SEM from the two experiments. The combined analyses show that AncERR has weak DNA binding cooperativity on EREpal and Ext_EREpal.

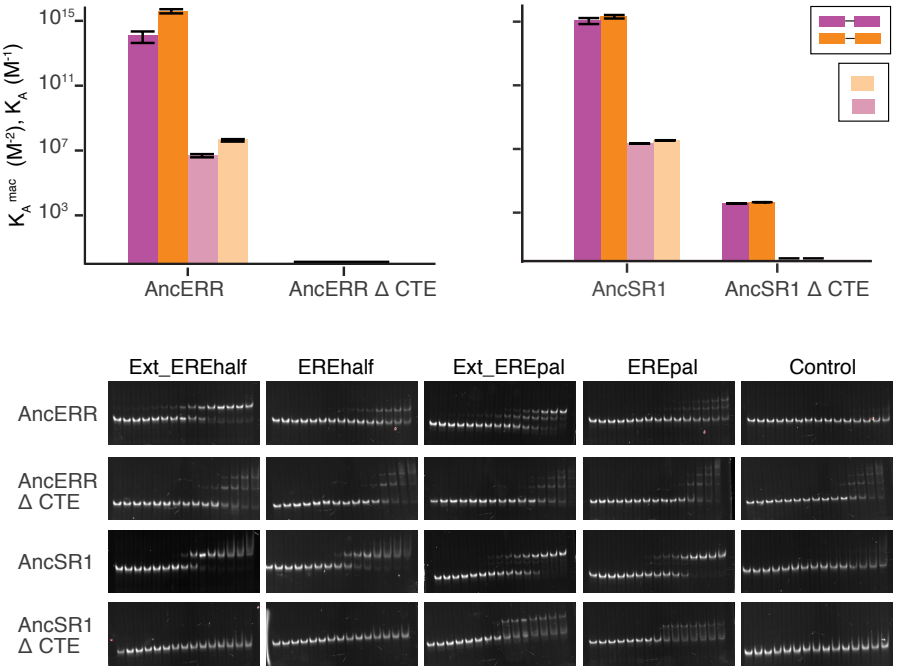


Figure 3.15: **The CTE is required for DNA binding in AncERR and AncSR1** The binding affinities of CTE truncation mutations of AncERR and AncSR1 (ΔCTE) are shown on the bar plots on the top, alongside the wild type protein constructs. Binding to all half-site and palindromic response elements was tested. Response elements are shown in the inset, colored according to Fig. 3.1E. Representative gels used for the estimation of K_A and $K_{A,Mac}$ are shown below the bar plots.

$$\theta_1 = \frac{(k1+k2)[P]}{1+(k1+k2)[P]+k1k2k12[P]^2}$$

$$\frac{v\frac{\partial u}{\partial [P]} - u\frac{\partial v}{\partial [P]}}{v^2} = 0$$

$$v\frac{\partial u}{\partial [P]} = u\frac{\partial v}{\partial [P]}$$

$$1+(k1+k2)[P]+k1k2k12[P]^2\frac{\partial (k1+k2)[P]}{\partial [P]} = (k1+k2)[P]\frac{\partial (1+(k1+k2)[P]+k1k2k12[P]^2)}{\partial [P]}$$

$$1+(k1+k2)[P]+k1k2k12[P]^2(k1+k2) = (k1+k2)[P]((k1+k2)+2k1k2k12[P])$$

$$1+(k1+k2)[P]+k1k2k12[P]^2 = (k1+k2)[P]+2k1k2k12[P]^2$$

$$1 = k1k2k12[P]^2$$

$$P = \frac{1}{k1k2k12}^{\frac{1}{2}} = (k1k2k12)^{\frac{-1}{2}}$$

$$\theta_1 max = \frac{(k1+k2)(k1k2k12)^{\frac{-1}{2}}}{1+(k1+k2)(k1k2k12)^{\frac{-1}{2}}+(k1k2k12)(k1k2k12)^{-}1}$$

$$\theta_1 max = \frac{(k1+k2)}{(k1+k2)+2(k1k2k12)^{\frac{1}{2}}}$$

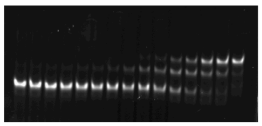
$$if\ k1 = k2$$

$$\theta_1 max = \frac{2k1}{2k1+2\sqrt{(}k1^2k12)}$$

$$\theta_1 max = \frac{2k1}{2k1+2k1\sqrt{(}k12)} = \frac{1}{(1+\sqrt{(}k12))}$$

$$k12 = (\frac{1}{\theta_1 max}-1)^2$$

AncERR , Ext_EREhalf/ Ext_EREPal



ω = 3.16 (max = 0.36),
delG = -0.68 kcal/mol

Figure 3.16: **Derivation of alternate method to estimate cooperative binding from EMSA monomer bands**

Figure 3.16: (continued) θ_1 is the monomer bound function derived from the Statistical thermodynamic modeling approach of Senear and Brenowitz; k_1 and k_2 are the half-site binding affinities; k_{12} is the cooperative constant, ω. The cooperative constant is derived by setting the derivative of θ_{1max}, the monomer bound function, to zero — that is, when the function reaches its maximum. This allows an estimation of the cooperative constant. An example EMSA gel of AncERR binding to Ext_EREPal is shown, and the cooperative constant estimated from the value of the maximum pixel intensity of the monomer bound fractions.

CHAPTER 4

CONCLUSION

The projects described in this thesis have been motivated by my interest in doing interdisciplinary work in molecular functional evolution. Particularly my interest in combining rigorous experimental work with computational modeling to shed light on new biological mechanisms. Using this integrative functional synthesis approach, I have aimed to investigate the mechanisms and dynamics by which genes and proteins evolve novel functions. Originally trained as a Computational biologist, learning how to do experiments to pursue mechanistic questions in protein evolution has been a very fulfilling experience in my graduate career.

4.1 Molecular spandrels revisited

In Chapter one, I have investigated the impact of unincorporated evolutionary complexity on inferences of gene adaptation. I integrated mechanistic knowledge of the mutational process, both at the level of the non-independence of mutations, and their associated transversion enrichment, into the classic branch-site test of adaptation. When the new branch-site test was run on protein coding gene sequences simulated with the revised models, I found that complex multi-mutational processes evolving purely under genetic drift are spuriously inferred to be under positive selection by the classic branch-sites test. Failure to integrate mechanistic insight of the mutational process can therefore leads to pervasive false inferences of positive selection in genes, thereby distorting our view of the relative importance of positive selection and drift in molecular evolution. Multinucleotide mutations confound the branch-site test, behaving like molecular spandrels that appear to have all the signatures consistent with true adaptation.

I showed that the majority of genes claimed to be under positive selection on the human and fly lineages are artifacts of unincorporated neutral mutational processes, and suggest that

many published inferences of positive selection on protein-coding genes in other species would similarly be artifacts. Neutral mutational processes are a dominant force in shaping genic divergences in mammalian and fly genomes, radically altering our view of the importance of positive selection in these genomes.

The study provides an example of what not to do in molecular evolution that is, sequence patterns consistent with adaptation can be explained by forces other than adaptation, such as other neutral forces or unincorporated evolutionary complexity. Chapter 2 describes how better mechanistic models of mutation can be used to develop better hypotheses about functional evolution and adaptation, and a convincing argument is made to combine computational inferences of adaptation with experiments. The models I describe however, do not include some other forms of evolutionary complexity such as site heterogeneity. Multinucleotide mutations at the boundary of codons are also not included. More work would be needed in the future to expand these models. I have made all the source code for this project publicly available on my Github account - https://github.com/aartivnkt/Multinucleotide-mutations-modeling

4.2 Evolution of new DNA specificity was achieved only through changes in half-site affinity and not cooperativity

In the experimental project outlined in Chapter two, I investigated using an integration of evolutionary, biochemical and genetic approaches, the molecular mechanisms by which specific mutations that fixed in protein evolution change protein functions. Using ASR combined with in vitro and in vivo characterizations of the genetic, biochemical and biophysical properties of ancestral proteins, I showed that the preference for a specific flanking sequence was lost on the SR lineage from an ERR-like ancestor. This transition was enabled through a gain in the 6-bp half-site affinity on the SR lineage, with little to no changes to the cooperative binding. Even though there are a many ways to evolve new DNA specificities —

through a change in half-site affinity, cooperativity, or both, not all energetic parameters needed to evolve concomitantly.

Non-overlapping DNA-binding specificities evolved in paralogous SR and ERR transcription factors through tinkering with thermodynamic basis of specificity, without a radical reorganization of DNA-protein and protein-protein interfaces, or even the evolution of novel interfaces. Only six mutations were sufficient to evolve this loss of flanking sequence preference, suggesting a relatively simple genetic basis compared to other published transitions in the same protein family. Overall, the results provide insight into thermodynamic basis of evolutionary novelty — SRs evolved new DNA-specificity by tinkering with half-site affinities, such that the overall free energy of binding of SRs on new derived response elements was still comparable to that of ancestral SRR on its target elements. Evolution can therefore drastically expand the repertoire of molecular functions through tinkering with the biophysical properties of proteins.

This study provides a mechanistic explanation for how flanking sequence preferences are lost in transcription factors. Extant transcription factors recognize and discriminate between response elements with subtle differences in sequences flanking core motifs. But how and why does this discrimination evolves has not been well understood. This project has focused on the SRR family to describe one such transition. The proteins investigated in the project have a modular structure, with the DNA binding domain interacting with the core half-site, and the CTerminus Extension interacting with the flanking sequence. However, the genetic experiments suggest that the evolution of new DNA specificity is a non-modular feature of SRRs.

4.3 Multiple approaches are needed to estimate and model the evolution of cooperative binding

This project has focused on the relationship of single-site affinities with cooperative binding, their interaction and evolution, a poorly understood area of research with broad implications for evolution of gene regulation.

I have described three methods to do so, using EMSAs and ITC. Cooperative binding increases the affinity and specificity of protein complexes, and it appears as though transcription factors have tremendous evolutionary mobility to tune and rewire single-site and cooperative binding energies. How these energies are partitioned over evolution, or even rewired to evolve novel DNA specificities and new modes of gene regulation is an open question. I have shown that a simple assay like EMSA is very useful in estimating DNA-induced cooperative binding by transcription factors. The relative amounts of monomers and dimers bound to a palindromic DNA element, as well as unbound DNAs, can be directly visualized and quantified from the gel. Strongly cooperative proteins should rapidly transition from unbound DNA to dimer-bound DNA as protein concentration increases, whereas less cooperative proteins will show a transition through monomeric binding. This makes it possible to quantify the fraction of palindromic DNA bound to 0,1 and 2 proteins.

The first method I present uses the elegant statistical thermodynamic framework proposed by Senear and Brenowitz in 1991 to perform a global fit of the models to half-site and two-site EMSA data. This approach jointly estimates the half-site binding affinity KA, and cooperativity constant, ω. Though I employed fluorescent probes that lack the sensitivity and precision of traditional radioactive probes, I have found that performing EMSA under the right conditions yields useful data for estimating cooperativity.

The second method I present only requires knowledge of the monomer bound fraction from the EMSA gel. The unique strength of EMSA is the visualization and quantification of the monomer function. Newer technologies based on fluorescence polarization or ITC do not

have this resolution, as one only estimates the total fraction bound with these techniques. Using a very simple approach of differentiating proteins based on their monomer function maxima, I deduced a very simple expression for the cooperative binding constant that depends on the inverse of the monomer bound function. I believe this approach can be used for a quick preliminary understanding on the cooperative behavior of an unknown protein, provided the monomer bounds can be quantified, and the EMSAs are performed under the right conditions.

The third method I present uses ITC for estimating equilibrium ΔG estimates. The large number of free parameters in the two-site ITC binding model, together with the experimental difficulties associated with estimating free energies of proteins with low enthalpies of binding make ITC a difficult experiment. In sum, it seems necessary to use multiple methods to increase the confidence in the estimates of ω

The statistical thermodynamic models I have implemented for analyzing K_A and ω from EMSA data are publicly available on my Github account:

https://github.com/aartivnkt/Cooperative-Binding-Modeling.

Overall, the work described in this thesis provides a detailed, mechanistic insight into the gradual nature of the evolutionary process, in which both radically novel biophysical mechanisms and selectively advantageous alleles are rare occurrences. The approaches outlined here use mechanistic insight to inform better phylogenetic models, and phylogenetic reconstruction of ancestral proteins to inform and explicitly test hypotheses about the biophysical mechanisms of evolution. In general, this mutually beneficial combination of biochemistry and phylogenetics holds great promise to provide long sought answers to the most basic questions in evolutionary biology.

REFERENCES

[1] DP Anderson, DS Whitney, V Hanson-Smith, A Woznica, W Campodonico-Burnett, BF Volkman, N King, JW Thornton, and KE Prehoda. Correction: Evolution of an ancient protein function involved in organized multicellularity in animals. *Elife*, 5:e14311, 2016.

[2] DW Anderson, AN McKeown, and JW Thornton. Intermolecular epistasis shaped the function and evolution of an ancient transcription factor and its dna binding sites. *Elife*, 4:e07864, 2015.

[3] M Anisimova and Z Yang. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol*, 24(5):12191228, 2007.

[4] ME Arana, M Seki, RD Wood, IB Rogozin, and TA Kunkel. Low-fidelity dna synthesis by human dna polymerase theta. *Nucleic Acids Res*, 36(11):38473856, 2008.

[5] ZJ Assaf, S Tilk, J Park, ML Siegal, and DA Petrov. Deep sequencing of natural and experimental populations of drosophila melanogaster reveals biases in the spectrum of new mutations. *Genome Res*, 27(12):19882000, 2017.

[6] M Averof, A Rokas, KH Wolfe, and PM Sharp. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science*, 287(5456):12831286, 2000.

[7] G Badis, 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. Diversity and complexity in dna recognition by transcription factors. *Science*, 324(5935):17201723, 2009.

[8] CR Baker, LN Booth, TR Sorrells, and AD Johnson. Protein modularity, cooperative binding, and hybrid regulatory states underlie transcriptional network diversification. *Cell*, 151(1):8095, 2012.

[9] CR Baker, V Hanson-Smith, and AD Johnson. Following gene duplication, paralog interference constrains transcriptional circuit evolution. *Science*, 342(6154):104108, 2013.

[10] MF Barber and NC Elde. Nutritional immunity. escape from bacterial iron piracy through rapid evolution of transferrin. *Science*, 346(6215):13621366, 2014.

[11] LA Barrera, A Vedenko, JV Kurland, JM Rogers, SS Gisselbrecht, EJ Rossin, J Woodard, L Mariani, KH Kock, S Inukai, T Siggers, L Shokri, R Gordn, N Sahni, C Cotsapas, T Hao, S Yi, M Kellis, MJ Daly, M Vidal, DE Hill, and ML Bulyk. Survey of variation in human transcription factors reveals prevalent dna binding changes. *Science*, 351(6280):14501454, 2016.

[12] RD Barrett and HE Hoekstra. Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet*, 12(11):767780, 2011.

[13] GA Bazykin, FA Kondrashov, AY Ogurtsov, S Sunyaev, and AS Kondrashov. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*, 429(6991):558562, 2004.

[14] MF Berger and ML Bulyk. Protein binding microarrays (pbms) for rapid, high-throughput characterization of the sequence specificities of dna binding proteins. *Methods Mol Biol*, 338:245260, 2006.

[15] H Berglund, M Wolf-Watz, T Lundbck, S van den Berg, and T Hrd. Structure and dynamics of the glucocorticoid receptor dna-binding domain: comparison of wild type and a mutant with altered specificity. *Biochemistry*, 36(37):1118811197, 1997.

[16] S Besenbacher, P Sulem, A Helgason, H Helgason, H Kristjansson, A Jonasdottir, A Jonasdottir, OT Magnusson, U Thorsteinsdottir, G Masson, A Kong, DF Gudbjartsson, and K Stefansson. Multi-nucleotide de novo mutations in humans. *PLoS Genet*, 12(11):e1006315, 2016.

[17] JD Bloom. An experimentally determined evolutionary model dramatically improves phylogenetic fit. *Mol Biol Evol*, 31(8):19561978, 2014.

[18] JI Boucher, JR Jacobowitz, BC Beckett, S Classen, and DL Theobald. An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases. *Elife*, 3, 2014.

[19] JT Bridgham, SM Carroll, and JW Thornton. Evolution of hormone-receptor complexity by molecular exploitation. *Science*, 312(5770):97101, 2006.

[20] JT Bridgham, GN Eick, C Larroux, K Deshpande, MJ Harms, ME Gauthier, EA Ortlund, BM Degnan, and JW Thornton. Protein evolution by molecular tinkering: diversification of the nuclear receptor superfamily from a ligand-dependent ancestor. *PLoS Biol*, 8(10), 2010.

[21] ML Bulyk. Analysis of sequence specificities of dna-binding proteins with protein binding microarrays. *Methods Enzymol*, 410:279299, 2006.

[22] BW Busser, L Shokri, SA Jaeger, SS Gisselbrecht, A Singhania, MF Berger, B Zhou, ML Bulyk, and AM Michelson. Molecular mechanism underlying the regulatory specificity of a drosophila homeodomain protein that specifies myoblast identity. *Development*, 139(6):11641174, 2012.

[23] GA Cary, AM Cheatle Jarvela, RD Francolini, and VF Hinman. Genome-wide use of high- and low-affinity tbrain transcription factor binding sites during echinoderm development. *Proc Natl Acad Sci U S A*, 114(23):58545861, 2017.

[24] C Casola and MW Hahn. Gene conversion among paralogs results in moderate false detection of positive selection using likelihood methods. *J Mol Evol*, 68(6):679687, 2009.

[25] YF Chan, ME Marks, FC Jones, G Villarreal, MD Shapiro, SD Brady, AM Southwick, DM Absher, J Grimwood, J Schmutz, RM Myers, D Petrov, B Jnsson, D Schluter, MA Bell, and DM Kingsley. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *Science*, 327(5963):302305, 2010.

[26] YF Chan, ME Marks, FC Jones, G Villarreal, MD Shapiro, SD Brady, AM Southwick, DM Absher, J Grimwood, J Schmutz, RM Myers, D Petrov, B Jnsson, D Schluter, MA Bell, and DM Kingsley. Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a pitx1 enhancer. *Science*, 327(5963):302305, 2010.

[27] M Chatzou, C Magis, JM Chang, C Kemena, G Bussotti, I Erb, and C Notredame. Multiple sequence alignment modeling: methods and applications. *Brief Bioinform*, 17(6):10091023, 2016.

[28] AM Cheatle Jarvela, L Brubaker, A Vedenko, A Gupta, BA Armitage, ML Bulyk, and VF Hinman. Modular evolution of dna-binding preference of a tbrain transcription factor provides a mechanism for modifying gene regulatory networks. *Mol Biol Evol*, 31(10):26722688, 2014.

[29] JM Chen, DN Cooper, and C Frec. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: 0.4single-nucleotide substitution mutation rate. *Hum Mutat*, 35(3):392394, 2014.

[30] JM Chen, C Frec, and DN Cooper. Complex multiple-nucleotide substitution mutations causing human inherited disease reveal novel insights into the action of translesion synthesis dna polymerases. *Hum Mutat*, 36(11):10341038, 2015.

[31] RG Christensen, MS Enuameh, MB Noyes, MH Brodsky, SA Wolfe, and GD Stormo. Recognition models to predict dna-binding specificities of homeodomain proteins. *Bioinformatics*, 28(12):i849, 2012.

[32] BE Clifton and CJ Jackson. Ancestral protein reconstruction yields insights into adaptive evolution of binding specificity in solute-binding proteins. *Cell Chem Biol*, 23(2):236245, 2016.

[33] KD Connaghan, Q Yang, MT Miura, AD Moody, and DL Bain. Homologous steroid receptors assemble at identical promoter architectures with unique energetics of cooperativity. *Proteins*, 82(9):20782087, 2014.

[34] C Crane-Robinson, AI Dragan, and PL Privalov. The extended arms of dna-binding domains: a tale of tails. *Trends Biochem Sci*, 31(10):547552, 2006.

[35] M Currat, L Excoffier, W Maddison, SP Otto, N Ray, MC Whitlock, and S Yeaman. Comment on ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens and microcephalin, a gene regulating brain size, continues to evolve adaptively in humans. *Science*, 313(5784):172; author reply 172, 2006.

[36] D Darriba, GL Taboada, R Doallo, and D Posada. Prottest 3: fast selection of best-fit models of protein evolution. *Bioinformatics*, 27(8):11641165, 2011.

[37] RW De Angelis, Q Yang, MT Miura, and DL Bain. Dissection of androgen receptor-promoter interactions: steroid receptors partition their interaction energetics in parallel with their phylogenetic divergence. *J Mol Biol*, 425(22):42234235, 2013.

[38] N De Maio, I Holmes, C Schltterer, and C Kosiol. Estimating empirical codon hidden markov models. *Mol Biol Evol*, 30(3):725736, 2013.

[39] AM Dean and JW Thornton. Mechanistic approaches to the study of evolution: the functional synthesis. *Nat Rev Genet*, 8(9):675688, 2007.

[40] T Devamani, AM Rauwerdink, M Lunzer, BJ Jones, JL Mooney, MA Tan, ZJ Zhang, JH Xu, AM Dean, and RJ Kazlauskas. Catalytic promiscuity of ancestral esterases and hydroxynitrile lyases. *J Am Chem Soc*, 138(3):10461056, 2016.

[41] S Dimitrieva and M Anisimova. Unraveling patterns of site-to-site synonymous rates variation and associated gene properties of protein domains and families. *PLoS One*, 9(6):e95034, 2014.

[42] MW Dimmic, JS Rest, DP Mindell, and RA Goldstein. rtrev: an amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J Mol Evol*, 55(1):6573, 2002.

[43] AI Dragan, Z Li, EN Makeyeva, EI Milgotina, Y Liu, C Crane-Robinson, and PL Privalov. Forces driving the binding of homeodomains to dna. *Biochemistry*, 45(1):141151, 2006.

[44] AI Dragan, JR Liggins, C Crane-Robinson, and PL Privalov. The energetics of specific binding of at-hooks from hmga1 to target dna. *J Mol Biol*, 327(2):393411, 2003.

[45] 12 Genomes Consortium Drosophila, AG Clark, MB Eisen, DR Smith, CM Bergman, B Oliver, TA Markow, TC Kaufman, M Kellis, W Gelbart, VN Iyer, DA Pollard, TB Sackton, AM Larracuente, ND Singh, JP Abad, DN Abt, B Adryan, M Aguade, H Akashi, WW Anderson, CF Aquadro, DH Ardell, R Arguello, CG Artieri, DA Barbash, D Barker, P Barsanti, P Batterham, S Batzoglou, D Begun, A Bhutkar, E Blanco, SA Bosak, RK Bradley, AD Brand, MR Brent, AN Brooks, RH Brown, RK Butlin, C Caggese, BR Calvi, A Bernardo de Carvalho, A Caspi, S Castrezana, SE Celniker, JL Chang, C Chapple, S Chatterji, A Chinwalla, A Civetta, SW Clifton, JM Comeron, JC Costello, JA Coyne, J Daub, RG David, AL Delcher, K Delehaunty, CB Do, H Ebling, K Edwards, T Eickbush, JD Evans, A Filipski, S Findeiss, E Freyhult, L Fulton, R Fulton, AC Garcia, A Gardiner, DA Garfield, BE Garvin, G Gibson, D Gilbert, S Gnerre, J Godfrey, R Good, V Gotea, B Gravely, AJ Greenberg, S Griffiths-Jones, S Gross, R Guigo, EA Gustafson, W Haerty, MW Hahn, DL Halligan, AL Halpern, GM Halter, MV Han, A Heger, L Hillier, AS Hinrichs,

I Holmes, RA Hoskins, MJ Hubisz, D Hultmark, MA Huntley, DB Jaffe, S Jagadeeshan, WR Jeck, J Johnson, CD Jones, WC Jordan, GH Karpen, E Kataoka, PD Keightley, P Kheradpour, EF Kirkness, LB Koerich, K Kristiansen, D Kudrna, RJ Kulathinal, S Kumar, R Kwok, E Lander, CH Langley, R Lapoint, BP Lazzaro, SJ Lee, L Levesque, R Li, CF Lin, MF Lin, K Lindblad-Toh, A Llopart, M Long, L Low, E Lozovsky, J Lu, M Luo, CA Machado, W Makalowski, M Marzo, M Matsuda, L Matzkin, B McAllister, CS McBride, B McKernan, K McKernan, M Mendez-Lago, P Minx, MU Mollenhauer, K Montooth, SM Mount, X Mu, E Myers, B Negre, S Newfeld, R Nielsen, MA Noor, P OGrady, L Pachter, M Papaceit, MJ Parisi, M Parisi, L Parts, JS Pedersen, G Pesole, AM Phillippy, CP Ponting, M Pop, D Porcelli, JR Powell, S Prohaska, K Pruitt, M Puig, H Quesneville, KR Ram, D Rand, MD Rasmussen, LK Reed, R Reenan, A Reily, KA Remington, TT Rieger, MG Ritchie, C Robin, YH Rogers, C Rohde, J Rozas, MJ Rubenfield, A Ruiz, S Russo, SL Salzberg, A Sanchez-Gracia, DJ Saranga, H Sato, SW Schaeffer, MC Schatz, T Schlenke, R Schwartz, C Segarra, RS Singh, L Sirot, M Sirota, NB Sisneros, CD Smith, TF Smith, J Spieth, DE Stage, A Stark, W Stephan, RL Strausberg, S Strempel, D Sturgill, G Sutton, GG Sutton, W Tao, S Teichmann, YN Tobari, Y Tomimura, JM Tsolas, VL Valente, E Venter, JC Venter, S Vicario, FG Vieira, AJ Vilella, A Villasante, B Walenz, J Wang, M Wasserman, T Watts, D Wilson, RK Wilson, RA Wing, MF Wolfner, A Wong, GK Wong, CI Wu, G Wu, D Yamamoto, HP Yang, SP Yang, JA Yorke, K Yoshida, E Zdobnov, P Zhang, Y Zhang, AV Zimin, J Baldwin, A Abdouelleil, J Abdulkadir, A Abebe, B Abera, J Abreu, SC Acer, L Aftuck, A Alexander, P An, E Anderson, S Anderson, H Arachi, M Azer, P Bachantsang, A Barry, T Bayul, A Berlin, D Bessette, T Bloom, J Blye, L Boguslavskiy, C Bonnet, B Boukhgalter, I Bourzgui, A Brown, P Cahill, S Channer, Y Cheshatsang, L Chuda, M Citroen, A Collymore, P Cooke, M Costello, K DAco, R Daza, G De Haan, S DeGray, C DeMaso, N Dhargay, K Dooley, E Dooley, M Doricent, P Dorje, K Dorjee, A Dupes, R Elong, J Falk, A Farina, S Faro, D Ferguson, S Fisher, CD Foley, A Franke, D Friedrich, L Gadbois, G Gearin, CR Gearin, G Giannoukos, T Goode, J Graham, E Grandbois, S Grewal, K Gyaltsen, N Hafez, B Hagos, J Hall, C Henson, A Hollinger, T Honan, MD Huard, L Hughes, B Hurhula, ME Husby, A Kamat, B Kanga, S Kashin, D Khazanovich, P Kisner, K Lance, M Lara, W Lee, N Lennon, F Letendre, R LeVine, A Lipovsky, X Liu, J Liu, S Liu, T Lokyitsang, Y Lokyitsang, R Lubonja, A Lui, P MacDonald, V Magnisalis, K Maru, C Matthews, W McCusker, S McDonough, T Mehta, J Meldrim, L Meneus, O Mihai, A Mihalev, T Mihova, R Mittelman, V Mlenga, A Montmayeur, L Mulrain, A Navidi, J Naylor, T Negash, T Nguyen, N Nguyen, R Nicol, C Norbu, N Norbu, N Novod, B ONeill, S Osman, E Markiewicz, OL Oyono, C Patti, P Phunkhang, F Pierre, M Priest, S Raghuraman, F Rege, R Reyes, C Rise, P Rogov, K Ross, E Ryan, S Settipalli, T Shea, N Sherpa, L Shi, D Shih, T Sparrow, J Spaulding, J Stalker, N Stange-Thomann, S Stavropoulos, C Stone, C Strader, S Tesfaye, T Thomson, Y Thoulutsang, D Thoulutsang, K Topham, I Topping, T Tsamla, H Vassiliev, A Vo, T Wangchuk, T Wangdi, M Weiand, J Wilkinson, A Wilson, S Yadav, G Young, Q Yu, L Zembek, D Zhong, A Zimmer, Z Zwirko, DB Jaffe, P Alvarez,

W Brockman, J Butler, C Chin, S Gnerre, M Grabherr, M Kleber, E Mauceli, and I MacCallum. Evolution of genes and genomes on the drosophila phylogeny. *Nature*, 450(7167):203218, 2007.

[46] DA Drummond and CO Wilke. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell*, 134(2):341352, 2008.

[47] RC Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004.

[48] GN Eick, JK Colucci, MJ Harms, EA Ortlund, and JW Thornton. Evolution of minimal specificity and promiscuity in steroid hormone receptors. *PLoS Genet*, 8(11):e1003072, 2012.

[49] SF Field, MY Bulina, IV Kelmanson, JP Bielawski, and MV Matz. Adaptive evolution of multicolored fluorescent proteins in reef-building corals. *J Mol Evol*, 62(3):332339, 2006.

[50] AD Foote, Y Liu, GW Thomas, T Vina, J Alfldi, J Deng, S Dugan, CE van Elk, ME Hunter, V Joshi, Z Khan, C Kovar, SL Lee, K Lindblad-Toh, A Mancia, R Nielsen, X Qin, J Qu, BJ Raney, N Vijay, JB Wolf, MW Hahn, DM Muzny, KC Worley, MT Gilbert, and RA Gibbs. Convergent evolution of the genomes of marine mammals. *Nat Genet*, 47(3):272275, 2015.

[51] LC Francioli, PP Polak, A Koren, A Menelaou, S Chun, I Renkens, of the Netherlands Consortium Genome, CM van Duijn, M Swertz, C Wijmenga, G van Ommen, PE Slagboom, DI Boomsma, K Ye, V Guryev, PF Arndt, WP Kloosterman, PI de Bakker, and SR Sunyaev. Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*, 47(7):822826, 2015.

[52] M Fumagalli, M Sironi, U Pozzoli, A Ferrer-Admetlla, A Ferrer-Admettla, L Pattini, and R Nielsen. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet*, 7(11):e1002355, 2011.

[53] MD Gearhart, SM Holmbeck, RM Evans, HJ Dyson, and PE Wright. Monomeric complex of human orphan estrogen related receptor-2 with dna: a pseudo-dimer interface mediates extended half-site recognition. *J Mol Biol*, 327(4):819832, 2003.

[54] AB Georges, BA Benayoun, S Caburet, and RA Veitia. Generic binding sites, generic dna-binding domains: where does specific promoter recognition come from. *FASEB J*, 24(2):346356, 2010.

[55] WH Gharib and M Robinson-Rechavi. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in gc. *Mol Biol Evol*, 30(7):16751686, 2013.

[56] A Glaser-Schmitt and J Parsch. Functional characterization of adaptive variation within a cis-regulatory element influencing drosophila melanogaster growth. *PLoS Biol*, 16(1):e2004538, 2018.

[57] N Goldman and Z Yang. A codon-based model of nucleotide substitution for protein-coding dna sequences. *Mol Biol Evol*, 11(5):725736, 1994.

[58] R Gordn, N Shen, I Dror, T Zhou, J Horton, R Rohs, and ML Bulyk. Genomic regions flanking e-box binding sites influence dna binding specificity of bhlh transcription factors through dna shape. *Cell Rep*, 3(4):10931104, 2013.

[59] C Goudot, C Etchebest, F Devaux, and G Lelandais. The reconstruction of condition-specific transcriptional modules provides new insights in the evolution of yeast ap-1 proteins. *PLoS One*, 6:e20924, 2011.

[60] S Guindon, JF Dufayard, V Lefort, M Anisimova, W Hordijk, and O Gascuel. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. *Syst Biol*, 59(3):307321, 2010.

[61] S Guindon and O Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5):696704, 2003.

[62] MV Han, JP Demuth, CL McGrath, C Casola, and MW Hahn. Adaptive evolution of young gene duplicates in mammals. *Genome Res*, 19(5):859867, 2009.

[63] MJ Harms and JW Thornton. Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol*, 20(3):360366, 2010.

[64] MJ Harms and JW Thornton. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet*, 14(8):559571, 2013.

[65] MJ Harms and JW Thornton. Historical contingency and its biophysical basis in glucocorticoid receptor evolution. *Nature*, 512(7513):203207, 2014.

[66] K Harris and R Nielsen. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res*, 24(9):14451454, 2014.

[67] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. *J Mol Evol*, 22(2):160174, 1985.

[68] S Henikoff and JG Henikoff. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, 89(22):1091510919, 1992.

[69] GKA Hochberg and JW Thornton. Reconstructing ancient proteins to understand the causes of structure and function. *Annu Rev Biophys*, 46:247269, 2017.

[70] B Horard and JM Vanacker. Estrogen receptor-related receptors: orphan receptors desperately seeking a ligand. *J Mol Endocrinol*, 31(3):349357, 2003.

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

[72] R Huang, F Hippauf, D Rohrbeck, M Haustein, K Wenke, J Feike, N Sorrelle, B Piechulla, and TJ Barkman. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proc Natl Acad Sci U S A*, 109(8):29662971, 2012.

[73] T Hrd, K Dahlman, J Carlstedt-Duke, JA Gustafsson, and R Rigler. Cooperativity and specificity in the interactions between dna and the glucocorticoid receptor dna-binding domain. *Biochemistry*, 29(22):53585364, 1990.

[74] AH Jan, E Dubreucq, J Drone, and M Subileau. A glimpse into the specialization history of the lipases/acyltransferases family of cplip2. *Biochim Biophys Acta*, 1865(9):11051113, 2017.

[75] P Jiang and M Rausher. Two genetic changes in cis-regulatory elements caused evolution of petal spot position in clarkia. *Nat Plants*, 4(1):1422, 2018.

[76] AD Johnson. The rewiring of transcription circuits in evolution. *Curr Opin Genet Dev*, 47:121127, 2017.

[77] SD Johnston, X Liu, F Zuo, TL Eisenbraun, SR Wiley, RJ Kraus, and JE Mertz. Estrogen-related receptor alpha 1 functionally binds as a monomer to extended half-site sequences including ones contained within estrogen-response elements. *Mol Endocrinol*, 11(3):342352, 1997.

[78] A Jolma, 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. Dna-binding specificities of human transcription factors. *Cell*, 152(1-2):327339, 2013.

[79] A Jolma, Y Yin, KR Nitta, K Dave, A Popov, M Taipale, M Enge, T Kivioja, E Morgunova, and J Taipale. Dna-dependent formation of transcription factor pairs alters their binding specificity. *Nature*, 527(7578):384388, 2015.

[80] R Joshi, JM Passner, R Rohs, R Jain, A Sosinsky, MA Crickmore, V Jacob, AK Aggarwal, B Honig, and RS Mann. Functional specificity of a hox protein mediated by the recognition of minor groove structure. *Cell*, 131(3):530543, 2007.

[81] L Kannan and WC Wheeler. Maximum parsimony on phylogenetic networks. *Algorithms Mol Biol*, 7(1):9, 2012.

[82] M Kazemian, H Pham, SA Wolfe, MH Brodsky, and S Sinha. Widespread evidence of cooperative dna binding by transcription factors in drosophila development. *Nucleic Acids Res*, 41(17):82378252, 2013.

[83] SL Kosakovsky Pond, B Murrell, M Fourment, SD Frost, W Delport, and K Scheffler. A random effects branch-site model for detecting episodic diversifying selection. *Mol Biol Evol*, 28(11):30333043, 2011.

[84] C Kosiol and N Goldman. Different versions of the dayhoff rate matrix. *Mol Biol Evol*, 22(2):193199, 2005.

[85] C Kosiol, I Holmes, and N Goldman. An empirical codon model for protein sequence evolution. *Mol Biol Evol*, 24(7):14641479, 2007.

[86] C Kosiol, T Vinar, RR da Fonseca, MJ Hubisz, CD Bustamante, R Nielsen, and A Siepel. Patterns of positive selection in six mammalian genomes. *PLoS Genet*, 4(8):e1000144, 2008.

[87] C Lanave, G Preparata, C Saccone, and G Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):8693, 1984.

[88] AM Larracuente, TB Sackton, AJ Greenberg, A Wong, ND Singh, D Sturgill, Y Zhang, B Oliver, and AG Clark. Evolution of protein-coding genes in drosophila. *Trends Genet*, 24(3):114123, 2008.

[89] SQ Le and O Gascuel. An improved general amino acid replacement matrix. *Mol Biol Evol*, 25(7):13071320, 2008.

[90] TH Little, Y Zhang, CK Matulis, J Weck, Z Zhang, A Ramachandran, KE Mayo, and I Radhakrishnan. Sequence-specific deoxyribonucleic acid (dna) recognition by steroidogenic factor 1: a helix at the carboxy terminus of the dna binding domain is necessary for complex stability. *Mol Endocrinol*, 20(4):831843, 2006.

[91] S Liu, 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 OConnell, JO McInerney, EW Born, L Daln, R Dietz, L Orlando, C Sonne, G Zhang, R Nielsen, E Willerslev, and J Wang. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157(4):785794, 2014.

[92] LA Loeb and RJ Monnat. Dna polymerases and human disease. *Nat Rev Genet*, 9(8):594604, 2008.

[93] P Lopez, D Casane, and H Philippe. Heterotachy, an important process of protein evolution. *Mol Biol Evol*, 19(1):17, 2002.

[94] I Macindoe, L Glockner, P Vukasin, FA Stennard, MW Costa, RP Harvey, JP Mackay, and M Sunde. Conformational stability and dna binding specificity of the cardiac t-box transcription factor tbx20. *J Mol Biol*, 389(3):606618, 2009.

[95] L Mariani, K Weinand, A Vedenko, LA Barrera, and ML Bulyk. Identification of human lineage-specific transcriptional coregulators enabled by a glossary of binding modules and tunable genomic backgrounds. *Cell Syst*, 5(3):187201.e7, 2017.

[96] T Matsuda, K Bebenek, C Masutani, F Hanaoka, and TA Kunkel. Low fidelity dna synthesis by human dna polymerase-eta. *Nature*, 404(6781):10111013, 2000.

[97] JH McDonald and M Kreitman. Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351(6328):652654, 1991.

[98] AN McKeown, JT Bridgham, DW Anderson, MN Murphy, EA Ortlund, and JW Thornton. Evolution of dna specificity in a transcription factor family produced a new gene regulatory module. *Cell*, 159(1):5868, 2014.

[99] N Mekel-Bobrov, SL Gilbert, PD Evans, EJ Vallender, JR Anderson, RR Hudson, SA Tishkoff, and BT Lahn. Ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens. *Science*, 309(5741):17201722, 2005.

[100] VS Melvin, C Harrell, JS Adelman, WL Kraus, M Churchill, and DP Edwards. The role of the c-terminal extension (cte) of the estrogen receptor alpha and beta dna binding domain in dna binding and interaction with hmgb. *J Biol Chem*, 279(15):1476314771, 2004.

[101] DW Mount. A test of the markov model of evolution in proteins. *CSH Protoc*, 2008:pdb.ip58, 2008.

[102] B Murrell, S Weaver, MD Smith, JO Wertheim, S Murrell, A Aylward, K Eren, T Pollner, DP Martin, DM Smith, K Scheffler, and SL Kosakovsky Pond. Gene-wide identification of episodic selection. *Mol Biol Evol*, 32(5):13651371, 2015.

[103] B Murrell, JO Wertheim, S Moola, T Weighill, K Scheffler, and SL Kosakovsky Pond. Detecting individual sites subject to episodic diversifying selection. *PLoS Genet*, 8(7):e1002764, 2012.

[104] SV Muse and BS Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol*, 11(5):715724, 1994.

[105] S Nakagawa, SS Gisselbrecht, JM Rogers, DL Hartl, and ML Bulyk. Dna-binding specificity changes in the evolution of forkhead transcription factors. *Proc Natl Acad Sci U S A*, 110(30):1234912354, 2013.

[106] KR Nitta, A Jolma, Y Yin, E Morgunova, T Kivioja, J Akhtar, K Hens, J Toivonen, B Deplancke, EE Furlong, and J Taipale. Conservation of transcription factor binding specificities across 600 million years of bilateria evolution. *Elife*, 4, 2015.

[107] M Nozawa, Y Suzuki, and M Nei. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc Natl Acad Sci U S A*, 106(16):67006705, 2009.

[108] EA Ortlund, JT Bridgham, MR Redinbo, and JW Thornton. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science*, 317(5844):15441548, 2007.

[109] CO Pabo and RT Sauer. Transcription factors: structural families and principles of dna recognition. *Annu Rev Biochem*, 61:10531095, 1992.

[110] SK Pond and SV Muse. Site-to-site variation of synonymous substitution rates. *Mol Biol Evol*, 22(12):23752385, 2005.

[111] SL Pond, SD Frost, and SV Muse. Hyphy: hypothesis testing using phylogenies. *Bioinformatics*, 21(5):676679, 2005.

[112] K Pougach, A Voet, FA Kondrashov, K Voordeckers, JF Christiaens, B Baying, V Benes, R Sakai, J Aerts, B Zhu, P Van Dijck, and KJ Verstrepen. Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nat Commun*, 5:4868, 2014.

[113] PL Privalov, AI Dragan, C Crane-Robinson, KJ Breslauer, DP Remeta, and CA Minetti. What drives proteins into the major or minor grooves of dna. *J Mol Biol*, 365(1):19, 2007.

[114] W Qian, JR Yang, NM Pearson, C Maclean, and J Zhang. Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet*, 8(3):e1002603, 2012.

[115] VA Risso, JA Gavira, DF Mejia-Carmona, EA Gaucher, and JM Sanchez-Ruiz. Hyperstability and substrate promiscuity in laboratory resurrections of precambrian -lactamases. *J Am Chem Soc*, 135(8):28992902, 2013.

[116] MJ Roeske, EM Camino, S Grover, M Rebeiz, and TM Williams. Cis-regulatory evolution integrated the bric--brac transcription factors into a novel fruit fly gene regulatory network. *Elife*, 7, 2018.

[117] IB Rogozin, F Belinky, V Pavlenko, SA Shabalina, DM Kristensen, and EV Koonin. Evolutionary switches between two serine codon sets are driven by selection. *Proc Natl Acad Sci U S A*, 113(46):1310913113, 2016.

[118] R Rohs, X Jin, SM West, R Joshi, B Honig, and RS Mann. Origins of specificity in protein-dna recognition. *Annu Rev Biochem*, 79:233269, 2010.

[119] R Rohs, SM West, A Sosinsky, P Liu, RS Mann, and B Honig. The role of dna shape in protein-dna recognition. *Nature*, 461(7268):12481253, 2009.

[120] J Roux, E Privman, S Moretti, JT Daub, M Robinson-Rechavi, and L Keller. Patterns of positive selection in seven ant genomes. *Mol Biol Evol*, 31(7):16611685, 2014.

[121] S Rowan, T Siggers, SA Lachke, Y Yue, ML Bulyk, and RL Maas. Precise temporal control of the eye regulatory gene pax6 via enhancer-binding site affinity. *Genes Dev*, 24(10):980985, 2010.

[122] PC Sabeti, P Varilly, B Fry, J Lohmueller, E Hostetter, C Cotsapas, X Xie, EH Byrne, SA McCarroll, R Gaudet, SF Schaffner, ES Lander, HapMap Consortium International, KA Frazer, DG Ballinger, DR Cox, DA Hinds, LL Stuve, RA Gibbs, JW Belmont, A Boudreau, P Hardenbol, SM Leal, S Pasternak, DA Wheeler, TD Willis, F Yu, H Yang, C Zeng, Y Gao, H Hu, W Hu, C Li, W Lin, S Liu, H Pan, X Tang, J Wang, W Wang, J Yu, B Zhang, Q Zhang, H Zhao, H Zhao, J Zhou, SB Gabriel, R Barry, B Blumenstiel, A Camargo, M Defelice, M Faggart, M Goyette, S Gupta, J Moore, H Nguyen, RC Onofrio, M Parkin, J Roy, E Stahl, E Winchester, L Ziaugra, D Altshuler, Y Shen, Z Yao, W Huang, X Chu, Y He, L Jin, Y Liu, Y Shen, W Sun, H Wang, Y Wang, Y Wang, X Xiong, L Xu, MM Waye, SK Tsui, H Xue, JT Wong, LM Galver, JB Fan, K Gunderson, SS Murray, AR Oliphant, MS Chee, A Montpetit, F Chagnon, V Ferretti, M Leboeuf, JF Olivier, MS Phillips, S Roumy, C Salle, A Verner, TJ Hudson, PY Kwok, D Cai, DC Koboldt, RD Miller, L Pawlikowska, P Taillon-Miller, M Xiao, LC Tsui, W Mak, YQ Song, PK Tam, Y Nakamura, T Kawaguchi, T Kitamoto, T Morizono, A Nagashima, Y Ohnishi, A Sekine, T Tanaka, T Tsunoda, P Deloukas, CP Bird, M Delgado, ET Dermitzakis, R Gwilliam, S Hunt, J Morrison, D Powell, BE Stranger, P Whittaker, DR Bentley, MJ Daly, PI de Bakker, J Barrett, YR Chretien, J Maller, S McCarroll, N Patterson, I Peer, A Price, S Purcell, DJ Richter, P Sabeti, R Saxena, SF Schaffner, PC Sham, P Varilly, D Altshuler, LD Stein, L Krishnan, AV Smith, MK Tello-Ruiz, GA Thorisson, A Chakravarti, PE Chen, DJ Cutler, CS Kashuk, S Lin, GR Abecasis, W Guan, Y Li, HM Munro, ZS Qin, DJ Thomas, G McVean, A Auton, L Bottolo, N Cardin, S Eyheramendy, C Freeman, J Marchini, S Myers, C Spencer, M Stephens, P Donnelly, LR Cardon, G Clarke, DM Evans, AP Morris, BS Weir, T Tsunoda, TA Johnson, JC Mullikin, ST Sherry, M Feolo, A Skol, H Zhang, C Zeng, H Zhao, I Matsuda, Y Fukushima, DR Macer, E Suda, CN Rotimi, CA Adebamowo, I Ajayi, T Aniagwu, PA Marshall, C Nkwodimmah, CD Royal, MF Leppert, M Dixon, A Peiffer, R Qiu, A Kent, K Kato, N Niikawa, IF Adewole, BM Knoppers, MW Foster, EW Clayton, J Watkin, RA Gibbs, JW Belmont, D Muzny, L Nazareth, E Sodergren, GM Weinstock, DA Wheeler, I Yakub, SB Gabriel, RC Onofrio, DJ Richter, L Ziaugra, BW Birren, MJ Daly, D Altshuler, RK Wilson, LL Fulton, J Rogers, J Burton, NP Carter, CM Clee, M Griffiths, MC Jones, K McLay, RW Plumb, MT Ross, SK Sims, DL Willey, Z Chen, H Han, L Kang, M Godbout, JC Wallenburg, P LArchevque, G Bellemare, K Saeki, H Wang, D An, H Fu, Q Li, Z Wang, R Wang, AL Holden, LD Brooks, JE McEwen, MS Guyer, VO Wang, JL Peterson, M Shi, J Spiegel, LM Sung, LF Zacharia, FS Collins, K Kennedy, R Jamieson, and

J Stewart. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449(7164):913918, 2007.

[123] J Santiago-Ortiz, DS Ojala, O Westesson, JR Weinstein, SY Wong, A Steinsapir, S Kumar, I Holmes, and DV Schaffer. Aav ancestral reconstruction library enables selection of broadly infectious viral variants. *Gene Ther*, 22(12):934946, 2015.

[124] H Saribasak, RW Maul, Z Cao, WW Yang, D Schenten, S Kracker, and PJ Gearhart. Dna polymerase generates tandem mutations in immunoglobulin variable regions. *J Exp Med*, 209(6):10751081, 2012.

[125] DR Schrider, JN Hourmozdi, and MW Hahn. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol*, 21(12):10511054, 2011.

[126] JW Schwabe, D Neuhaus, and D Rhodes. Solution structure of the dna-binding domain of the oestrogen receptor. *Nature*, 348(6300):458461, 1990.

[127] DF Senear and M Brenowitz. Determination of binding constants for cooperative site-specific protein-dna interactions using the gel mobility-shift assay. *J Biol Chem*, 266(21):1366113671, 1991.

[128] VB Seplyarskiy, GA Bazykin, and RA Soldatov. Polymerase activity is linked to replication timing in humans: Evidence from mutational signatures. *Mol Biol Evol*, 32(12):31583172, 2015.

[129] D Shriner, DC Nickle, MA Jensen, and JI Mullins. Potential impact of recombination on sitewise approaches for detecting positive natural selection. *Genet Res*, 81(2):115121, 2003.

[130] MA Siddiq, GK Hochberg, and JW Thornton. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr Opin Struct Biol*, 47:113122, 2017.

[131] T Siggers, J Reddy, B Barron, and ML Bulyk. Diversification of transcription factor paralogs via noncanonical modularity in c2h2 zinc finger dna binding. *Mol Cell*, 55(4):640648, 2014.

[132] M Sironi, R Cagliani, D Forni, and M Clerici. Evolutionary insights into host-pathogen interactions from mammalian sequence data. *Nat Rev Genet*, 16(4):224236, 2015.

[133] ST Smale. Dimer-specific regulatory mechanisms within the nf-b family of transcription factors. *Immunol Rev*, 246(1):193204, 2012.

[134] MD Smith, JO Wertheim, S Weaver, B Murrell, K Scheffler, and SL Kosakovsky Pond. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol Biol Evol*, 32(5):13421353, 2015.

[135] TR Sorrells, LN Booth, BB Tuch, and AD Johnson. Intersecting transcription networks constrain gene regulatory evolution. *Nature*, 523(7560):361365, 2015.

[136] F Spitz and EE Furlong. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet*, 13(9):613626, 2012.

[137] A Stamatakis. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22(21):26882690, 2006.

[138] TN Starr, LK Picton, and JW Thornton. Alternative evolutionary histories in the sequence space of an ancient protein. *Nature*, 549(7672):409413, 2017.

[139] TN Starr and JW Thornton. Epistasis in protein evolution. *Protein Sci*, 25(7):12041218, 2016.

[140] K Stefflova, D Thybert, MD Wilson, I Streeter, J Aleksic, P Karagianni, A Brazma, DJ Adams, I Talianidis, JC Marioni, P Flicek, and DT Odom. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530540, 2013.

[141] PA Steindel, EH Chen, JD Wirth, and DL Theobald. Gradual neofunctionalization in the convergent evolution of trichomonad lactate and malate dehydrogenases. *Protein Sci*, 25(7):13191331, 2016.

[142] TA Steitz. Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Q Rev Biophys*, 23(3):205280, 1990.

[143] AB Stergachis, E Haugen, A Shafer, W Fu, B Vernot, A Reynolds, A Raubitschek, S Ziegler, EM LeProust, JM Akey, and JA Stamatoyannopoulos. Exonic transcription factor binding directs codon choice and affects protein evolution. *Science*, 342(6164):13671372, 2013.

[144] JE Stone, SA Lujan, TA Kunkel, and TA Kunkel. Dna polymerase zeta generates clustered mutations during bypass of endogenous dna lesions in saccharomyces cerevisiae. *Environ Mol Mutagen*, 53(9):777786, 2012.

[145] GD Stormo, Z Zuo, and YK Chang. Spec-seq: determining protein-dna-binding specificity by sequencing. *Brief Funct Genomics*, 14(1):3038, 2015.

[146] E Sucena and DL Stern. Divergence of larval morphology between drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. *Proc Natl Acad Sci U S A*, 97(9):45304534, 2000.

[147] Y Suzuki. False-positive results obtained from the branch-site test of positive selection. *Genes Genet Syst*, 83(4):331338, 2008.

[148] K Tamura and M Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. *Mol Biol Evol*, 10(3):512526, 1993.

[149] SA Tishkoff, FA Reed, A Ranciaro, BF Voight, CC Babbitt, JS Silverman, K Powell, HM Mortensen, JB Hirbo, M Osman, M Ibrahim, SA Omar, G Lema, TB Nyambo, J Ghori, S Bumpstead, JK Pritchard, GA Wray, and P Deloukas. Convergent adaptation of human lactase persistence in africa and europe. *Nat Genet*, 39(1):3140, 2007.

[150] JM Vanacker, K Pettersson, JA Gustafsson, and V Laudet. Transcriptional targets shared by estrogen receptor- related receptors (errs) and estrogen receptor (er) alpha, but not by erbeta. *EMBO J*, 18(15):42704279, 1999.

[151] S Whelan and N Goldman. Estimating the frequency of events that cause multiple-nucleotide changes. *Genetics*, 167(4):20272043, 2004.

[152] JC Wilgenbusch and D Swofford. Inferring evolutionary trees with paup*. *Curr Protoc Bioinformatics*, Chapter 6:Unit 6.4, 2003.

[153] TE Wilson, TJ Fahrner, and J Milbrandt. The orphan receptors ngfi-b and steroidogenic factor 1 establish monomer binding as a third paradigm of nuclear receptor-dna interaction. *Mol Cell Biol*, 13(9):57945804, 1993.

[154] MA Wouters, K Liu, P Riek, and A Husain. A despecialization step underlying evolution of a family of serine proteases. *Mol Cell*, 12(2):343354, 2003.

[155] F Yang, D Tang, Y Bai, M Zhao, and Q Zhu. [recent progress in multiple sequence alignment]. *Sheng Wu Yi Xue Gong Cheng Xue Za Zhi*, 27(4):924928, 2010.

[156] Z Yang. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol*, 11(9):367372, 1996.

[157] Z Yang. Paml 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, 24(8):15861591, 2007.

[158] Z Yang and M dos Reis. Statistical properties of the branch-site test of positive selection. *Mol Biol Evol*, 28(3):12171228, 2011.

[159] Z Yang and R Nielsen. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*, 19(6):908917, 2002.

[160] Z Yang and B Rannala. Molecular phylogenetics: principles and practice. *Nat Rev Genet*, 13(5):303314, 2012.

[161] M Yavartanoo and JK Choi. Encode: A sourcebook of epigenomes and chromatin language. *Genomics Inform*, 11(1):26, 2013.

[162] F Yu, RS Hill, SF Schaffner, PC Sabeti, ET Wang, AA Mignault, RJ Ferland, RK Moyzis, CA Walsh, and D Reich. Comment on ongoing adaptive evolution of aspm, a brain size determinant in homo sapiens. *Science*, 316(5823):370, 2007.

[163] A Zandvakili and B Gebelein. Mechanisms of specificity for hox factor activity. *J Dev Biol*, 4(2), 2016.

[164] W Zhai, R Nielsen, N Goldman, and Z Yang. Looking for darwin in genomic sequences–validity and success of statistical methods. *Mol Biol Evol*, 29(10):28892893, 2012.

[165] J Zhang. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. *Mol Biol Evol*, 16(6):868875, 1999.

[166] J Zhang. Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol*, 21(7):13321339, 2004.

[167] J Zhang, R Nielsen, and Z Yang. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol*, 22(12):24722479, 2005.

[168] W Zhu, DN Cooper, Q Zhao, Y Wang, R Liu, Q Li, C Frec, Y Wang, and JM Chen. Concurrent nucleotide substitution mutations in the human genome are characterized by a significantly decreased transition/transversion ratio. *Hum Mutat*, 36(3):333341, 2015.

[169] RA Zirngibl, JS Chan, and JE Aubin. Estrogen receptor-related receptor alpha (erralpha) regulates osteopontin expression through a non-canonical erralpha response element in a cell context-dependent manner. *J Mol Endocrinol*, 40(2):6173, 2008.