THE UNIVERSITY OF CHICAGO

COMPUTATIONAL APPROACHES TOWARDS THE INTEGRATION OF FUNCTIONAL AND COMPARATIVE DATASETS IN THE EVOLUTION OF GENE REGULATION

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 ECOLOGY AND EVOLUTION

BY
ROBERT K. ARTHUR

CHICAGO, ILLINOIS
DECEMBER 2015

Horatio:

O day and night, but this is wondrous strange!

Hamlet:

And therefore as a stranger give it welcome.

There are more things in heaven and earth, Horatio,

Than are dreamt of in your philosophy.

—Shakespeare, Hamlet

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I owe enormous appreciation to the many kind people who helped me finish this dissertation. I cannot properly thank all the people who deserve it, because there are too many and their contributions are too immense. It would take me another dissertation's worth of text to detail them properly. In lieu of that, please accept my deep gratitude even if your name is not specifically mentioned.

I am grateful to my committee, who helped me through every stage of this process. As a group, you challenged me to think clearly and present my work logically, and never accepted anything but the best I had to give. For that alone, I am thankful.

Each member of my committee proved especially helpful at one time or another. I am thankful to Joe Thornton for helping my design experiments to test Zeus' function, and pushing me to be precise. I am thankful to Martin Kreitman for walking me through the logic of several concepts in molecular evolution, and blessing my attempts to measure selective regimes in cis-regulatory elements. I owe Manyuan Long much gratitude for teaching me about duplicate genes, and, more generally, how to be a better scientist.

I am especially grateful to my two advisors for their patience and willingness to teach me. It was only through their help that I have progressed as far as I have. Both accepted me warmly into their labs, even though I often split my time between tasks and subjects. I believe I am a much better scientist and thinker for having been exposed to the diverse influences and subjects that my coadvisors' two labs permitted me to experience.

I was blessed to find a community of students and friends at the University of Chicago the likes of which I had never encountered before. I will always treasure the people I met and the relationships I formed, and it is only with deep regret that I find myself moving on.

Finally, I thank my family, who have always trusted and supported me.

ABSTRACT

The evolution of gene regulatory sequences is a complex subject, fraught with difficulties. In the following dissertation, I apply novel, functional datasets to better understand questions within the field of regulatory evolution. By computationally integrating data from these emerging techniques, I argue that we can better understand how selection operates on regulatory genes and elements. In Chapter 1, I introduce the outline of the subject matter, and sketch an argument for bringing genomic datasets to bear. In Chapter 2, I probe the relationship between function and evolutionary rate, using a meta-analysis of existing datasets which had tested the extent to which single-nucleotide substitutions could change gene expression. In Chapter 3, I examine function and conservation through the lens of the architecture of cis-regulatory elements. By combining ChIP-seq and DNase-seq datasets, I show that we can interrogate the arrangement of bound sites within elements, and that this information bears on the probability that such elements will be found in different species. In Chapter 4, I turn to another layer of the regulatory apparatus, the histone modification H3K27me3. I show that the evolution of this modification is coupled to gene duplication in a variety of interesting ways. One gene which shows an especially rapid pattern of H3K27me3 evolution is Zeus, whose protein-coding evolution I study further in Chapter 5. In collaboration with Ben Krinsky, we find evidence for a cis-trans coevolutionary process, driven by the emergence of Zeus. Finally, I conclude in Chapter 6 with an overview of the research, some emerging lessons from it, and future frontiers which remain to be addressed.

Chapter 1: Integrating Functional and Evolutionary Data to Resolve Questions in Gene Regulation

Abstract

Gene regulation is the process by which regulatory sequences drive and maintain spatiotemporally-specific patterns of gene expression. In this chapter, I summarize the literature on gene regulatory sequences and their diverse patterns of evolution. I argue that novel techniques to probe biochemical function on a genome-wide scale provide a uniquely powerful means to identify regulatory sequences as well as study their diverse regimes of molecular evolution. While these new technologies offer substantial promise, they also pose unconventional statistical and computation challenges which must be addressed. I conclude by summarizing the dissertation's research chapters, each of which integrates functional datasets to better understand evolutionary patterns within gene regulation.

Introduction

Understanding the evolution of gene regulation is a difficult but important task. A considerable portion of eukaryotic species genomes' is regulatory in nature, whether by virtue of encoding regulatory proteins or as cis-regulatory elements (Kellis et al. 2014). The relative speed with which cis-regulatory sequence evolves suggests that it is an important contributor to adaptation (King and Wilson 1975). Owing to the modular nature of cis-regulatory elements, some have suggested that these elements may harbor a special role in shaping morphological evolution (Carroll 2008). For these reasons, it is vital to better understand the ways in which regulatory elements evolve.

Cis-regulatory elements are not the only contributors to regulatory evolution, however. Other research has shown that transcription factors themselves are some of the most quickly evolving protein-coding genes in the genome (Bustamante et al. 2005; Begun et al. 2007). Since transcription factors bind many thousands of sites throughout the genomes of most multicellular eukaryotes, their rapid evolution may entail correspondingly rapid evolution at target sites.

Chief among the problems in understanding the evolution of gene regulation is the uncertainty inherent in the relationship between functionality and evolutionary rate. While many cis-regulatory elements show patterns of molecular evolution consistent with purifying selection (Dermitzakis et al. 2003), others exhibit rapid turnover of functional sequences (Ludwig et al. 2000). These two contrasting observations suggest that the evolution of cis-regulatory sequence is complex.

Given that *in vivo* studies have demonstrated a fitness cost to mutating cis-regulatory sequences (Parsch et al. 2000), there is undoubtedly a relationship coupling biochemical function at the nucleotide level to the selective forces acting upon those nucleotides. However, extensive prior work shows that this relationship is complicated and subject to redundancy (Barolo 2011). Previous research has shown that sites can be functionally conserved without corresponding nucleotide conservation, and vice versa (Emberly et al. 2003; Balhoff and Wray 2005). Individual binding sites may arise and die, and compensatory events may occur in which the loss of one site is buffered by the gain of another (Bullaughey 2011). As a result of this process, methods of evolutionary analysis which rely upon nucleotide identity may be confounded (Lusk and Eisen 2010). Conservation is not necessarily a prerequisite for function.

Moreover, cis-regulatory sequences in the genome may be under a complicated mix of selective forces which is heterogeneous both through time and across nucleotides (Rockman et al. 2003; He et al. 2011). Transcription factor binding sites (TFBSs)—arguably the core unit of cis-regulatory function—constitute short, degenerate fragments, often surrounded on either side by nonfunctional sequences. As a result, even adjacent sites in the genome may be under very different levels of constraint, if one is within a TFBS and one outside. Since sites vary with respect to the regime and strength of selection, inappropriately grouping sites together can create a misleading average pattern which is representative of no single site.

As a result of this heterogeneity, it is vitally important to collect data on the precise location of cis-regulatory elements and the particular functional sequences within them. Various experimental approaches have been leveraged to that effect. Many researchers have utilized reporter assays to carefully dissect regulatory regions, attaching putative enhancers to reporter genes in order to study the degree and pattern of expression such sequences drive (Tomancak et al. 2002). Further work in this vein characterizes individual functional sequences such as transcription factor binding sites, mutating sites one by one to determine how they affect function (Maniatis et al. 1987).

While these approaches are extremely powerful, the need to select regulatory elements to study is an important limitation. Since investigators have to pre-specify the genes and elements to study, there is risk that the selected elements constitute a non-random and potentially biased subset of the genome’s collection (Tabor et al. 2002). Such biases could interfere with our ability to generalize the observed patterns across all regulatory elements. Similarly, the manual nature of these approaches limits their throughput, preventing a wide sampling of the genome’s regulatory architecture.

In recent years, modern techniques have permitted the efficient gathering of genome-wide functional data, thereby circumventing any need to select candidate regions of the genome. With the advent of next-generation sequencing, many laboratory techniques have been devised to selectively enrich pools of DNA or RNA which can then be sequenced. In essence, sequencing has been harnessed as a method of high-throughput and even genome-wide quantitation. (Such quantitation was possible before sequencing via microarray hybridization, but these chip-based approaches were severely limited in resolution [Park 2009], which is a particularly important limitation given the relatively small size of regulatory elements and TFBSs).

For example, Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) allows a researcher to interrogate every nucleotide in the genome with respect to its binding by some factor. In contrast to earlier, candidate-based methods, ChIP-seq can identify a majority of the bound regions within the genome. Another major technical innovation is RNA sequencing, or RNA-seq. By sequencing the pool of RNA within a cell (or collection thereof), one can collect highly accurate, quantitative estimates of the abundance of each transcript.

By combining quantitative estimates of gene expression with corresponding assays for regulatory activity or transcription factor binding, one can connect patterns of regulatory activity to the expression of the genes which they drive. In concert, performing these assays across species allows the interrogation of not only gene expression divergence, but potentially the regulatory mechanisms of that divergence (Ni et al. 2012). These and other techniques open a novel path towards understanding how function relates to evolutionary patterns across the genome in an unbiased and direct way.

The technological innovations of sequencing-based methods also create substantial computational, technical and statistical problems. One category of problem is scale-related: the genome is very large, and the size of the inference problems can impose a significant multiple testing burden (Marchini et al. 2005; Storey and Tibshirani 2003). Another relates to the application of these data-collection methods on multicellular, developing organisms, which represent a mix of spatiotemporal gene expression profiles (Arthur et al. 2014). A final issue concerns making reliable inferences across species, where annotation can be uneven and genome quality is a serious concern. All of these problems (and others besides) must be addressed before progress can be made in correctly using these datasets.

Summary of Research

In each of the next four chapters, I take a kind of functional data and try to better understand how the biochemical function it describes affects, and is affected by, patterns of sequence evolution. I address the computational challenges that are raised in each type of data and the kinds of limitations that they impose. In spite of the difficulties, significant steps forward are possible in our understanding, and I explore the means by which these data can be made useful.

The several broad categories of problems I noted above can temper the power of genome-scale datasets, but also create an opportunity to carefully leverage these data. In this dissertation, Each of the four data chapters introduces different functional datasets and how they change our perception of the function/conservation relationship in regulatory sequence. I summarize each chapter below.

In Chapter Two, I gather published datasets measuring the extent to which different sites affect gene expression. Using mutagenesis studies, we can directly compare the extent to which a site affects gene expression and the degree to which it is conserved. We find that there is a relationship: sites which affect expression are more likely to be conserved between species.

In the third chapter, I investigate how functional and nucleotide level conservation are linked in cis regulatory elements. By utilizing previously published ChIP-seq data across a number of species, as well as a novel computational approach, we are able to ask how architectural features of enhancers and binding sites are related to functional conservation. Using a number of such features, we build a classification model which can correctly predict functional conservation with high accuracy across a number of different kinds of cis-regulatory elements.

In Chapter Four, I study a different layer of gene regulation: chromatin. An important recent insight in molecular biology is that the genome is spatially structured into chromatin domains which can repress or activate genes over large spans (Kharchenko et al. 2011). In this chapter, I examine the functional conservation of different domains of a single chromatin mark, H3K27me3. I find a strong dichotomy in rate of functional evolution that is coupled to rate of molecular evolution. Specifically, I find that duplicated genes evolve much more rapidly with respect to chromatin state, and that chromatin state can affect the probability of retention of duplicate genes.

In the final data chapter, I study the manner in which a novel, rapidly evolving gene called Zeus has integrated into an existing network of transcriptional activity. In collaboration with Benjamin Krinsky, we find strong evidence that the rapid molecular evolution of the gene was coupled to a wave of positive selection on the motif associated with the gene. Altogether, our results uncover a coevolutionary process operating in cis- and trans-, reflecting the complex relationship between function and conservation across sites.

Finally, I argue in the conclusion that there are several insights to be gained by synthesizing the knowledge from each of these works. By bringing together multiple kinds of data in an informed fashion, one can better approach important questions within the evolution of gene regulation. As the power of our experimental approaches increases further still, new prospects for future research will open, several of which I describe in the concluding section.

Chapter 2: Evidence That Purifying Selection Acts on Promoter Sequences

Abstract

We tested whether functionally important sites in bacterial, yeast, and animal promoters are more conserved than their neighbors. We found that substitutions are predominantly seen in less important sites and those that occurred tended to have less impact on gene expression than possible alternatives. These results suggest that purifying selection operates on cis-regulatory sequences.

Introduction

The study of cis-regulatory evolution presents “challenges beyond those typically encountered in analyses of coding sequence evolution” (Wray et al. 2003). We are currently unable to infer regulatory function from primary sequences and, consequently, do not have a clear understanding of a relationship between function and conservation. Whereas it is clear that many cis-elements are under selective constraint (Bergman and Kreitman 2001; Dermitzakis et al. 2003; Andolfatto 2005; Hahn 2007; Loots and Ovcharenko 2010), in some instances sites known to be functional in one species have been lost in closely related species (Ludwig et al. 2000; Dermitzakis and Clark 2002; Moses et al. 2006; Doniger and Fay 2007; Bradley et al., 2010). Genome annotation approaches, such as “phylogenetic footprinting” (Blanchette and Tompa 2002; Zhang and Gerstein 2003) and “phylogenetic shadowing” (Bofelli et al. 2003), rely on greater conservation of functional sites compared to surrounding sequences, yet this supposition may not always be true (Emberly et al. 2003; Balhoff and Wray 2005). Indeed, positive selection may drive turnover of binding sites (Rockman et al. 2003; He et al. 2011).

Although evidence suggests that fitness costs of mutations in non-coding regions may be relatively low (Kryukov et al. 2005; Chen et al. 2007; Raijman et al. 2008), few studies have explicitly tested the relationship between functions of individual nucleotides and the fitness costs of mutations at these sites (Shultzaberger et al. 2010).

Our knowledge of the forces driving the evolution of cis-elements largely comes from sequence comparisons between and within species, often without specific reference to the function of individual nucleotides within these elements (Wong and Nielsen 2004; Bush and Lahn 2006; Casillas et al. 2007). Yet regulatory functions and constraints are not uniformly distributed within cis-elements.

Binding energy of transcription factor binding sites can be experimentally measured and computationally modeled (Djordjevic et al. 2003; Maerkl and Quake 2007; Weindl et al. 2007; Zhao et al. 2009). Modeling and comparative sequence analyses suggest that selection effects on binding sites may be mediated by their binding energy (Mustonen and Lässig 2005; Mustonen et al. 2008). Within transcription factor binding sites, substitutions occur at position-specific rates (Tanay et al. 2004; Kim et al. 2009). Specifically, the degree of conservation of individual nucleotides is proportional to their information content, likely because sites that make direct contact with transcription factors tend to be highly conserved (Mirny and Gelfand 2002). For this reason, it is tempting to use binding energy as a proxy for the functional consequences of mutation at a site.

However, the relationship between them is not well understood (Mirny and Gelfand 2002). In some instances there exists a correlation between binding energy and substitution rate (Brown and Callan 2004), but this may not always be the case (Kotelnikova et al. 2005).

Furthermore, nonbinding nucleotides may exert some effect on transcription (Mirny and Gelfand 2002; Wozniak and Hughes 2008) and potentially fitness.

A comprehensive understanding of the evolution of cis-regulatory sequences will require the synthesis of knowledge concerning binding energy, function, and fitness consequences of individual mutations within these elements. Because such data are not generally available, it would be desirable to ascertain whether a relationship exists between functions of specific nucleotides within cis-elements and their rates of evolution. Such analyses would constitute a critical link between functional studies and comparative sequence analysis.

Methods

Data

Functional data were derived from published articles reporting studies of promoter mutagenesis (Table S1 of Arthur and Ruvinsky 2011). A "complete" data set for a given position would contain the information on the consequences of changing the wild type nucleotide to every one of the three alternatives. Altogether, our data set contained 165 base pairs examined in such a way. There were also 209 nucleotides with "incomplete" data, i.e. situations when information was only available for one or two of the three possible substitutions. Although high-throughput mutagenesis data are available for several additional promoters (Patwardhan et al. 2009), we found that these data were inconsistent with the results of single-gene studies, even on the same promoter (data not shown). We therefore did not include them in the present analysis.

For each cis-regulatory element for which mutagenesis data were available, we identified orthologous sequences in a number of closely related species (Arthur and Ruvinsky 2011). In each broad taxonomic group (bacteria, yeast, animals), we endeavored to align sequence from a set of species of roughly equivalent phylogenetic distance (measured by the metric of substitutions per base pair). In counting substitutions, we took into account the phylogenetic relationship of the species being compared. For instance, substitutions in sister species that could be parsimoniously attributed to the common ancestor of these species, were counted once, not twice.

Statistical analyses

We calculated “mutation cost index” for all experimentally characterized mutations; it is a measure of the extent to which a mutation alters promoter function. Expression levels of all mutagenized promoters were normalized to the expression levels of the wild type promoter and then subtracted from one (in rare instances when the mutant promoter drove higher expression, the inverse of the normalization ratio was recorded instead). For a mutation that reduced gene expression to α (normalized to the level of the wild type promoter), this index was defined as $1-\alpha$, therefore mutation cost index can range from 0 (no alteration of expression level) to 1 (complete abrogation of promoter function). Similarly, every nucleotide within a promoter can be said to have a “site index”, computed as a sum of mutation cost indexes of all three possible substitutions. Site index can range from 0 (all mutations are inconsequential to promoter function) to 3 (all mutations at the site abolish expression).

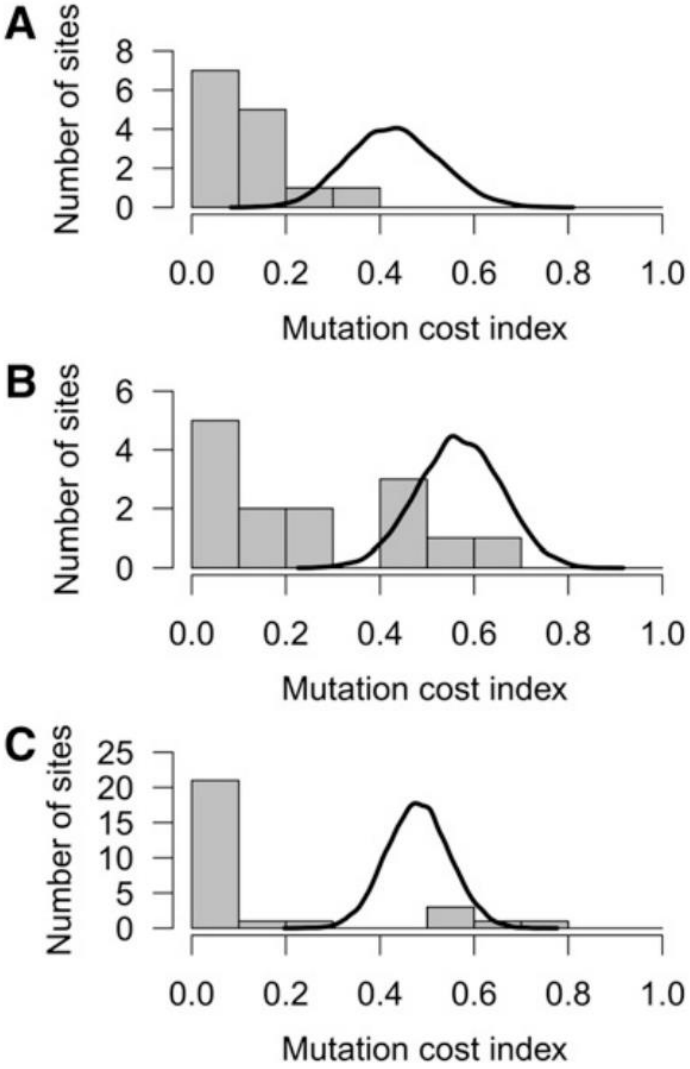
To test whether mutation cost index was significantly lower for substitutions than for all possible mutations, as would be expected under purifying selection, we performed sampled randomization tests in which artificial datasets were generated by randomly sampling from the set of all mutations (Sokal and Rohlf 1995). Each artificially-generated set matched the substitution data in the number of mutations, but differed in the specific mutations sampled. In the most general version of the test the artificially-generated sets were randomly drawn from all experimentally tested mutations.

We performed three variations of this test, each of which constrained certain characteristics of the sampled sets. In the first, artificial sets were constructed to have the same frequency of nucleotides as that of the sites that sustained substitutions. In the second, artificial sets were matched in nucleotide frequencies to that of derived nucleotides (i.e. those to which substitutions changed the ancestral nucleotides). In the third, the numbers of transition and transversion mutations were matched between the set of substitutions and the artificially-generated sets. Sampled randomization tests were performed separately for bacteria, yeast, and animals. We performed at least 10,000 tests and calculated the fraction of instances in which the mean of an artificial set was lower than or equal to that of the substitutions set. This ratio, which represents the probability that the observed set of substitutions would occur by chance alone, constituted the reported p-value. We chose this method because the distributions of mutation cost indexes were highly non-normal and the substitutions represented a subset of all mutations. The sampled randomization test makes no assumptions about the underlying data. It reports the likelihood that the observed dataset resembles a randomly chosen dataset in regards to certain summary statistics (Sokal and Rohlf 1995). All statistical analyses were performed in the R statistical programming language (http://www.r-project.org).

Results

We explicitly tested whether functionally important nucleotides within promoters evolve under the same regime as their neighbors. A number of studies have been published in which individual nucleotides in a given promoter were replaced, while holding all other nucleotides constant (e.g. Maniatis et al. 1987). Most commonly, mutagenized promoters were fused to reporter genes to compare their levels of expression to wild type promoters. These tests measured the impact of each nucleotide substitution on the level of expression. For example, at a given site, an A could be a wild type nucleotide, while mutations to C, G, and T could reduce gene expression to 10, 30, and 60% of the wild type level, respectively. Combining these functional data with analysis of orthologous promoters could establish a relationship between function and rates of evolution. We assembled a dataset of 13 such studies (Arthur and Ruvinsky 2011), conducted on organisms from three distinct phylogenetic groups: animals (4), yeast (5), and bacteria (4). Together, these papers reported mutagenesis of 315 nucleotides and examined expression levels of 972 constructs (animals: 136 nucleotides, 350 constructs; yeast: 79, 275; bacteria: 100, 347). Of all these experimentally tested nucleotides, 53 were inferred to have sustained substitutions (Figure S1). While limited in size, we believe this dataset is a near-exhaustive collection of published articles reporting experiments of this type.

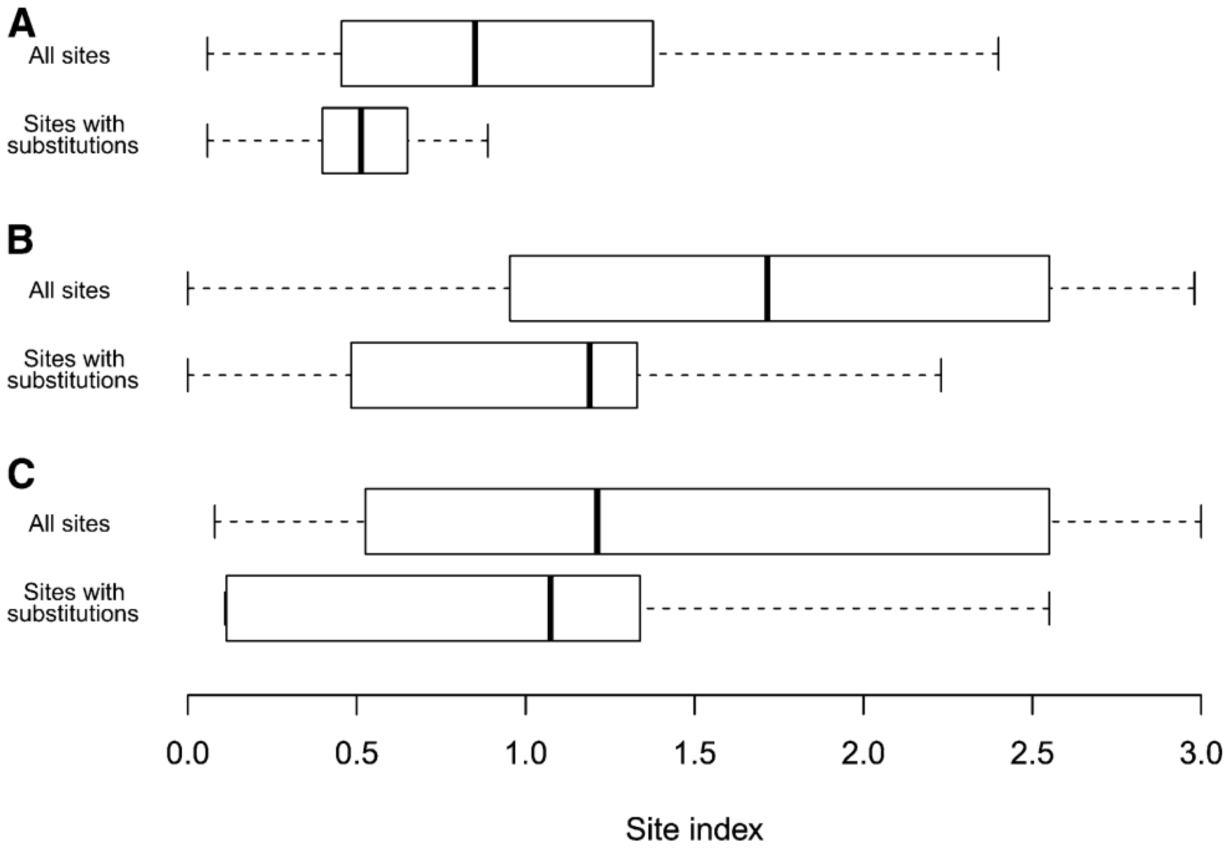
Figure 2.1. Mutation cost index of substitutions. Substitutions (gray bars) in (A) bacterial, (B) yeast, and (C) animal promoters have significantly milder effects on levels of gene expression than datasets with randomly selected mutations (the means of which are shown as black curves). We performed sampled randomization tests in which artificial datasets were generated by randomly sampling from the set of all mutations (Sokal and Rohlf, 1995). We performed at least 10,000 tests and calculated the fraction of instances in which the mean of an artificial set was lower than or equal to that of the substitutions set.



Milder mutations were preferentially fixed during evolution

It may be expected that the effects on gene expression of substitutions that accumulated during evolution would be less severe than the effects of average mutations that could have occurred within these promoters. We tested this hypothesis (Figure 1). From the published data, we calculated mutation cost indexes. We found that the mutations corresponding to substitutions have less severe impacts on expression than average mutations (bacteria: p = 6 x 10-4; yeast: p = 4 x 10-4; animals: p < 10-5). Therefore, among the substitutions that did occur, there was a substantial bias in favor of changes with lower impact on gene expression. This implies that purifying selection has acted to maintain the levels of gene expression.

Figure 2.2. Site index of substitutions compared to all sites. Substitutions in (A) bacterial, (B) yeast, and (C) animal promoters tended to occur at sites with less severe impact on expression. "Site indexes" of fixed substitutions are shown in gray.



Substitutions preferentially occurred at sites with milder effects on expression

Results in Figure 1 suggest that, in general, mutations with lower effects on promoter function tended to become fixed. Two distinct scenarios could account for this trend. First, the milder fixed substitutions could be distributed relatively evenly across sites. Alternatively, they may preferentially occur at a particular subset of sites. We used the functional data described above to test the hypothesis that substitutions are more common at sites where mutations have less severe effect on gene expression levels (Figure 2). One measure of functional importance of a site is an index defined as a sum of mutation cost indexes (α) of all three possible mutations that could occur at this nucleotide. Site indexes were significantly lower for positions with substitutions compared to all sites for which experimental mutagenesis data were available (bacteria: p = 1.1 x 10-3; yeast = 6.4 x 10-3; animals: p = 3.4 x 10-3). Therefore, in all three groups, substitutions preferentially occurred at sites that were less disruptive of gene expression.

Mutational biases are not sufficient to account for the trends reported above. First, in the promoter sequences we analyzed there was no systematic difference in nucleotide composition between sites that sustained substitutions and those that did not (Arthur and Ruvinsky 2011). Second, correcting for multiple hypothesis testing, there were no significant differences in mutation cost indexes between mutations involving different wild type nucleotides (Figure S2). Finally, we repeated sampled randomization tests holding constant the number of (i) wild type (ii) derived nucleotides, and (iii) transitions and transversions. All of these modified tests showed significant differences between mutation cost indexes of substitutions compared to all possible mutations (Table S3).

Discussion

Our results suggest that purifying selection acts on promoter sequences in bacteria, yeast, and animals because we saw fewer than expected substitutions that corresponded to mutations of substantial effect. While these findings are concordant with previous reports of sequence conservation in cis-elements (Andolfatto 2005; Casillas et al. 2007; Molina and van Nimwegen 2008), they add an important functional explanation for the observed patterns. An additional reason for the relative abundance of mutations of smaller effect is that they could be more likely to be beneficial and therefore be fixed by directional selection. Positive selection has been shown to act on cis-regulatory elements (Rockman et al. 2005; Haygood et al. 2007) and it may drive transcription factor binding site turnover (Rockman et al. 2003; He et al. 2011). The inference of both positive and negative selection may not be contradictory, as it has been shown that both types of selection operate on gene regulatory elements in a variety of species (Haddrill et al. 2008; Torgerson et al. 2009). At least some regulatory regions are evolving under stabilizing selection (Ludwig et al. 2000; Loisel et al. 2006).

Five caveats should be noted. First, it is generally not known how changes in the level of expression translate into measures of fitness. However, our conclusions do not require a particular relationship, but merely a positive correlation between the extent to which a mutation changes expression of a gene and its fitness consequences. Available data suggest that such a correlation is likely (Shultzaberger et al. 2010). Second, the set of mutagenized sites was not random in all studies – in some cases experimenters chose sites in which to induce mutations in a way presumably biased in favor of nucleotides expected to have more dramatic effects on gene expression. Third, functional effects of mutations in promoters are highly context-specific. Therefore, fitness consequences of mutations are contingent on the backgrounds on which they occur and may have changed substantially over time (Bullaughey 2011). Fourth, functions of mutated promoters were tested either in cell lines (animals) or under laboratory conditions (bacteria and yeast). This leaves open a possibility that in vivo or under different environmental conditions, mutations seen in the laboratory as “functionally silent” may have substantial impact on fitness. Furthermore, although a point mutation of a given nucleotide may not have caused an appreciable change in expression level, the site may still be under selection, because its deletion could cause a substantial decrease in gene expression (Patwardhan et al. 2009). It appears unlikely, however, that mutations which abrogate or substantially reduce expression are selectively neutral. Finally, all sequences analyzed in this study were derived from proximal promoter elements. The arrangement and composition of functional sites may be different between promoters and other cis-regulatory elements. Therefore, different types of cis-sequences may evolve under different selective regimes. Nonetheless, the results presented here highlight the value of functional data for understanding regulatory evolution.

Acknowledgements

We are grateful to Kevin Bullaughey for numerous improving suggestions and help with data analysis. We thank Bin He and Marty Kreitman for critical reading and helpful suggestions and Chan Hee Choi for help during an early stage of this study. This work was made possible by grant support from the NSF (IOS-0843504) and the NIH (P50 GM081892) to IR and by an NIH training grant (T32 GM007197) to RKA.

Chapter 3: Architectural Characteristics Determine Functional Conservation of Cis-Regulatory Elements

Abstract

Functional conservation of different *cis*-regulatory elements varies widely across the genomes of all organisms, and cannot be determined from sequence conservation alone. Different levels of functional conservation between cis-regulatory elements may be a result of varying levels of selection, architectural constraints, or other factors. Using published comparative ChIP-seq datasets combined with DNase-seq, we show that architectural constraints are a dominant factor in determining functional conservation across species. Enhancers and bound sites with evidence of binding by more factors are the most likely to be conserved between species, and conservation can be accurately predicted based on a handful of architecture-related features. To verify that our results are generalizable, we perform out-of-sample validation and show that models trained with one factor's data can predict conservation status with data from another transcription factor. Sites with more architectural constraint also show stronger signatures of sequence conservation in multiple alignments with other species. Our results indicate that the local arrangement of binding sites is a substantial constraint on the evolution of cis-regulatory sequence, and can inform our knowledge of which regions are most likely to retain functionality across species.

Introduction

Cis-regulatory sequences vary in terms of the extent of their functional conservation, or their ability to perform similar biochemical functions in different species (for clarity, we term this “retention”). Some elements appear retained over millions of years (Visel et al. 2008; Cuddapah et al. 2012), while others fail to drive expression or retain transcription factor occupancy between even closely related species (Moses et al. 2006; Ulirsch et al. 2014).

Predicting conserved occupancy or activity across species is of interest for two reasons. The first is academic in nature: if certain characteristics can reliably predict retention across closely related species, then it would imply that those characteristics are important to understanding the evolution of function within cis-regulatory sequences. The second reason is practical. Investigators working in a comparative genomics framework may lack the resources or ability to gather data in certain non-model organisms, but may wish to predict whether a certain segment of DNA within that organism retains its function of binding a transcription factor or driving expression. If reliable predictions could be made concerning conserved occupancy/activity, these would be of great practical utility.

Unfortunately, making such predictions with regards to cis-regulatory elements is difficult. Sequence conservation within a cis-regulatory element is only modestly useful (Schmidt et al. 2012) in determining the extent to which an element remains functional in different species. Individual enhancers and binding sites may remain functional despite considerable sequence turnover, and a lack of sequence turnover does not necessarily indicate functionality (Ahituv et al. 2007).

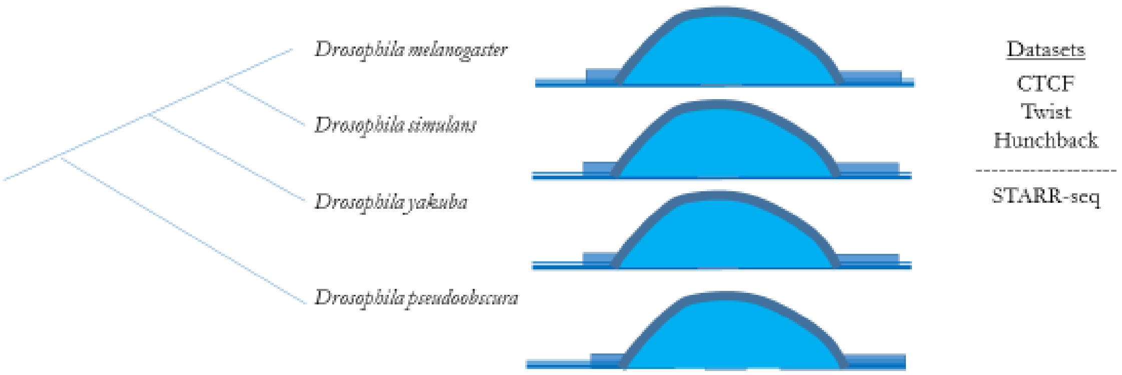
There are at least two general classes of explanation for the complex relationship between function and sequence conservation within cis regulatory elements. One reason is that different elements may be under different selective regimes. Some elements might be evolving under strong purifying selection (Siepel and Arbiza 2014), driving retention, while others are replaced frequently while evolving under neutrality.

The second explanation relates to the architecture of individual elements, or the structure of the constituent binding sites. Previous work has shown that individual sites and enhancers can be replaced, even while undergoing purifying selection, by going through an intermediate stage in which either of two functionally redundant analogs may be lost (Bullaughey 2011). This process, which we term compensation, has been shown to be plausible within realistic population genetic models and has been observed in vivo (Bullaughey 2013; Ludwig et al. 2000). Compensation, whether of sites or whole enhancers, can yield the false impression that sequences are not under selection, due to their replacement over time.

The process of compensation depends crucially on the probability of a new region of equivalent functionality arising. In turn, this probability is a product of the number of nucleotides constituting a bound site, as well as the extent of the region in which they can arise. Longer sites, and those under spatial constraint, are less likely to arise by chance (Bullaughey 2011). We refer to the arrangement of functional sequences within a cis-regulatory element as its architecture.

Accordingly, we hypothesized that cis regulatory elements with certain architectures—containing a greater number of spatially constrained sequences—ought to be replaced less frequently. To examine this question, we gathered published ChIP-seq datasets in Drosophila (Supplementary Table 1). In each experiment, the binding sites for a single transcription factor were located genome-wide in multiple, closely related species, allowing us to know whether a site was retained, or exhibited conserved occupancy, between species (Figure 1). We designated a site as retained if and only if it was present in all of the examined species in a given experiment (see Methods; the species examined are detailed in Supplementary Table 2).

Figure 3.1. Collected cross-species ChIP- and STARR-seq datasets. Diagram depicting the experimental design of experiments which we collected. In each case, ChIP- or STARR-seq was performed across multiple species, allowing us to interrogate the relationship between conservation of biochemical function and various architectural characteristics.



We also made use of one STARR-seq experiment performed in multiple species of Drosophila. STARR-seq is a technique to annotate regions of the genome which drive their transcription and thus may be enhancers. This dataset therefore encompasses a different kind of cis-regulatory activity from that discovered in ChIP-seq experiments.

To interrogate the architecture of each binding site or putative enhancer, we performed a DNase-seq footprint analysis (Madrigal and Krajewski 2012; Piper et al. 2013) on a series of publically available transcription factor binding and enhancer datasets (He et al. 2011; Ni et al.

2012; Paris et al. 2013; Arnold et al. 2014; Kharchenko et al. 2011; Ling et al. 2010; Thomas et al. 2011; Schmidt et al. 2010; Kutter et al. 2011) (Supplementary Table 3). Footprint analysis relies on the distribution of DNase I-seq reads beneath a ChIP-seq peak. Specifically, reads tend to stack around, but not within, the nucleotides which are directly bound by proteins. This region, over which DNA is protected from digestion, constitutes the footprint. Whereas ChIP-seq results form broad regions of TF binding enrichment which are typically several hundred nucleotides in length, footprint analysis narrows the actual bound sites to single nucleotide resolution. A comprehensive identification of these footprints allows us to examine the size and arrangement of bound loci beneath a ChIP-seq peak, properties which we term collectively the "architecture."

Methods

Enhancer/Binding site calls

To examine conservation of cis-regulatory elements, we used data from a number of published studies, summarized in Table S1. In each case, we made use of the author's own peak/enhancer calls, rather than attempting to recreate our own. To perform genomic manipulations on these calls, we used the software Bedtools, v.2.16.2 (Quinlan and Hall 2010).

DNase Data and Processing

In order to perform digital footprinting analysis at regulatory elements, we downloaded double-hit DNase I-seq data from a number of published sources (Ling et al. 2010; Kharchenko et al. 2011; Thomas et al. 2011). The sources are summarized in Table S2.

In processing DNAse data, we aligned reads in each case using bwa version 0.7.5 (Li and Durbin 2009), with default settings (as follows: -n .04 -k 2 -M 3 -O 11 -E 4). We utilized the following genome versions: *Drosophila melanogaster*, UCSC genome version dm3; *Mus musculus*, UCSC version mm9.

DNase Footprinting

In order to examine the architecture of cis-regulatory elements, we utilized DNase-seq data. DNase-seq has been employed in the past to determine regions of open or active chromatin. However, DNase-seq also produces a single nucleotide-resolution signature of transcription factor occupancy (the so-called "footprint"). We made use of footprint analysis software called Wellington (Piper et al. 2013) in order to examine the architecture of cis-regulatory elements in a high-throughput, genome-wide fashion. We ran Wellington with the following arguments: footprint sizes 5-30, every two base pairs (-fp 5,30,2), reporting all sites with a p-value of .1 or lower (-pv -1).

For each experiment, we took the intersection between footprints called in two separate replicates. We found that footprints strongly overlapped between replicates (typically more than 90% of footprints overlapped; Supplementary Table 5). We also called footprints with more stringency (employing a p-value cutoff of .001), but found that the resulting conservation predictions were less accurate, suggesting that the default p-values were too conservative.

LiftOver

To map regions between species, we employed the software liftOver (Karolchik et al. 2014) with default settings (minMatch=.95). We used the precomputed chain files provided by UCSC.

For mammalian enhancer validations, we found that the default minMatch parameter resulted in too few sites being mapped between species. Correspondingly, we reduced the minMatch parameter to .5.

Conservation

We used UCSC's precomputed phastCons scores, aggregated across genomic intervals by converting to bigWig files which were then summarized using the bwtool software (Pohl and Beato 2014).

Classifier Training and Testing

To train and test our machine learning tools, we adopted a classification approach. Sites were divided into "conserved" and "diverged" groups. We designated a site as retained if and only if it was present in all of the species examined in each study (Supplementary Table 2). Sites retained in three or more species were typically retained across the phylogeny. For example, 222 CTCF sites were retained between *D. melanogaster*, *D. simulans,* and *D. pseudoobscura*. Of those 222, 209 (94%) were also retained within *D. yakuba*. As a result, varying the threshold for a site to be considered retained did not significantly alter our results. We also built classification models which attempted to predict the total phylogenetic distance over which a site was retained, and results were similar for these regression-based classification models.

After investigating the performance of a number of different machine learning models, we chose to use neural networks, as they exhibited the best, most robust performance. In particular, we utilized the implementation in the nnet package in R. After significant tuning, we chose to build models with 50 neurons and a decay rate of .1. Models were trained with 10-fold cross-validation, and the reported accuracy and AUC values are based on the averaged models from this procedure.

We also tested additional machine learning tools, but found that neural networks performed the best. We trained Support Vector Machines using the R statistical programming language and the package e1071 (http://cran.r-project.org/web/packages/e1071/index.html). We also utilized another machine learning tool, Random Forests (Breiman 2001) from the randomForest package.

To compute AUC, we used the R package pROC (Robin et al. 2011). Reported confidence intervals on AUC represent the results of 1000 bootstraps.

To determine which features significantly improved classification performance, we made use of two separate approaches. The first relied upon the 95% confidence intervals on the value of the AUC supplied by the pROC package. In the second, we randomly permuted the row labels of individual features, and tested whether the permuted variables increased classification accuracy at all. In all cases, both approaches showed similar results.

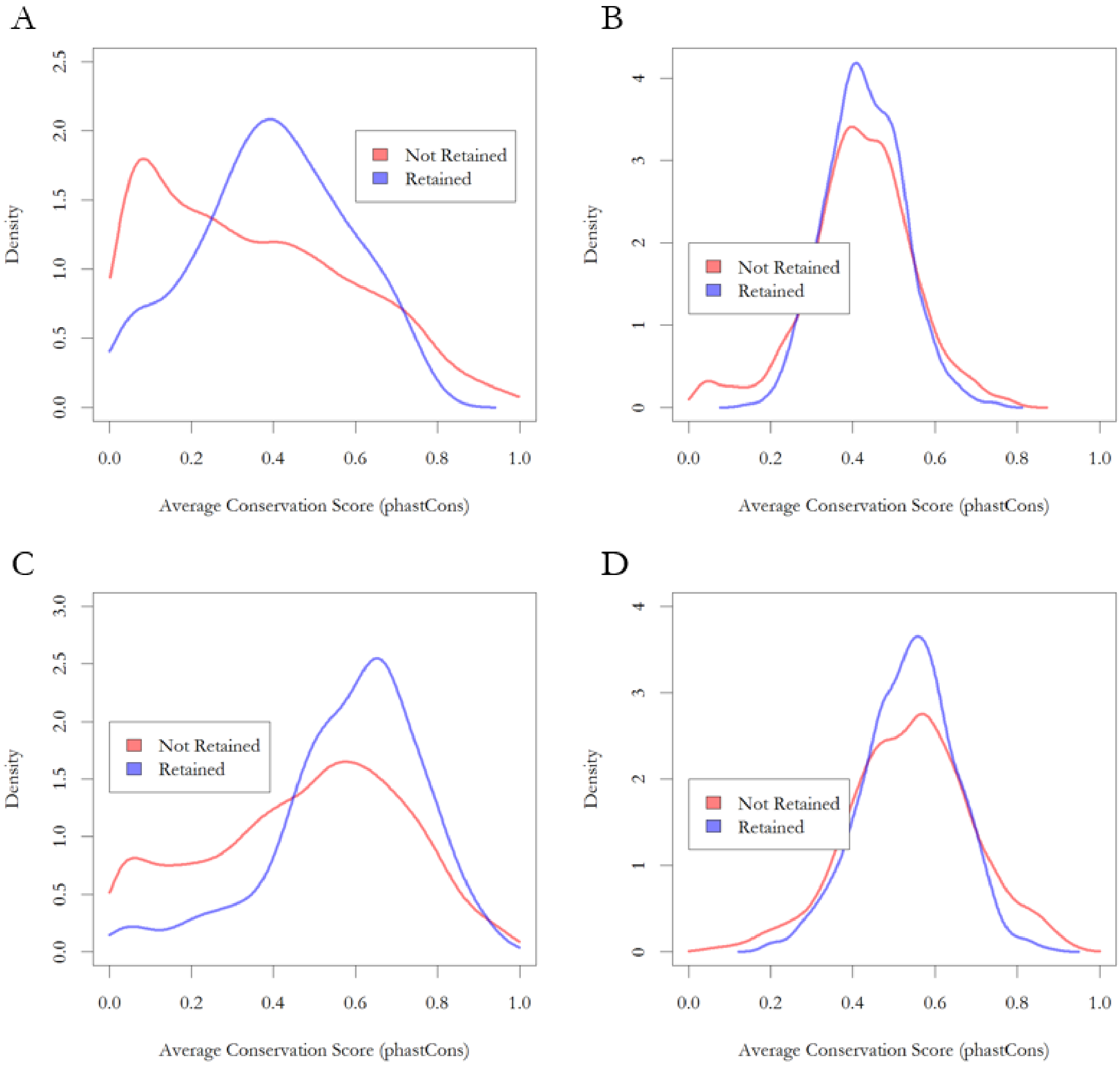
To compare the difference in means between two populations, we used permutation tests.

Results

Nucleotide Conservation

Before turning to architectural features, we examined the relationship between nucleotide conservation (as measured by phastCons scores (Karolchik et al. 2014)) and retention. We found that molecular conservation differed between sites of retained and diverged function. Sites with retained function possessed, on average, more conserved sequences for CTCF (Fig. 2) and STARR-seq datasets (CTCF: p<.005; Hunchback: p<.005; STARR-seq: p<.005), while there was no significant difference for Twist sites (p=.5).

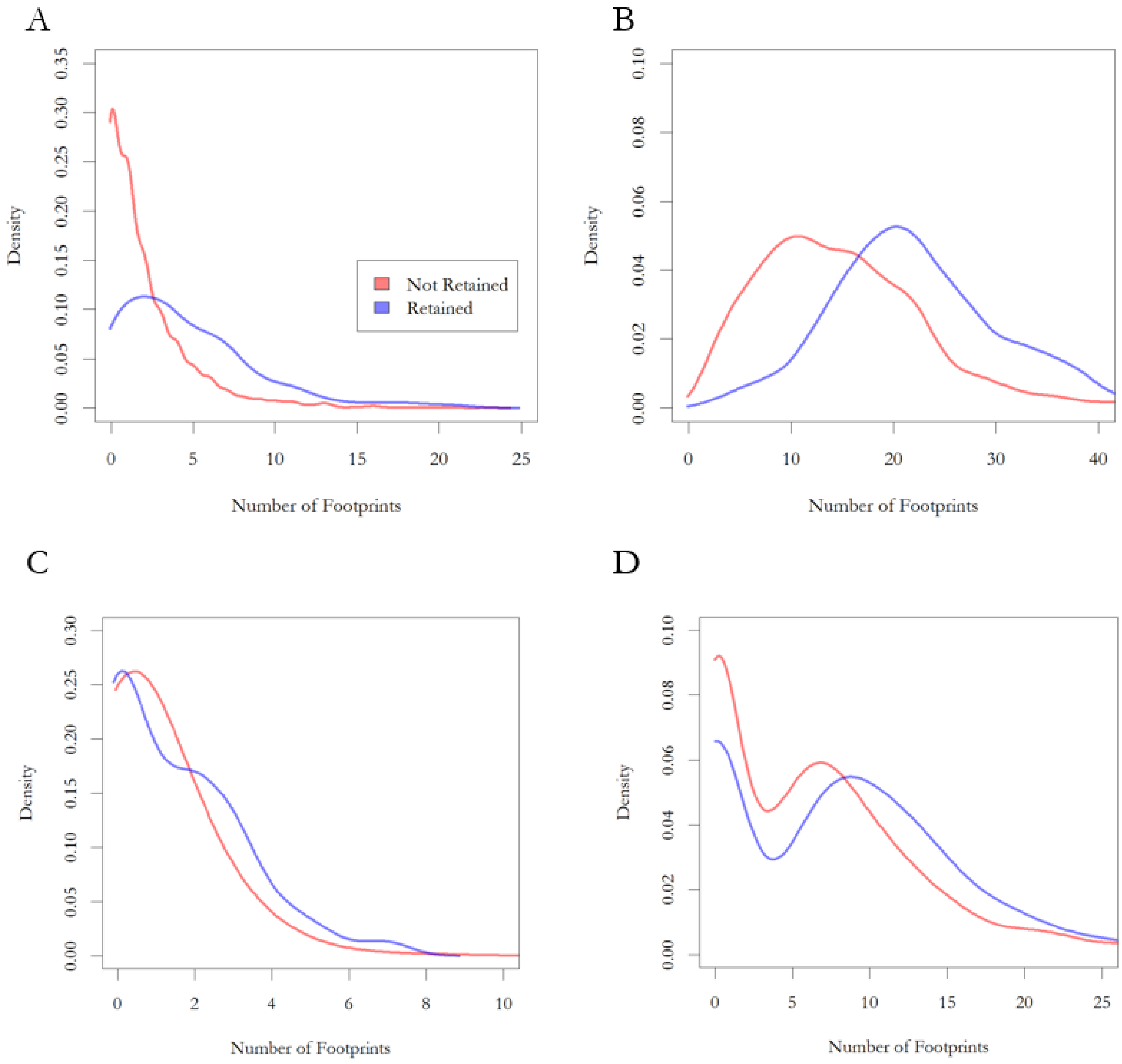
Figure 3.2. Summary of differences between functionally-conserved and diverged sites with respect to sequence conservation and architectural differences. A-D) Density plots showing the difference between conserved and diverged CTCF sites with respect to phastCons conservation scores. Conserved sites tend to show higher phastCons scores, showing that there is a weak but significant relationship between sequence and functional conservation. The red line represents the density of sites which are not retained between species, while the blue line represents the density of sites which are retained across the phylogeny.



Number of Footprints

We next examined predictions concerning architectural features that may be linked to functional conservation. We hypothesized that binding sites and enhancers possessing more footprints would feature greater levels of cooperative binding, thus constraining the extent to which replacement sites could arise. In agreement with our hypothesis, we found a strong and significant difference in the number of footprints per site and retention, for all factors (Fig. 3; p<.001).

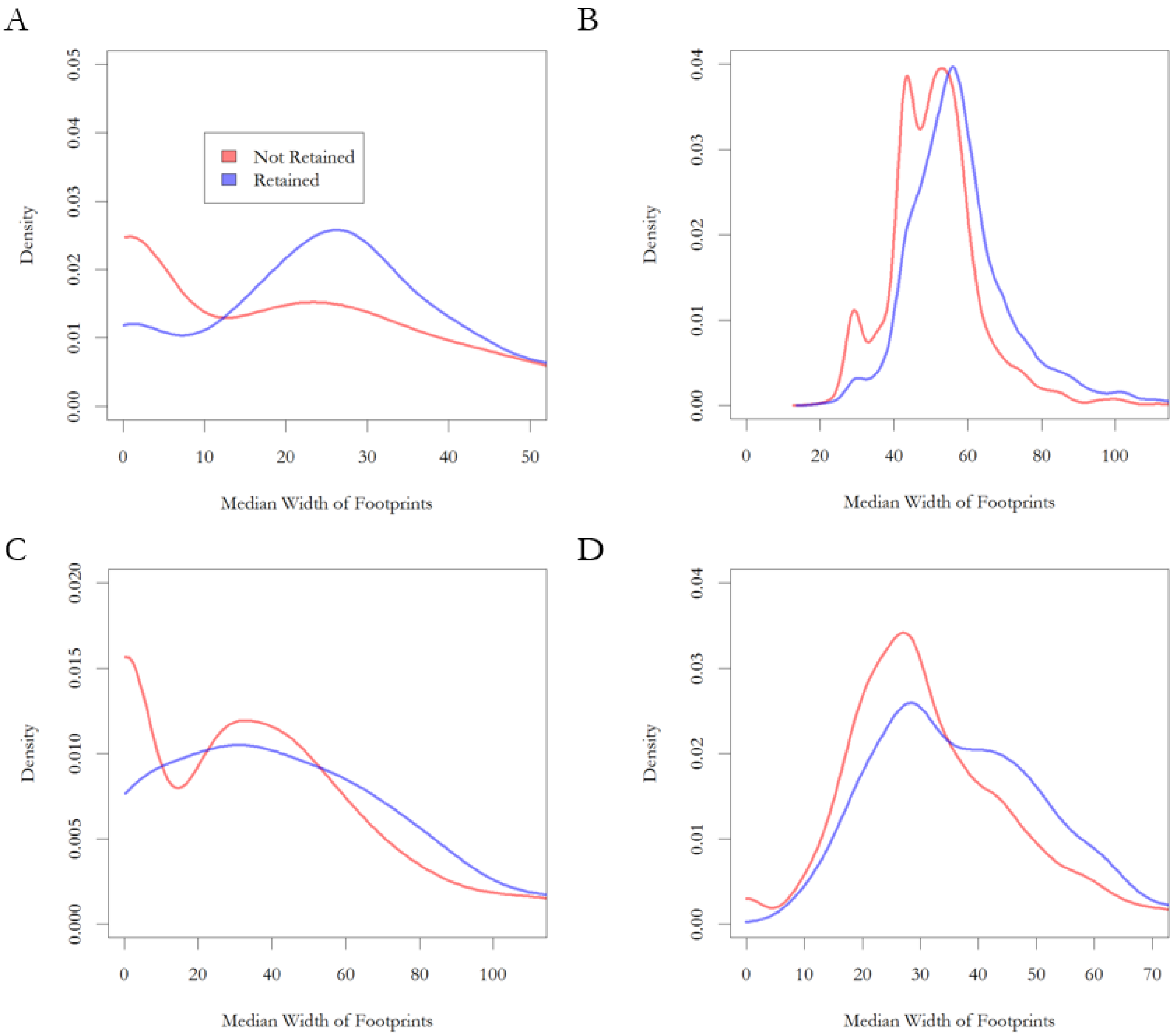
Figure 3.3. Retained sites tend to possess a greater number of footprints. A-D) As in Figure 2, but depicting the number of footprints per peak or enhancer. A) phastCons scores for CTCF. B) phastCons scores for STARR-seq enhancers. C) phastCons scores for Hunchback. D) phastCons scores for Twist.



Width of Footprints

Another prediction is that longer footprints would be less likely to arise by chance, thus limiting the probability of compensation (Bullaughey 2011). Such longer footprints may be associated with more cooperative binding events, as well. Accordingly, we found that conserved loci tended to be associated with longer footprints (Fig. 4; p<.001).

Figure 3.4. Retained sites tend to possess wider footprints. A-D) As in Figure 2, but depicting the median width of footprints per peak or enhancer.



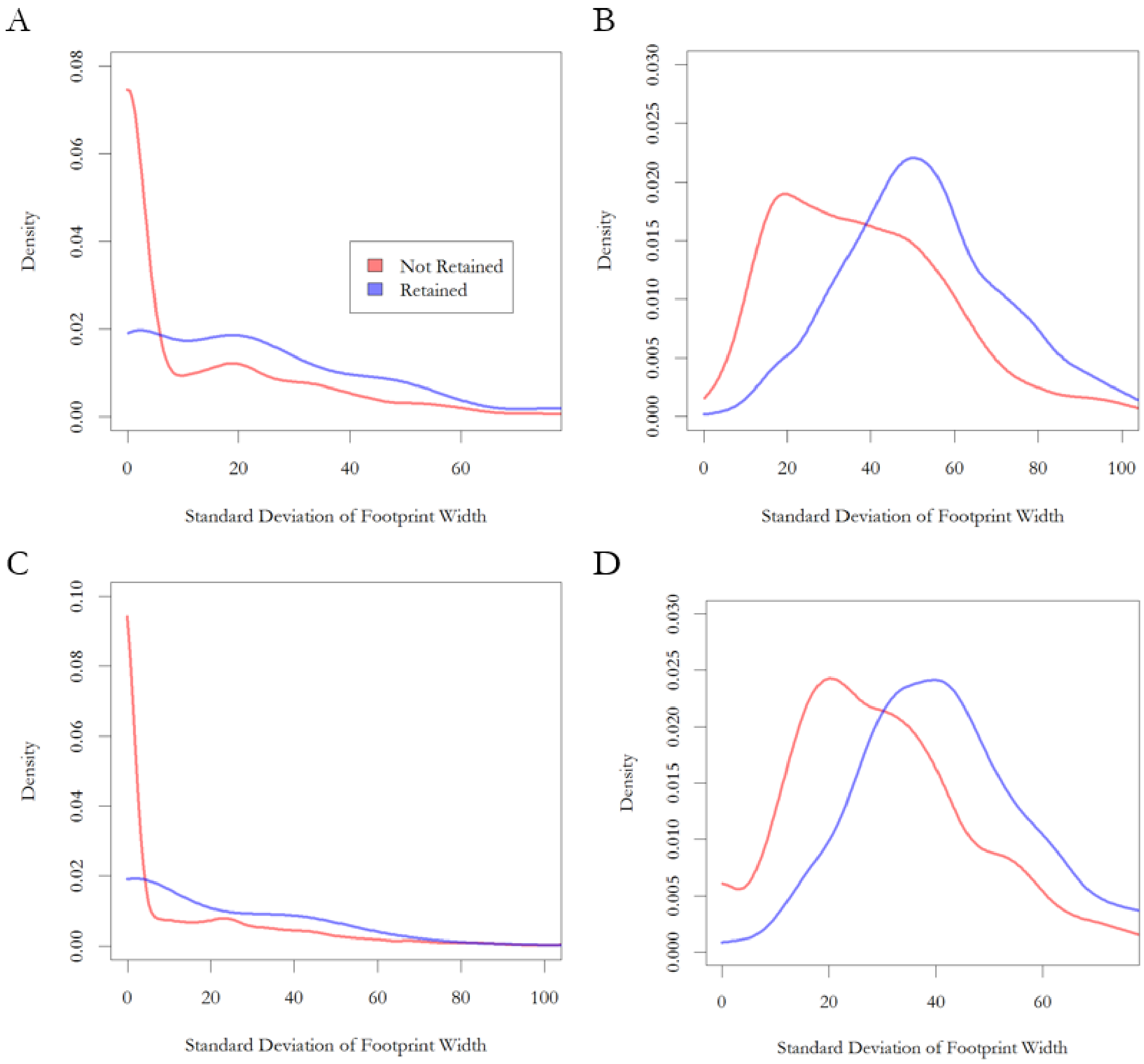
Regulatory Complexity

While a greater number of footprints should indicate a greater number of distinct transcription factors on average, in some cases all of the footprints may be for the same factor (a homotypic cluster of binding sites). As a result, we hypothesized that another useful set of features might relate to the number of distinct transcription factors binding at each element. We term this the regulatory complexity.

To measure this property, we used multiple, distinct approaches. We first examined the sequences underlying each footprint within a cis-regulatory element, reasoning that the variability between such sequences might be indicative of greater regulatory complexity. To quantify their variability, we used the average pairwise Levenshtein distance, and found that this metric differed strongly between conserved and nonconserved peaks/enhancers (p<.001).

Another metric of potential interest is the variation in size of footprints within a given peak or enhancer. While footprints contain no inherent information as to the identity of the bound transcription factor, footprints of different size are unlikely to be bound by the same transcription factor. As a result, we used the standard deviation in footprint sizes within a peak/enhancer, and found that, as expected, retained sites showed a much greater standard deviation (Fig. 5; p<.001).

Figure 3.5. Retained sites tend to possess a greater standard deviation in the size of footprints. A-D) As in Figure 2, but depicting the standard deviation in width per peak or enhancer.

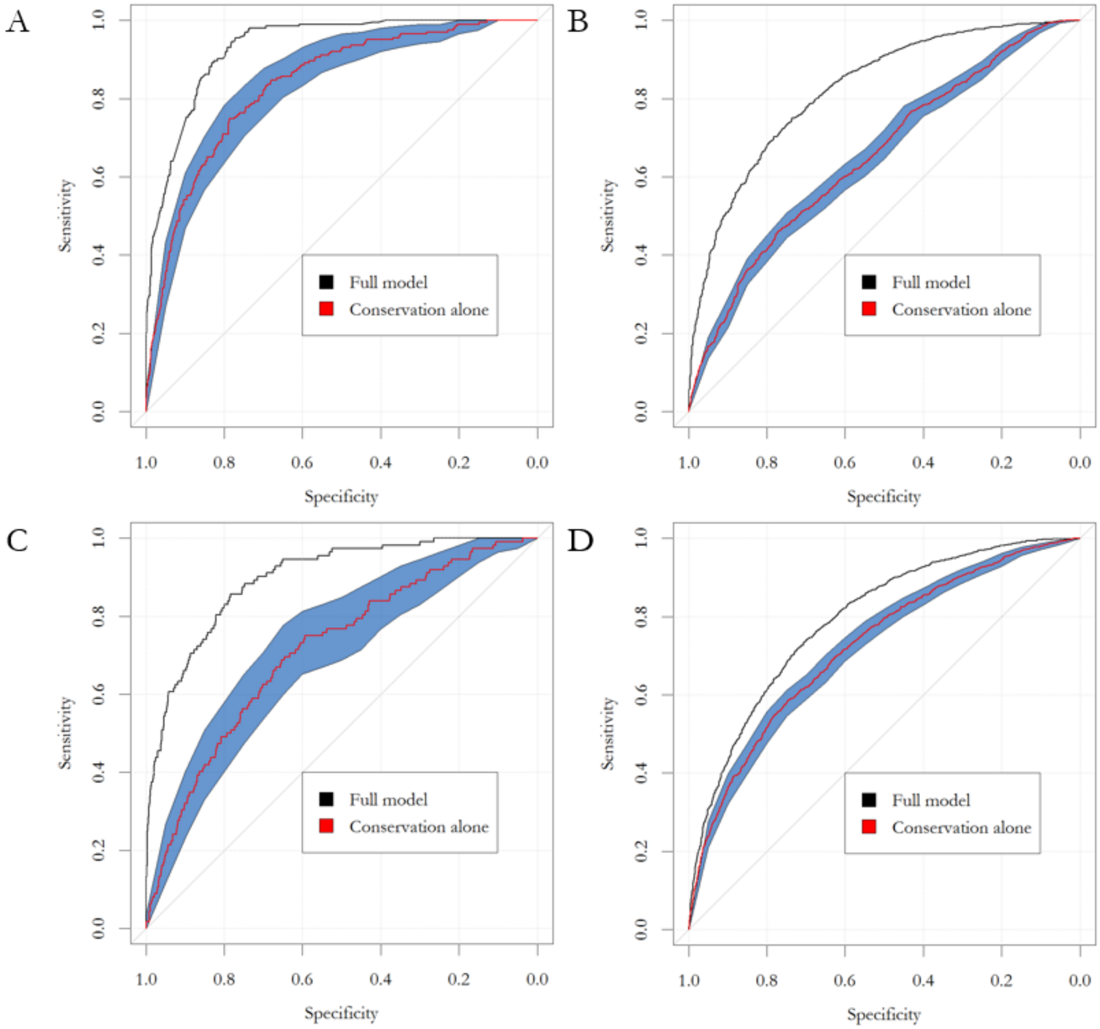


Machine Learning

Given that architectural features prove strongly associated with retention of individual cis-regulatory elements, we pursued further analyses to determine the extent to which architectural features could predict conservation. As described above, accurate models which can predict conserved occupancy across species could be practically useful in comparative genomics contexts. We adopted a machine learning approach using artificial neural networks (see Methods), building models to predict conserved occupancy/activity based on a handful of features related to the size and number of footprints within each cis-regulatory element (Supplementary Table 4). In each case, we performed 10-fold cross-validation to guard against overfitting.

First, we developed a basic model to predict conserved occupancy, the only features of which were 1) the level of nucleotide conservation and 2) the strength of binding or expression, as determined by the number of reads within the peak or enhancer (see Methods for details; Fig. 5). This base model scored AUC values significantly exceeding the performance of a random classifier for all ChIP and STARR datasets (Fig. 6A-D; p<.05), indicating that simple features like nucleotide conservation and binding/expression strength provide some guide as to whether an enhancer or binding site will retain the capability to bind a transcription factor or drive expression.

Figure 3.6. Architectural characteristics improve predictions of cross-species cis-regulatory element retention. Receiver-operator curves illustrating conservation classifier accuracy for each of three representative datasets. For each plot, two curves are shown: in red, a classifier was trained based on nucleotide conservation and binding site/enhancer strength (see Supplemental Information); in black, a model was trained using those characteristics as well as footprinting-based architectural features. In each case, the model incorporating architectural features obtains substantially better classification accuracies. A) ROC curves based on data from CTCF ChIP-seq experiments (Ni et al. 2012). B) ROC curves based on data from Twist experiments (He et al. 2011). C) ROC curves based on data from STARR-seq experiments (Arnold et al. 2014). D) Barplot illustrating prediction accuracy for each full model. Bars represent the maximum attained accuracy (defined as the number of correct predictions over the total number of predictions), using the full model (consisting of architectural and sequence features together). Black dots show the minimum accuracy, which would be obtained if one were to guess the most common conservation status (conserved/diverged) for each site.

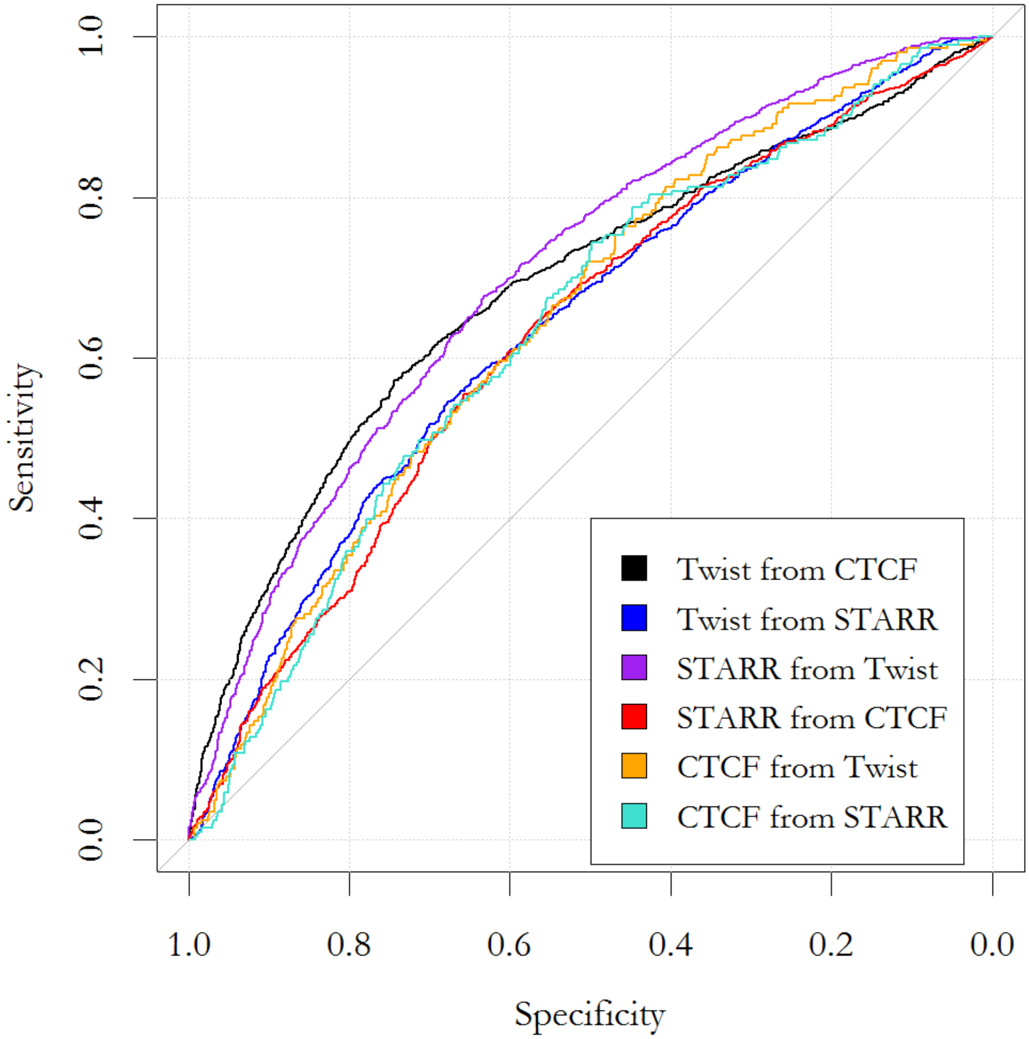


Next, we integrated information about architecture. After extensive testing, we determined that the classifier functioned best when incorporating architectural features measuring the number of footprints, median width of footprints, and standard deviation in the width of footprints. Area under the curve (AUC) values improved dramatically with the addition of footprinting-derived architectural data (for all datasets, $p<.05$), suggesting that architecture is a significant factor in determining whether a site will be replaced. Notably, no simple function of DNase coverage was able to predict site retention as well as the architectural parameters we identified ($p<.05$; see Supplemental Information). Using the full model, which combined sequence conservation, binding site strength, and architectural features, we were able to predict occupancy across species with better than 90% accuracy for a majority of our datasets.

We also found that sequence conservation was associated with particular characteristics of footprints on a site by site level. While conserved occupancy was a poor predictor of sequence conservation alone (CTCF: $r=.076$; $p=.0004$), combined with the number of footprints and the median size of the footprints within a binding site, sequence conservation was more accurately regressed ($r=.259$; $p<2.2\text{x}10^{-16}$). These results suggest that the relationship between sequence and retention is mediated in part by the architecture of individual cis-regulatory elements.

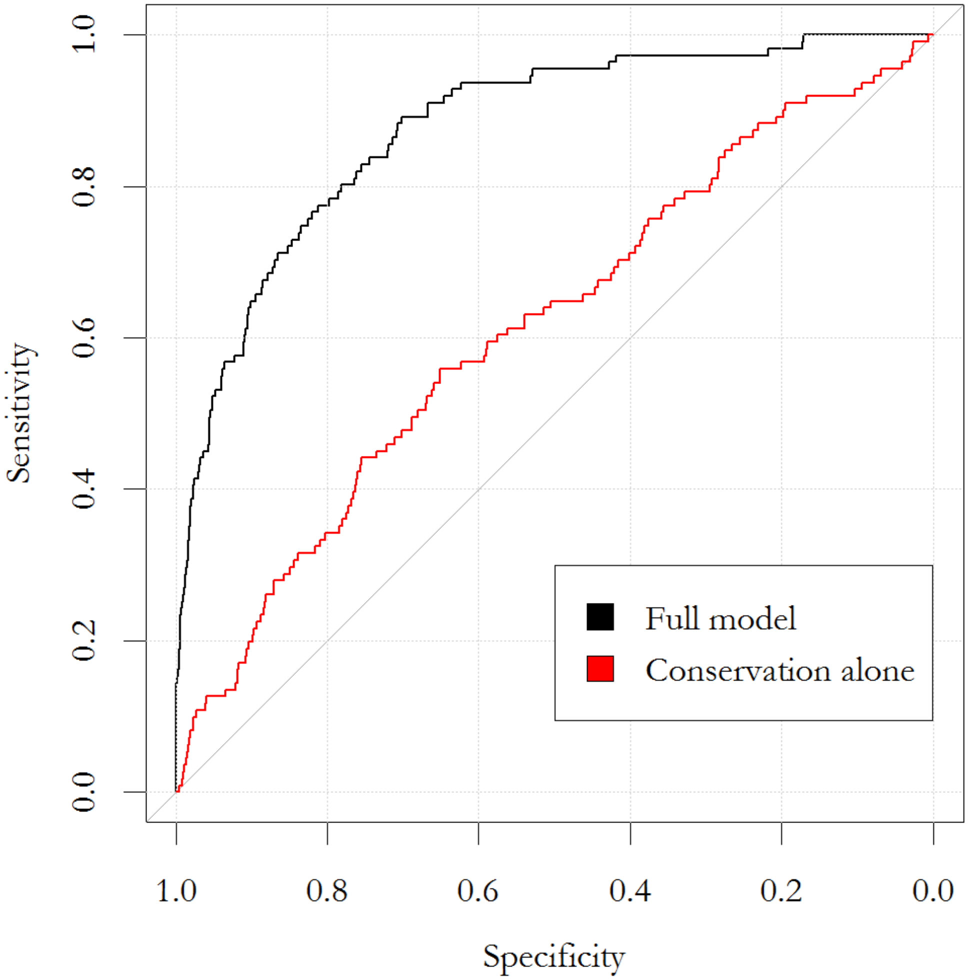
To verify that our predictions were generalizable, we trained models on the data from one ChIP-seq experiment, and tested them on the data from another ChIP or STARR experiment. In each case, we found that models retained some of their performance across assays, remaining better than a random classifier (Fig. 7; $p<.05$). However, models trained on one dataset typically performed less well on other datasets than equivalent models trained and tested on the same factor (Supplementary Figure 4), suggesting that there are factor-specific architectural components which predict retention.

Figure 3.7. Receiver-operator curves showing that classifier performance generalizes readily between experiments. A number of ROC curves are depicted (described in the legend): in each, the conservation classifier was trained on one experiment's data, and generated predictions for another experiment's data. In all cases, the classifiers obtain better than random performance, indicating that the features we identified generalize across different experiments and transcription factors.



In order to ensure the generality of our results across species and techniques, we performed a further out-of-sample validation analysis using enhancers identified in mammalian species (Villar et al. 2015). Villar et al. (2015) identified enhancers in a completely orthogonal manner to STARR-seq, relying upon ChIP-seq performed against the enhancer-associated histone modification H3K27ac. Despite the methodological, technical, and phylogenetic differences, we are able to use the same features to predict conservation with high accuracy (Fig. 8). In fact, even models trained on Drosophila ChIP-seq datasets can predict functional conservation of mammalian enhancers (AUC: .6; $p<.05$).

Figure 3.8. A receiver-operator curve showing validation with mammalian enhancer data. We used data from (Villar et al. 2015) and (Ling et al. 2010) to validate our classification strategy. A classifier trained on architectural features and conservation was able to successfully and accurately predict functional conservation of enhancer regions between human and mice, indicating the generality of our approach.



Discussion

In summary, we have shown that functional conservation of binding sites and enhancers can be predicted with great accuracy from architectural features, regardless of sequence conservation. Using a series of validation methods, we have shown that the architectural features we identified are generally applicable to both binding sites and enhancers, even across distantly-related clades. Our results show that architectural constraints are a dominant factor in the evolution of cis-regulatory function, and suggest a method to identify functional regions even in species without detailed functional genomics data.

Acknowledgments

RKA was supported by an NSF Graduate Research Fellowship as well as an NIH Genetics and Regulation Training Grant (T32GM007197). IR and RK were supported by the NSF (IOS-0843504) and the NIH (P50 GM081892). We are grateful to Kevin Bullaughey for many helpful discussions in the early stages of this project. Finally, we thank members of the Ruvinsky lab for useful comments and advice at all stages in the development of this manuscript.

Chapter 4: Evolution of H3K27me3-Marked Chromatin Is Linked to Gene Expression Evolution and to Patterns of Gene Duplication and Diversification

Abstract

Histone modifications are critical for the regulation of gene expression, cell type specification and differentiation. However, evolutionary patterns of key modifications that regulate gene expression in differentiating organisms have not been examined. Here we mapped the genomic locations of the repressive mark histone 3 lysine 27 trimethylation (H3K27me3) in four species of *Drosophila,* and compared these patterns to those in *C. elegans*. We found that patterns of H3K27me3 are highly conserved across species, but conservation is substantially weaker amongst duplicated genes. We further discovered that retropositions are associated with greater evolutionary changes in H3K27me3 and gene expression than tandem duplications, indicating that local chromatin constraints influence duplicated gene evolution. These changes are also associated with concomitant evolution of gene expression. Our findings reveal the strong conservation of genomic architecture governed by an epigenetic mark across distantly related species and the importance of gene duplication in generating novel H3K27me3 profiles.

Introduction

While transcriptional regulation has long been recognized as a significant target of evolutionary change (King and Wilson 1975), the specific mechanisms behind regulatory divergence have been difficult to dissect (Carroll 2008). In most cases, research has focused on recognizable cis-elements where transcription factors bind in a sequence-specific manner (Borneman et al. 2007; Bradley et al. 2009; Ni et al. 2012 Dowell 2010). Changes to either a transcription factor's DNA-binding properties or transcription factor binding sites (TFBS) enable the evolution of differential regulation, and numerous examples of this phenomenon have been observed (Shultzaberger et al. 2012; Ludwig et al. 2005).

Whereas changes in cis regulatory elements have been implicated in phenotypic and specifically morphological changes (Carroll 2008; Stern and Orgogozo 2009), other components of transcriptional regulation such as the chromatin environment have been scarcely explored (Lenhard et al. 2012). Histone modifications, which are chemical alterations of the histone spools upon which DNA is threaded, constitute one of the best-described elements of chromatin state (Zhou et al. 2011). These modifications can act directly or indirectly (through recruited enzymes) to alter DNA accessibility, thereby controlling other DNA-protein interactions (Campos and Reinberg 2009). Unlike TFBSs, however, histone modifications are not necessarily easily localizable to particular sequence elements, making their evolution difficult to study (Delest et al. 2012).

Histone 3 lysine 27 trimethylation (H3K27me3) is one of the best known histone modifications, in terms of both its biogenesis and effects. H3K27me3 is associated with complex cis-regulatory elements called Polycomb Response Elements (PREs; Delest et al. 2012). Unlike TFBSs, PREs are compound regulatory elements composed of multiple, sometimes partially redundant, sequence elements (Delest et al. 2012). Although the exact biochemical binding site composition and structural constraints upon PREs are not completely resolved, a group of coordinately binding proteins called Polycomb Group (PcG) factors recruit Polycomb Repressive Complexes, which are necessary to deposit and maintain H3K27me3 and have been associated genome-wide with H3K27me3 domains (Lanzuolo and Orlando 2012).

The effect of H3K27me3 is unambiguously and strongly repressive (Kharchenko et al. 2011; Filion et al. 2010; Li et al. 2014). H3K27me3-associated loci have been proposed to congregate in silenced foci called Polycomb factories, which inhibit transcription by preventing access to RNA Polymerase II and other trans-factors (Bantignies et al. 2011; Sexton et al. 2012). This repressive state is both temporally and spatially dynamic (Akkers et al. 2009; Filion et al. 2010; Kharchenko et al. 2011; Negre et al. 2011).

Accordingly, the PcG proteins and H3K27me3 have been shown to be necessary for differentiation and maintenance of cell type identity in organisms across eukaryotes (Bernstein et al. 2006; Feng and Jacobsen 2011). Dysregulation of H3K27me3 has also been implicated in the genesis and progression of cancer (Chi et al. 2010; Ellinger et al. 2012). These results indicate that H3K27me3 plays a role in the creation and maintenance of cell-type specific programs of transcriptional control for a wide variety of species and cell fates.

Given its close association with the process of differentiation, the extent to which H3K27me3 domains are conserved across species is of great interest (Cain et al. 2011; Wilson et al. 2008; Shubin et al. 2009). We sought to examine the rate of evolutionary change in H3K27me3 between four species of the well-characterized *Drosophila* clade, the mechanisms by which such change might occur, and the possible consequences of H3K27me3 evolution for nearby gene expression.

In order to investigate the evolution of H3K27me3 patterns across the genome, we performed chromatin immunoprecipitation followed by sequencing (ChIP-seq) for H3K27me3 in four species, *D.melanogaster*, *D. simulans*, *D. yakuba* and *D. pseudoobscura*, with divergence times ranging from less than 5 million years (Myr) to more than 35 Myr (Tamura et al. 2004; Obbard et al. 2012). To determine whether and how H3K27me3 changes might alter gene expression, we also performed RNA-sequencing (RNA-seq) in all species. For all analyses we used white prepupae for ease of developmental synchronization between species and because there has been extensive previous work characterizing genome wide regulatory evolution at this developmental stage (Rifkin et al. 2003; Gu et al. 2004; Ni et al. 2012). We find that for single-copy orthologous genes, H3K27me3 signal is strongly conserved in even distantly related species. However, duplicated gene orthologs exhibit much greater divergence in H3K27me3. Moreover, different kinds of duplicates appear to have very different rates of H3K27me3 and expression divergence, indicating functional distinctions in the epigenetic consequences of different gene duplication mechanisms.

Methods

ChIP-seq

Strains were maintained at room temperature until collection at the white pre-pupa stage. ChIP-seq was carried out in duplicate using the standard modENCODE protocols. We mapped 36 base pair reads to the appropriate reference genome using Bowtie (v.0.12.7; Langmead et al. 2009). *Caenorhabditis elegans* H3K27me3 ChIP-seq data was generated as described in Landt et al. (2012). Data from Negre et al. 2011 is available under accession numbers GSE27111 and GSE23537.

RNA-seq

Single-end reads were trimmed and mapped using Tophat, and FPKM values were called with Cufflinks. *Caenorhabditis elegans* RNA-seq data was processed as described in Li et al. 2014.

Quantitative analysis

We used BEDTools to count the number of reads occurring in either 1) the exons of each gene as determined by the reference annotation or 2) regular sliding windows 1kb in width. Genic H3K27me3 was normalized by the number of exons (see Supplementary Methods). We compared orthologous genes using the list of modENCODE orthologs, and we compared orthologous windows by using liftOver to map windows between species.

Pseudogene location analysis

In order to examine the H3K27me3 signal occurring in the locations of recently duplicated genes/pseudogenes, we extracted the two flanking regions (both 500bp) around each recently duplicated gene or pseudogene, which were then lifted into the *D. simulans* genome using liftOver. To limit our analysis to very recent, confident duplication events, we only analyzed the comparison of *D. melanogaster-D. simulans*, and considered only cases in which both flanking regions mapped, with a distance of less than 5kb between them. We examined H3K27me3 level in the region between flanks as representative of the ancestral state (see Supplemental Methods).

Statistical analysis

To call individual intervals as significantly diverged in different species, we used DESeq (Anders and Huber 2010) on the control-subtracted read counts within bins. To estimate a False Discovery Rate (FDR), we employed the q-value package in R (Storey and Tibshirani 2003). To estimate the significance of differences between two populations' sample means, we used permutation tests as described in Sokal and Rohlf (1995) (pg. 808; called therein "sampled randomization" tests). All statistical analysis was performed in R.

Data Access

Drosophila RNA-seq data and ChIP-seq data from this study have been submitted to the NCBI Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE49945. Caenorhabditis ChIP-seq data have been submitted to GEO under accession numbers GSE49724 and GSE49738.

Results

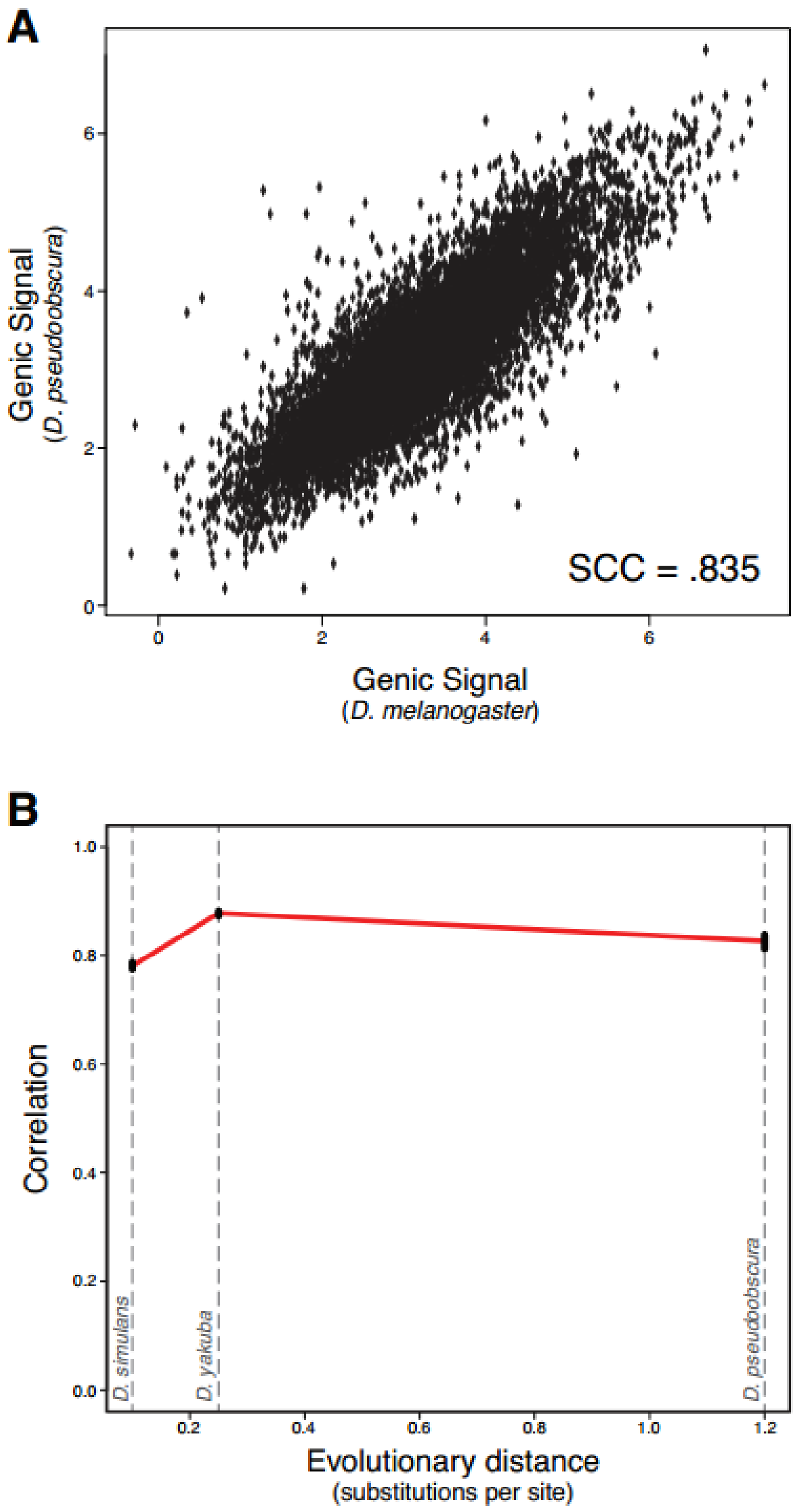
Evolutionary stasis of H3K27me3 levels across single-copy genes in **Drosophila**

We examined H3K27me3 signal at the white pre-pupal stage of development, a tightly-defined, 20 minute window at the beginning of pupariation. Two biological replicates were collected for each experiment and one control. H3K27me3 presents unique challenges for ChIP-seq analysis. Unlike transcription factors and certain other histone modifications, H3K27me3 forms broad patterns of enrichment that are of indeterminate length, and therefore simple peak-based analysis is not a suitable analysis method (Xiao et al. 2012; genome-wide analysis has shown these to be up to ~100kb in *Drosophila* [Papp and Muller 2006; Kharchenko et al. 2011; Negre et al. 2011]). Because of the many distinct cell types of a *Drosophila* puparium, it is also not appropriate to treat H3K27me3 as a binary modification of chromatin (Zang et al. 2009). Therefore we chose to analyze the data quantitatively, using the number of sequence reads falling within pre-specified intervals (e.g. the exons of a gene) as an indication of the overall level of H3K27me3 signal within a region (Arthur et al. 2014). This measure is both highly correlated between biological replicates (Arthur et al. 2014) and robustly anti-correlated with gene expression, suggesting that the metric is biologically meaningful. Using the modENCODE-curated list of orthologous genes in *Drosophila* (Boyle et al., in prep.), we directly compared each gene in the *melanogaster* genome to its ortholog in *D. simulans, D. yakuba*, and *D. pseudoobscura* (abbreviated Sim, Yak, and Pse). In total, we compared 12,017 orthologs from

D. melanogaster to *D. simulans*; 11,018 from *D. melanogaster* to *D. yakuba*; and 11,881 from *D. melanogaster* to *D. pseudoobscura*.

Figure 4.1 Strong conservation of H3K27me3 in Drosophila orthologs. (A) Example graph of orthologous gene conservation for the comparison of *melanogaster* to *pseudoobscura*. Each dot is a single-copy orthologous gene pair, and the position on the X-axis represents the log *melanogaster* genic signal (see Methods) while the y-axis represents log *pseudoobscura* genic signal; each is the mean of two experiments. The overall rank correlation coefficient between species is .835. **(**B) Overall trend of single-copy ortholog conservation within *Drosophila*. Each point is the Spearman rank correlation of one pairwise between-species comparison, plotted against the evolutionary distance from *Drosophila melanogaster* (in substitutions per neutral site). Black bars represent bootstrapped 95% confidence intervals. From left to right, the species are *D. simulans* (SCC = .781), *D. yakuba* (SCC = .878), and *D. pseudoobscura* (SCC = .835).

Figure 4.1 Continued.



Overall we found high conservation of H3K27me3 signal between *Drosophila* genomes, ranging from Spearman correlation coefficients (SCC) of 0.78 to 0.88 (Fig. 1A; a representative locus is depicted in Fig. S4 of Arthur et al. 2014). Given that any two pairs of biological replicate experiments show a SCC of ~0.95 (Arthur et al. 2014), the observed between-species correlations indicate relatively slow evolutionary change for H3K27me3 patterns across the genome. Simulations showed that a relatively small level of ortholog misidentification or misannotation error (10-20%; similar to published estimates [Creevey et al. 2011]) is sufficient to explain the observed decrease in correlation between species relative to the technical variation within species. Notably, we do not see evidence for a linearly decreasing trend of conservation over phylogenetic distance within the three examined pairwise comparisons (Fig.1B), in contrast to what has been observed for transcription factor binding sites in similarly low numbers of species (Dowell 2010; Ni et al. 2012; He et al. 2011).

Fold change in H3K27me3 of single-copy orthologs showed a weak correlation with accelerated sequence evolution. For example, change in H3K27me3 is positively associated with elevated d_N/d_S (SCC = 0.045, permutation test, $p < 10^{-5}$) in a comparison of *D. melanogaster* and *D. simulans*. These results are consistent with observations of mammalian stem cells (Xiao et al. 2011). Surprisingly, there is no strong correlation between change in H3K27me3 signal and change in gene expression among single-copy orthologous genes (SCC: Sim: .06; Yak: .043; Pse: -.027). As noted above, this observation is consistent with most differences between single-copy orthologs being a result of technical variation. Alternatively, differences in H3K27me3 signal might be compensated for by other mechanisms of transcriptional regulation, resulting in little ultimate difference in gene expression.

Conservation of H3K27me3 between* Drosophila *and* Caenorhabditis *orthologs

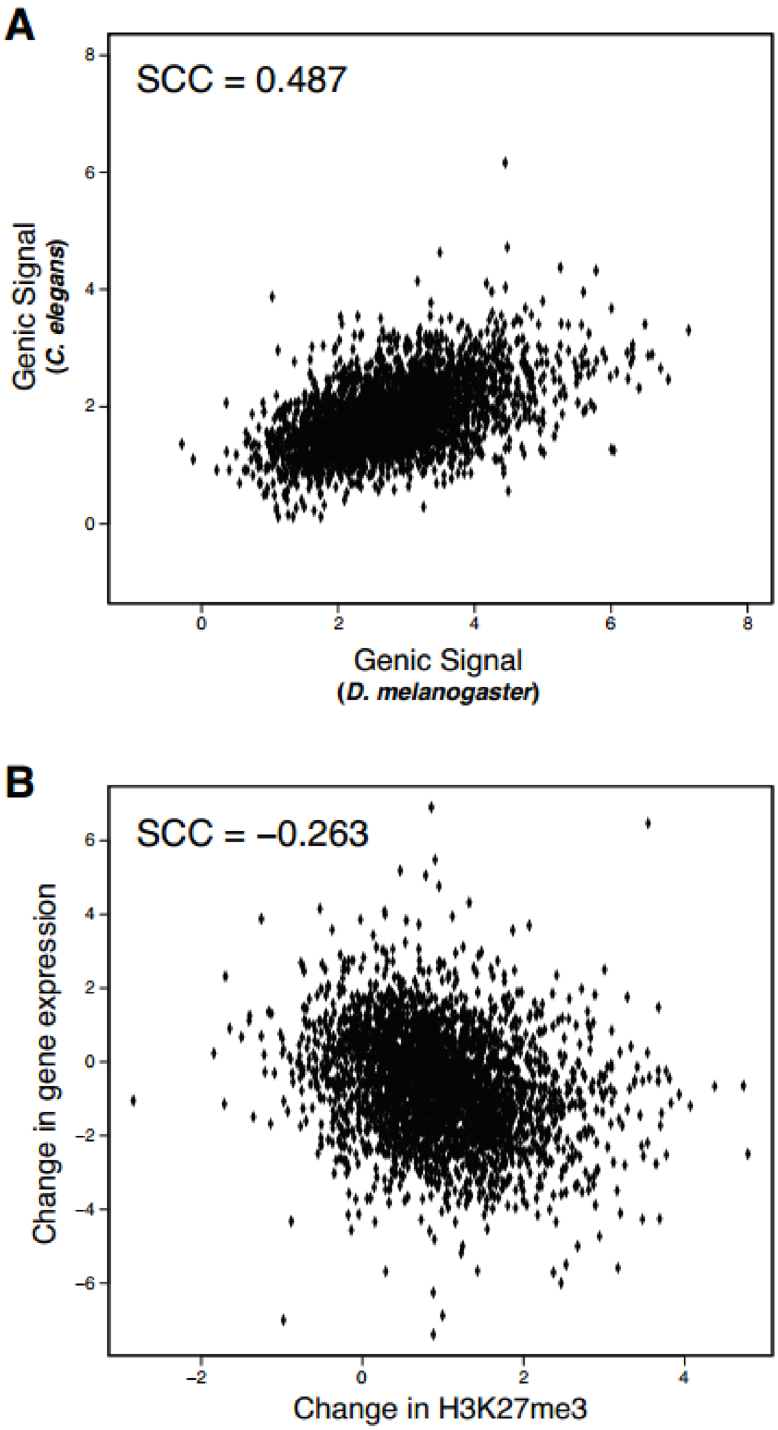
Given the extremely high conservation of the H3K27me3 epigenetic mark between *Drosophila* orthologs, we investigated whether H3K27me3 signal is conserved at greater phylogenetic distances. We chose to examine pairwise conservation between *D. melanogaster* and the nematode worm *Caenorhabditis elegans* (abbreviated Cel), whose most recent common ancestor dates to the Cambrian divergence (>500 Myr; Nei et al. 2001). We compared our *D. melanogaster* prepupal data with corresponding H3K27me3 data generated by the modENCODE Consortium at the L3 stage in worms (results were similar when comparing embryonic stages as well [Arthur et al. 2014]). There were a total of 3,157 single-copy orthologs between these two species.

We found substantial evidence of H3K27me3 conservation between these two highly diverged species (Fig. 2A; SCC = .487); H3K27me3 is nearly as conserved as gene expression for these genes (SCC = .497). Assuming the lowest observed rate of loss in correlation observed within *Drosophila* and a linear relationship between divergence time and loss of correlation, one would expect no significant correlation given the phylogenetic distance separating worm and fly. H3K27me3 has previously been compared between cell lines from human, mouse, and pig, three species separated by 50-100 million years (Xiao et al. 2012). To our knowledge this is the first study to show significant conservation of an epigenetic mark between species as diverged as *D. melanogaster* and *C. elegans*. There is a moderate, significant negative correlation between gain of H3K27me3 and loss of expression (Fig. 2B; SCC = -.263). This result stands in contrast to the relative lack of correlation between expression change and H3K27me3 change within *Drosophila*, and may be a consequence of greater fold change when comparing *D. melanogaster*-

C. elegans (median, absolute $\log_2$ fold change: 0.699) than *D. melanogaster* to, for example, *D. pseudoobscura* (median, absolute $\log_2$ fold change: 0.469).

Figure 4.2 Conservation of H3K27me3 extends to *Caenorhabditis elegans*. (A) Substantial conservation of H3K27me3 between *D. melanogaster* and *C. elegans*. As in Fig. 1A, each dot represents a gene, with x-axis position corresponding to log H3K27me3 signal in D. melanogaster and y-axis position corresponding to log H3K27me3 signal in *C. elegans*. The Spearman correlation coefficient is .487 between them. **(**B) Change in H3K27me3 levels is associated with changes in gene expression. Each dot is an individual single-copy orthologous gene, where the x-axis is the change in log H3K27me3 and the y-axis is change in log expression. The overall Spearman correlation coefficient is -.263.

Figure 4.2 Continued



Evolution of H3K27me3 is more rapid in gene duplicates

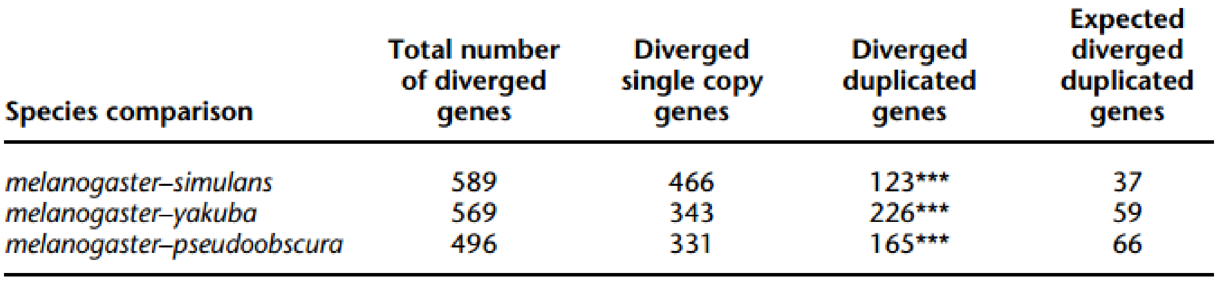
Next we sought to examine how H3K27me3 signal might vary in the aftermath of gene duplication events. Gene duplication has been found to be a significant driver of regulatory and gene expression divergence in other interspecific comparisons (Kaessmann et al. 2009; Gu et al. 2004). While we found above that H3K27me3 is highly conserved among single-copy orthologs, we hypothesized that gene duplication might allow the creation of novel H3K27me3 regulation among duplicated gene orthologs. We therefore compared H3K27me3 levels for each newly duplicated gene to its corresponding single copy ortholog in another genome.

Because duplicated genes tend to share substantial nucleotide similarity, it is possible that some ChIP-seq reads might have difficulty mapping accurately to paralogs, either by mapping to the wrong paralog or failing to map uniquely and thus being discarded. To investigate the extent of this issue, we performed read simulation studies (Huang et al. 2012) which showed that most paralogs are accurately mappable, and we excluded from further analysis those genes which are not mappable (see Supplemental Methods and Fig. S7 of Arthur et al. 2014). As a further precaution, we also performed read mapping in which up to two valid alignments are accepted for the pairwise comparison of *D. melanogaster* to *D. yakuba*; all results remained significant.

To call individual genes, either duplicated or single-copy orthologs, as significantly diverged in H3K27me3 signal (relative to the orthologous copy in the other genome), we quantified the degree of H3K27me3 divergence and compared to a permutation-based null (see Supplemental Methods). Significantly evolved genes (FDR of .05) showed greater absolute differences in expression between species (Permutation test: Sim: $p = .11$; Yak: $p < 10^{-5}$; Pse: $p < 10^{-5}$), and higher d_N/d_S (Permutation test: Sim: $p < 10^{-5}$).

H3K27me3 conservation differs systematically depending upon the occurrence of gene duplications. Duplicated gene orthologs have generally lower conservation of H3K27me3 signal, relative to the single-copy orthologs (Fig. 3A-C; this is true even after accounting for differences in number of duplicated vs. single-copy orthologs). In each species comparison, duplicated genes constitute 3-4x more of the H3K27me3-diverged set than expected based on their overall frequency in the compared genomes (Table 1; Fisher's Exact Test: Sim: $p = 3x10^{-12}$; Yak: $p < 2.2x10^{-16}$; Pse: $p = 6.9x10^{-11}$). Based on these results, we infer that H3K27me3 signal is more labile following gene duplication events.

Table 4.1. Duplicated genes are more likely to be significantly diverged than expected
Using a permutation-based approach, we called individual genes as significantly evolved with respect to H3K27me3 (both single-copy and duplicated gene orthologs; see Methods). We find that genes which have undergone duplication are much more likely to be significantly diverged than expected by chance. From left to right, columns are: the species being compared; the total number of genes called as significantly diverged; the number of single-copy orthologous genes called as diverged; the number of duplicated genes called as diverged; and the expected number of duplicated genes called as diverged. Asterisks represent statistically significant overrepresentation of duplicated genes relative to expectation (p < .0001). Duplicated genes are more likely to be H3K27me3-diverged than expected, based on their proportion in each genome (Fisher's Exact Test: Sim: p = 3x10-12; Yak: p < 2.2x10-16; Pse: p = 6.9x10-11).

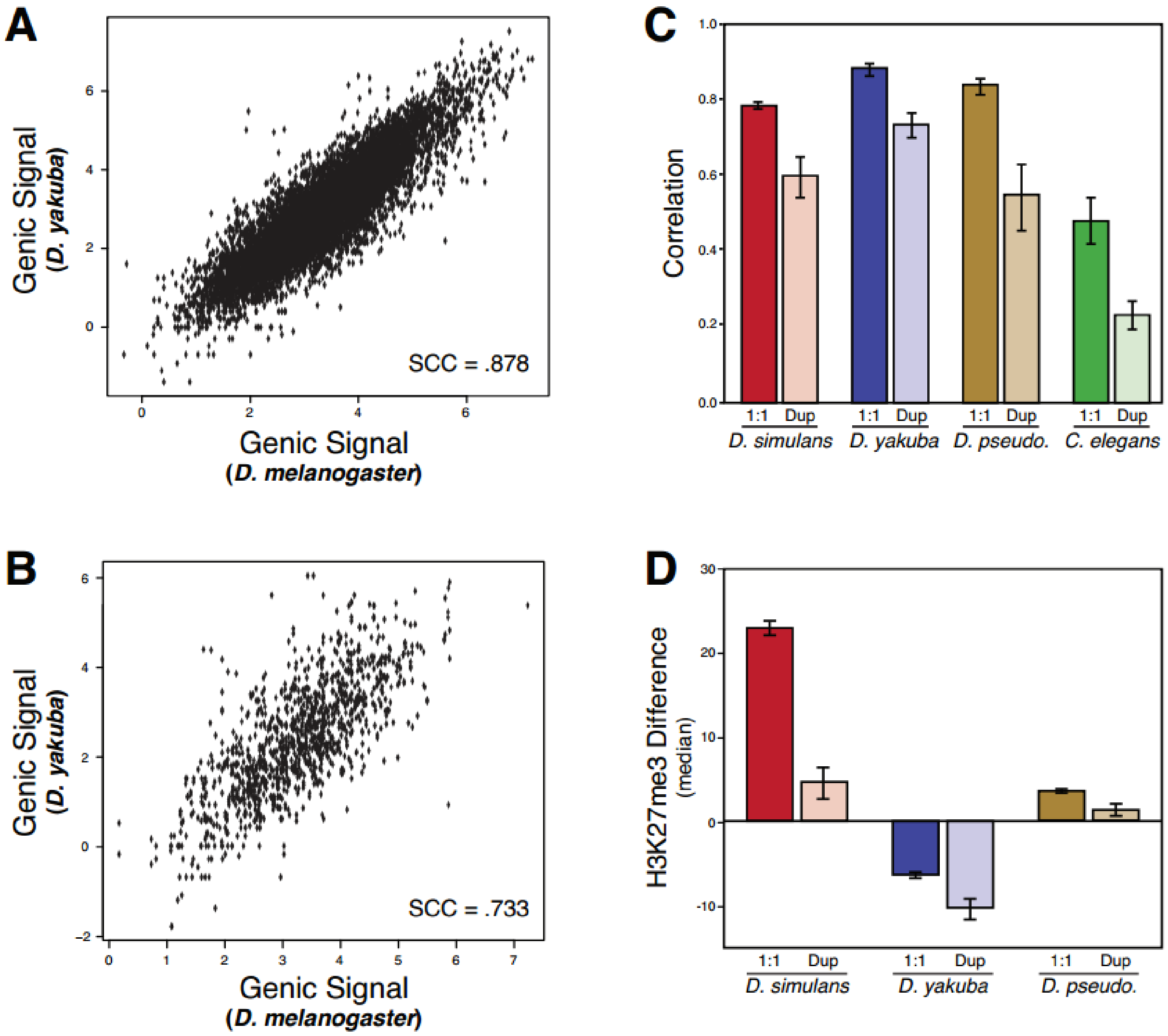


Species comparison	Total number of diverged genes	Diverged single copy genes	Diverged duplicated genes	Expected diverged duplicated genes
melanogaster–simulans	589	466	123***	37
melanogaster–yakuba	569	343	226***	59
melanogaster–pseudoobscura	496	331	165***	66

While duplicated gene orthologs tended to have less confidence in orthology assignment, we did not find a significant association between ortholog bootstrap confidence and magnitude of change in H3K27me3, indicating that poor ortholog identification is not responsible for the observed trend. In addition, excluding large ortholog families which had experienced many duplications did not change the trend significantly (Arthur et al. 2014). To further verify that read mismapping within paralogs was not the cause of our results, we analyzed the promoters of each orthologous gene similarly. Because promoter sequences evolve more rapidly than exonic sequence, promoters should suffer less from read mismapping. We found that, as with exonic H3K27me3 signal, promoters' methylation signal conservation was less for duplicated gene orthologs than for single-copy orthologs (Arthur et al. 2014).

Figure 4.3 Duplicated genes have less conserved H3K27me3 signal.
(A) As in 1A, single-copy orthologous gene H3K27me3 conservation between *melanogaster* and *yakuba*. (B) As in 1A/3A, but depicting duplicated gene orthologs. (C) Overall pattern of H3K27me3 conservation in different gene sets. Each pair of bars represents one species comparison; bars on the left are the correlation coefficient of single-copy orthologs; on the right, the correlation coefficient of duplicated gene orthologs. Each bar represents a single pairwise comparison's Spearman rho, compared against *Drosophila melanogaster*: red is for *simulans*, purple is for *yakuba*, blue is for *pseudoobscura*, and grey is for *C. elegans*. Overall correlation decreases in each case. Lines are 95% bootstrapped confidence intervals for each correlation; we account for the differences in sample size (single-copy orthologs are more common than duplicated gene orthologs) by resampling the number of duplicated gene orthologs in each case. (D) Duplicated gene orthologs are more likely to lose H3K27me3. Each pair of barplots is one species (*simulans*, *yakuba*, *pseudoobscura*); the left in each pair is the median H3K27me3 change in duplicated genes relative to the single gene in the other genome; while the right is the median in single-copy orthologs. Each pair shows that duplicated genes are biased towards the loss of H3K27me3 signal. For each pair (by permutation test): *simulans*, p < 10-5; *yakuba*, p < 10-5; and *pseudoobscura*, p < 10-5.

Figure 4.3 Continued



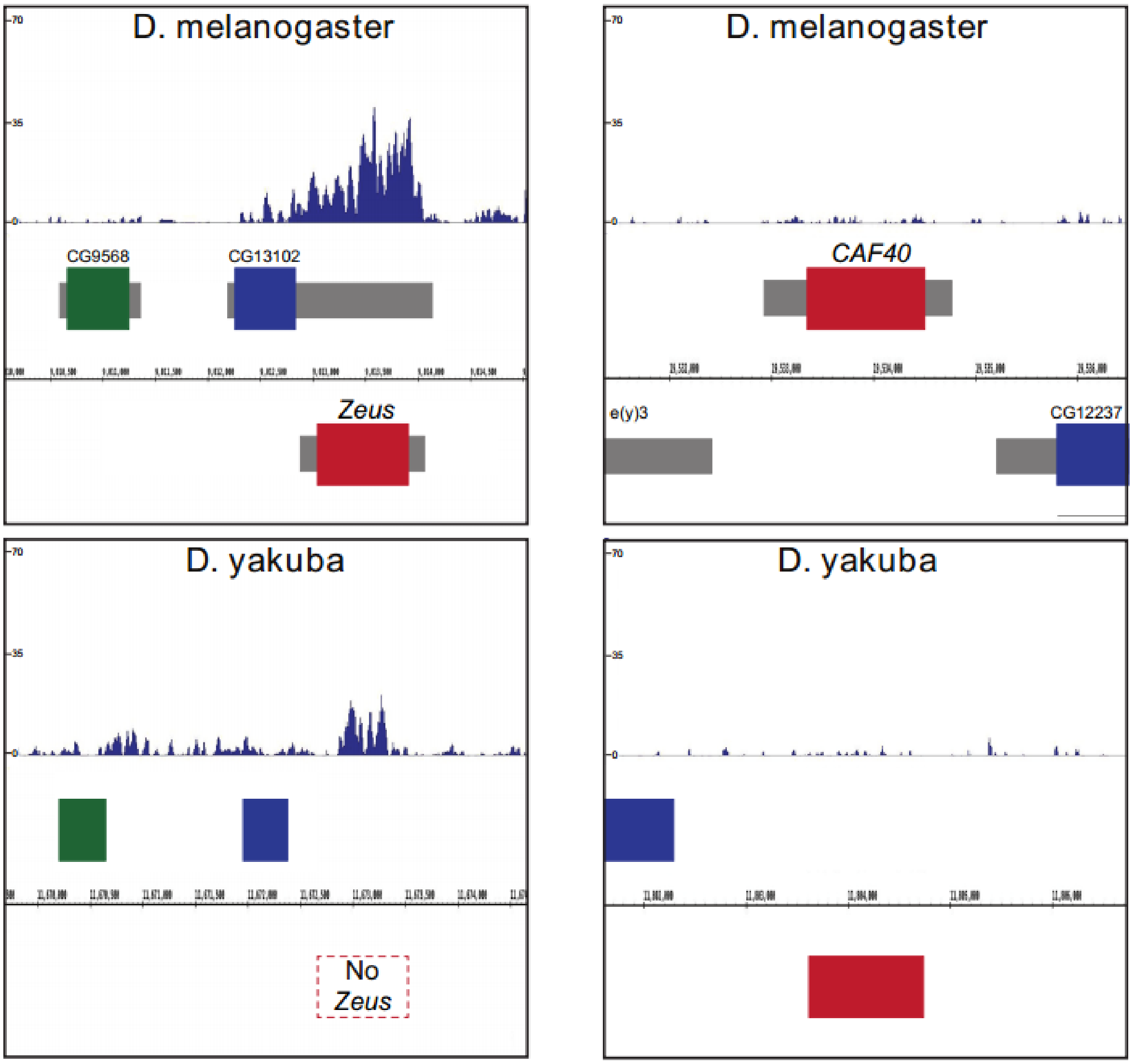
While change in H3K27me3 was associated with accelerated sequence evolution in single-copy orthologs, this effect was much more pronounced in duplicated gene orthologs (SCC = .099 for duplicated genes, vs. SCC = .045 for single-copy orthologs; this difference is significant [by bootstrap, p <.05]). Similarly, duplicated gene orthologs showed stronger correlations between gain of H3K27me3 and loss of expression (Permutation test: Sim, $p < 10^{-4}$; Yak, $p < 10^{-4}$; Pse, $p < 10^{-4}$). The more robust associations of H3K27me3 evolution, gene expression evolution, and d_N/d_S in duplicated gene orthologs may indicate a tighter relationship between H3K27me3 evolution and neofunctionalization in duplicated genes than in single-copy orthologs.

Comparing duplicated genes to each other (rather than orthologous genes in another genome), we found that duplicated gene paralogs are tightly correlated in the direction and magnitude of H3K27me3 evolution after a duplication event. Evolution of H3K27me3 in duplicated gene paralogs showed significant positive correlations, indicating that both duplicates tended to undergo similar epigenetic changes after their duplication (Sim: 0.264, Yak: 0.362, Pse: 0.348, Cel: 0.447). The positive correlation indicates that the duplication event allows not only the new gene to alter its epigenetic state, but also the parent.

The median H3K27me3 change for duplicated orthologs is significantly lower than single-copy orthologous gene H3K27me3 differences in all cases (Fig. 3D), indicating an overall loss of H3K27me3 following duplication (Permutation test: Sim, $p < 10^{-5}$; Yak, $p < 10^{-5}$; Pse, $p < 10^{-5}$). However, we observed that there were significant positive correlations between gene duplicate age and evolutionary change of H3K27me3 levels, indicating that gene duplicates regain lost H3K27me3 signal as they age (SCC: Yak: .109, $p < 10^{-5}$; Pse: .064, $p = .015$). The loss of H3K27me3 was also related to an average increase in expression of both duplicates (Permutation test: Sim, $p < 10^{-5}$; Yak, $p < 10^{-5}$; Pse, $p < 10^{-5}$). We examined expression patterns in newly-duplicated genes using the Berkeley Drosophila Genome Project's *in situ* database (Tomancak et al. 2002) and found that such genes are enriched for tissue specific expression (Fisher's Exact Test, $p = 0.0006$). Since these recent duplicates generally lose H3K27me3, our results imply that such genes may acquire tissue-specificity through DNA-binding trans-regulators.

A notable exception to this pattern is the gene *Zeus* (FBgn0032089; also called *Rcd-1r*). Previously, *Zeus* has been shown to be a recent duplicate of *CAF40* (FBgn0031047; also called *Rcd-1*) which has acquired an essential role in gonadal development (Chen et al. 2012). *Zeus* duplicated sometime before the common ancestor of *D. melanogaster* and *D. simulans*; it is not present in *D. yakuba*. *Zeus* has acquired a substantial increase in H3K27me3 in *D. melanogaster* relative to both its paralog, *CAF40*, and the location into which it duplicated in the *D. yakuba* genome (Fig. 4).

Figure 4.4 The novel gene duplicate *Zeus* has undergone rapid gain of H3K27me3. (A) A genome browser snapshot indicating H3K27me3 signal at the *Zeus* (*Rcd-1r*) locus in D. melanogaster. The first window indicates the normalized, input-subtracted H3K27me3 signal pooled from two separate biological replicates in *D. melanogaster*. Below, boxes indicates protein-coding genes: CG9568 (green), CG13102 (blue), and *Zeus* (red). The next window is as above, but in an outgroup species (*D. yakuba*) which does not possess the *Zeus* retroposition. Browser snapshots are aligned such that orthologous genes match position. Note that H3K27me3 level around *Zeus* is substantially higher in *D. melanogaster* relative to the equivalent region in *D. yakuba*. (B) As above, but focusing on the parental gene *CAF40* (*Rcd-1*; red), flanked on the left by *e(y)3* (blue), and on the right by CG12237 (blue). Note that *CAF40* possesses little to no H3K27me3 signal, in strong contrast to *Zeus*. The sequence identity between Zeus and CAF40 is 71% (Quezada-Diaz et al. 2010).

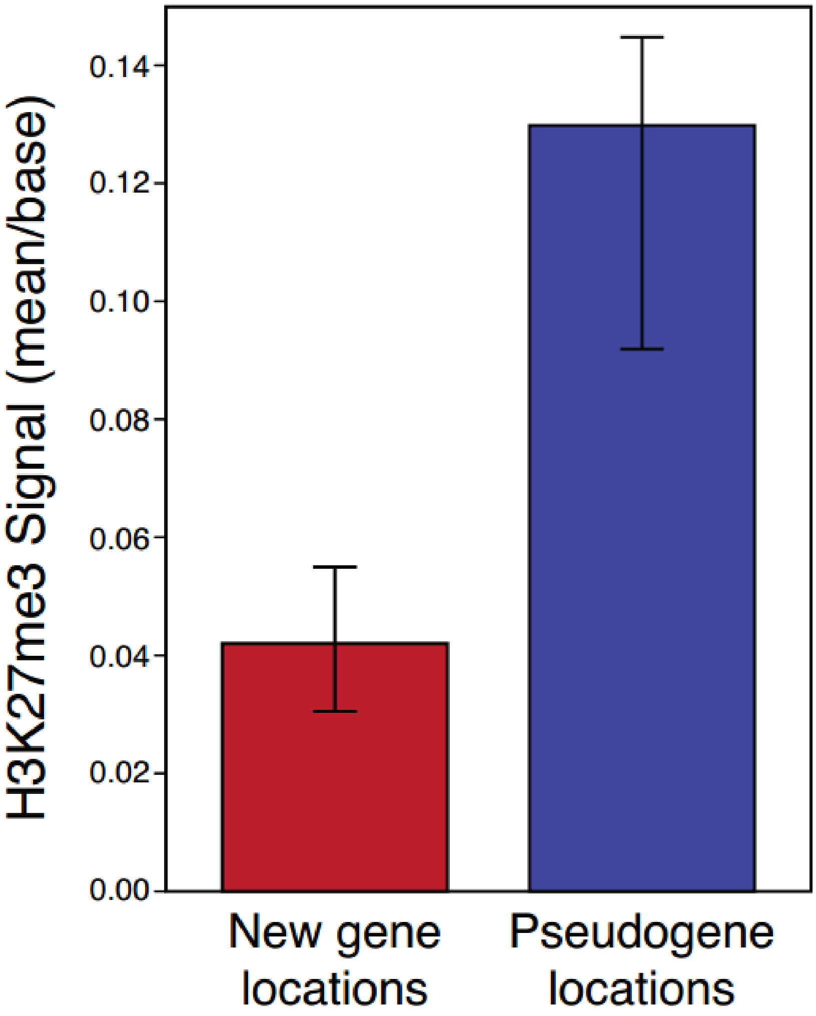


Duplicated genes relocate to regions of low H3K27me3 signal

Novel genes, once duplicated, must rapidly acquire unique functionality or else suffer pseudogenization (Force et al. 1999). However, to acquire functionality a gene must first be expressed and exposed to selection. We predicted that newly duplicated genes in repressive chromatin environments (bearing high H3K27me3 signal) would be less expressed and thus more likely to become pseudogenes.

We examined the relative H3K27me3 profiles in *D. simulans* for the locations into which both *D. melanogaster*-specific novel genes and pseudogenes had duplicated. Using synteny, we located the regions in *D. simulans* that novel genes moved into to infer the ancestral H3K27me3 levels before the duplication occurred (see Supplemental Methods of Arthur et al. 2014). Importantly, this analysis relies on the assumption that the extant *D. simulans* H3K27me3 signal is, on average, similar to the H3K27me3 signal in the ancestor of *D. melanogaster* in which the gene duplication event occurred. We found that locations into which pseudogenes had duplicated were characterized by three-fold more H3K27me3 signal than the locations into which novel, protein-coding genes had duplicated (Permutation test, p = .0015; Fig. 5). This finding is consistent with our hypothesis, and suggests that chromatin-mediated silencing of novel gene expression can prevent the acquisition of function in newly-duplicated genes, adding a new dimension to the Duplication-Degeneration-Complementation model (Force et al. 1999). Furthermore, this result provides a mechanistic explanation for the observed bias towards loss of H3K27me3 in recent duplicated genes: only those duplicates which move into regions of low H3K27me3 signal are likely to remain functional.

Figure 4.5 Location into which duplicates move affect eventual pseudogenization fate. Regions in *D. simulans* genome into which new genes and pseudogenes had moved in *D. melanogaster* show differences in mean H3K27me3 signal. On the left, functional new genes' duplication locations show significantly less H3K27me3 signal than pseudogenes' equivalent locations. Bars represent bootstrapped 95% confidence intervals. Note that for the purposes of this analysis, we assume that extant *D. simulans* H3K27me3 signal is representative of the H3K27me3 signal occurring in the ancestor of *D. melanogaster* in which new gene duplications occurred (see Methods).



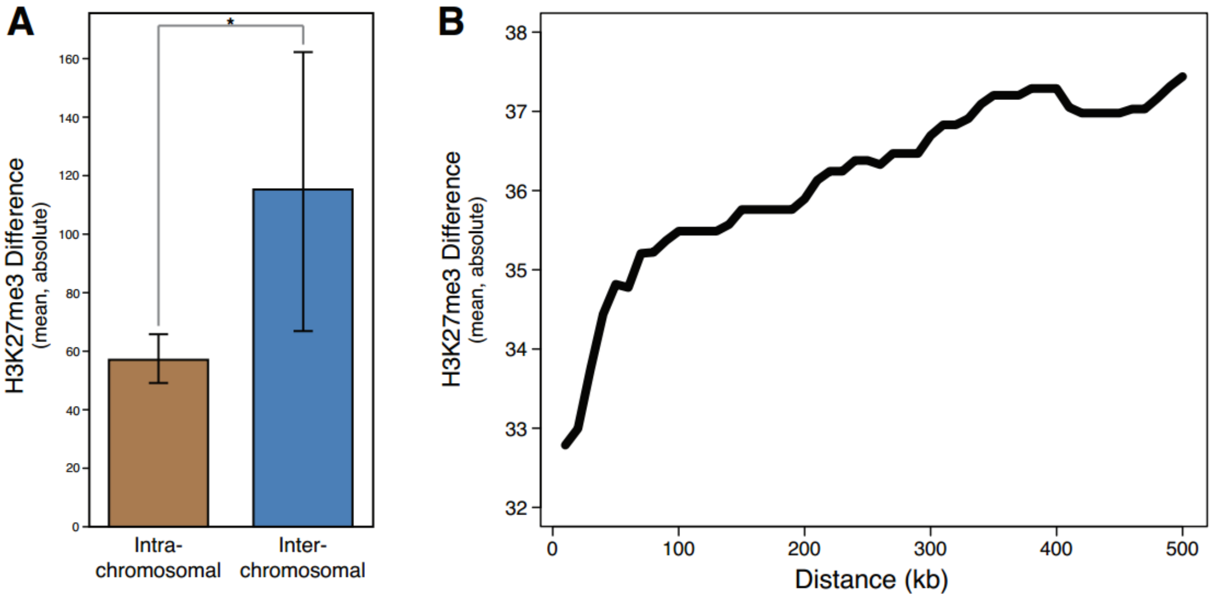
Patterns of H3K27me3 Evolution Depend on Duplication Mechanism

Gene duplication events are known to occur via diverse mechanisms (Hastings et al. 2009), which lead to different consequences for the resulting duplicated gene's location. Because H3K27me3 is spatially localized (Kharchenko 2010; Negre et al. 2011), we might expect that localized gene duplication events (e.g. tandem duplication) would lead to the newly duplicated gene falling within or close to the parent gene's H3K27me3 domain. Alternatively, a mechanism such as retroposition, which is able to deposit novel genes as far from their parent as different chromosomes (Kaessmann et al. 2009), might be able to completely change a novel gene's epigenetic regulation by delivering the novel gene to a location with an entirely different chromatin profile.

We examined whether the distribution of H3K27me3 changes varies by the location of the duplicated gene relative to its parent. We found that interchromosomal translocations show greater H3K27me3 divergence than intrachromosomal translocations (Fig. 6A; Sim, $p < 10^{-5}$; Yak, $p < 10^{-5}$). As expected, interchromosomal translocations also show greater divergence in gene expression (Sim, $p = .017$; Yak, $p < 10^{-5}$). These results show that interchromosomal duplications drive greater change in H3K27me3 signal and concomitant divergence in gene expression.

The hypothesis that localization affects H3K27me3 divergence also predicts that intrachromosomal duplicates are more likely to undergo more extensive H3K27me3 evolution the further they are from the parent gene. Indeed, we see evidence within the intrachromosomal population that the distance between the old gene and the new gene is related to the subsequent H3K27me3 divergence (Fig. 6B). There is a significant correlation between distance from the parental gene and the magnitude of H3K27me3 divergence (Yak: SCC = 0.094; Permutation test, p = .001), but it is worth noting that this effect saturates after ~100kb, suggesting that moving beyond that threshold does little to further change H3K27me3 regulation. This saturation effect coincides with a loss of H3K27me3 autocorrelation in approximately the same range (data not shown).

Figure 4.6 H3K27me3 changes in paralogs differ depending on duplicate location.
(A) Within *Drosophila simulans*, interchromosomal translocations have significantly greater H3K27me3 changes (in terms of absolute value; permutation test, p < 10-5). To accurately identify chromosomal translocations, we limited comparison to the major chromosome arms (2R, 2L, 3R, 3L, and X), and considered any gene which moved between chromosome arms an interchromosomal translocation.
(B) On the same chromosome, duplicated gene orthologs' H3K27me3 changes increase with distance from the parent gene. On the y-axis, the cumulative mean change in H3K27me3 signal as a function of distance (on the x-axis). There is a significant correlation between distance and magnitude of H3K27me3 change (SCC = 0.103, p = .002).



Discussion

Previous studies have discovered relatively slow evolution of histone modifications (Wilson et al. 2008; Cain et al. 2011). Our results reinforce the slow evolution of a histone modification, but we find very different evolutionary regimes between single-copy orthologs and orthologs which have undergone duplication events. While single-copy orthologs exhibit relatively little divergence in terms of genic H3K27me3, duplicated genes evolve comparatively rapidly.

These differences are exaggerated in the most distant comparison we make, between *Drosophila melanogaster* and *Caenorhabditis elegans*. Orthologous genes in these species show substantial similarity in H3K27me3 signal, even though the two species are deeply diverged. Meanwhile, orthologous gene sets in which duplication events have occurred are significantly less conserved.

Among duplicated genes, we find interesting dynamics of H3K27me3 change. Gene duplicates show highly correlated gain or loss of H3K27me3 signal, indicating that both parent and duplicate undergo similar changes to their epigenetic status following duplication. Duplicated genes are biased with respect to the H3K27me3 change they are subject to: both duplicates are more likely to lose methylation than single-copy orthologous genes. By examining the regions into which functional new genes localize relative to pseudogenes, we show that new genes tend to move to regions of low H3K27me3 signal. This result comports with the Duplication-Degeneration-Complementation theory (Force et al. 1999), and suggests the importance of chromatin state as a determinant of novel gene fate. We speculate that the removal of H3K27me3 could allow paralogs to acquire new tissue specificity via trans regulatory factors instead of histone modifications (Huerta-Cepas et al. 2011).

We find evidence that gene duplications on different chromosomes and at greater distances are more likely to acquire H3K27me3 signal and gene expression changes. The novel gene *Zeus* exemplifies this pattern, and has gained a substantial level of H3K27me3 relative to its parent gene *CAF40*. These results strongly imply that gene duplication mechanisms which can move the resulting duplicate further from the original gene are associated with more H3K27me3 evolution in the duplicate, which could have important consequences for understanding the eventual fate of gene duplicates. We do not know to what extent these differences would manifest for other histone marks, which tend to be relatively more compact than H3K27me3 domains (Kharchenko et al. 2011; Negre et al. 2011).

Significant evolution of H3K27me3 signal was associated with evolution of gene expression and sequence evolution (as measured by d_N/d_S). In a separate comparison, independent of genes, a bin-based approach also found that evolution of H3K27me3 was also associated with an elevated rate of sequence substitution between species (see Supplemental Material of Arthur et al. 2014). It appears that evolution of this mark is associated with evolution in the underlying sequence, although the causal basis of that association is ambiguous in the present study (Xiao et al. 2012). Whether H3K27me3 evolution allows the sequence to evolve, or whether H3K27me3 divergence is the result of new Polycomb regulation due to the creation or destruction of PcG binding sites is unknown.

To this end, future work is necessary to clarify exactly how the evolution of H3K27me3 is mediated genetically. Although many of the key enzymes responsible for deposition and maintenance of H3K27me3 are known, it is likely that we do not have a complete catalog of all the elements of all PREs (Delest et al. 2012), nor do we know the exact spatial or orientation requirements of the involved binding sites (if indeed there are any). However, the extreme conservation of H3K27me3 across distantly related species leads to the question of how a cis-regulatory element composed of relatively degenerate sequence motifs could persist functionally over such a long period of time (Fisher et al. 2006; Ruvinsky and Ruvkun 2003).

It is worth noting caveats which apply to the above analysis. Our results are gathered at only one strictly defined time period, which was chosen for its ease of use and developmental significance. Genes for which we observe no change in H3K27me3 signal may instead be highly diverged at other developmental times. Furthermore, it is possible that the dynamics of histone modification evolution differ depending on the stage of development (Kalinka et al. 2010). Additionally, we believe the observed differences in H3K27me3 reflect a lower bound estimate of the possible differences between species. Because of the heterogeneous nature of the sample material, it may be difficult to detect small-scale, tissue-specific differences in the epigenetic markings over loci. It may well be that orthologous genes undergo many changes of this nature between the species examined.

In summary, we have shown that H3K27me3 signal diverges more rapidly in duplicated genes. When diverged, duplicated genes most often show correlated loss of H3K27me3 relative to the single ortholog in the other genome, but the extent to which this mark changes depends on how far the duplicates are moved from the parent gene.

These results are indicative of an interesting dichotomy in the regulation of epigenetic states. Whereas single gene orthologs remain locked into their epigenetic status, duplicates are able to undergo rewiring of H3K27me3 signal and resulting gene expression. Our results indicate an interplay between gene duplication and the evolution of chromatin state as a mechanism for generating evolutionary novelty.

Acknowledgements

The modENCODE Project was supported by an NIH grant U01HG004264 awarded to KPW. RKA was supported by an NSF Graduate Research Fellowship and an NIH training grant (T32GM007197). We thank Jason Lieb and his lab for use of the *C. elegans* H3K27me3 ChIP-seq data. We are grateful to Benjamin Krinsky and Manyuan Long for interesting discussions on gene duplication. We thank Kacy Gordon, Aashish Jha, and Xiaochun Ni for helpful comments and critical review of our manuscript. We are indebted to many members of the White and Ruvinsky labs for useful criticism, wonderful discussions, and technical help. Finally, we appreciate the comments of three anonymous reviewers whose feedback greatly improved the paper.

Chapter 5: Recent Cis-Trans Coevolution Driven by the Emergence of a Novel Gene in Drosophila

Abstract

Newly evolved genes can rapidly acquire crucial molecular functions. However, the way in which a new regulatory protein acquires and co-evolves with its targets in the genome is not understood. Here, we investigate the genome-wide binding profiles of two orthologues of the young Drosophila gene, Zeus, compared to the ancestral binding pattern inferred from a pre-duplication orthologue of its parental gene, Caf40. Zeus binding rapidly co-evolved with a previously unknown DNA motif under the influence of positive selection. Furthermore, while both versions of Zeus acquired targets at male-biased and testis-specific genes, each Zeus protein has specialized binding on different chromosomes, a pattern echoed in the evolution of the associated motif. Our results thus reveal that over short evolutionary timescales, young regulatory genes can drive dynamic, genome-wide processes including substantial re-wiring of the transcriptional networks into which they integrate.

Introduction

The origin of new genes can lead to the evolution of new and crucial functions in myriad biological processes, including gene regulation (Chen et al. 2010; Chen et al. 2012). Regulatory elements can now be investigated on a genome-wide scale in order to compare patterns of conservation and divergence of gene regulation in multiple, closely related species (Stefflova et al. 2013). This presents a unique opportunity to investigate the evolution of new regulatory genes and as well as their effects on bound regulatory elements.

The gene Zeus (CG9573) is a young gene that arose via a retroposition event approximately 5 million years ago in the lineage leading to *Drosophila melanogaster* and its closest relatives. Zeus subsequently underwent a period of rapid molecular evolution (Quezada-Díaz et al. 2010), and evolved specific roles in the development of Drosophila sperm and testis (Chen et al. 2012). In contrast, its parental gene Caf40, conserved across eukaryotes, is ubiquitously expressed and is essential for viability in Drosophila melanogaster (Chen et al. 2012). On the molecular level, it had been previously inferred that Caf40 has nucleic acid binding properties, and thus might act as a regulator (Garces et al. 2007). By performing chromatin immunoprecipitation followed by microarray analysis (ChIP-chip), it was discovered that both Zeus and Caf40 from *D. melanogaster* bind to several hundred sites throughout the genome, and that Zeus has acquired many novel regulatory targets in the genome, consistent with neofunctionalization following duplication (Chen et al. 2012).

The Zeus locus arose after the divergence of the lineages leading to *D. melanogaster* and *D. yakuba*, but prior to the divergence of *D. melanogaster* and one of its closest sister species, *D. simulans*. The Zeus protein subsequently acquired a large number of species-specific substitutions along these two lineages (Chen et al. 2012) (Fig. 1A). To understand patterns of lineage specific regulatory evolution of Zeus, as well as its initial divergence from the ancestral state of Caf40, we have characterized the genome-wide binding profiles of Zeus from *D. melanogaster* and its sister species *D. simulans*, as well as Caf40 from *D. yakuba*. We selected *D. yakuba* (pre-duplication) Caf40 as the best proxy from which to infer ancestral Caf40 binding, because the *D. yakuba* and *D. melanogaster* proteins differ in only 4 positions.

Methods

ChIP-seq Data Production

ChIP-seq experiments were performed using standard modEncode protocols (www.modencode.org) after collecting adults in each species. Sequencing data was generated by the High-Throughput Genome Analysis Core (HGAC) at the Institute for Genomics and Systems Biology.

Genomic Data Analysis

ChIP-seq reads were mapped with BWA (Li and Durbin 2009), using default parameters, to the most recent UCSC genome versions. Motif discovery was performed with DREME (Bailey 2011).

Population Genetics

We used SNP calls from the DGRP (Mackay et al. 2012), filtering variants with minor allele frequency less than .05 to remove weakly deleterious variation.

Results

We engineered transgenic lines of D. melanogaster (w1118) containing FLAG-tagged *D. melanogaster* Zeus, *D. simulans* Zeus and *D. yakuba* Caf40, and performed ChIP-seq on each of these lines (Fig. 1A). This allows us to directly compare binding properties of the three proteins in a common genome. We term these three proteins Dmel Zeus, Dsim Zeus, and Dyak Caf, respectively. We obtained reproducible ChIP-seq signals between replicates (Supplementary Fig. 1; Supplementary Table 1 of Krinsky et al. [submitted]), and observed Zeus binding affinity correlated with temporal patterns of expression in the testis (Krinsky et al. [submitted]). We observed enrichment of ChIP signal primarily at the transcription start site and within the exons of bound genes, refining previously hypothesized Zeus and Caf40 binding preferences (Chen et al. 2012). To assess the differences in binding among the three proteins, we calculated signal enrichment for each gene based on the enrichment of reads within 700bp of the transcription start site. Principal component analysis on the gene-by-gene signal revealed that replicates corresponding to each protein (Dmel Zeus, Dsim Zeus, Dyak Caf40) formed distinct clusters (Fig. 1C), demonstrating significant differences in binding preferences between proteins. Moreover, Dsim Zeus sites were more highly diverged from Dyak Caf40 than Dmel Zeus sites (Krinsky et al. [submitted]; pairwise Euclidean distance, p<.001), consistent with the reported pattern of protein-coding sequence divergence (Fig. 1A).

Figure 5.1. Design and results of ChIP-seq experiments. A) Depiction of *Zeus/Caf-40* phylogeny, with experimental design. *Zeus* originated from a gene duplication event 4-6 million years ago, before the split of *D. melanogaster* and *D. yakuba*. We sampled two copies of Zeus (*D. melanogaster* and *D. simulans*) as well as a single copy of Caf-40 from *D. yakuba*, which represents the ancestral, pre-duplication state of the protein. All three proteins were introduced into the D. melanogaster genome with 3x FLAG tags attached, in order to eliminate problems with variable antibody affinity. Numbers indicate the volume of nonsynonymous (before the slash) and synonymous (after the slash) changes. B) ChIP-seq signal is concentrated at the transcriptional start site of genes. The scaled, read depth normalized mean signal is shown for each protein's average ChIP-seq binding pattern in the vicinity of the TSS. The central line represents the Transcription Start Site. Signal is highest at this point, falling off rapidly upstream of the TSS, and remaining strong into the exons of genes (downstream). The three proteins showed no detectable difference in signal as a function of distance from the TSS. C) A graph of the first two principal components of ChIP-seq read counts revealed reproducible clustering of replicates of the same protein, while different proteins showed differentiation. D) A bar plot showing the median normalized read counts over TSSs for each chromosome, indicating differences in chromosome-level affinity of the three proteins. Both copies of Zeus show increased affinity relative to Caf-40 on chromosomes X and 4, albeit to different degrees.

Figure 5.1 Continued

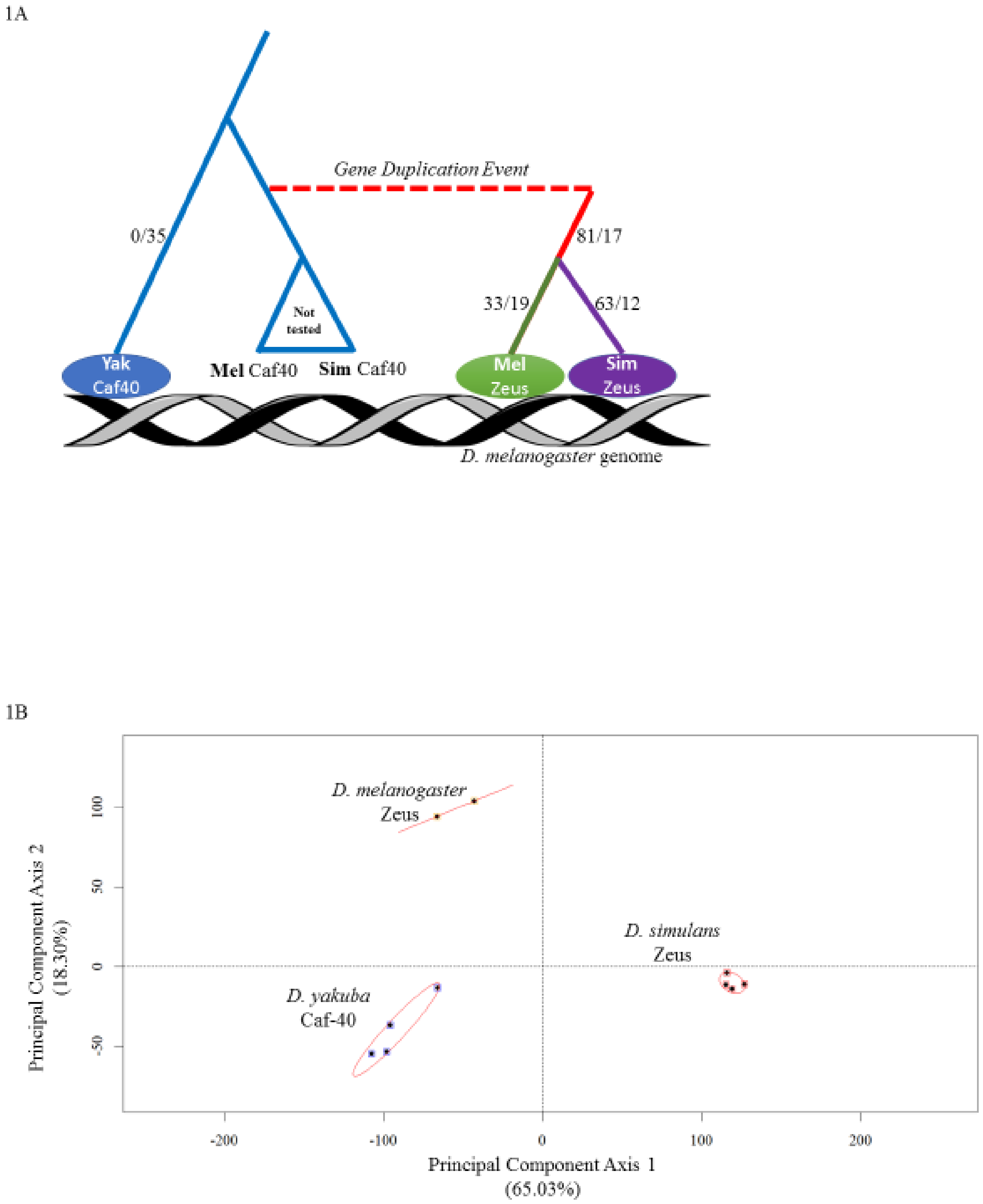
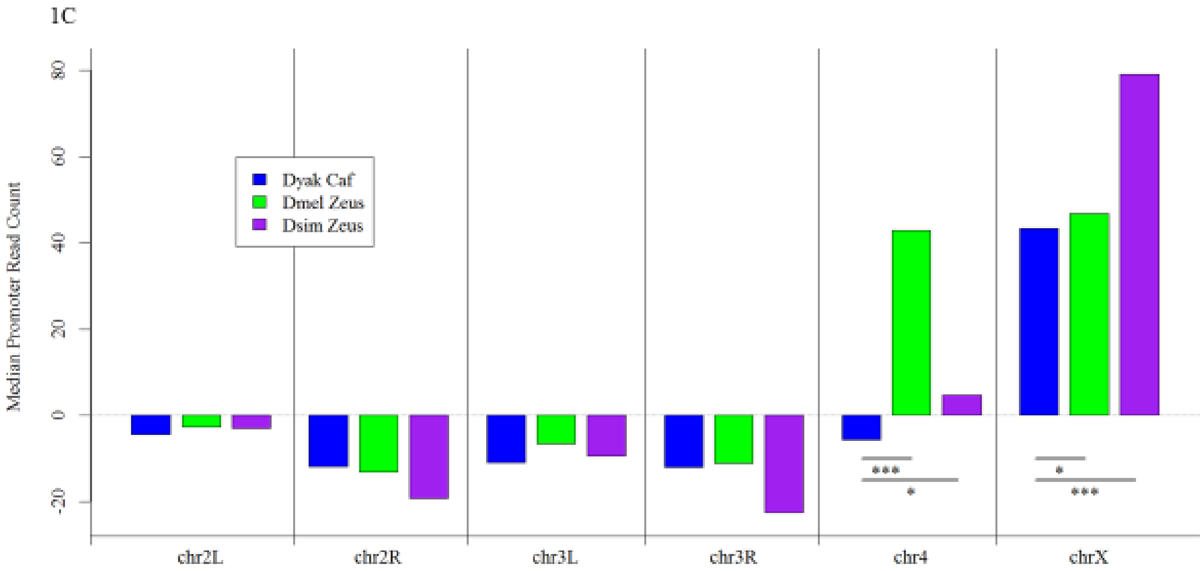


Figure 5.1 Continued



Based on previous ChIP-chip results showing that Zeus preferentially binds the X-chromosome, and because of the known roles of Zeus in regulating sex-specific functions, we compared the chromosomal distribution of reads (Conrad and Akhtar 2012). We found ChIP-seq read enrichment for all three proteins on the X-chromosome relative to the autosomes (Fig. 1D), but both Zeus orthologs showed significantly higher X vs. autosome signal enrichment compared to Dyak Caf40 (Permutation test: p<.05). Dsim Zeus showed particularly strong X-chromosome enrichment (Permutation test: p<.001).

Both Zeus proteins also exhibited a bias for the 4th (dot) chromosome, which has been hypothesized to be an ancestral sex chromosome (Vicoso and Bachtrog 2013). Interestingly, Dmel Zeus was strongly enriched for signal on the 4th chromosome (p<.001), whereas Dsim Zeus was mildly, but significantly, enriched (p<.05).

Both the X and dot chromosomes are enriched for female-biased genes (Gnad and Parsch 2006) (Fisher's Exact Test: p = 2.728x10-14). The chromosomal distribution of sites thus suggests a scenario in which Zeus gained affinity for sex-biased genes on the X-chromosome and the 4th chromosome as part of its testis-specific neofunctionalization, and then subsequently evolved differences in chromosome-level binding between *D. melanogaster* and *D. simulans*. Genes on *D. melanogaster*'s 4th chromosome were found to be more highly female-biased than *D. simulans*, explaining the significant species-specific difference in affinity (Ranz et al. 2003) (See Supplementary Methods; Permutation test, p<.05).

Caf40 is among the most conserved nucleic-acid binding proteins across eukaryotes, from metazoans to flowering plants (Hoffmann and Valencia 2004). We reasoned that the extensive protein-coding (trans-) divergence of Caf-40 and Zeus may have driven evolution of conserved, bound cis-regulatory elements (Ross et al. 2013). We therefore searched for motifs for each protein using DREME (Bailey, 2011). A single, highly specific motif (ACTGCTT) was enriched in all three proteins' binding sites, which we term the Caf-40 and Zeus-Associated Motif (CAZAM).

We first noted that the genome-wide frequency of the CAZAM differed between Drosophila species with and without the Zeus gene. The three species of sequenced Drosophilids with both the Zeus and Caf40 genes had significantly lower overall CAZAM frequencies than sequenced species with only Caf40, which remained true after correcting for genome size (Fig. 2A; using phylogenetic ANOVA of Garland et al. 1993). No randomly constructed motifs were similarly unevenly distributed among the genomes (Krinsky et al. [submitted]).

Figure 5.2. Results of analysis of the CAZAM. (A) Plot showing the difference in mean frequency of the CAZAM in species with (top bar) and without (bottom bar) the *Zeus* duplication. (B) Schematic illustrating our modification of the McDonald-Kreitman test. We substitute central and flanking sites for dN and dS, respectively, allowing us to measure selection on all identified instances of the motif. (C) Observed alpha (α) values (interpreted as the proportion of adaptive substitutions) for different comparisons, with bootstrapped 99% confidence intervals. (D) Comparison of estimated alpha values from motifs which map to the X chromosome (left) versus those which map to the autosomes (right).

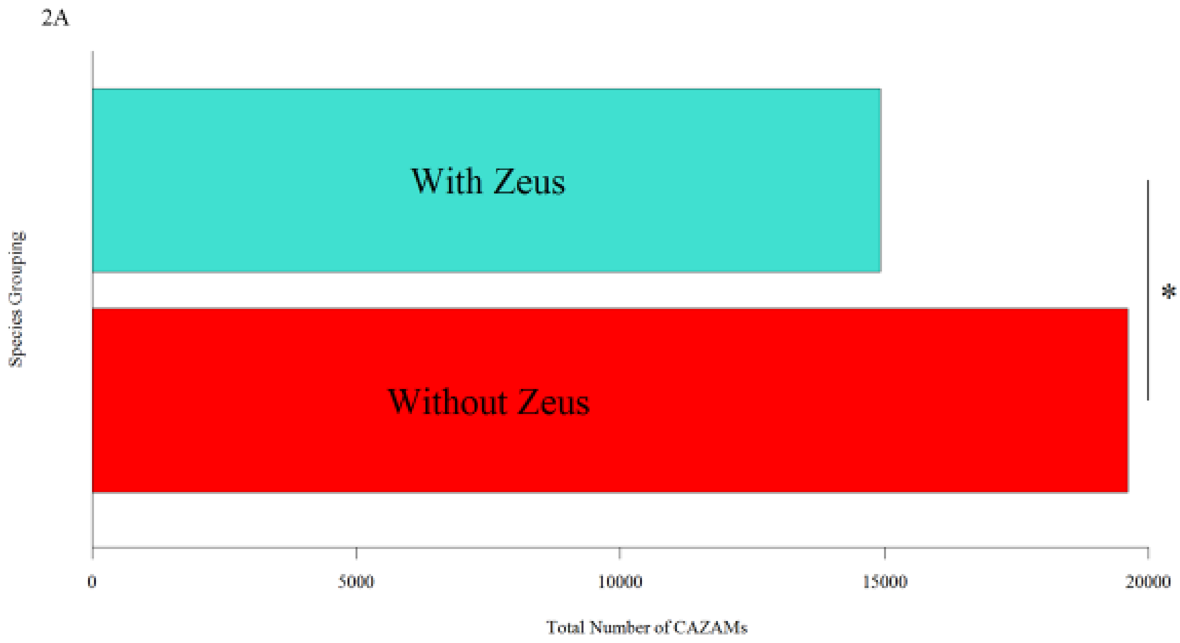


Figure 5.2 Continued

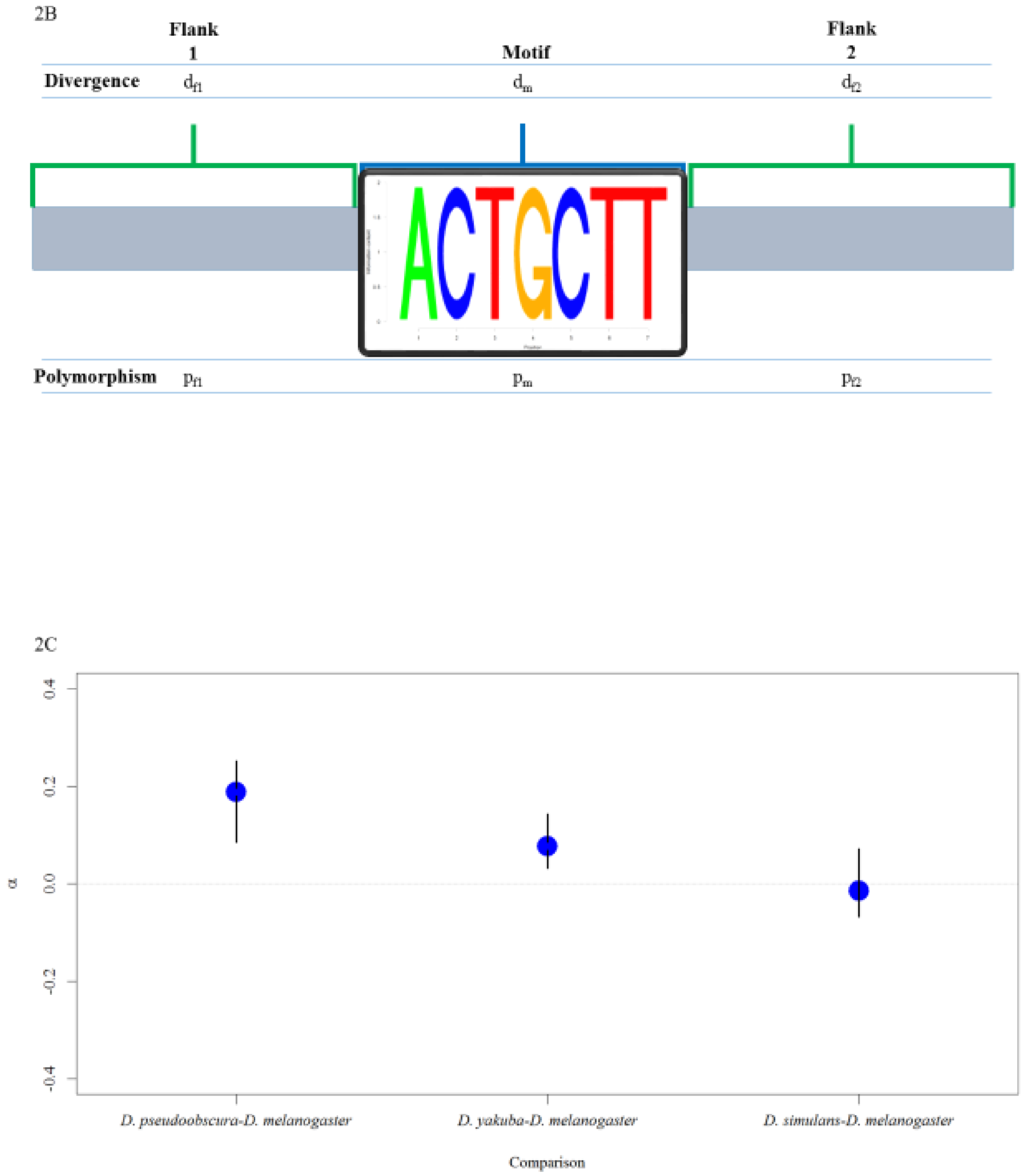
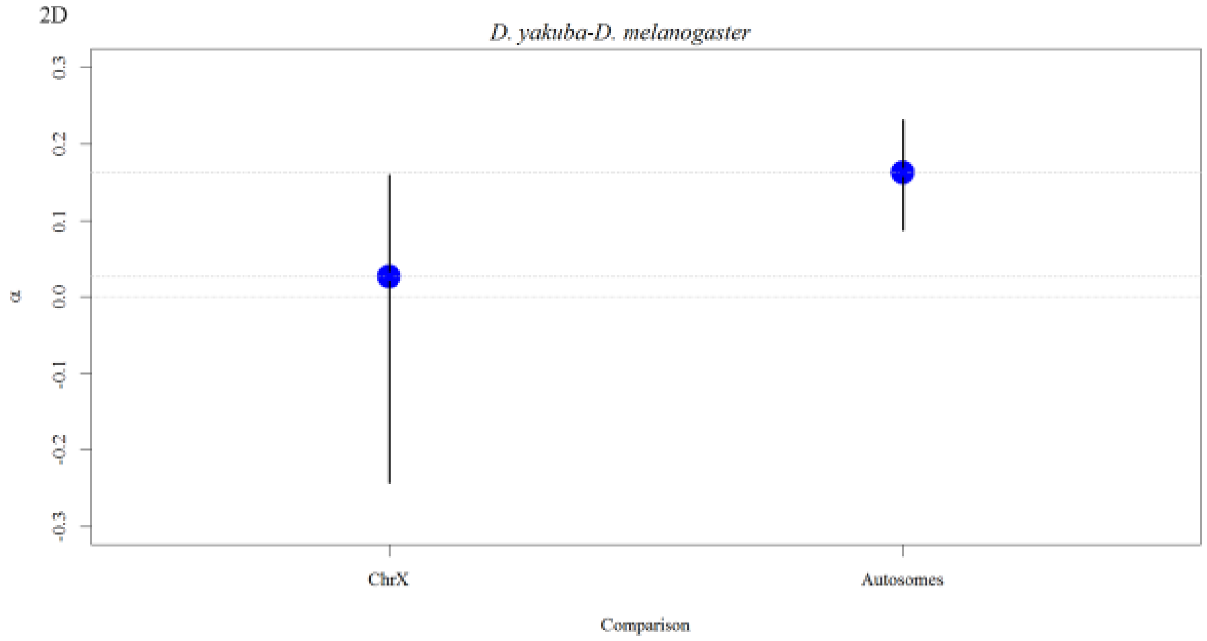


Figure 5.2 Continued



We further found that the chromosomal distribution of the motif was radically different among the genomes with and without Zeus. After the origination of Zeus, the frequency of CAZAMs on the X-chromosome did not change appreciably, while motifs decreased on the autosomes (Supplementary Table 2). The fraction of motifs within 1kb of exons, on both X and autosomes, increased dramatically as well (97.8% in *D. simulans* vs. 84.0% in *D. yakuba*; Fisher's Exact Test: p < 1x10-7). These results suggest selection acted to remove and redistribute thousands of copies of the motif following the gene duplication event, perhaps as a result of a new selective regime driven by the emergence of Zeus.

To directly assay for positive selection, we modified the framework of the McDonald-Kreitman test so that it could apply to motif-level analyses (Jordan and McDonald 1998) (Figure

2B; see Krinsky et al. [submitted]). Because we posited, based on the overall difference in motif frequency between species, that there was selection to remove the motif from the genomes of species after Zeus duplicated from Caf40, we identified motif instances in pre-duplication species (*D. yakuba, D. pseudoobscura*) as well as one post-duplication species (*D. simulans*) and mapped their syntenic locations into *D. melanogaster*.

Using divergence data from whole-genome multiple alignments and D. melanogaster, and polymorphism data for D. melanogaster (Mackay et al. 2012), we found that there was significant evidence of positive selection on instances of the CAZAM following the gene duplication event (Figure 2C-D; *D. pseudoobscura-D. melanogaster*: Bootstrap test: p<.01; *D. yakuba-D. melanogaster*: p<.01). Selection was significantly stronger on intergenic motifs than on exonic motifs, consistent with our findings that all three proteins bound near exonic regions and that there was redistribution of the motif following the duplication event (Krinsky et al. [submitted]; p<.01). In contrast, performing the same comparison between *D. simulans* and *D. melanogaster* revealed no significant signature of positive selection, suggesting that selection acted after the duplication event, but decreased by the time the *D. melanogaster* and *D. simulans* lineages diverged (Long et al. 2013)(Figure 2D).

Because we determined earlier that Zeus binding shows a strong chromosome-specific bias consonant with its role in testes development (Ellegren and Parsch, 2007), we posited that regimes of selection differed across chromosomes. Correspondingly, we found evidence of stronger selection to remove CAZAMs from autosomal chromosomes than from the X chromosome (Fig. 2E; comparing intergenic motifs in *D. yakuba* and *D. melanogaster*; Permutation test: p<.01). We conclude, based on motif frequency differences and evidence of positive selection, that widespread selection driven by the origination of Zeus shaped both the abundance and distribution of the motif.

To confirm that this signature of selection was related specifically to the appearance of Zeus, we performed a series of control analyses. We examined several motifs that were shuffled versions of the CAZAM, and found significant McDonald-Kreitman test results for none of them after multiple testing correction (Bootstrap tests: p>.5; see Supplementary Information). We examined sequence divergence of the CAZAM between two species (*D. pseudoobscura* and *D. yakuba*) without Zeus, and found no significant signal of increased divergence at motif sites relative to flanks (Krinsky et al. [submitted]; Bootstrap test: p>.5). Our results therefore suggest that it was the origination of Zeus that led to widespread positive selection specifically on the CAZAM.

Discussion

Our results show that Zeus, a novel nucleic-acid binding factor in Drosophila, underwent a regime of rapid neofunctionalization ultimately leading to specialized binding to different chromosomes in different species. This trans-evolution drove strong positive selection to rearrange the chromosomal distribution of the motif associated with both Zeus and Caf40 binding, suggesting a co-evolutionary process of neofunctionalization occurring in both cis and trans.

With regard to the specific molecular mechanism by which Zeus might be regulating downstream targets, initial studies of Caf40 (also known as Rcd-1) suggested that it regulates target genes through direct interaction with the genome due to the fact that it contains six armadillo-type repeats, implicated in DNA binding. However, more recent work suggests that Caf40 in Drosophila may also act indirectly with nucleic acids as a member of the larger Ccr4-

Not complex, which has roles in mRNA processing and degradation. Our data show that via ChIP, we can indeed discover signals of Caf40 and Zeus binding that illuminate their evolutionary histories, although we cannot discount the possibility that the signals we detect could in fact be due to indirect interactions with the genome mediated through protein-protein binding. For example, extensive studies of transcription factors (TF) binding have revealed that interactions between TFs and the genome are mediated through a highly complex and variable suite of direct and indirect interactions between TFs and cofactors.

Although we do not have definitive evidence that Zeus functions within a protein complex, we are confronted by the tantalizing possibility that Zeus has undergone rapid evolutionary changes both in terms of its protein-protein and protein-nucleic acid interactions. Further study of the detailed biochemistry of Zeus should shed light on its precise mechanism of action. We also noted from our analysis that both Dmel and Dsim Zeus show enriched binding on the X and 4th chromosomes, consistent with the putative role of Zeus in the downregulation of female-biased genes. This finding is consonant with the fact that these two chromosomes are heavily hetrochomatinized, and that Zeus may also be involved in chromatin dynamics.

Regarding the evolution of the CAZAM, one might suggest that there are several important caveats that apply to our version of the McDonald-Kreitman test. Because of the genome-wide nature of our test, we examined many motifs which are likely not bound by either Caf-40 or Zeus due to occlusion by closed chromatin or DNA-bound factors. In addition, extending the M-K test to a genome-wide scale aggregated many unlinked motifs, which can have adverse and unpredictable effects upon the bias of the test. However, by creating an empirical null distribution of sequences resembling, but different from, the CAZAM, many of the potential issues with the modified M-K test can be reduced. If the test were overly liberal in detecting selection, we would expect to see selection on the permuted CAZAM sequences, as well as in pairs of species that did not differ in terms of the presence of Zeus. Instead, we find that the null hypothesis is rejected only for the specific motif we found in our ChIP-seq data, and only in the particular case in which one compares two species across a specific phylogenetic node that corresponds to the origination of Zeus.

Our results shed light on the fate of newly-arisen functional gene duplicates. From our studies of Zeus, we have demonstrated that novel regulatory proteins may cause positive selection to drive substantial rewiring of the transcriptional networks into which they integrate through changes both in the protein itself and the global cis-regulatory environment. These global changes, in turn, can have important phenotypic consequences (e.g. the development and function of the reproductive system), even over relatively short evolutionary time scales.

Acknowledgements

BHK was supported by an NIH Genetics and Regulation Training Grant (T32GM007197) and a Department of Education GAANN Fellowship (P200A090309/P200A120178). RKA was supported by an NSF Graduate Research Fellowship and an NIH Genetics and Regulation Training Grant (T32GM007197). We thank Steffen Lemke, Al Handler, Lijia Ma, and Daniel Matute for technical advice and for providing experimental materials and reagents. We also thank Rebecca Spokony, Kacy Gordon, Aashish Jha, Sidi Chen, and Grace Yuh Chwen Lee for helpful comments and critical review of drafts of our manuscript. We are grateful to Joe Thornton, Martin Kreitman, and Ilya Ruvinsky for useful advice. We are indebted to innumerable members of the White and Long labs for beneficial criticism, advice, and support.

Chapter 6: Functional Data Informs Our Understanding of Regulatory Evolution

Abstract

In this dissertation, I confront several of the major problems in our understanding of the evolution of gene regulation. I argue that by integrating functional datasets with a detailed understanding of molecular biology, we can shed light upon the evolution of gene regulation. In the final chapter, I conclude with a brief summary of the dissertation, as well as some still-unanswered questions and prospects for future research.

The varied evolutionary regimes occurring within cis-regulatory sequences pose a challenging set of questions. Detecting selection within such sequences is confounded by the birth/death processes of individual binding sites (Bullaughey 2011). The dynamics of these birth and death events can obscure canonical patterns of nucleotide identity indicative of selection.

Novel techniques allow the interrogation of regulatory function with exceptional resolution, at genome-wide scale, and across species. In combination, such techniques can be leveraged to understand the ways in which evolution in regulatory sequences drives gene expression evolution, as well as patterns of nucleotide divergence and polymorphism. These techniques offer great promise for understanding the evolution of gene regulation.

To use the novel techniques and datasets effectively, however, we have to address their many challenges and limitations. In each chapter, I dealt with a host of computational, statistical, and analytical difficulties which made the task of generating knowledge from the data difficult.

Broadly, it is important to be aware of how the data is generated and be familiar with the biological nuances of each technique. Uncritical usage of genomic datasets like RNA-seq risks severe misinterpretation, examples of which are abundant (Li et al. 2011).

For example, in Chapter 4, I analyzed ChIP-seq data for a histone modification, H3K27me3, which is differentially marked across tissues. Existing methods for analysis of this data assumed a binary, "presence/absence" model for the mark. As I argued, however, a continuous analysis is more appropriate under conditions of tissue type heterogeneity, such as were present in examining whole-embryo data. In this case, there was a mismatch between the existing methods of analysis and the technique used to collect the data, a dilemma which required resolution before evolutionary questions could be approached.

In addition to being aware of the ways in which the data is collected and how that relates to the underlying biological reality, I derived several broader lessons from the study of the evolution of gene regulation presented in this dissertation.

First, owing to the intricacies of the gene regulatory apparatus, there may not be any simple mapping of function onto fitness. Instead, we need to embrace the inherent complexity of the evolution of gene regulation, and understand that there are multiple layers of regulatory activity which interact with each other in various ways (e.g. cis and trans, in Chapter 5). In doing so, we may require more sophisticated models (like machine learning tools) and paradigms, as well as work which bridges subfields (chromatin and transcription factor biology, for example).

Secondly, it will be impossible to functionally annotate the whole genome in all individual tissues, at all time points, under all conditions, in all species. So we must develop methods to shortcut the process, perhaps by detecting sites under functional constraint without the classic signals of purifying selection (i.e. an absence of sequence substitutions). I believe

Chapter 3 is a way to build towards this goal, because it allows us to make predictions about functional conservation in another species using data from only one species.

Future Directions

We are still early in the process of deciphering the regimes and patterns of evolution which affect cis-regulatory sequences. By leveraging current and future technologies in an informed, sophisticated fashion, we will be able to build a better understanding of the complex relationship between function and conservation.

Sequencing-based technologies such as ChIP-seq are still young and maturing rapidly. Innovations like ChIP-exo and ChIP-nexus (Rhee and Pugh 2011; He et al. 2015), both of which couple the antibody pulldown to an exonuclease digestion step, offer substantial improvements in performance, especially with regards to resolution. Combining ChIP with more widely-available DNase-seq data, as I did in Chapter 3, allowed me similar resolution, but without knowledge of the identity of the bound factor at each footprint. By performing multiple ChIP-exo experiments for different factors, one would be able to fully characterize the architecture of cis-regulatory elements, including the identity of each of the bound factors. Such a scheme could overcome one of the major, significant challenges in understanding the evolution of the cis-regulatory sequences: simply identifying the causative nucleotides at single-base resolution.

There are also computational frontiers which demand exploration. For example, naïve approaches to detecting selection are likely to fail when applied to complex cis-regulatory sequence (Bullaughey 2011).   As a result, a new class of model is required which takes into account the particular evolutionary dynamics of regulatory elements (specifically, the tendency to turn over functional sequence). Novel algorithms are being developed in this vein, which rely upon similarity of short, word (k-mer) profiles between cis-regulatory elements (Gordon et al., in press; Kazemian et al. 2014). However, current approaches suffer problems of poor specificity, partially driven by the sheer size of the genome. There remains the possibility of much progress.

In the long run, many challenging problems in the evolution of gene regulation stem from the particular shape of the fitness landscape in gene regulation. In particular, gene regulatory networks are of such high dimensionality that there are multiple solutions to most regulatory problems (Bullaughey 2013; Dawid et al. 2010). This fact underlies the compensatory birth/death process discussed in Chapter 3, where multiple arrangements of binding sites can drive the same patterns of expression. It also informs our understanding of gene expression, in which multiple regulatory solutions can drive the same spatiotemporal patterns of expression, leading to turnover of enhancers over time (Kalay and Wittkopp 2010).

Despite the ubiquity of fitness landscapes like this, we lack good evolutionary models for understanding the dynamics of populations evolving upon them (Bullaughey 2011). A crucial step forward in that pursuit will be developing theories for how populations and species migrate between peaks on such landscapes, and how to infer selective regimes under such conditions. Finally, experimental approaches integrating new technologies must be developed to test those theories.

References

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of Ultraconserved Elements Yields Viable Mice. PLoS Biology 5: e234.

Akkers RC, van Heeringen SJ, Jacobi UG, Janssen-Megens EM, Françoijs K-J, Stunnenberg HG, Veenstra GJC. 2009. A Hierarchy of H3K4me3 and H3K27me3 Acquisition in Spatial Gene Regulation in Xenopus Embryos. Developmental Cell 17: 425–434.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biology 11: R106.

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–1152.

Anisimova M, ed. 2012. Evolutionary genomics: statistical and computational methods. Humana Press : Springer, New York.

Arnold CD, Gerlach D, Spies D, Matts JA, Sytnikova YA, Pagani M, Lau NC, Stark A. 2014. Quantitative genome-wide enhancer activity maps for five Drosophila species show functional enhancer conservation and turnover during cis-regulatory evolution. Nature Genetics 46: 685–692.

Arnold CD, Gerlach D, Stelzer C, Boryn LM, Rath M, Stark A. 2013. Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. Science 339: 1074–1077.

Arthur RK, Ma L, Slattery M, Spokony RF, Ostapenko A, Negre N, White KP. 2014. Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. Genome Research 24: 1115–1124.

Arthur RK, Ruvinsky I. 2011. Evidence That Purifying Selection Acts on Promoter Sequences. Genetics 189: 1121–1126.

Bailey TL. 2011. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics 27: 1653–1659.

Balhoff JP, Wray GA. 2005. Evolutionary analysis of the well characterized endo16 promoter reveals substantial variation within functional sites. Proceedings of the National Academy of Sciences 102: 8591–8596.

Bantignies F, Roure V, Comet I, Leblanc B, Schuettengruber B, Bonnet J, Tixier V, Mas A, Cavalli G. 2011. Polycomb-Dependent Regulatory Contacts between Distant Hox Loci in Drosophila. Cell 144: 214–226.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN, et al. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. PLoS Biology 5: e310.

Bergman CM. 2001. Analysis of Conserved Noncoding DNA in Drosophila Reveals Similar Constraints in Intergenic and Intronic Sequences. Genome Research 11: 1335–1345.

Bergman CM, Carlson JW, Celniker SE. 2005. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. Bioinformatics 21: 1747–1749.

Bernstein BE, Mikkelsen TS, Xie X, Kamal M, Huebert DJ, Cuff J, Fry B, Meissner A, Wernig M, Plath K, et al. 2006. A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells. Cell 125: 315–326.

Blanchette M. 2002. Discovery of Regulatory Elements by a Computational Method for Phylogenetic Footprinting. Genome Research 12: 739–748.

Blankenberg D, Taylor J, Nekrutenko A, The Galaxy Team. 2011. Making whole genome multiple alignments usable for biologists. Bioinformatics 27: 2426–2428.

Boffelli D. 2003. Phylogenetic Shadowing of Primate Sequences to Find Functional Regions of the Human Genome. Science 299: 1391–1394.

Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, Seringhaus MR, Wang LY, Gerstein M, Snyder M. 2007. Divergence of Transcription Factor Binding Sites Across Related Yeast Species. Science 317: 815–819.

Bradley RK, Li X-Y, Trapnell C, Davidson S, Pachter L, Chu HC, Tonkin LA, Biggin MD, Eisen MB. 2010. Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species ed. G.A. Wray. PLoS Biology 8: e1000343.

Breiman L. 2001. Random forests. Machine learning 45: 5–32.

Brown CT, Callan CG. 2004. Evolutionary comparisons suggest many novel cAMP response protein binding sites in Escherichia coli. Proceedings of the National Academy of Sciences 101: 2404–2409.

Bullaughey K. 2011. Changes in Selective Effects Over Time Facilitate Turnover of Enhancer Sequences. Genetics 187: 567–582.

Bullaughey K. 2013. Multidimensional adaptive evolution of a feed-forward network and the illusion of compensation: multidimensional adaptation. Evolution 67: 49–65.

Bush EC, Lahn BT. 2006. The Evolution of Word Composition in Metazoan Promoter Sequence. PLoS Computational Biology 2: e150.

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, Tanenbaum DM, White TJ, Sninsky JJ, Hernandez RD, et al. 2005. Natural selection on protein-coding genes in the human genome. Nature 437: 1153–1157.

Cain CE, Blekhman R, Marioni JC, Gilad Y. 2011. Gene Expression Differences Among Primates Are Associated With Changes in a Histone Epigenetic Modification. Genetics 187: 1225–1234.

Campos EI, Reinberg D. 2009. Histones: Annotating Chromatin. Annual Review of Genetics 43: 559–599.

Carroll SB. 2008. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell 134: 25–36.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila. Molecular Biology and Evolution 24: 2222–2234.

Chen F, Mackey AJ, Vermunt JK, Roos DS. 2007. Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes ed. C. Fairhead. PLoS ONE 2: e383.

Chen S, Ni X, Krinsky BH, Zhang YE, Vibranovski MD, White KP, Long M. 2012. Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. The EMBO Journal 31: 2798–2809.

Chen S, Zhang YE, Long M. 2010. New Genes in Drosophila Quickly Become Essential. Science 330: 1682–1685.

Chi P, Allis CD, Wang GG. 2010. Covalent histone modifications — miswritten, misinterpreted and mis-erased in human cancers. Nature Reviews Cancer 10: 457–469.

Conrad T, Akhtar A. 2012. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nature Reviews Genetics 13: 123–134.

Creevey CJ, Muller J, Doerks T, Thompson JD, Arendt D, Bork P. 2011. Identifying Single Copy Orthologs in Metazoa ed. A. Siepel. PLoS Computational Biology 7: e1002269.

Cuddapah S, Roh T-Y, Cui K, Jose CC, Fuller MT, Zhao K, Chen X. 2012. A Novel Human Polycomb Binding Site Acts As a Functional Polycomb Response Element in Drosophila ed. R. Jothi. PLoS ONE 7: e36365.

Dawid A, Kiviet DJ, Kogenaru M, de Vos M, Tans SJ. 2010. Multiple peaks and reciprocal sign epistasis in an empirically determined genotype-phenotype landscape. Chaos: An Interdisciplinary Journal of Nonlinear Science 20: 026105.

Delest A, Sexton T, Cavalli G. 2012. Polycomb: a paradigm for genome organization from one to three dimensions. Current Opinion in Cell Biology 24: 405–414.

Dermitzakis ET, Bergman CM, Clark AG. 2003. Tracing the Evolutionary History of Drosophila Regulatory Regions with Models that Identify Transcription Factor Binding Sites. Molecular Biology and Evolution 20: 703–714.

Dermitzakis ET, Clark AG. 2002. Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. Molecular Biology and Evolution 19: 1114–1121.

Djordjevic M. 2003. A Biophysical Approach to Transcription Factor Binding Site Discovery. Genome Research 13: 2381–2390.

Doniger SW, Fay JC. 2007. Frequent Gain and Loss of Functional Transcription Factor Binding Sites. PLoS Computational Biology 3: e99.

Dowell RD. 2010. Transcription factor binding variation in the evolution of gene regulation. Trends in Genetics 26: 468–475.

Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. Nature Reviews Genetics 8: 689–698.

Ellinger J, Kahl P, von der Gathen J, Heukamp LC, Gütgemann I, Walter B, Hofstädter F, Bastian PJ, von Ruecker A, Müller SC, et al. 2012. Global Histone H3K27 Methylation Levels are Different in Localized and Metastatic Prostate Cancer. Cancer Investigation 30: 92–97.

Emberly E, Rajewsky N, Siggia ED. 2003. Conservation of regulatory elements between two species of Drosophila. BMC bioinformatics 4: 57.

Feng S, Jacobsen SE. 2011. Epigenetic modifications in plants: an evolutionary perspective. Current Opinion in Plant Biology 14: 179–186.

Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, et al. 2010. Systematic Protein Location Mapping Reveals Five Principal Chromatin Types in Drosophila Cells. Cell 143: 212–224.

Fisher S, Grice EA, Vinton RM, Bessling SL, McCallion AS. 2006. Conservation of RET Regulatory Function from Human to Zebrafish Without Sequence Similarity. Science 312: 276–279.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531–1545.

Garces RG, Gillon W, Pai EF. 2007. Atomic model of human Rcd-1 reveals an armadillo -like-repeat protein with in vitro nucleic acid binding properties. Protein Science 16: 176–188.

Garland T, Dickerman AW, Janis CM, Jones JA. 1993. Phylogenetic Analysis of Covariance by Computer Simulation. Systematic Biology 42: 265–292.

Gnad F, Parsch J. 2006. Sebida: a database for the functional and evolutionary analysis of genes with sex-biased expression. Bioinformatics 22: 2577–2579.

Gu Z, Rifkin SA, White KP, Li W-H. 2004. Duplicate genes increase gene expression diversity within and between species. Nature Genetics 36: 577–579.

Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and Negative Selection on Noncoding DNA in Drosophila simulans. Molecular Biology and Evolution 25: 1825–1834.

Hahn MW. 2006. Detecting natural selection on cis-regulatory DNA. Genetica 129: 7–18.

Hastings PJ, Lupski JR, Rosenberg SM, Ira G. 2009. Mechanisms of change in gene copy number. Nature Reviews Genetics 10: 551–564.

Haygood R, Fedrigo O, Hanson B, Yokoyama K-D, Wray GA. 2007. Promoter regions of many neural- and nutrition-related genes have experienced positive selection during human evolution. Nature Genetics 39: 1140–1144.

He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nature Genetics 43: 414–420.

He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of in vivo transcription factor binding footprints. Nature Biotechnology 33: 395–401.

Hoffmann R, Valencia A. 2004. A gene network for navigating the literature. Nature Genetics 36: 664–664.

Huang W, Li L, Myers JR, Marth GT. 2012. ART: a next-generation sequencing read simulator. Bioinformatics 28: 593–594.

Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldon T. 2011. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Briefings in Bioinformatics 12: 442–448.

Jordan IK, McDonald JF. 1998. Interelement Selection in the Regulatory Region of the copia Retrotransposon. Journal of Molecular Evolution 47: 670–676.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA-based gene duplication: mechanistic and evolutionary insights. Nature Reviews Genetics 10: 19–31.

Kalay G, Wittkopp PJ. 2010. Nomadic Enhancers: Tissue-Specific cis-Regulatory Elements of yellow Have Divergent Genomic Positions among Drosophila Species ed. A. Kopp. PLoS Genetics 6: e1001222.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. Nature 468: 811–814.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. Nucleic Acids Research 42: D764–D770.

Kazemian M, Suryamohan K, Chen J-Y, Zhang Y, Samee MAH, Halfon MS, Sinha S. 2014. Evidence for Deep Regulatory Similarities in Early Developmental Programs across Highly Diverged Insects. Genome Biology and Evolution 6: 2301–2320.

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. 2014. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences 111: 6131–6138.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 471: 480–485.

Kim J, He X, Sinha S. 2009. Evolution of Regulatory Sequences in 12 Drosophila Species ed. T. Gojobori. PLoS Genetics 5: e1000330.

King DC. 2005. Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Research 15: 1051–1060.

King M, Wilson A. 1975. Evolution at two levels in humans and chimpanzees. Science 188: 107–116.

Kotelnikova EA, Makeev VJ, Gelfand MS. 2005. Evolution of transcription factor DNA binding sites. Gene 347: 255–263.

Kryukov GV. 2005. Small fitness effect of mutations in highly conserved non-coding regions. Human Molecular Genetics 14: 2221–2229.

Kutter C, Brown GD, Gonçalves Â, Wilson MD, Watt S, Brazma A, White RJ, Odom DT. 2011. Pol III binding in six mammals shows conservation among amino acid isotypes despite divergence among tRNA genes. Nature Genetics 43: 948–955.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research 22: 1813–1831.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

Lanzuolo C, Orlando V. 2012. Memories from the Polycomb Group Proteins. Annual Review of Genetics 46: 561–589.

Lenhard B, Sandelin A, Carninci P. 2012. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nature Reviews Genetics. http://www.nature.com/doifinder/10.1038/nrg3163 (Accessed May 21, 2015).

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li JJ, Huang H, Bickel PJ, Brenner SE. 2014. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. Genome Research 24: 1086–1101.

Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG. 2011. Widespread RNA and DNA Sequence Differences in the Human Transcriptome. Science 333: 53–58.

Ling G, Sugathan A, Mazor T, Fraenkel E, Waxman DJ. 2010. Unbiased, Genome-Wide In Vivo Mapping of Transcriptional Regulatory Elements Reveals Sex Differences in Chromatin Structure Associated with Sex-Specific Liver Gene Expression. Molecular and Cellular Biology 30: 5531–5544.

Loisel DA, Rockman MV, Wray GA, Altmann J, Alberts SC. 2006. Ancient polymorphism and functional variation in the primate MHC-DQA1 5′ cis-regulatory region. Proceedings of the National Academy of Sciences 103: 16331–16336.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nature Reviews Genetics 4: 865–875.

Loots GG, Ovcharenko I. 2010. Human Variation in Short Regions Predisposed to Deep Evolutionary Conservation. Molecular Biology and Evolution 27: 1279–1288.

Ludwig MZ, Bergman C, Patel NH, Kreitman M. 2000. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403: 564–566.

Lusk RW, Eisen MB. 2010. Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in Drosophila Enhancers ed. G.S. Barsh. PLoS Genetics 6: e1000829.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, Casillas S, Han Y, Magwire MM, Cridland JM, et al. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178.

Madrigal P, Krajewski P. 2012. Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. Frontiers in Genetics 3.

Maerkl SJ, Quake SR. 2007. A Systems Approach to Measuring the Binding Energy Landscapes of Transcription Factors. Science 315: 233–237.

Maniatis T, Goodbourn S, Fischer J. 1987. Regulation of inducible and tissue-specific gene expression. Science 236: 1237–1245.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nature Genetics 37: 413–417.

Mirny LA, Gelfand MS. 2002. Using Orthologous and Paralogous Proteins to Identify Specificity-determining Residues in Bacterial Transcription Factors. Journal of Molecular Biology 321: 7–20.

Molina N, van Nimwegen E. 2007. Universal patterns of purifying selection at noncoding positions in bacteria. Genome Research 18: 148–160.

Moses AM, Pollard DA, Nix DA, Iyer VN, Li X-Y, Biggin MD, Eisen MB. 2006. Large-Scale Turnover of Functional Transcription Factor Binding Sites in Drosophila. PLoS Computational Biology 2: e130.

Mustonen V, Kinney J, Callan CG, Lässig M. 2008. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. Proceedings of the National Academy of Sciences 105: 12376–12381.

Mustonen V, Lassig M. 2005. Evolutionary population genetics of promoters: Predicting binding sites and functional phylogenies. Proceedings of the National Academy of Sciences 102: 15936–15941.

Nègre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, Kheradpour P, Eaton ML, Loriaux P, Sealfon R, et al. 2011. A cis-regulatory map of the Drosophila genome. Nature 471: 527–531.

Nei M, Xu P, Glazko G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. Proceedings of the National Academy of Sciences 98: 2497–2502.

Ni X, Zhang YE, Nègre N, Chen S, Long M, White KP. 2012. Adaptive Evolution and the Birth of CTCF Binding Sites in the Drosophila Genome ed. H.S. Malik. PLoS Biology 10: e1001420.

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating Divergence Dates and Substitution Rates in the Drosophila Phylogeny. Molecular Biology and Evolution 29: 3459–3473.

Okulski H, Druck B, Bhalerao S, Ringrose L. 2011. Quantitative analysis of polycomb response elements (PREs) at identical genomic locations distinguishes contributions of PRE sequence and genomic environment. Epigenetics & Chromatin 4: 1–16.

Papp B, Muller J. 2006. Histone trimethylation and the maintenance of transcriptional ON and OFF states by trxG and PcG proteins. Genes & Development 20: 2041–2054.

Paris M, Kaplan T, Li XY, Villalta JE, Lott SE, Eisen MB. 2013. Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression ed. P. Wittkopp. PLoS Genetics 9: e1003748.

Park PJ. 2009. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10: 669–680.

Parsch J, Russell JA, Beerman I, Hartl DL, Stephan W. 2000. Deletion of a conserved regulatory element in the Drosophila Adh gene leads to increased alcohol dehydrogenase activity but also delays development. Genetics 156: 219–227.

Patwardhan RP, Lee C, Litvin O, Young DL, Pe'er D, Shendure J. 2009. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nature Biotechnology 27: 1173–1175.

Piper J, Elze MC, Cauchy P, Cockerill PN, Bonifer C, Ott S. 2013. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Research 41: e201–e201.

Pohl A, Beato M. 2014. bwtool: a tool for bigWig files. Bioinformatics 30: 1618–1619.

Quezada-Díaz JE, Muliyil T, Río J, Betrán E. 2010. Drcd-1 related: a positively selected spermatogenesis retrogene in Drosophila. Genetica 138: 925–937.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Raijman D, Shamir R, Tanay A. 2008. Evolution and Selection in Yeast Promoters: Analyzing the Combined Effect of Diverse Transcription Factor Binding Sites. PLoS Computational Biology 4: e7.

Ranz JM. 2003. Sex-Dependent Gene Expression and Evolution of the Drosophila Transcriptome. Science 300: 1742–1745.

Rhee HS, Pugh BF. 2011. Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell 147: 1408–1419.

Rifkin SA, Houle D, Kim J, White KP. 2005. A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature 438: 220–223.

Rifkin SA, Kim J, White KP. 2003. Evolution of gene expression in the Drosophila melanogaster subgroup. Nature Genetics 33: 138–144.

Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, Müller M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics 12: 77.

Rockman MV, Hahn MW, Soranzo N, Goldstein DB, Wray GA. 2003. Positive Selection on a Human-Specific Transcription Factor Binding Site Regulating IL4 Expression. Current Biology 13: 2118–2123.

Rockman MV, Hahn MW, Soranzo N, Zimprich F, Goldstein DB, Wray GA. 2005. Ancient and Recent Positive Selection Transformed Opioid cis-Regulation in Humans. PLoS Biology 3: e387.

Romero IG, Ruvinsky I, Gilad Y. 2012. Comparative studies of gene expression and the evolution of gene regulation. Nature Reviews Genetics 13: 505–516.

Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof A, Mellone BG, Malik HS. 2013. Stepwise Evolution of Essential Centromere Function in a Drosophila Neogene. Science 340: 1211–1214.

Ruvinsky I, Ruvkun G. 2003. Functional tests of enhancer conservation between distantly related species. Development 130: 5133–5142.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O’Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW, et al. 2008. Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. Genetics 179: 1601–1655.

Schluter D, Price T, Mooers AO, Ludwig D. 1997. Likelihood of Ancestor States in Adaptive Radiation. Evolution 51: 1699.

Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages. Cell 148: 335–348.

Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, Marshall A, Kutter C, Watt S, Martinez-Jimenez CP, Mackay S, et al. 2010. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. Science 328: 1036–1040.

Segal E, Raveh-Sadka T, Schroeder M, Unnerstall U, Gaul U. 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451: 535–540.

Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. Cell 148: 458–472.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. Nature 457: 818–823.

Shultzaberger RK, Malashock DS, Kirsch JF, Eisen MB. 2010. The Fitness Landscapes of cis-Acting Binding Sites in Different Promoter and Environmental Contexts ed. D.S. Guttman. PLoS Genetics 6: e1001042.

Siepel A, Arbiza L. 2014. Cis-regulatory elements and human evolution. Current Opinion in Genetics & Development 29: 81–89.

Sokal RR, Rohlf FJ. 1995. Biometry: the principles and practice of statistics in biological research. 3rd ed. W.H. Freeman, New York.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. Cell 154: 530–540.

Stern DL, Orgogozo V. 2009. Is Genetic Evolution Predictable? Science 323: 746–751.

Stern DL, Orgogozo V. 2008. The loci of evolution: how predictable is genetic evolution? Evolution 62: 2155–2177.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100: 9440–9445.

Tabor HK, Risch NJ, Myers RM. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics 3: 391–397.

Tamura K. 2003. Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. Molecular Biology and Evolution 21: 36–44.

Tanay A, Sharan R, Kupiec M, Shamir R. 2004. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proceedings of the National Academy of Sciences 101: 2981–2986.

Thomas S, Li X-Y, Sabo PJ, Sandstrom R, Thurman RE, Canfield TK, Giste E, Fisher W, Hammonds A, Celniker SE, et al. 2011. Dynamic reprogramming of chromatin accessibility during Drosophila embryo development. Genome Biol 12: R43.

Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, Lewis SE, Richards S, Ashburner M, Hartenstein V, Celniker SE, et al. 2002. Systematic determination of patterns of gene expression during Drosophila embryogenesis. Genome Biol 3: 0081–0088.

Torgerson DG, Boyko AR, Hernandez RD, Indap A, Hu X, White TJ, Sninsky JJ, Cargill M, Adams MD, Bustamante CD, et al. 2009. Evolutionary Processes Acting on Candidate cis-Regulatory Regions in Humans Inferred from Patterns of Polymorphism and Divergence ed. D.J. Begun. PLoS Genetics 5: e1000592.

Ulirsch JC, Lacy JN, An X, Mohandas N, Mikkelsen TS, Sankaran VG. 2014. Altered Chromatin Occupancy of Master Regulators Underlies Evolutionary Divergence in the Transcriptional Landscape of Erythroid Differentiation ed. M. Snyder. PLoS Genetics 10: e1004890.

Vasanthi D, Mishra RK. 2008. Epigenetic regulation of genes during development: A conserved theme from flies to mammals. Journal of Genetics and Genomics 35: 413–429.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. Nature 499: 332–335.

Villar D, Berthelot C, Aldridge S, Rayner TF, Lukk M, Pignatelli M, Park TJ, Deaville R, Erichsen JT, Jasinska AJ, et al. 2015. Enhancer Evolution across 20 Mammalian Species. Cell 160: 554–566.

Visel A, Prabhakar S, Akiyama JA, Shoukry M, Lewis KD, Holt A, Plajzer-Frick I, Afzal V, Rubin EM, Pennacchio LA. 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nature Genetics 40: 158–160.

Weindl J, Hanus P, Dawy Z, Zech J, Hagenauer J, Mueller JC. 2007. Modeling DNA-binding of Escherichia coli 70 exhibits a characteristic energy landscape around strong promoters. Nucleic Acids Research 35: 7003–7010.

Wilson MD, Barbosa-Morais NL, Schmidt D, Conboy CM, Vanes L, Tybulewicz VLJ, Fisher EMC, Tavare S, Odom DT. 2008. Species-Specific Transcription in Mice Carrying Human Chromosome 21. Science 322: 434–438.

Wong WSW. 2004. Detecting Selection in Noncoding Regions of Nucleotide Sequences. Genetics 167: 949–958.

Wozniak CE, Hughes KT. 2008. Genetic Dissection of the Consensus Sequence for the Class 2 and Class 3 Flagellar Promoters. Journal of Molecular Biology 379: 936–952.

Wray GA. 2003. The Evolution of Transcriptional Regulation in Eukaryotes. Molecular Biology and Evolution 20: 1377–1419.

Xiao S, Xie D, Cao X, Yu P, Xing X, Chen C-C, Musselman M, Xie M, West FD, Lewin HA, et al. 2012. Comparative Epigenomic Annotation of Regulatory DNA. Cell 149: 1381–1392.

Yuan Y, Norris C, Xu Y, Tsui K-W, Ji Y, Liang H. 2012. BM-Map: an efficient software package for accurately allocating multireads of RNA-sequencing data. BMC genomics 13: S9.

Zang C, Schones DE, Zeng C, Cui K, Zhao K, Peng W. 2009. A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics 25: 1952–1958.

Zhang Z, Gerstein M. 2003. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. Journal of biology 2: 11.

Zhao Y, Granas D, Stormo GD. 2009. Inferring Binding Energies from Selected Binding Sites ed. C. Workman. PLoS Computational Biology 5: e1000590.

Zhou VW, Goren A, Bernstein BE. 2011. Charting histone modifications and the functional organization of mammalian genomes. Nature Reviews Genetics 12: 7–18.