The Evolution of a Transcription Factor: Divergence in DNA Binding Behavior of the Sex-Determination Gene *hermaphrodite* in the Genus *Drosophila*

by

Colin Walsh Brown

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Molecular and Cell Biology

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Michael B. Eisen, Chair
Professor Michael R. Botchan
Professor Steven E. Brenner
Professor Thomas W. Cline

Fall 2010

The Evolution of a Transcription Factor: Divergence in DNA Binding Behavior of the Sex-Determination Gene *hermaphrodite* in the Genus *Drosophila*

Abstract

The Evolution of a Transcription Factor: Divergence in DNA Binding Behavior of the Sex-Determination Gene *hermaphrodite* in the Genus *Drosophila*

by

Colin Walsh Brown
Doctor of Philosophy in Molecular and Cell Biology

University of California, Berkeley

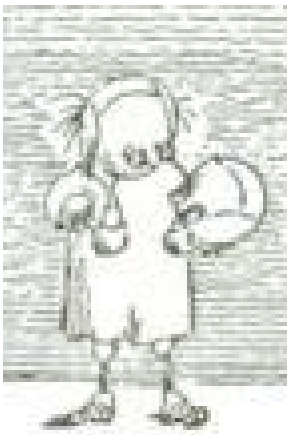
Professor Michael B. Eisen, Chair

Changes in transcriptional regulatory networks are thought to underlie most morphological change observed across taxa, but the general principles of how such networks change and why remain unknown. While many studies have focused on evolutionary changes in *cis*-acting components of transcriptional networks (enhancers and transcription factor binding sites), changes in the transcription factor proteins which bind these sites have been mostly overlooked. In this thesis I identify the putative *Drosophila* transcrtiption factor *hermaphrodite* (*her*) as having a highly divergent DNA-binding domain, and examine its DNA-binding profile in two fruitfly species, *Drosophila melanogaster* and *Drosophila pseudoobscura* . I find evidence for a large-scale divergence in the distribution of HER binding, as well as a possible difference in DNA-binding preference between these two species. This establishes *Drosophila her* as an excellent system for the study of regulatory evolution through changes in transcription factors.

Contents

List of Figures

List of Tables



That evening, something emerged at the edge of the dump, not far from the puddle which had by now dried up, and this something, a creature of pure accident, was Mymosh the Selfbegotten, who had neither mother nor father, but was son unto himself, for his father was Coincidence, and his Mother –Entropy. And Mymosh rose up from the garbage dump, totally oblivious of the fact that he had about one chance in a hundred billion jillion raised to the zillionth power of ever existing, and he took a step, and walked until he came to the next puddle, which had not as yet dried up, so that, kneeling over it, he could easily see himself.

- Stanislaw Lem "Mymosh the Selfbegotten"

Acknowledgments

I would like to thank my fiancee Bonnie, my parents, and dear Marty Brown-Real for their endless support. I would also like to thank Mike Eisen, my thesis committee members Mike Botchan, Tom Cline, and Steven Brenner, and all of the members of the Eisen Lab past and present for help in designing, executing, and interpreting the experiments.

Specifically, I would like to thank Mike Eisen, Dan Pollard, and Venky Iyer for generating the seqeunce and orthology data used in Chapter 2. Mike Eisen, Xiao-Yong Li, Mark Stapleton, Malik Francis, Melissa Harrison, Leath Tonkin, Mathilde Paris, Tommy Kaplan, and Jacqueline Villalta for help with the work in Chapters 3 and 4.

Chapter 1

Introduction

1.1 The Rise of Regulatory Thinking in Studies of Evolution

The seemingly endless variety of the living world is both a wonder and an everyday fact — an infinity that we confront everywhere we look. The desire to understand this variety runs deep in humankind, and explanations for it run far back in history, from the Biblical Book of Genesis to Aristotle and the "Great Chain of Being." The advent of *The Origin of Species* [1] and evolutionary thinking provided a framework for thinking about the problem that was both appealing in its apparent simplicity and staggering in its explanative power — that the morphological diversity that we observe is the result of a historical process of selection acting on heritable phenotypic diversity within a population. However, this explanation begs an almost equally problematic set of questions — what are the origins of variation, and how is it inherited? What are the mechanisms of selection?

Subsequent years of biological inquiry began to provide some answers for these questions, but just as often opened up entirely new and more challenging problems. The re-discovery of Mendel's work in the early part of the century provided an explanation for the mechanisms of inheritance, and led Thomas Hunt Morgan to propose in *A Critique of the Theory of Evolution* [2] that genetic mutations could act as a source of novel, heritable traits, which when combined with the force of selection could act to generate diversity. The work of population geneticists in the following decades extended this work to provide more detailed mechanistic explanations for the relationship between traits and selective forces, and how nonselective forces such as drift could lead to fixation of traits. However, "genes" were still abstract units of inheritance, and the mechanistic relationship between mutations, genes, and phenotype was poorly understood.

With the rise of molecular biology, however, a clearer picture of the mechanisms governing variation and the generation of phenotypes began to emerge. Two discoveries in the 1960s are of particular importance. The first was the recognition (summarized by Jacob and Monod [3]) that regulatory and structural mutations, and therefore regulatory and structural elements themselves, are separable, and that regulatory mutations are often defined by pleiotropic effects on multiple parts of a metabolic pathway. The second was the realization made by Zuckerkandl and Pauling [4] that biological macromolecules could carry phylogenetic information, and therefore that changes in these "semantophoretic"

macromolecules could explain differences among species, at least at the level of physiology. Zuckerkandl and Pauling also point out that these two observations imply that changes in regulatory genes or operator sequences can have important effects on other genes without altering their sequence. These discoveries implied for the first time a direct relationship between changes in physical entities (protein and DNA molecules) and changes in organismal phenotype

By the 1970s, evidence was accumulating that regulatory mutations may have important conseqeunces for organismal evolution, although the nature of these mutations, and of eukaryotic gene regulation itself, were still being debated (e.g. [5]) . In a classic study in 1975, King and Wilson [6] surveyed available evidence regarding the rate of amino acid changes in a variety of proteins between chimpanzees and humans, and concluded that the observed rates of change were insufficient to explain the drastic differences between the two species. The suggestion was then that mutations affecting the expression of these genes, rather than mutations affecting the structure or function of the genes' protein products, were the major driving force behind the divergence of these two species. Although the question of the relative importance of structural vs. regulatory mutations is still a matter of debate (e.g. [7]), the idea that mutations affecting regulatory systems are the source of much of the evolutionary variation that we see has come to be extremely influential, particularly with the increasing availability of complete genome sequences.

Even as far back as King and Wilson, however, there was recognition that "regulatory mutations" could occur in two different types of elements: the *cis*-acting operators and promoters, which exist proximal to the regulated gene, and the genes encoding repressor and activator proteins, which act on these elements in *trans*. From the beginning, it was clear that these two types of mutations could have very different consequences for the regulatory systems they participate in, and that the mutational and selective processes acting on them are likely to be quite different. This then begs the question of which type of regulatory mutation is more important for the process of evolution. In one sense, the answer seems obvious; given the amazing variety of evolutionary mechanisms, it is likely that both *cis*- and *trans*-regulatory mutations have played a role at some point during evolutionary history in generating diversity. However, the debate about *relative* importance of the two categories of regulatory evolution still rages, and has, if anything, become more heated in recent years. As an increasing number of studies have provided beautifully detailed examples of morphological evolution via changes in *cis*-regulatory sequences [8–11], some researchers have argued, based on both these experiments as well as theoretical reasoning, that *cis*-regulatory mutations are not only generally more common, but also are responsible for many of the most fundamental changes that have taken place over the course of evolution [8,12]. Changes in *trans*-regulatory elements, on the other hand, are generally assumed to be rare, and unlikely to play an important part in most examples of evolutionary change.

This assumption is based upon the argument that changes in a transcription factor will result in excessive pleiotropic consequences, and will therefore be subject to strong purifying selection in almost all cases. Although recently a number of mechanisms have been proposed to explain how transcription factors might overcome negative pleiotropy [7,13,14], there are very few convincing examples of these processes in action. In particular, there has never been a detailed analysis of evolutionary divergence in a metazoan transcription factor

in a context where target sets can be compared between divergent species. Such an analysis could allow the examination of not only the protein-level changes in the transcription factor, but also the changes in target sites that it binds to; in this way, one could gain insight into the balance of selective penalties and advantages that allowed a divergent transcription factor to solve the "pleiotropy problem." My aim in this thesis is to establish such a case, first by conducting an unbiased search for divergent transcription factors in a well-studied group of closely related species, the 12 sequenced fly species of the genus *Drosophila*, and then using well-established biochemical methods for studying the binding of this factor in species where there is evidence for protein divergence.

1.2 Gene Regulatory Networks in Eukaryotes

Although many modes of gene regulation exist in eukaryotic cells, including regulation of mRNA production, mRNA splicing, mRNA stability, translational regulation, and post-translational modification, the most common and best understood is the process of transcriptional regulation. The complex nature of eukaryotic transcriptional regulation means that there are many types of elements that could be exploited to alter gene expression in an evolutionary context. The fundamental *cis*-acting element of transcription is the core promoter, usually defined as the minimal DNA sequence element necessary for *in vitro* initiation of transcription [15–18]. Although there is not a single well-defined sequence feature that defines the core promoter, it often consists of one or more instances of several common sequence elements, including the TATA box, the Inr element, and the DPE element. These elements generally direct binding of the TBP (TATA binding protein)-containing TFIID complex through direct or indirect sequence interactions. TFIID then guide assembly of the pre-initiation complex consisting of RNA Polymerase II and a variety of cofactors. Although the basal transcription machinery is often thought of as a stereotyped set of factors common to most genes and promoter types, it is now becoming clear that its composition can vary considerably among cell types, and can have a role in determining specific gene expression [18, 19]

However, the core promoter and the basal machinery alone are, in general, insufficient to initiate transcription. The spatially and temporally specific activation of transcription is instead carried out by a different type of *cis*-acting element, the promoter or enhancer region, which generally consists of a clustered array of short DNA binding sites for sequence-specific transcription factors [11, 20]. The binding of transcription factors to these sequences is "read out" in a still poorly understood way to either activate or repress transcription from the basal promoter, which may be located thousands of base pairs away [11, 18, 20]. Several features of enhancers are relevant to the topic of regulatory evolution. First, they are generally considered to be modular, that is, a single enhancer can act as discrete unit which drives a specific pattern of expression, even outside of its native context. The overall expression pattern of a gene is then the sum of the effects of the individual enhancer elements [11,21]. This means that the effects of mutations in enhancers will be localized only to the expression domain of the enhancer. In addition, enhancers display some degree of robustness to mutations — since transcription factor binding sites are for the most part short and degenerate, it is possible for mutations to create new binding sites

within enhancers that can then compensate for mutations occurring in other binding sites, a process known as turnover [22].

The transcription factors that bind to enhancers are highly modular proteins, most of which belong to a remarkably small group of domain families [15, 23, 24]. Most transcription factors share two common structural features: a DNA-binding domain, which is responsible for determining the DNA target specificity, and a transcriptional effector domain, which can interact with transcriptional coregulator proteins, other transcription factors, or directly with elements of the basal transcription machinery to exert control over transcriptional initiation [15,24]. Transcription factors usually recognize a family of related target sequences, which are bound with different affinities based on their similarity to a preferred set of sequence features that is determined by the structure of the transcription factor's DNA-binding domain [23, 25]. DNA-binding domains are generally alpha-helical, and most often recognize DNA bases through interactions between individual residues at the binding interface and the chemical features of the nucleotide exposed in the major groove of the DNA molecule, although many exceptions to this exist (e.g. [23]). The specific mechanisms of DNA recognition vary widely across different domain types. Although in several cases models exist for predicting binding behavior based on sequence alone [26–29], it is not in general possible to determine what the binding specificity of a given factor is *a priori*, or even what the effects of a given substitution within a binding domain of known specificity will be.

Finally, the division of gene regulatory networks into *cis*- and *trans*- acting components is in some sense a simplification. With respect to a single transcription start, enhancer sequences can be said to act in *cis* and transcription factor proteins in *trans*; however, regulatory networks are by their nature hierarchical, and a gene that is regulated by a particular enhancer may itself code for a transcription factor protein with a large number of downstream targets [30]. Thus, while mutations in an enhancer sequence are in some sense restricted to a single gene and a single regulatory domain, they may have a wide range of indirect effects through that gene's downstream targets.

1.3 The *cis*-regulatory paradigm

The past several decades of research into the regulatory and developmental mechanisms of morphological evolution have yielded a growing number of examples of evolutionary change occurring through alterations in gene regulation; many of these involved changes in *cis*-acting components of regulatory pathways. Although the number of such examples is still relatively limited, a school of thinking has developed in the field in which *cis* regulatory changes are viewed as the primary mechanism underlying most regulatory changes, and most evolutionary changes in morphology in general [8, 9, 12]. The reasoning for this viewpoint involves both a number of theoretical arguments and extensions from empirical evidence.

The theoretical arguments for the primacy of *cis* regulatory changes in evolution mostly follow from what we now know about the structure of developmental enhancers and regulatory networks. The most common of these arguments is that *cis* regulatory changes are by their nature less subject to the selective constraints resulting from pleiotropic

effects. Pleiotropy is thought to be one of the primary forces constraining many aspects of evolutionary change, and population genetic models predict that even a modest amount of pleiotropy can result in strong purifying selection [31]. By definition, a *cis* -regulatory element only effects the transcription of one, or at most a handful, of genes in its immediate vicinity [11,18,20]. Although the effects become much more complicated if the *cis*-regulatory region controls a gene that itself is pleiotropic, for cases of *cis* regulatory regions controlling structural genes, mutations in *cis* will have a limited effect. In addition, the modularity of enhancers in terms of both their structure on the DNA molecule and the expression domain they control can limit pleiotropy [8–12]. Since enhancers are discrete units and control a discrete expression pattern, the effects of mutations can in some sense be better "targeted" so as to alter only a small portion of a gene's expression pattern while avoid changing the most critical expression domains.

Another set of arguments in favor of the primacy of *cis*-regulatory evolution is based on the idea that *cis* regulatory elements respond to mutation and selection differently than coding sequences. One such argument put forth by Wray [10] is that the effects of individual mutations in *cis*-regulatory regions are likely to be much smaller, and consequently the range of available phenotypic states that can be sampled is more continuous than corresponding changes in coding regions, which are assumed to have much more drastic and discrete effects. This is may be particularly advantageous for mutations affecting continuous traits, or when the selective forces change over time, where the small effect size could allow for "fine tuning" of phenotype. Another argument, also proposed by Wray [10], is that selection acts more efficiently on changes in *cis*-regulatory elements. This is based on the observation that *cis*-regulatory mutations are by their nature codominant, and therefore the phenotypic effects are visible to selection when the mutation is heterozygous.

The empirical evidence supporting the primacy of *cis*-regulatory sequences in evolution consists mainly of two classes of observations. The first is in the same vein as King and Wilson, in which it is observed that the divergence in protein sequences (transcription factors in this case) is insufficient to explain the amount of regulatory and morphological diversity seen in extant species. The classic example of this is the Hox genes, which are conserved so thoroughly that, at least in some cases, they can be substituted for one another across vast phylogenetic distances with little loss of function [32]. All regulatory evolution that involves these proteins, then, must be the result of changes in *cis*-acting sequences either downstream or upstream of the invariant "toolkit" transcription factor [8, 12]. The second line of evidence in favor of *cis* regulation consists of a rapidly expanding collection of detailed case studies that examine the molecular basis of a visibly differing phenotype across closely related species. The most complete of these studies have been conducted in flies, where *cis*-regulatory mutations have been found to underlie differences among species in abdominal pigmentation [33–36], pheromone expression [37], wing pigmentation [38, 39] and larval bristle patterns [40,41]. Similar cases have also been identified in sticklebacks [42] and humans (e.g. [43–45]). The conclusion is that these cases form a representative sample of regulatory evolution in general, implying that *cis*-regulatory mutations underlie traits varying across much broader evolutionary timescales.

1.4 Evolution of Gene Regulation through changes in Transcription Factor Proteins

One of the primary pieces of evidence used to support the *cis*-regulatory paradigm of developmental evolution is that the molecular "parts list", that is, the proteins that participate in the central processes of development, are common to most animal phyla. The rapidly increasing number of sequenced animal genomes has borne this out to a remarkable degree, and it is now becoming clear that a large fraction of the transcription factors and signaling molecules that make up the developmental "toolkit" used by most metazoans were present in the last common ancestor of bilaterians [46–48] and possibly even before the emergence of all metazoan phyla [49,50]. Many of these genes show a remarkable degree of conservation with respect to both primary sequence and their position in the ontogenic cascade, raising the intriguing possibility that some core developmental modules present in the last bilaterian ancestor are still in use across the kingdom today.

However, the focus on these extraordinarily conserved components masks a second, equally interesting point that has become apparent from large-scale sequencing projects. This is that the "toolkit" components make up only a fraction of the total complement of regulatory genes present in the genome, and that many transcription factors and other signaling molecules are not conserved across phyla. Although most major transcription factor families had already been established before the diversification of metazoans [50, 51], the distribution of the member proteins from these families across eukaryotic clades is far from uniform. Table 1.1 shows the distribution of member proteins from a variety of major transcription factor binding domain families within sequenced eukaryotic genomes. Although many of these proteins represent non-toolkit orthologs shared among these genomes, it is clear that the transcription factor complement is quite variable, even within vertebrates and mammals. Carroll [46] lists a core developmental toolkit of 36 TFs for *D. melanogaster*. Even if this number has increased by an order of magnitude for species with more complicated developmental programs, there are still hundreds of lineage-specific transcription factors with unknown or poorly characterized function that could represent a largely unexplored source of evolutionary variation. In addition, recent studies have shown that the size of several transcription factor families shows a high degree of correlation with both genome size [52] and organismal complexity (as measured by number of cell types) [53], and many major cladogenic events, such as the Cambrian explosion [46, 47]. Also, the diversification of vertebrates [54–56], bony fishes [57], and mammals [58] has been accompanied by major transcription factor protein family expansions. This implies that major changes in body plan might be most easily accomplished by the large-scale rearrangements in regulatory networks resulting from the appearance of novel transcription factor proteins [46,47,57,59].

In addition to these major macroevolutionary events, there is now evidence from a variety of sources that change in transcription factor proteins is an ongoing process, with potentially important consequences for regulatory evolution. In addition to a number of relatively recent duplication and family expansion events (reviewed in [59]; see also [60,61] and Section 1.4.2 below), a number of studies have also found evidence for increased rates of evolution in transcription factor orthologs. Bustamante *et al.* [62] examined sequences of more than 11,000 human genes in 38 individuals, and found evidence of diversifying

selection (as measured by the ratio of synonymous to nonsynonymous amino acid substitutions) in transcription factors as a group, as well as within individual TF families such as the KRAB-box zinc fingers, nuclear hormone receptors, and homeodomain-containing proteins. Human transcription factors also show an enrichment of derived alleles at positions conserved in other mammals [63], some of which correspond to positions that may have functional consequences for DNA-binding, and human TFs also show evidence of diversifying selection for expression pattern [64]. Few systematic studies of transcription factor divergence have examined non-human clades, although there is strong evidence that diversification of bacterial transcriptional systems is in part driven by divergence of transcription factor proteins [65].

This genomic evidence suggests that differences in the transcription factor complement of genomes and within TF proteins themselves have been important evolutionary processes, both in the deep divergence of development and form and in more recent evolution of novel phenotypes. It seems clear, therefore, that an understanding of the mechanisms of change within transcription factors should be a key part of our overall picture of developmental and regulatory variation. However, in spite of the recent drive towards understanding the evolution of development in terms of changes in genetic regulatory networks, there has been relatively little attention paid to the role that *trans*-acting factors can play in this process. While much progress has been made in understanding transcription factor evolution at the level of protein sequence, there were, until very recently, almost no case studies of transcription factor evolution approaching the level of detail of the examples of *cis*-regulatory evolution outlined above. While there is still very little known about how sequence changes in transcription factors are tied to transcriptional network rearrangements and ultimately to changes in phenotype, a number of recent studies have begun to shed light on these issues. In addition, while the theoretical arguments supporting the primacy of *cis*-regulatory change (see 1.3) have continued to be influential in spite of the growing evidence for the importance of *trans*-change, new perspectives are beginning to emerge in the literature. The following sections outline both the theoretical arguments for the possibility and importance of *trans*-regulatory evolution and the growing body of empirical evidence supporting these arguments.

1.4.1 Transcriptional Network Structure and *trans*-Regulatory Evolution

The short, degenerate nature of transcription factor binding sites means that most TF proteins bind to a relatively large set of targets, many of which are likely to be of functional importance to the organism [66, 67]. The fact that these binding sites can in principle regulate genes active in a variety of different pathways has led to the expectation that most transcription factors will display a high degree of pleiotropy. Indeed, induced mutations in transcription factors tend to have phenotypic consequences distributed across multiple characters, and often lead to profound disruptions in development and homeostasis [31]. Natural mutations that occur in TF coding regions are therefore also expected to have deleterious effects for most genes that the TF regulates, even if the mutation results in adaptive changes at some regulated loci [8–12]. Rates of change ("evolvability") in transcription factor proteins are therefore expected to be generally low, and this has led some in the field to conclude that transcription factor evolution is of relatively small importance

compared to *cis*-regulatory changes as a mechanism for generating evolutionary diversity.

Nonetheless, the patterns of transcription factor evolution discussed above and the case studies discussed in sections 1.4.4 and 1.4.5 indicate that TFs can and do evolve. Given the expectations about pleiotropy, then, how are changes in TF proteins able to occur at all? There is little in the way of experimental evidence that could give a definitive answer (indeed, providing such evidence is one of the goals of this thesis), although a number of recent theoretical and modeling studies can give some hints about possible mechanisms. A primary point is that the pleiotropic effects of a mutation in a given TF are dependent on the network context in which the TF resides, and there is growing evidence that biological networks possess a number of features that could potentially mitigate the negative consequences of highly pleiotropic TF mutations. As more complete transcriptional networks have been mapped, for example, it has become apparent that many of them possess a highly hierarchical structure — that is, they tend to be stratified into top-level regulators that initiate transcriptional responses, "core" TFs and signaling molecules that receive inputs from top-level regulators and show many interconnections to other regulators, and bottom-level TFs that control batteries of structural genes [59, 68, 69]. In developmental systems, these levels have temporal implications, with the top-level regulators ("kernel" or "initial" TFs) specifying segment or organ identities early in the developmental progression [47, 68, 70, 71] core regulators acting as "plugin" signaling modules that are often shared among a large number of regulatory cascades, and bottom-level TFs acting as terminal differentiation genes, integrating inputs from the higher levels and directing cell-type-specific structural gene expression [47, 68].

This structural configuration has two main consequences for TF evolution. First, the severity of phenotypic effect from a TF mutation is expected to be related more to its position in the hierarchical network than the simple number of regulatory interactions it participates in (i.e. number of targets / binding sites) [72]. Mutations in top-level regulators will have effects that propagate through lower levels of the network, potentially resulting in effects on a large number of phenotypic characters, the classical definition of pleiotropy. Mutations in lower-level or terminal regulators, on the other hand, will show a much lower degree of propagation, and may only influence a handful of related characters. Indeed, many of the most commonly cited cases of TFs conserved across long evolutionary timescales are those involved in top-level specification of developmental programs [69–71, 73]. Low-level regulators, then, may represent a category of TFs which are particularly prone to protein changes as well as *cis*-regulatory changes [74].

On the other hand, the hierarchical nature of transcriptional networks means that TFs mutations could also provide a pathway to large-scale network rearrangements. Mutations in top-level TFs could result in large-scale, but coordinated, changes that could drive major shifts in phenotype [13, 59, 75]. Although mutations of this type are expected to be generally deleterious, and therefore will rarely rise to fixation, two additional features of transcriptional networks mean that they may occur with more frequency than we might intuitively expect. First, most biological networks tend be modular, often consisting of clusters of highly connected regulators and structural genes showing few connections between clusters [76]. Consequently, inputs from a high- or core-level module to other modules will generally consist of a few connections, the loss or gain of which can be accomplished

without disrupting internal functions of the target module [68, 76]. Second, it is possible that at least some transcriptional regulatory networks may be structured in such a way as to be robust to even large-scale changes resulting from the gain or loss of a highly connected regulator. Two recent modeling studies [77, 78] found that in regulatory networks of sufficient complexity, mutations in central genes are often neutral with respect to the ultimate regulatory phenotype. Even when mutations that significantly change phenotype occur, the networks can be structured by selective [77] or neutral [78] factors in such a way that deleterious effects are minimized and the phenotypic change tends to shift the network towards the selective optimum. It may be the case, then, that even highly connected TFs could undergo evolutionary changes with minimal effects on other parts of the network (i.e. other modules) and with favorable phenotypic consequences.

An interesting observation related to the role of network architecture in TF evolution is the apparent existence of adaptive "hotspots", genes that appear to be frequent targets of evolutionary change in networks [79, 80], resulting in parallel evolution of phenotypes or even different phenotypes in the same traits [81, 82]. While one class of hotspots appears to consist of structural genes that act in terminal specifiers (e.g. the pigmentation gene Mc1r [83, 83, 84] and the skin appendage gene *Eda / Edar* [85–88]), another class has been identified that may correspond to so-called "input-output" switch genes [68, 79, 80]. An example is the *ovo/shavenbaby* locus in *Drosophila*, which specifies trichome formation on the larval cuticle [89]. Interspecific differences in trichome patterning between *Drosophila melanogaster* and *Drosophila sechellia* have been shown to result from three independent changes in the regulatory region of *svb/ovo*, which has been interpreted by Stern and coworkers [79, 80, 89] as evidence that the *svb/ovo* locus is particularly prone to accumulating evolutionarily relevant changes. Interestingly, *svb/ovo* resides at a bottleneck in the signaling cascade that governs trichome pattern, integrating patterning gene inputs with its *cis*-regulatory regions and in turn regulating a large set of downstream terminal differentiation genes specifying trichome development. Stern *et al* then hypothesize that this position in the network is responsible for the 'hotspot' behavior exhibited by *svb/ovo* since the expression of the gene is sufficient to direct trichome development without interfering in the function of other pathways.

Although *svb/ovo* and a number of similar cases [41,42,90,91] involve *cis*-regulatory changes, a similar idea may apply to transcription factors evolving via coding sequence changes as well. If one of these input-output genes were to undergo a segmental duplication, for example, it would likely carry with it most of the parent gene's *cis*-regulatory information (and consequently its expression pattern). Divergence of the duplicates' target sets could then result in a new set of terminal differentiation genes being expressed in a pattern similar to that of the parent. Similarly, if expression of an input-output TF were restricted to only a few tissues or cells (as may be the case for some neuronal transcription factors [92]), negative pleiotropic consequences could be reduced to the point that changes in the gene's target preference or transcriptional output function could occur. Alternatively, the modular 'plugin' developmental programs directed by input-output genes could facilitate changes in upstream transcription factors by providing a mechanism to switch between cell differentiation programs with a minimum of side effects. The *Drosophila* gene *poils-au-dos* (*pad*), for example, is a patterning transcription factor that can regulate ex-

pression of the input-output gene *scute*, a locus that is responsible for much of the inter- and intra-specific variation in bristle number observed in flies and other insects [79,80,90,93]. A variant of *pad* has been identified that results in extra notum bristles in a moroccan population of *Drosophila melanogaster* [80]. This variant has undergone a deletion that removes the DNA-binding domain from the final protein [93], which appears to release *scute* from repression by *pad*. Although the frequency of this variant is likely to be quite low, and it does apparently have some deleterious effects, the fact that these individuals are viable and fertile enough to survive for any period of time indicates that patterning TFs of this type may be more prone to fix mutations that are relevant to regulatory evolution.

1.4.2 Evolutionary Forces Driving Changes in Transcription Factors

Although the mechanisms discussed above point to possible paths through which change in TF proteins could reduce or avoid pleiotropic consequences, they do not address the actual evolutionary forces that could cause functional TF divergence. These can be broadly divided into selective (Darwinian) and non-selective forces (e.g. drift). Although selective explanations are speculated about in many of the well-established examples of regulatory evolution (e.g. [9, 10]), there are few cases in which clear adaptive explanations for regulatory change have been identified (e.g. [83, 94]). As such explanations are likely to be highly dependent on ecological, developmental, and genetic context, it is difficult to speculate about all but the most general mechanisms of adaptive change in regulatory networks, especially as they pertain to *trans*-regulatory change.

The most likely case would involve maintenance of a neutral or deleterious *trans*-acting mutation; this allele would contribute to standing regulatory variation in the population and therefore be subject to selection by ecological or other factors acting on the phenotype(s) it influences. Although it is possible that a new mutation could be immediately adaptive, a more likely scenario is that it will initially be neutral or slightly deleterious. The ultimate likelihood that this mutation could eventually sweep to fixation would therefore depend on its ability to persist in the population. This is in turn a function of the extent of any deleterious effects, the probability of mutation, and its dominance relationship with other alleles. Deleterious effects would likely be a function of pleiotropy as discussed above. Mutation probability (i.e. reversion or conversion to another allele) is largely a function of the mutational target size — although for point mutations in a DNA-binding domain this is likely quite small, it could be larger in cases of domain gain or loss or duplication (see section 1.4.3). Dominance would likely be related to the exact nature of the mutation involved. Null or hypomorphic alleles, or those which reduce the number of target genes relative to other alleles, would likely be recessive, and could therefore be maintained in the population for longer periods by heterozygosity. Gain of function alleles, which expand or shift the set of targets to incorporate new sites, would likely be codominant and therefore visible to selection in heterozygotes. Similarly, alleles that involve deletion of either the DNA-binding or effector domain, or a change in the transcriptional effect of the factor with no alteration in binding, would likely be dominant negative or incompletely dominant, and would also be visible to selection in heterozygotes. In either of these last two cases, pleiotropy and other negative effects would need to be negligible in order for the mutation to persist. In any case, the ultimate fixation of the allele would require that either ecological conditions shift

such that any deleterious effects of the allele are outweighed by adaptive benefits before it is eliminated by drift or purifying selection, or that a novel allele has an immediate adaptive benefit.

Selective scenarios such as these are intuitively appealing, but it is also possible that non-adaptive scenarios could lead to fixation of transcription factor mutations. Genetic drift, for example, could lead to fixation of even deleterious (pleiotropic) TF alleles, particularly in species with small effective population sizes. The role of drift in shaping genomic and organismal complexity is still an issue of much debate (see e.g. [95, 96]), but there is a growing body of evidence indicating that developmental regulatory networks may be shaped to a large extent by neutral forces. One set of examples comes from studies of binding site turnover in *cis*-regulatory regions [22, 97–99]. In these cases, stabilizing selection acts to maintain the expression pattern of a developmental enhancer, while the TF binding sites that make up the enhancer appear and disappear in a process governed mainly by random mutations drifting to fixation [100]. An analogous case involving *trans*-changes would be a transcription factor that largely maintains its target set while changing binding specificity or protein-protein interactions, possibly through a co-evolutionary mechanism. Such cases have now been observed for several transcription factors in yeast (e.g. [101,102]; see sections 1.4.4 and 1.4.5). Although a similar case has yet to be identified in multicellular eukaryotes, the smaller effective population size in these organisms relative to unicellular fungi indicates that drift may play an even larger role in the transcriptional regulatory evolution of metazoans and plants [95, 103].

Further support for the idea that drift could play a major role in regulatory evolution can be found in several recent comparative, genome-wide chromatin immunoprecipitation studies. These indicate an unexpectedly large degree of variability in transcription factor binding [104]. For example, a recent examination of binding by the yeast pseudohyphal transcription factors Ste12 and Tec1 [105] across three yeast species using ChIP-chip found that only 20% of binding events are conserved across all three species, and 50-61% of binding events are species specific. Similar comparative analysis of binding by the hepatocyte transcription factors FOXA2, HNF1A, HNF4A, and HNF6 between human and mouse [106] found that 41-89% of binding events were species specific, and analysis of HNF4A and CEBPA binding among five vertebrate species [107] found that binding conservation drops as low as 2% between human and chick. In both cases most of the network rearrangement is due to gain and loss of binding sites from *cis*-regulatory regions rather than changes in *trans*. Although these studies do not directly address the ultimate causes of the observed large-scale rearrangements in binding, it is unlikely that adaptive mechanisms could explain such a large number of independent changes. It is much more likely that all or most of this change results from drift, with any purifying selection acting to maintain a core set of targets where deleterious effects are particularly strong. If drift is able to elicit such massive rearrangements of network topology by changes in *cis*regulation, it is reasonable to assume that similar mechanisms could drive changes in *trans*-regulatory components as well. At the very least these observations lend support to the idea that a large portion of binding by transcription factors is nonfunctional (see also [66, 67]), and therefore that pleiotropy is likely to be much lower than what might be expected based on binding alone.

1.4.3 Mutational Mechanisms of Transcription Factor Evolution

Most biological phenomena involve multiple levels of causation — developmental events, for example, can be said to result first from a series of biochemical interactions, but also from the evolutionary history of the organism, which ultimately encoded those interactions in the genome. Even evolutionary causation can be further subdivided: evolutionary change requires both mutational processes to generate variation, and a selective (or neutral) processes to drive the mutation to fixation. Evolutionary changes in transcription factors are no different. The mutational processes that generate TF diversity should in principle be the same as those that act on other types of proteins — duplication, domain gain and loss, and diversification by point mutations. Indeed, many examples of each of these processes acting on transcription factors have emerged in recent years.

By far the most common mode of transcription factor evolution is the classical duplication/divergence model, first outlined by Ohno [108] and later extended by Lynch and Conery [109]. This process begins with the creation of a duplicate copy of the DNA sequence containing the transcription factor, either through a tandem duplication caused by recombination error, or transpositional segmental duplications resulting from processes such as chromosomal rearrangement or retrotransposition. The duplicate proteins will then undergo a short period of functional redundancy, which can lead to a relaxation of purifying selection. This in turn allows mutations to accumulate in one or both proteins, which then can follow a number of different evolutionary paths. By far the most common outcome is nonfunctionalization, in which one of the duplicates accumulates sufficient deleterious mutations for it to lose its function; purifying selection then returns in full force to the remaining functional duplicate while the nonfunctional duplicate becomes a pseudogene and is rapidly degraded. In rare cases, one or both copies can accumulate mutations that lead to functional divergence, and purifying selection will then act to retain the two divergent genes. In one scenario (neofunctionalization) one of the duplicates retains the ancestral function while the other diverges to take on a novel function. In subfunctionalization, the ancestral function becomes subdivided between the two duplicates. The most extreme examples of duplication occur when an organism acquires an extra copy of its entire genome through meiotic error or interspecific hybridization; if the organism survives the initial duplication event, this can generate a large pool of redundant proteins which can then diverge in function [110, 111].

At the most trivial level, the fact that transcription factors tend to utilize a small number of binding domain types, and tend to cluster in large protein families, demonstrates the importance of duplication as a mechanism for generating transcription factor diversity. Most important transcription factor families have experienced some amount of expansion within the metazoans, although the number of duplications and their pattern of phylogenetic distribution can vary widely among the various TF families and subfamilies [51, 112]. Although in some cases the majority of functional diversity within a protein family can be traced to phylogenetically deep duplication events (e.g. [113]), there are also many well-established cases of more recent lineage-specific expansions, indicating that family expansion is an ongoing process. Lineage-specific expansions in these protein families can occur by repeated duplications of single genes, such as the hundreds of duplicates of the nuclear hormone receptor HNF4 found in nematodes [114, 115], the ZAD C_2H_2 zinc-finger family

in *Drosophila* [116, 117], and some KRAB zinc finger subfamilies in mammals [118–120], or through larger segmental duplications involving multiple factors, such as the *RHox* cluster in mouse [121] as well as clusters of KRAB zinc finger genes in mammals [118]. Whole genome duplications believed to have also been linked to expansions of many TF families, such as those believed to be responsible for the expansion of Hox genes [54, 122–124], basic leucine zippers [125], and nuclear receptors [126] preceding the divergence of vertebrates and during the divergence of teleost fishes.

Another mutational mechanism that can lead to transcription factor diversification is the duplication, loss, or rearrangement of functional domains within orthologous proteins. Transcription factors generally require at least two types of domains in order to function: a DNA-binding domain to direct the factor to the correct DNA sequences, and an effector or transactivation domain that produces a positive or negative effect on transcription through protein-protein interactions, either directly with the basal transcriptional machinery or through interactions with coregulators. Gain or loss of these domains can occur by a number of different genetic mechanisms [127]. Loss of domains is likely to be more common, with duplicated genes being particularly susceptible to losses due to relaxed selective constraint. Loss could occur by point mutations that result in a premature termination codon upstream of the domain, eliminate splice sites, or which disrupt key functional residues within a domain [120, 127]. Errors in recombination [128] or DNA repair following double-strand breaks [129] can lead to domain swapping among proteins, or domain loss. Transposons can also play a role in domain rearrangements by generating double-strand breaks or by directly contributing domains [130]. Domain gains are likely to be relatively rare, but could take place via small segmental duplications generated by one of the above mechanisms [127]. Internal duplications of this type can result in large arrays of repeated domains, as are often observed for some types of DNA-binding domains [56, 119, 120, 131, 132]. Direct recruitment of noncoding sequence or non-domain protein sequence into domains is rare, although examples do exist [133]; this is likely to be most important for short linear peptide motifs [134] that regulate protein-protein interactions (see 1.4.4).

Gain or loss of effector domains or protein-protein interaction domains has been observed during the evolution of many transcription factor families, and has in some cases been tied directly to functional divergence of these proteins (e.g. [135, 136]; see section 1.4.4). Gains and losses of DNA-binding domains have not been studied in great detail, although examples of both have been found in mammalian C_2H_2 zinc finger proteins, in which orthologs have been shown to expand or contract the arrays of zinc fingers that govern their DNA-binding preference [56, 58, 119, 120, 131, 132]. Loss of DNA binding domain has also been observed for some helix-loop-helix transcription factors, such as the *Id* protein and its orthologs, which have lost the ability to bind DNA but retained their dimerization domains, allowing them to act as negative regulators of other HLH proteins [137]. HLHs also show evidence of a general pattern of evolution by domain shuffling [137].

The accumulation of point mutations also occurs in many cases of transcription factor evolution, but due to the fact that the effects of such mutations are difficult to predict without detailed structural knowledge their importance in a given case can be difficult to assess. Short linear motifs can easily be created and destroyed by point mutations, and may undergo a turnover process similar to that observed for transcription factor binding

sites in DNA sequences [134]. Larger domains contain sites that exhibit a range of selective constraint, from almost completely neutral to completely conserved [138]. DNA-binding domains, for example, contain highly conserved sites that are important for maintaining the overall structural integrity of the domain, as well as sites that directly contact the DNA bases or backbone [139]. While mutations in structural sites are likely to completely disrupt the domain, mutations in DNA-contacting sites can modulate the specificity of the factor to varying degrees ([140–142]; see section 1.4.5). Similarly, mutations at specific sites in protein-protein interaction domains can alter a factor's dimerization partners, as has been observed in the coiled-coil domain of bZIP [143–145] and HLH [146] family TFs. In either type of domain, the importance of mutations that do not affect residues participating directly in binding can be extremely difficult to assess, even for domain families where a wealth of structural information is available. However, many cases have been identified in which mutations in residues of this type have apparent functional and evolutionary importance (e.g. [147–151]); it is possible that they could participate in secondary interaction with binding residues [152], act to reposition or reshape the binding interface (e.g. [152–154], or alter the distribution of isoforms in the structural ensemble [155]. Finally, microinsertions and deletions that alter the length of simple sequence repeats (SSRs) may have functional consequences for transcription factors [13]. Repeats of glutamine, alanine, proline, and glycine have been shown to influence transcriptional output driven by transcription factors [156–160] and may act like evolutionary rheostats that can tune the output of a TF through even small changes in repeat length [161, 162].

1.4.4 Rewiring Transcriptional Networks by Evolution of Transcription Factor Protein-Protein Interaction Domains

Transcription factors can participate in many types of protein-protein interactions, with a variety of functional outcomes. Many transcription factors, such as bZIP [163], HLH [137,164,165], and nuclear receptors [137] TFs, are known to form homo- and heterodimeric complexes when binding DNA. Higher-order multimeric complexes involving proteins from different domain families are known to exist as well [166], and evidence for cooperative binding has been shown for a number of TFs [167–169]. The particular binding partners that are involved in a multimeric complex can determine to a large extent the binding specificity of the factor [145,146], and hence the target genes it is able to regulate. Protein-protein interactions are also key to the transcriptional output function of a TF. Although it may be possible for some repressors to regulate expression only by occluding other factors with a similar binding specificity [170], in most cases TFs need to participate directly in interactions with other proteins to carry out their function. This may involve direct interaction with the basal transcription machinery [171]; interaction with transcriptional cofactor complexes such as the Mediator complex, which can recruit RNA Pol II to initiate transcription [15, 18, 20]; with chromatin remodeling and histone modification complexes such as Swi/Snf and p300 [15, 18, 20]; or, in the case of repressors, short-range interaction with activator factors through a variety of mechanisms [172].

Most known protein-protein interactions occur through discrete structural domains or short, unstructured peptides called short linear interaction motifs (SLiMs) [134]. Both of these structures are known to be evolutionarily labile [127,134,138] — domains can be gained

or lost, shuffled between proteins by a variety of mechanisms, and can be altered by point mutations (see section 1.4.3). SLiMs are short enough to be easily created and destroyed by a few point mutations [134]. Change in transcription factor protein-protein interactions therefore represents a wealth of potential evolutionary pathways for *trans*evolution in transcription networks. It has been proposed that protein-protein interactions may be particularly prone to evolutionary modification because they are less subject to negative pleiotropic consequences than other types of changes [13, 75, 104, 173], primarily due to the fact that novel protein-protein interactions can evolve independently of the DNA-binding activity of the TF. This leads to a number of related models for how protein-protein interactions could lead to evolutionary change in gene regulatory networks [13, 75, 173]. In the most basic scenario, a TF (TF A) can expand or shift its sets of target genes by first acquiring a novel protein-protein interaction activity with a second TF (TF B); this interaction would then position TF A at *cis*-regulatory elements bound by TF B without direct binding by TF A. If the presence of TF A at TF B regulatory elements provides a selective advantage, binding sites for TF A could then evolve at positions flanking the pre-exisiting TF B binding sites in order to strengthen the interaction of TF A with the element. It is also possible to imagine non-selective scenarios in which TF A binding sites emerge by drift; if purifying selection is acting only to maintain the presence of TF B, it is possible that the presence of a TF A binding site and an A-B protein-protein interaction could allow the original TF B binding site to be weakened by point mutations until function of the element requires binding by both TF A and TF B. Alternatively, if TF B requires a cofactor TF C, the new TF A - TF B interaction could loosen the selective constraint on the TF A - TF C interaction, leading to co-option of TF C targets by TF B.

Genomic studies in yeast have identified a number of cases that seem to support these models as one valid pathway for TF change by protein-protein interaction evolution. The first of these involves the mating pathway genes that are differentially expressed in the **a** and $\boldsymbol{\alpha}$ mating types [101,174]. In *S. cerevisiae* and several closely related yeasts, **a**-specific genes (**a**sgs) are expressed by default in **a** cells, and are repressed in $\boldsymbol{\alpha}$ and **a**/$\boldsymbol{\alpha}$ diploid cells by the homeodomain repressor $\boldsymbol{\alpha}$2 [175,176]. However, this system appears to be a derived synapomorphy; in the inferred ancestral state (displayed by *Candida albicans* and other non-*sensu stricto* yeasts), the **a**sgs are repressed by default, and activated in **a** cells by the activator **a**2, which is not present in the species with $\boldsymbol{\alpha}$2-dependent repression [101,177,178]. The shift from **a**2 activation to $\boldsymbol{\alpha}$2-based repression of the **a**sgs has been shown to be caused by the gain of a protein-protein interaction between $\boldsymbol{\alpha}$2 and Mcm1, a MADS family TF that binds the **a**sgs in both species groups [177–179]. Structural modeling indicates that this interaction emerged as a result of a series of single-aa changes in the Mcm1-binding surface of $\boldsymbol{\alpha}$2, which resulted in stronger binding between Mcm1 and $\boldsymbol{\alpha}$2 in *S. cerevisiae* [101]. Intermediate forms are observed in *Kluyveromyces lactis*, which is predicted to have a weaker Mcm2-$\boldsymbol{\alpha}$2 interaction, and contains binding sites for both $\boldsymbol{\alpha}$2 and **a**2 flanking the Mcm1 binding sites in its **a**sgs. Interestingly, genome-wide ChIP studies of Mcm1 indicate that it also may have gained and lost a number of protein-protein interactions during yeast evolution [174]. In *K. lactis*, for example, Mcm1 shows high rates of co-occurence with binding sites for the ribosomal TF Rap1, and may have handed off an interaction with the arginine metabolism gene Arg81 to the Mcm1 paralog Arg80 following the whole-genome

duplication in the *sensu stricto* yeasts.

In addition to its *K. lactis*-specific interactions with Mcm1, there is evidence that Rap1 may also have undergone a major shift in cofactor interactions during evolution of the *sensu stricto* species [180–182]. In *S. cerevisiae*, Rap1 binds to promoters of ribosomal protein and glycolysis genes as well as telomeres [182]. At the ribosomal genes, Rap1 (together with Hmo1) recruits the ribosome-specific factors Ifh1 and Fhl1 to activate condition-specific expression [183–191]; however, in *C. albicans*, Rap1and Hmo1 do not bind to ribosomal genes and are not required for their expression [180–182]. Instead, two other factors, Tbf1 and Cbf1, execute an analogous function, recruiting Ifh1 and Fhl1 to ribosomal protein genes [181, 182]. A similar scenario has been noted for the galactose metabolism genes [192]. In *S. cerevisiae* and closely related species, these are regulated by the transcription factor Gal4, which activates expression in the presence of galactose and is repressed in other conditions by protein-protein interaction with Gal80 [193, 194]. In *C. albicans* and other yeast species more diverged from *S. cerevisiae*, the regulatory role of the Gal4 ortholog has been completely shifted such that it is no longer required for activation of the *GAL* genes, which is instead accomplished by another TF, Cph1 [192]. As with the Mcm1-$\boldsymbol{\alpha}$2 case discussed above, *K. lactis* represents an intermediate state, in which both Gal4 and Cph1 sites co-occur in *GAL* gene promoters. Although the shifts in the Rap1 and Gal4 regulons have not been traced to specific mutations in protein-protein interaction domains, there is substantial circumstantial evidence that such changes have at least taken place in coordination with these changes in both cases. For Rap1, for example, the differences in the ability of *C. albicans* and *S. cerevisiae* to recruit Fhl1/Ifh1 indicate at the very least an indirect shift in protein-protein interactions [181, 182]. Also, it is known that the ribosomal protein genes have strict requirements for stoichiometric expression [195], which would make a shift in regulation by stepwise changes in *cis*-regulatory sequences unlikely. For Gal4, the *C. albicans* ortholog is almost completely diverged outside of the DNA-binding domain, with no regions showing clear homology to the *S. cerevisiae* transctivation or Gal80-interaction domains [192]; it is therefore likely that protein-protein interactions involving Gal4 have diverged between these species.

Examples of TF evolution by gain and loss of protein-protein interactions can also be found in metazoan lineages, although these generally lack the level of genome-wide detail of the yeast studies. Two of the better-studied examples are the Hox transcription factors *Ultrabithorax* (*Ubx*) and *fushi tarazu* (*ftz*). *Ubx* from *Drosophila melanogaster* is known to have a variety of homeotic activities when ectopically expressed [196,197]. Ectopic expression in *Drosophila melanogaster* of *Ubx* from an onychophoran (a sister taxa to the arthropods) shows that many of these activities are conserved between the two genes [135, 198]. However, onychophoran *Ubx*, unlike its *Drosophila melanogaster* ortholog, is unable to drive embryonic segmental transformations or repress expression of the gene *Distalless* (*Dll*) [198]. These activities are directed by a QAQA domain that was gained in the ancestral copy of Ubx prior to the divergence of arthropods, and has been shown to function as a transcriptional repression domain [135]. The *Drosophila ftz* gene, on the other hand, has lost the ability to drive homeotic transformations that is observed for copies of *ftz* isolated from the grasshopper *Schistocerca gregargia* and the beetle *Tribolium castaneum* [136]. Instead, *Drosophila ftz* functions primarily as a segmentation gene, and causes loss of segmental

identity when ectopically expressed; this effect is reduced for *Tribolium* and *Schistocerca ftz*. These functional differences can be traced to two changes in the Ftz protein during insect evolution. The loss of homeotic activity in *Drosophila* Ftz correlates with loss of a YPMW motif upstream of the homeodomain, which is known to regulate interactions with the Hox cofactor Extradenticle (Exd) [198–201]. The enhancement of segment-specification activity in *Drosophila* is at least partially due to an LxxLL motif that mediates interactions with the Ftz cofactor Ftz-F1; this motif is conserved in *Tribolium* and absent in *Schistocerca*, coinciding with the partial conservation of segmentation activity in *Tribolium* and its almost complete absence in *Schistocerca* [136, 202–204]. Other examples of evolutionary changes in metazooan TF protein-protein interactions have emerged in recent years as well, including an apparent case of coevolution between the insect ecdysone receptor (EcR) and its binding partner Usp [153], and a novel interaction between mammalian *Hox11a* and *FOXO1A* which may be tied to the evolution of pregnancy in eutherian mammals [205, 206].

Finally, further evidence for the importance of protein-protein interaction domain changes for TF evolution comes from studies of large TF families. Extensive diversification of homo- and hetero-dimerization interactions has been observed for helix-loop-helix TFs [146, 165, 207], basic Leucine Zipper TFs [125, 165, 208, 209] and Nuclear Receptors [165, 207, 210]. Many of these interactions are lineage specific, indicating that evolutionary and functional flexibility exists in many of these dimerization networks. There is also evidence that this diversification in dimerization properties may underlie at least some family-level divergence in DNA-binding specificity [146]. Another line of evidence comes from cases of large-scale lineage-specific family expansions. These large sets of novel proteins are often characterized by a novel protein-protein interaction domain, as is the case for the vertebrate-specific family of C_2H_2 zinc finger TFs containing the KRAB domain [118, 211, 212], which acts as a transcriptional repression domain by recruiting histone deacetylases through a cofactor, KAP1 [213–215]. A similar pattern has been observed for lineage-specific ZF families containing the mammalian SCAN domain [58], which can direct formation of homo- and heterodimers among family members [58, 216], and the insect ZAD domain [116, 117], which also functions as a dimerization domain [217]. The causal relationship between family expansion and acquisition of protein-protein domains is still unclear, however. It is possible to imagine scenarios in which the acquisition of novel dimerization or transcriptional effector activity opens a new region of "regulatory space" (that is, the universe of possible linkages between DNA sequences and transcriptional outputs) that can then be explored by selective or neutral forces driving fixation and divergence of duplicate proteins. On the other hand, it is also possible that these novel domains could simply be the most conserved portions of proteins, which are more prone to duplications because of genomic location or other characteristics not related to function; indeed many of these family expansions occur in tightly linked arrays, which may be duplication hotspots [118, 211, 212].

1.4.5 Rewiring Transcriptional Networks by Evolution of Transcription Factor DNA-Binding Domains

The ability to bind DNA sequences with specificity and selectivity is the defining characteristic of regulatory transcription factors. A variety of structural domains participate in DNA binding, using many different types of amino acid residue - DNA base interactions

and binding interface architectures [23]. The member proteins of these domain families in turn recognize and bind to DNA sequences with an astonishing variety of specificities and affinities [218, 219], resulting in many billions of possible DNA-protein interactions. These protein-DNA interactions determine to a large extent the topological structure and dynamic function of gene regulatory networks. Given the now-widespread recognition that change in these networks is a key component of much morphological and physiological variation [8–12], it seems obvious that any complete picture of regulatory evolution must include an understanding of how and why transcription factor DNA-binding specificities change over time. However, the immense success of the *cis*-regulatory paradigm (see section 1.3) as both an experimental program and a theoretical organizing principle has resulted in transcription factors, and their DNA-binding specificity in particular, being treated as immutable building blocks of regulatory networks rather than possible sources of evolutionary variation [8]. Due to the use of computational tools for TF binding site discovery and orthology assignment, studies of *cis*regulation necessarily involve species that are closely related enough for *cis*-regulatory sequences to be alignable at the level of primary DNA sequence; in addition, it is often necessary to assume that the DNA-binding properties of the involved TFs can be transferred between orthologous proteins . Although studies employing these assumptions have certainly been successful in identifying the molecular basis for a large number of phenotypic traits, their validity over longer evolutionary time spans and for explaining general principles of regulatory evolution are dubious. In the preceding sections I have shown that many properties of TFs are evolutionarily labile (section 1.4.4), and that there are many theoretical reasons to expect that TFs generally may represent an important source of regulatory variation (sections 1.4.1, 1.4.2, and 1.4.3). The evolution of binding specificity is in some ways the most difficult phenomenon to explain theoretically, and is in fact often avoided in discussions of TF evolution (e.g. [13, 75, 104, 173]). However, there is now a significant body of empirical evidence indicating that changes in TF DNA-binding specificity can and do happen; in this section I will discuss this evidence and outline an experimental program for examining these changes in detail.

There are two major questions that arise when considering the evolution of transcription factor DNA-binding specificity. The first concerns TF change at the macroevolutionary level: what explains the patterns of DNA-binding specificity conservation (across species) and diversity (within conserved families) observed across large phylogenetic distances? The second concerns the generation of TF diversity at the species level: what are the forces that can lead to divergence of transcription factor-DNA binding specificity in orthologs and recently arisen paralogs, and what is the relationship between these changes and regulatory and morphological phenotypes? As with many evolutionary questions, the distinction between these two is mostly a matter of scale; presumably what we learn from studying TF evolution between species and populations reflects similar processes that acted in generating TF diversity more broadly.

Itzkovitz et. al. [220] proposed a hypothesis that serves as a convenient conceptual framework for addressing questions of large-scale TF DNA-binding specificity evolution. As discussed above, it has been noted that the total number of transcription factor proteins generally scales with the number of genes in the genome [52] and number of cell types [53]. Although determining the flow of causation in this relationship is difficult, it is likely that

it can be at least partially explained by the need for more complex "addressing" of TFs to their targets to accommodate the larger sequence search space and more complicated patterns of gene expression. Interestingly, Itzkovitz et. al. [220] show that the degree of scaling is not homogeneous across all TF families, and in fact most TF superfamilies seem to have an upper bound to the number of TFs that can be accommodated even in large genomes. Once this limit is reached, expansion of the family remains fairly flat. These limits seem to be related to the DNA-binding mechanism of the domain family, specifically the number of variable positions (and hence the potential number of target sequences) that it allows. HLH TFs, for example, have a fairly restricted binding mechanism that only allows for two variable positions [164,221,222] in each binding interface and can recognize a total of $\frac{4^4}{2}$ possible sequences (correcting for reverse complements and the ability of HLHs to dimerize); the maximum number of HLH TFs present in any sequenced genome is 186 in *Arabidopsis* [220]. At the other end of the spectrum are C_2H_2 zinc fingers, which have modular DNA-binding interfaces; each C_2H_2 domain can recognize 3 bases [26, 141], and there are examples of proteins which possess 30 or more of these binding domains [223]. Accordingly, the maximum number of C_2H_2 zfs is 735 in human [112]. Itzkovitz et. al. propose that this relationship is a result of the need to reduce overlap in the set of sequences bound by each TF in a given domain family; for domains that bind with fewer variable positions, there will be a smaller set of sequences available for binding, and consequently fewer TFs will be able to "fill" that space without overlaps in binding specificity. Once the sequence space available to a particular domain is saturated by binding specificities of domain family members, the need to avoid overlaps will impose a barrier to the fixation of new proteins since these would necessarily overlap an already occupied portion of "binding space".

There is, of course, likely to be a great deal more complexity than this relatively simple model implies. While subsequent family-wide studies of TF binding specificity have revealed a great deal of diversity in DNA-binding among TF family members [29, 142, 146, 224, 225], these studies also show that there can exist a great deal of overlap in binding specificities. Interestingly, there is some evidence that for a number of TFs overlap only occurs in highly-bound sequences; *in vitro* studies using protein-binding microarrays [29,225] have found that differences in binding affinity between closely related TFs were often more pronounced for weakly bound sequences, even when position-weight matrix binding models for the proteins were identical. The amount of possible sequence space that is actually accessible by mutation to a DNA-binding domain may also be restricted by structural factors [142]. The functional costs of TF binding specificity overlap are also unclear. While Itzkovitz et. al. provide evidence suggesting that TFs with close binding specificity tend to bind targets with similar function to mitigate functional consequences from binding errors [220], for some TFs binding overlap may play a role in repression [170,172]. It may also be the case that the primary DNA-binding specificity of the factor may be less important than the protein-protein interactions it participates in with other factors — many homeodomain proteins, for example, have relatively short binding motifs that are often closely related to one another and bind widely throughout the genome, and it has been suggested that combinatorial binding with other TFs may be important for restricting activation to a smaller set of targets [226]. Similarly, non-overlapping expression domains could prevent

negative effects from cross-binding [227], and differing chromatin environments across cell types could differentially restrict the available set of binding targets [66, 67, 228, 229].

In addition, the existence of an upper limit to family size does not speak to the forces that drive expansions of these protein families in the first place. Although larger and more complex genomes almost certainly require a larger repertoire of transcription factors in order to direct precise binding and intricate expression patterns, the evolutionary forces which created these complex regulatory systems are difficult to conceive and even harder to generalize. The complexity and strong structural conservation of most DNA-binding domains implies that most superfamilies are likely monophyletic (descendents of a single ancestor protein) [51]. The origin of ancient protein domains is still mysterious [230], however, given that many TF domains use fairly common structural motifs [23] such as coiled-coils, it is possible that some may have been exapted from domains with other functions that had incidental or nonspecific DNA-binding activity (e.g. [231]). After the establishment of an ur-TF, the family is presumably built by duplication / divergence mechanisms; however, whether this process is driven by an adaptive benefit of additional regulatory complexity or by neutral mechanisms is unclear. It is possible to imagine scenarios in which additional regulatory complexity could be adaptive by allowing physiological adaptation to environmental changes or novel morphological features. However, it is also possible to imagine situations in which drift could play a more significant role. If, for example, the DNA-binding specificities of a pair of paralogous TFs drifted apart, it is possible that their downstream targets, which were initially fully shared between the paralogs, could become split between the two proteins by accumulating mutations in TF binding sites which favor one paralog or the other.

A particularly productive mechanism for generating diversity in TF family DNA-binding may be through whole-genome duplications (WGDs). These have happened at several points during the evolution of animals, plants, and fungi, and often coincide with significant expansion in TF families [46, 47, 54–56, 58]. As WGDs produce a large number of initially redundant TFs along with a large set of redundant binding sites and target genes, these events have the potential to generate a pool of novel, weakly selected regulatory interactions. Drift or adaptive pressure could then gradually lead to fixation of a large amount of novel regulatory complexity. Many of the instances of TF-mediated regulatory variation in yeast, for example, coincide with the whole-genome duplication in yeast, which preceded divergence of the *sensu stricto* species (e.g. [101, 174, 180–182, 192] ; see section 1.4.4 and below). In addition, the fact that large TF family expansions and whole genome duplications often co-occur with novelties in the body plan indicates that this process may be a key part of many deep cladogenic events [55, 57].

In addition to these broad trends in TF evolution, there now exist a number of smaller scale case studies that involve only single or a few proteins examined over relatively short evolutionary timescales. These studies of lineage-specific shifts in DNA binding can potentially provide insight into the macroevolutionary processes of family expansion and diversification as well as the importance of lineage-specific TF differences for species-level morphological evolution. One of the more well-studied examples of this phenomenon involves the *Drosophila* homeodomain proteins *bicoid* (*bcd*) and *zerknullt* (*zen*) [232–237]. *Bcd* is a maternally deposited TF that acts as a morphogen to establish the initial anterior-

posterior axis of the *Drosophila* embryo [238], and *zen* plays a role in gene expression in extra-embryonic tissues such as the amnioserosa [239, 240]. Both of these genes are insect specific. Early in insect evolution, *zen* arose from a *Hox3*-like ancestor, which had Hox-like embryonic expression and possessed the hexapeptide (YPMW) motif found in many Hox genes. Concurrent with this duplication, *Zen* lost both its embryonic expression pattern and the hexapeptide motif, and acquired extra-embryonic expression and a novel peptide motif, which may be a protein-protein interaction domain [232,234,236]. Prior to divergence of the cyclorraphan dipterans (which includes the genus *Drosophila*), *zen* was duplicated to produce the ancestral *bcd* [233]. *Bcd* then quickly diverged, acquiring a novel maternal expression pattern [234], and undergoing a number of amino acid substitutions within its binding domain, including a Glu→Lys substitution at postition 50 of its homeodomain [233]. This change altered the binding specificity of *bcd* from the TAATTA consensus of *zen* to a novel TAATCC consensus [224].

The complex evolutionary history of *bcd* and *zen* illustrates a number of key points about TF evolution. First, duplication can be a major catalyst for dramatic sequence and functional changes, even those such as changes in protein-protein interactions or DNA-binding specificity that may have widespread pleiotropic effects. Second, *cis* and *trans* regulatory changes are often interrelated — although the exact timing of the expression changes relative to the sequence changes in *bcd* and *zen* is not known, it is possible that the changes in expression correlated with each duplication event helped to reduce the pleiotropic costs of functional changes in the duplicated proteins; this pattern is observed in other cases of TF divergence as well [135, 198]. Finally, the shift in binding specificity between *zen* and *bcd* illustrates that changes in TF DNA-binding domains can occur over relatively short evolutionary timescales as a result of relatively simple sequence changes. Interestingly, a recent examination of DNA-binding specificities for the entire complement of *Drosophila* homeodomain proteins by Noyes et al [224] demonstrated that changes of this type in duplicated homeodomains may be relatively common within homeodomains. By mapping experimentally determined binding specificities onto a phylogenetic tree built using primary amino acid sequence, Noyes et. al. were able to show that binding specificity is polyphyletic, and that 9 pairs of homeodomains (in addition to *zen* and *bcd*) have diverged in binding specificity following a duplication. Seven of these changes can be accounted for by mutations within DNA-contacting residues of the binding domain.

Lineage specific protein families, such as the insect ZAD [116, 117] and vertebrate KRAB [118, 211, 212] zinc finger families, and the nematode supernumerary HNF4-related nuclear receptors [115,210,241], provide further examples of divergence in TF DNA-binding specificity. Each of these families has undergone an explosive expansion within the lineages that contain it, and in many cases a significant number of the duplicated TFs appear to be species-specific; for example, only 134 of 256 HNF4 NRs present in *C. elegans* are conserved in the genomes of *C. briggsae* and *C. remanei* [241], and many KRAB zfs are found in human-specific expansions [242, 243]. There is evidence that at least some of the diversity in these families manifests in the form of divergent DNA-binding specificities. For example, the paralogous *Drosophila* ZAD TFs *Sry-δ* and *Sry-β* have been shown to have differential binding specificities [244] likely driven by changes in the C_2H_2 -zf binding domain [245]. Interestingly, these changes come in spite of apparent conservation of other functions, as *Sry-*

β can rescue *Sry-δ*-mediated activation of maternal *bcd* when β binding sites are substituted for δ binding sites in the *bcd* promoter [244]. KRAB proteins also show evidence of likely functional changes within their C_2H_2 -zf binding domains [132], and the DNA-binding P-boxes of *C. elegans* HNF4-like NRs are divergent at potentially specificity-determining positions [114, 115]. The reason for the flexibility of these particular domain families is unclear. All seem particularly prone to multiply by tandem duplications [114–117, 242, 243] and consequently exist in large, closely linked arrays; it is therefore possible that an increased propensity for duplication (possibly enhanced by the presence of long stretches of closely linked, repeated sequence) alone could explain these families. Also, the ZAD and KRAB families are characterized by novel protein-protein interaction domains ([213–215, 217], see section 1.4.4); it is possible that the addition of these domains enabled novel cofactor interactions that allowed new regions of DNA-binding space to be explored.

Although the evidence for the role of duplication and family expansion in the evolution of TF DNA-binding specificities is extensive, cases of binding domain evolution in the absence of duplication are still relatively rare in the literature. In some ways, DNA-binding specificity divergence in the absence of duplication is the most difficult type of TF evolution to explain theoretically. This is primarily because changes of this type will almost by definition cause a change in target genes, which would most likely lead to extensive pleiotropic consequences. Although it may be possible for a TF to expand the range of DNA sequences it binds without losing pre-existing interactions [246], it is much more likely that any changes in in the DNA-binding domain that alter specificity will result in a loss of binding to some sites.

In spite of the potential for detrimental pleiotropic consequences, however, it is now clear that changes of this type can and do occur. As was the case for divergence in protein-protein interactions, the most detailed studies of DNA-binding specificity divergence in orthologs come from yeast. The first of these involves the regulators of ribosomal gene expression Hmo1, Rap1, Cbf1, and Tbf1 [182] already discussed in Section 1.4.4. In addition to the possible changes in protein-protein interactions among these factors during the divergence of *S. cerevisiae* and *C. albicans*, Hmo1, Tbf1, and Rap1 appear to have undergone subtle shifts in DNA-binding specificity as well. Hmo1, for example, binds a sequence consisting of GGT repeats in *C. albicans*, but a more compact motif with a core consensus of GGCGG in *S. cerevisiae*. Rap1 also binds a more extended motif (consensus CATCCANACANCAATAG) in *C. albicans* relative to *S. cerevisiae* (consensus CACCC-NNACA); this may be related to changes in the telomeric repeat sequence in *C. albicans*, which is a known target of Rap1 in both species. The changes in Tbf1 are primarily due to its requirement for binding palindromic sequences in *C. albicans*, which is likely due to a novel ability to dimerize in this species.

Interestingly, for all three of these proteins the changes in DNA-binding specificity are accompanied by an almost complete shift in the set of targets the factors regulate — Hmo1 binds ribosomal genes in *S. cerevisiae*, but binds mostly carbohydrate metabolism and cell-cycle control genes in *C. albicans*, while ribosomal gene regulation by Rap1 in *S. cerevisiae* is taken over by Tbf1 in *C. albicans*, with Rap1 retaining only its role in binding telomeres [182]. In contrast, the remaining two examples of yeast TF DNA-binding specificity divergence both involve apparent coevolution of TFs with their binding sites. The first

of these involves the noncanonical C_2H_2 -zf Rpn4 [147], which is a regulator of proteasomal genes in *S. cerevisiae* [247, 248]. *S. cerevisiae* proteasomal gene promoters contain a motif GGTGGCAAAW, which is known to bind Rpn4; however, when proteasomal genes from *C. albicans* were examined for this sequence, none were found [147]. Instead, a different sequence with consensus GRAGGCAAAA was found in 25% of proteasomal promoters. Subsequent *in vitro* binding assays using Rpn4p proteins from each species showed that they differ in their binding affinity for these fragments, and that these differences are due to changes in the DNA-binding domain, although the specific causative changes could not be identified. A similar case involves the bZIP TF YAP1 [102], which belongs to a family of 8 related TFs in *S. cerevisiae* [249, 250]. These family members form two groups based on binding specificity: YAP3, YAP4, and YAP5 bind the YAE-A element consisting of adjacent palindromic half-sites (TTACGTAA), while YAP1, YAP2, YAP5, YAP7, and YAP8 bind a YAE-O element (TTACTAA) in which the half sites overlap [251–254]. The difference in binding specificity is determined in large part by the amino acid at position 12 in the DNA-binding domain — YAE-A factors have Arg or Asp, and YAE-O factors have Lys [252, 255]. Although *S. cerevisiae* YAP1 binds the YAE-O element strongly and has an Arg at position 12 of the DBD, the YAP1 ortholog from another yeast species,*Candida glabrata*, has a lineage specific R→K mutation at position 12, which switches its specificity to the YAE-A element [102]. This difference in binding specificity is accompanied by conservation of a core set of shared targets between *S. cervisiae* and *C. glabrata* YAP1, most of which have acquired YAE-A binding sites in their promoters. The proposed explanation in both cases is that the binding specificity of the TF is co-evolving with the binding sites upstream of its targets.

There are as yet no confirmed examples of DNA-binding specificity divergence between orthologous genes in metazoans, although there are hints that this process may be active here as well. Mammalian KRAB C_2H_2 -zinc fingers again provide some of the best examples. Although the DNA-binding specificity is not well characterized for most of these proteins, there is extensive sequence evidence available for a variety of mammalian species, and the general mechanisms of C_2H_2 -zf DNA binding are well understood [26, 141]. These data show evidence for divergence within DNA-binding domains of orthologous proteins, mainly through insertions and deletions of entire C_2H_2 -zf subunits [56, 119, 119, 120, 132]. Emerson et. al. [132], for example, found that of 137 human-cow-mouse C_2H_2 -zf ortholog triplets, 26 showed variation in numbers of zinc finger units in at least one species. Evidence for divergence by point mutations at DNA-contacting positions is weaker, although such changes can happen [132].

An especially interesting example which incorporates both of these mechanisms is the *Prdm9* gene, a KRAB-box C_2H_2 -zf TF that functions in proper meiotic progression [256]. *Prdm9* has been shown to act as a speciation gene in mice, where differing *Prdm9* alleles cause spermatogenic failure in hybrids between the subspecies *Mus musculus musculus* and *Mus musculus domesticus*, leading to hybrid sterility [257–260]. Alleles of *Prdm9* which differ in the number of C_2H_2 -zf repeats have also been shown to have differential effects on fertility — an allele with 13 fingers leads to hybrid sterility, while an allele with 14 fingers does not [260]. Because Dobzhansky-Muller incompatibilities of the type displayed by *Prdm9* require epistatic interactions between two loci, these causative

differences in finger number are thought to affect DNA-binding in such a way that it becomes incompatible with binding sites recognized by the other allele. This implies both divergence in binding specificity and co-evolution of DNA binding sites [256, 260]. Interestingly, *Prdm9* appears to have been maintained as a rapidly-evolving speciation gene for long periods of evolutionary time [256]. Sequences of *Prdm9* from 13 rodent species spanning 25my of divergence show that the *Prdm9* DNA-binding domain has undergone massive rearrangements, with complex shuffling, gain, and loss of C_2H_2 -zf repeats accompanied by positive selection for divergence at DNA-contacting residues. This repeated rearrangement and strong selection at *Prdm9* likely represents a case of concerted evolution, and a similar pattern of constant rearrangement extends to other mammals, fish, and invertebrates [256]. The specific effects of these changes on DNA-binding specificity have not been investigated, but given the extensive divergence in binding domains it is likely that *Prdm9* is changing its specificity quite rapidly. *Prdm9* may represent an extreme case, but other examples of transcription factors acting as speciation genes have been found as well [148, 261–263]. These cases do not have the obvious connection to DNA-binding specificity seen in *Prdm9*, but it is still possible that involvement in speciation could be an important force driving TF diversification.

Although these examples are intriguing, there are still no well-established examples of TF DNA-binding specificity divergence in metazoans. Ideally, such a case would involve a clear change in the DNA-binding domain of a TF between two species that are closely related enough to allow direct comparison of orthologous *cis*-acting sequences and target genes. This would allow the examination of not only changes in the inherent biochemical characteristics of the factor in question, but also of any changes in the binding sites that it targets. This would allow for the direct examination of the evolutionary processes, such as co-evolution or target gene shifts, that underlie the changes, and may even allow for the examination of broader effects of these changes on the transcription network. The goal of my thesis work has been to establish such a case using the genomic and biochemical tools afforded by the 12 sequenced *Drosophila* species [264]. In the following chapters, I will 1) identify cases of transcription factor gene divergence in 12 sequenced species of *Drosophila* and 2) test a case of transcription factor divergence that I identified, the sex determination gene *hermaphrodite* [265–268], by examining the genomewide binding of this factor in two different fly species, *Drosophila melanogaster* and *Drosophila pseudoobscura*. Finally, I will discuss preliminary evidence for a large-scale divergence of binding between the HER proteins in these two species, which may be accompanied by a shift in DNA-binding specificity.

Table 1.1: **Phylogenetic Distribution of Transcription Factors** (Data from [51] and * [115])

Species	**Homeodomain**	**C_2H_2 Zinc Finger**	**bZIP**	**bHLH**	**Hormone Receptor**
Homo sapiens	255	735	52	107	48
Pan troglodites	218	665	47	95	44
Mus musculus	276	641	51	106	47
Canis familiaris	215	322	48	90	46
Bos taurus	215	469	50	100	43
Gallus gallus	150	233	34	75	34
Danio rerio	295	388	68	123	67
Ciona intestinalis	79	100	23	36	12
Drosophila melanogaster	106	249	20	56	17
Caenorhabditis elegans	94	111	25	42	270*
Nematostella vectensis	162	213	35	72	18
Monosiga brevicolis	1	55	16	14	0

Chapter 2

A Computational Screen for Rapidly Evolving Transcription Factor DNA-binding Domains in 12 *Drosophila* Genomes

2.1 Abstract

Although many studies have explored the mechanisms of *cis*-regulatory evolution with a high level of molecular detail, there are still relatively few well-established examples of regulatory evolution occuring through changes in transcription factors. In this work I examine the complete complement of predicted transcription factor proteins (TFs) in 12 sequenced *Drosophila* genomes with the goal of identifying TFs with rapidly evolving DNA-binding domains. Although most TFs examined show a high degree of conservation throughout the 12 species, I identify several that exhibit elevated rates of amino acid substitution within their DNA-binding domains, and contain changes at residues likely to be of functional importance. These include the sex-determination gene *hermaphrodite* (*hermaphrodite*), which shows evidence of extensive divergence within its C_2H_2 zinc finger DNA-binding domain as well as a possible 3-fold duplication of a zinc finger domain in *Drosophila pseudoobscura* and *Drosophila persimilis*.

2.2 Introduction

Gene regulatory systems in metazoans consist of both *cis*-acting DNA elements and *trans*-acting proteins that bind them [18]. While evolutionary change in transcriptional regulatory systems could in principle take place in either *cis* regulatory sequences or transcription factor proteins, a number of convincing theoretical arguments have been made indicating that changes in *cis* elements will be much more common than coding changes in *trans*-acting proteins [8–11] (see Ch. 1). However, there are reasons to expect that, although they may be rare, changes in transcription factors represent an important component of regulatory evolution in many organisms (See Ch.1). There are many open questions about the process by which such changes might occur, such as how transcription factors can overcome negative pleitropy, what effects changes in protein sequences have on the number and distribution of target sites in the genome, and how the phenotypic changes that result from changes in binding are "read out" by selection.

In order to address these questions, I need to not only examine variation in the sequence of a transcription factor protein but also the sequences of the *cis*-acting DNA elements that it acts on. Therefore an experimental system is needed in which a transcription factor protein has been shown to vary across a closely related set of species, where synteny and gene content are largely conserved and meaningful comparisons could be made among potential sets of target genes. My goal in this work is to identify such a case, with the ultimate aim of examining the genome-wide binding of a variable transcription factor using biochemical methods.

The experimental approach taken in most studies of *cis*-regulatory change is to begin with a phenotype known to vary among a closely related group of species or subspecies [33–45] and dissect it using genetic and molecular biological tools. On the other hand, the approach in many studies of transcription factor evolution is to collect sequences for a single protein or protein family across a broad range of species [126, 163] and examine the phylogenetic relationship among these proteins using computational tools. Neither of these approaches is ideal if the goal is to study transcription factor evolution in detail, at

both the level of protein changes and target sites. The first approach is usually restricted to examining short evolutionary timescales where meaningful comparisons can be made among noncoding sequences, and is generally agnostic to the molecular basis of changes in visible phenotype. Because changes in transcription factors are expected to be rare, this approach is far more likely to identify *cis* changes. While the second approach is certainly able to identify significant changes in transcription factors, the wide range of species usually examined means that comparisons of target genes and binding sites will be difficult.

In order to maximize the chances of identifying a divergent transcription factor while still allowing for meaningful comparison of *cis*-acting targets, it is necessary to survey a wide range of transcription factor proteins within group of closely related species. The 12 sequenced fruitfly species of the genus *Drosophila* (Fig. 2.1) [264] are ideal for such an analysis. These species represent more than 40 million years of divergence, have a wide geographic distribution, and are divergent in a variety of phenotypic, behavioral, and life-history characteristics [269]. However, they are closely related enough that synteny and gene content are generally conserved [264], and orthology can be established for most genes.

In this work I examine the entire complement of known and predicted transcription factors from these 12 *Drosophila* species with the goal of identifying cases of transcription factors with rapidly evolving DNA binding domains. As expected, I find that most of these proteins are highly conserved. However, I was able to identify a number of potentially interesting cases of divergence within DNA-binding domains, although for the most part these do not involve changes in predicted DNA-contacting residues. An exception is the gene *hermaphrodite* (*her*), which shows a high degree of divergence within its 4x C_2H_2 zinc finger binding domain, including at least one potential specificity-altering change, as well as a large insertion in *Drosophila pseudoobscura* , which contains 3 additional zinc finger domains.

2.3 Materials and Methods

2.3.1 Identification and alignment of Orthologous DNA-binding Domains

In order to identify DNA binding domains, I used a set of 754 *D. melanogaster* transcription factors from the FlyTF database ([270]), representing 49 TF families. The amino acid sequence of each transcription factor from this set was searched using the appropriate domain model from the Pfam database [271] using HMMER [272] to locate the *Drosophila melanogaster* DNA-binding domain. The set of 1-1 orthologs for each *Drosophila melanogaster* protein were collected from the 12 drosophila genomes orthology data [264] generated using GLEANR [273]. These ortholog sets were aligned in amino acid space using MUSCLE [274], and the location of the *Drosophila melanogaster* DNA-binding domain was used as a guide to create an alignment slice representing the binding domain in the other species. Proteins from species that did not contain a full-length binding domain were excluded from further analysis.

2.3.2 Evolutionary Rate Calculation and Identification of Rapidly Evolving DNA-binding Domains

Amino acid substitution rates for each set of aligned, orthologous DNA-binding domains were calculated using the codeml program from the PAML package [275]. Rates were calculated using the tree topology from Fig. 2.1 and converted into pairwise rates for each protein in each species relative to *Drosophila melanogaster* . Rates for the proteins in each domain family were ranked for each species, and those with the top rates were inspected manually for evidence of changes that may impact binding specificity, protein-protein interactions, or domain structural integrity. Determination of significant changes was performed based on the structural literature for each domain type.

2.4 Results/Discussion

2.4.1 Most *Drosophila* Transcription Factor DNA-binding domains show high levels of conservation

I calculated DNA-binding domain amino acid substitution rates for a set of 754 *Drosophila* transcription factors belonging to 49 transcription factor families in order to identify candidates that may have divergent binding properties across the 12 sequenced *Drosophila* species. I chose to examine only DNA binding domains because these are well annotated relative to other domain types, and structural data is available for most families of binding domains, making identification of potentially important amino acid substitutions possible. I also restricted my analysis to 1-1 orthologs, although I expect that paralogs will be more evolvable than 1-1 orthologs; downstream analysis of close paralogs using available experimental techniques (e.g. ChIP-seq) would be much more difficult.

For each domain family, I constructed a matrix with rows representing binding domains and columns consisting of pairwise rates for each species. Rates for each family were ranked in each species, and the sequence alignments of the top hits examined manually for changes with potential functional relevance. Figure Fig. 2.2 shows the distribution of median pairwise AA substitution rates (relative to *Drosophila melanogaster*) for each of the 49 TF families in four representative fly species. In *D. simulans* (Fig. 2.2, upper left), almost all domain families are conserved, as expected for a species so closely related to *D. mel* . Examining more divergent species, many families begin to accumulate more changes (*D. ananassae* and *D. pseudoobscura*, Fig. 2.2, upper right and lower left). In *D. grimshawi* (Fig. 2.2, lower right), many families contain changes, but a small set (9 families) are still highly conserved. Clearly, many domains from these species contain changes; however, in almost all cases these fall within regions that have no known functional relevance (i.e. not in DNA-contacting residues, known protein-protein interaction sites, or regions important for structural integrity of the domain). As it is generally not possible to predict functional residues based only on sequence, in most cases I matched the changes manually against structural information about the domain from the literature, and found that they occured in regions which had no known role in binding or domain architecture.

2.4.2 Some *Drosophila* TF DNA-binding Domains Show Evidence of Important Changes Outside of DNA-contacting Residues

Most factors I examined were highly conserved, but I was able to identify a few cases with potentially significant changes within their DNA-binding domains, though almost all of these were not in known DNA-contacting positions. These included *kayak*(*kay*), a basic leucine zipper (bZIP) protein that is the *Drosophila* Jun homolog Fig. 2.4 [276]. This protein is among the most divergent bZIP domains in my dataset (Fig. 2.3) Although there are no significant changes within the DNA-contacting region of the domain (Fig. 2.4, blue box), there are a number of substitutions in the N-terminal region of the domain (Fig. 2.4, magenta box). This is the alpha-helical "zipper" region, which is responsible for specific homo- and heterodimerization of these proteins [277]. Specifically, *kay* shows a number of lineage-specific substitutions at the 'e' and 'g' positions of the helix (Fig. 2.4, orange stars), which interact with each other by steric forces in the coiled-coil of the dimerized transcription factor [143]. This may imply that *kay* has altered dimerization properties in different *Drosophila* species.

Some other potentially interesting examples involve changes in residues within binding domains that may be important for the structural integrity of the DNA-binding domain. One example of this is the *Drosophila* homolog of p53 [278], which contains a substitution at R395 to cysteine in *Drosophila pseudoobscura* and *Drosophila persimilis* and histidine in *Drosophila ananassae*. This position corresponds to R282 in human p53 [278], a buried residue that is thought be important for maintaining structural stability of the domain; this position is also one of the most frequently mutated residues in human cancers [279]. The GATA factor *srp* [280] also contains substitutions in a number of species at position 14 of its C-terminal zinc-finger domain, which is important for structural stability in the chicken GATA factor cGATA-1 [281]. Although the effects of structural changes on DNA-binding behavior of the domain are difficult to determine, it is possible that the changes I observe in these two proteins could influence the shape of the DNA-binding interface of the domain, and therefore binding specificity or affinity.

2.4.3 The Putative DNA-binding Domain of the Gene *hermaphrodite* is Highly Divergent

Because C_2H_2 -zinc finger proteins are known to have some unique evolutionary properties, including changes which alter the number and function of arrays of DNA-binding domains [132,256], I gave special attention to proteins from this family. Figure Fig. 2.5 shows an example distribution of pairwise AA substitution rates for C_2H_2 -zinc finger domains from *D. pseudoobscura*. I identified a number of C_2H_2 zinc fingers with elevated rates of evolution within their binding domains; however, most of these did not involve changes in the canonical DNA-contacting positions (-1,2,3,6 in the C_2H_2 α-helix) [26]. Of those which did contain potentially specificity-altering changes, most were unnamed genes with little information available regarding possible functions. I ultimately chose to examine the gene *hermaphrodite* (*her*), a 4-finger C_2H_2 zinc finger because in addition to having some dramatic changes in its DNA binding domain (Fig. 2.6, Table 2.1), it had been studied previously and implicated as having a role in the core somatic sex determination path-

way [265–268]. The discovery of divergent transcription factor in the sex determination is intriguing, as the elements of this pathway [282–284], as well as a variety of sex-specific traits [35, 285] are known to be quite variable across insect species.

The DNA-binding domain of *her* contains 4 C_2H_2 zinc fingers near its N-terminus that appear to be arranged in an array (Fig. 2.6), although they do not contain the typical TGEKP linker regions observed in other C_2H_2 zinc fingers such as *Kruppel* and Zif268 [26]. Within species of the *melanogaster* group (*D. melanogaster*,*D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta* and *D. ananassae*), the *her* binding domain shows a large number of changes at non-contacting residues (Fig. 2.6, green shaded residues) and a Y→F substitution in *D. yakuba* and *D. erecta*. In species of the *obscura* group (*D. pseudoobscura* and *D. persimilis*, an even larger number of substitutions are observed outside of DNA-contacting residues in all four fingers, and the fourth zinc finger domain contains an A→G substitution at the 2 position in the binding α-helix. All of these species share a well-conserved region near the C-terminus (Fig. 2.6) that does not resemble a known domain. It is possible that this region may function as a transcriptional effector or other type of protein-protein interaction domain.

In addition, the *obscura* species *her* genes contain a $>$ 800bp insertion just 3' of the DNA binding domain. This region contains 3 additional C_2H_2 zinc finger domains (Fig. 2.6), with longer stretches of intervening sequence. Examination of the DNA sequences by dot plot shows that this insertion is most likely the result of a 3-fold tandem duplication of a region including the fourth shared zinc finger in the DNA-binding domain (Fig. 2.7). Although it is known that the number of zinc fingers can vary widely across proteins, cases of multiple duplication of this type are rare, and usually happen only in paralogs [132]. *her* is the only example I was able to identify in flies, although it is possible that others may exist since my analysis was restricted to 1-1 orthologs.

The only species with a clear ortholog for *her* outside of the sophophora (which includes the *melanogaster* and *obscura* groups) is *Drosophila willistoni*. The *her* ortholog in this species shows an even greater degree of divergence within its four zinc finger domains (Fig. 2.6), including an S→T substitution at the 2 position of finger 2, and an F→Y substitution at position 2 in finger 3. *Drosophila willistoni her* also contains a fifth zinc finger in the region 3' of the shared DNA-binding domain; although this finger aligns with the N-terminal most *obscura*-specific finger, it is not clear if the two are orthologous. In addition, the C-terminal region that is conserved in the sophophora does not appear to be present in *Drosophila willistoni hermaphrodite* .

Hermaphrodite does not appear to be conserved in the *Drosophila* genus in species more divergent from *Drosophila melanogaster* than *Drosophila willistoni*, or in any other metazoan clade. The fuzzy reciprocal BLAST approach used to call orthologs in the 12 *Drosophila* genomes [264] returns a protein that matches *her* only to the conserved structural positions of the C_2H_2 zinc fingers, and is located on a different chromosome; this may represent an extremely divergent *her* or an unrelated protein that happens to have C_2H_2 zinc finger motifs. In addition, although synteny around the *her* locus appears to be conserved at least to *Drosophila grimshawi*, the region including the *her* coding sequence shows poor sequence conservation beyond *Drosophila willistoni* (Fig. 2.8). Analysis of this region from *Drosophila virilis*, *Drosophila mojavensis*, and *Drosophila grimshawi* using the

computational gene-finding tool GeneWise [286] fails to find any potential coding sequences.

It therefore appears that the *her* gene has undergone a number of potentially interesting rearrangements of its DNA-binding domain during evolution of the *Drosophila*. Although sequence data from more species will be needed before a detailed picture of the evolution of *her* can be constructed, a speculative model is as follows. The *her* gene most likely arose by duplication of an unknown zinc finger protein in a common ancestor of both *Drosophila willistoni* and the Sophophora species. This protein probably had either 4 or 5 zinc finger domains, and lacked the C-terminal conserved domain observed in the Sophophora. The C-terminal domain was likely acquired in a common ancestor of the Sophophora species; one possibility is that the acquisition of this domain allowed *her* to gain transcription regulatory function. The extra zinc fingers observed in *Drosophila pseudoobscura* and *Drosophila persimilis* then likely arose by duplication following the divergence of the *obscura* and *drosophila* groups.

The variety of changes in the *her* binding domain that may have consequences for DNA-binding of *her* made it the best candidate for a detailed evolutionary and functional analysis among all of the hundreds of transcription factors examined. I ultimately chose to explore in particular the differences between the *Drosophila melanogaster* and *Drosophila pseudoobscura her* genes, since these show the most drastic differences in binding domain structure within the shortest amount of evolutionary time. In chapters 3 and 4 I will examine the DNA binding behavior of *her* in these two species experimentally.

Table 2.1: Pairwise Amino Acid substitution rates for conserved *her* C_2H_2 -zinc finger domains

Domain	Dsim	Dsec	Dyak	Dere	Dana	Dpse
CG4694-PA her zf-C2H2 dom0	0.041072	0.041072	0.123344	0.123446	0.383365	0.717898
CG4694-PA her zf-C2H2 dom1	0.101895	0.101895	0.129704	0.21039	0.467387	0.758517
CG4694-PA her zf-C2H2 dom2	8e-06	8e-06	0.08373	0.08373	0.2346	0.495754
CG4694-PA her zf-C2H2 dom3	8e-06	8e-06	8e-06	8e-06	0.131828	0.283962

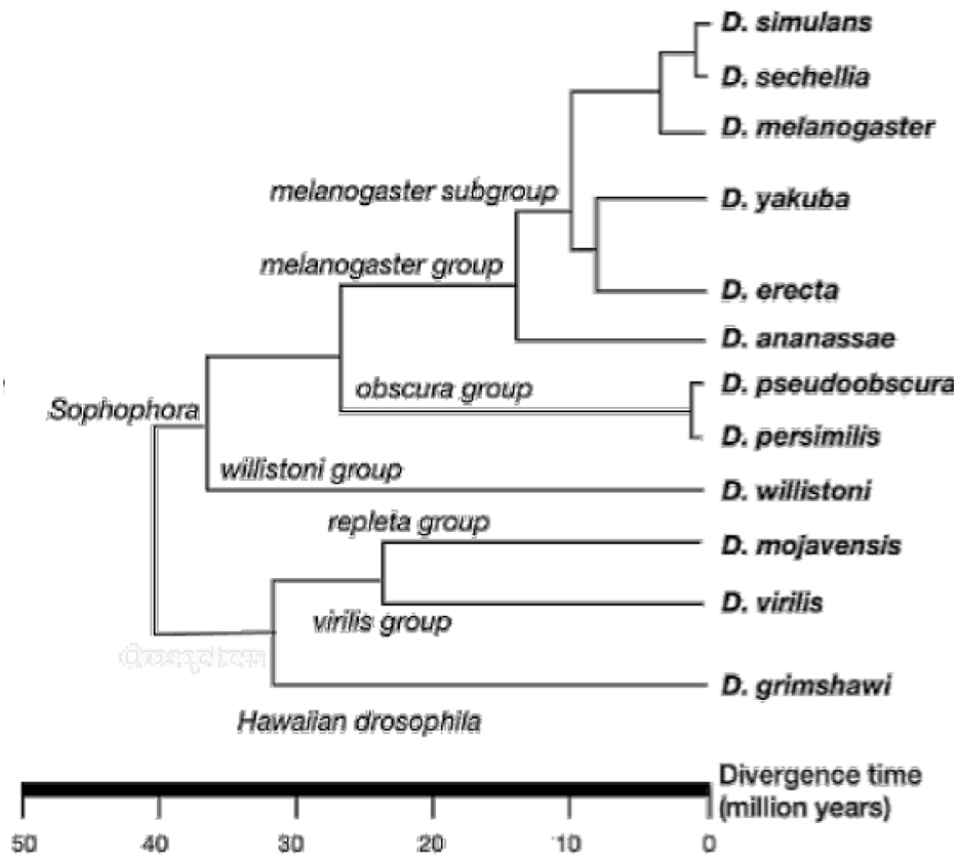


Figure 2.1: **Phylogeny of the 12 *Drosophila* Species Examined in this Study:** Figure adapted from FlyBase (http://www.flybase.org).

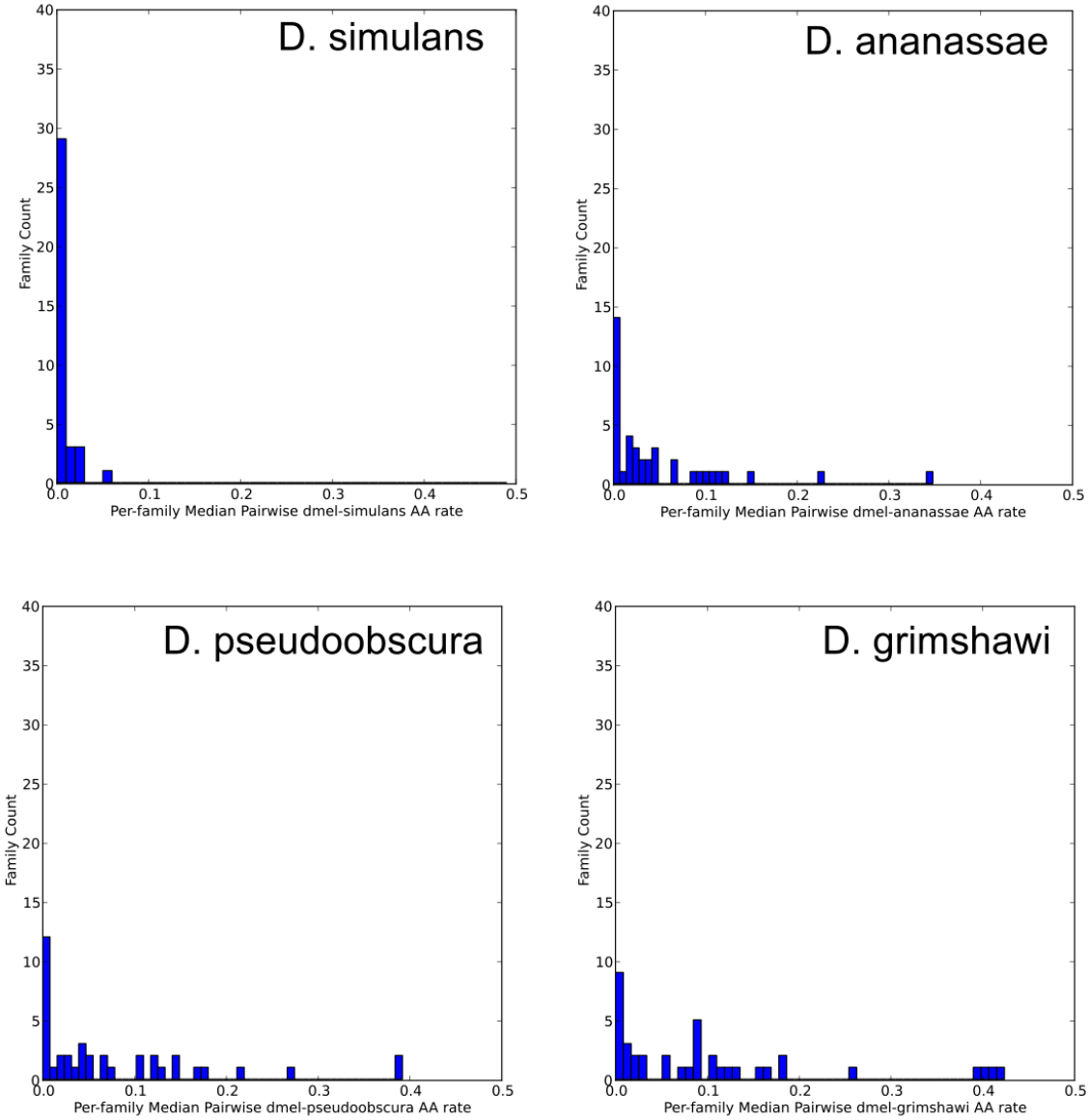


Figure 2.2: **Median pairwise AA substitution rates for TF Families:** Plots show histograms of median AA substitution rates calculated for the set of proteins in each TF family. In *D. simulans* (top left), the majority of families have no substitutions, as indicated by the mass of density at 0; in contrast, for *D. grimshawi* (lower right), the most divergent species examined, almost all families have accumulated some substitutions, although several families are still at 0

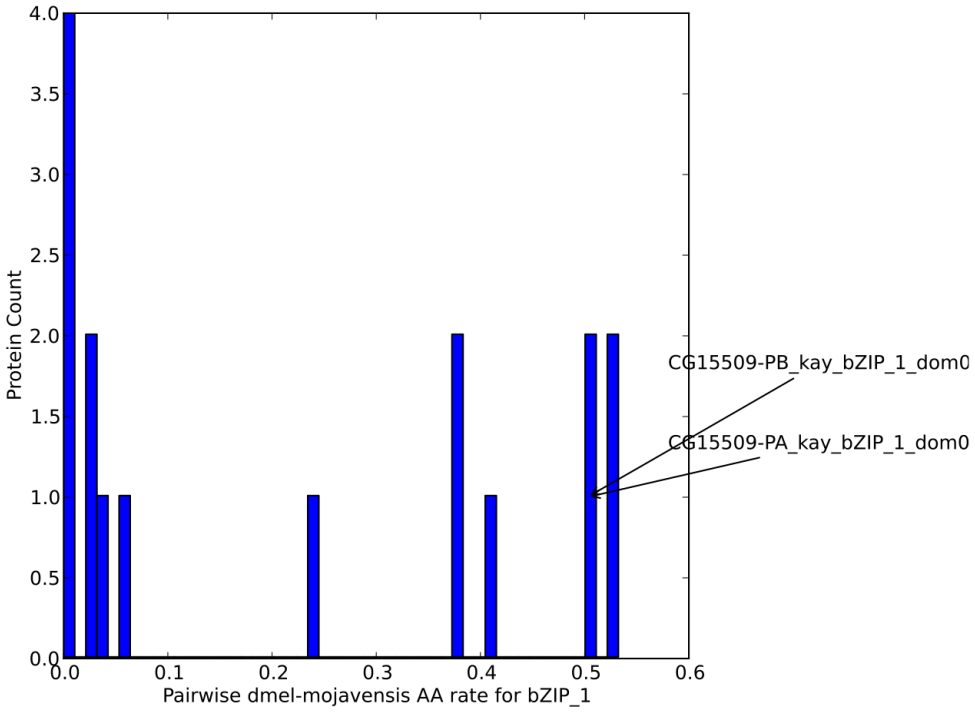


Figure 2.3: **Pairwise AA Substitution Rates for bZIP-1 Domains in *D. mojavensis*:** Plot shows the histogram of pairwise AA substitution rates (calculated using the codeml program from the PAML package [275]) between orthologous domains from *D. mojavensis* and *Drosophila melanogaster* . *D. moj* domains orthologous to the *kay* bZIP domain is indicated on the plot.

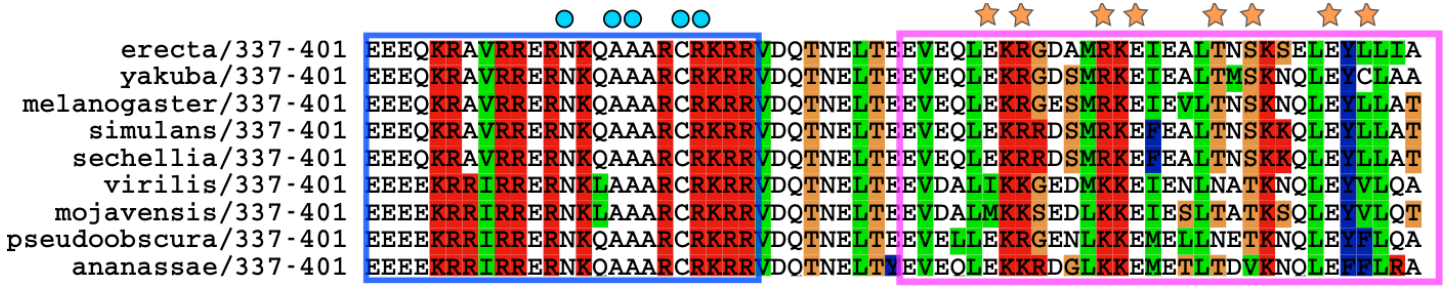


Figure 2.4: **Alignment of the *kayak* (dFos) bZIP DNA-binding domain in 9 *Drosophila* species:** Amino acid alignment generated using MUSCLE [274] of the region corresponding to the *Drosophila melanogaster* DNA binding domain as called by HMMER [272]. Blue box shows the basic region, magenta box shows the leucine zipper region. Aqua circles indicate DNA-contacting residues; orange stars indicate the 'e' and 'g' positions of the zipper alpha helix, which interact to determine dimerization specificity.

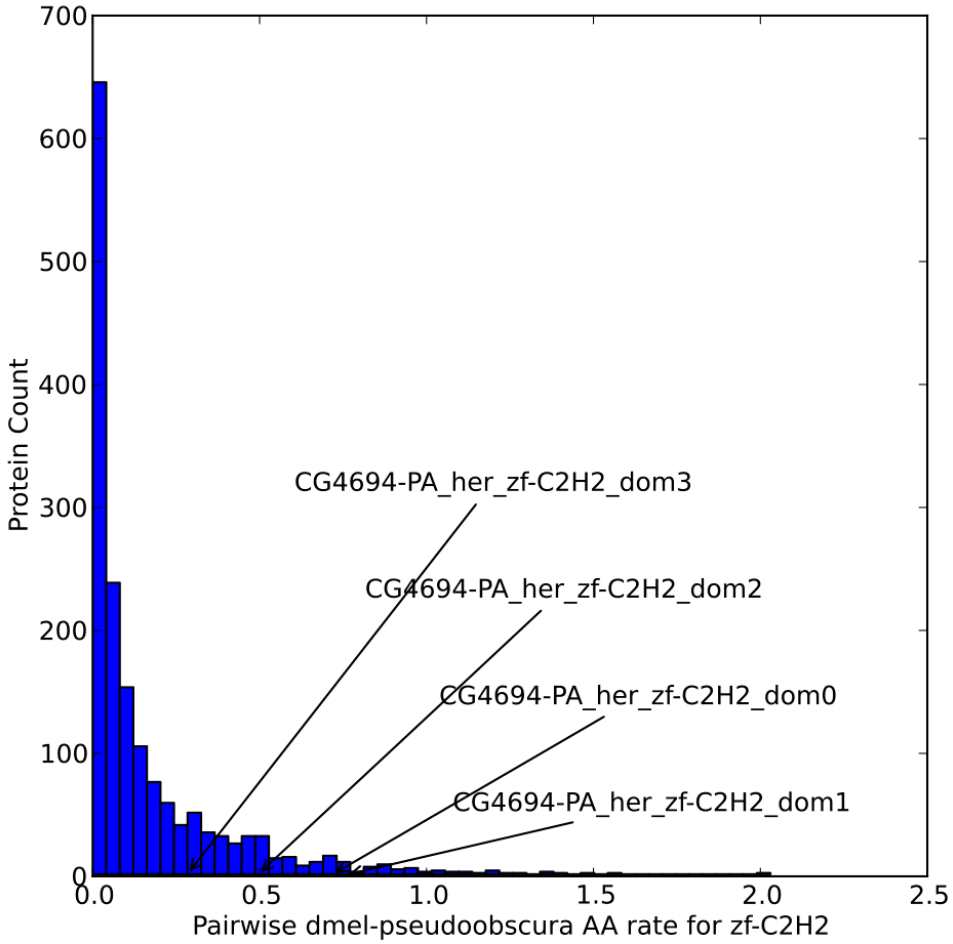


Figure 2.5: **Pairwise Amino Acid Rates for C_2H_2 -zinc Finger Domains Between *Drosophila melanogaster* and *Drosophila pseudoobscura* :** Plot shows the histogram of pairwise AA substitution rates (calculated using the codeml program from the PAML package [275]) between orthologous domains from *Drosophila pseudoobscura* and *Drosophila melanogaster* . *D. pse* domains orthologous to the four *her* C_2H_2 -zf domains are indicated on the plot. *her* domain 1 shows the highest rates of the four fingers.

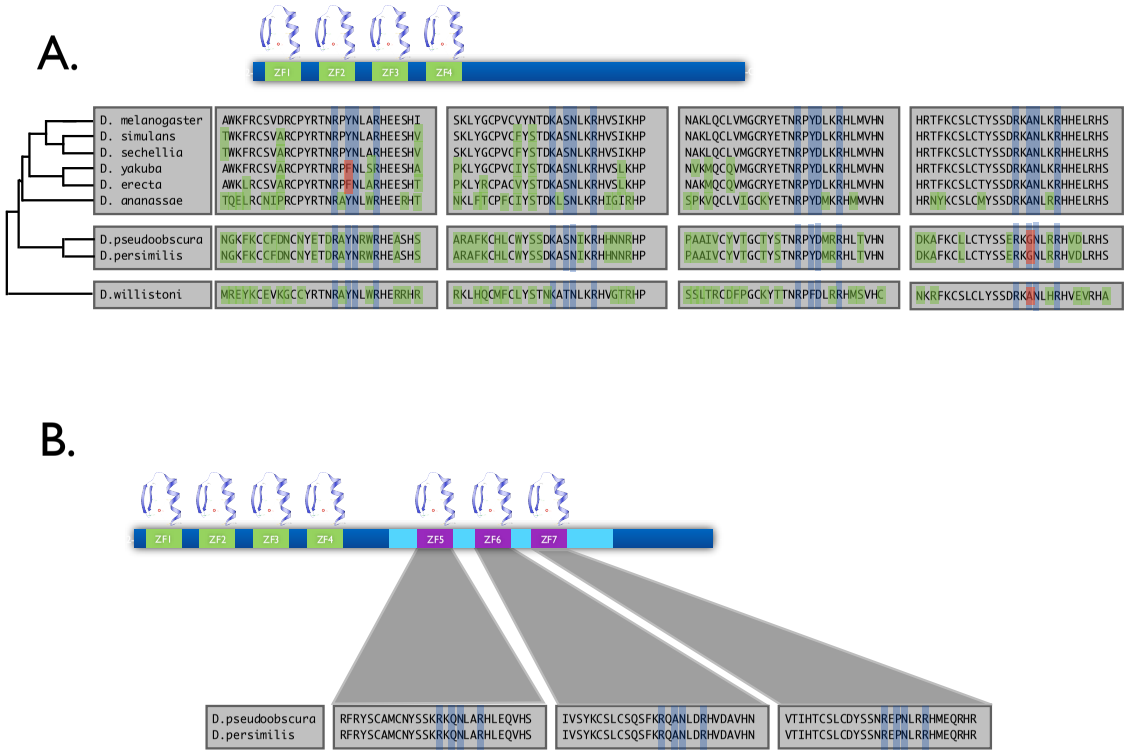


Figure 2.6: **Sequence Evolution in the *her* DNA-binding Domain:**(**A**.) Amino acid sequence alignments of the 4 shared C2H2 zinc fingers of the *her* DNA-binding domain from 9 Drosophila species. Alignments generated using MUSCLE. Cartoon at top shows position of domains near N-terminus of the protein. In alignments, blue shaded residues are putative DNA-contacting residues [26], green shaded residues represent changes relative to *Drosophila melanogaster* , red shaded residues represent changes relative to *Drosophila melanogaster* in DNA-contacting residues. Cladogram at left not to scale. (**B**) C2H2-zinc finger motifs present in a large insertion specific to *Drosophila pseudoobscura* and *Drosophila persimilis*. Cartoon shows position of insertion (blue) and fingers (purple) relative to the 4 shared zinc fingers at left. Sequences are alignments of the C2H2 zinc finger domains between *Drosophila pseudoobscura* and *Drosophila persimilis*.

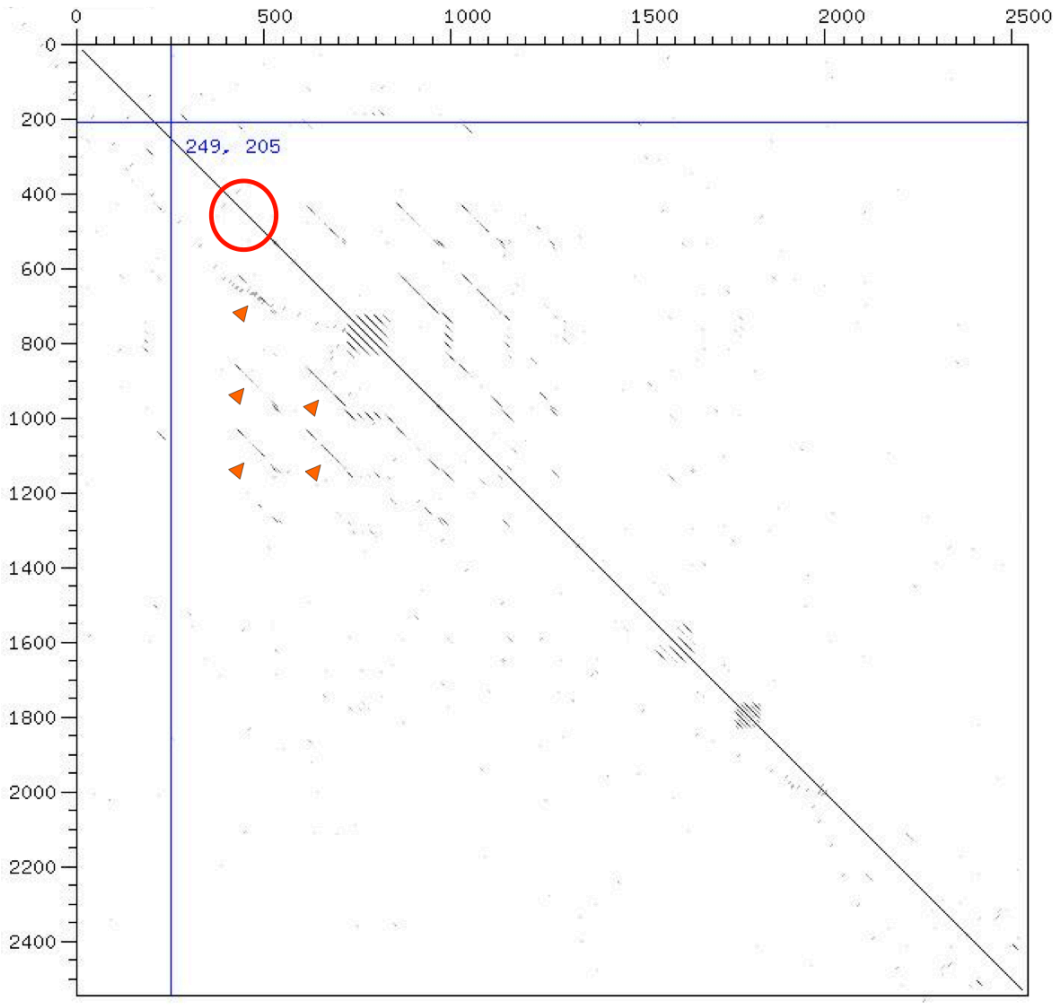


Figure 2.7: **Dot Plot of self vs. self for *Drosophila pseudoobscura her* , Showing Duplications of the 4th Zinc Finger:** Shows a dot plot of the *Drosophila pseudoobscura her* inserted region plotted against itself to show repetitive sequences. Red circle denotes the position of the 4th zinc finger (shared among all 9 species); orange triangles point to upper-left to lower-right diagonal lines, indicating a match (repeated sequence). The three tandem repeats correspond approximately with the positions of the inserted *Drosophila pseudoobscura* zinc fingers. Plot created using Dotter [287].

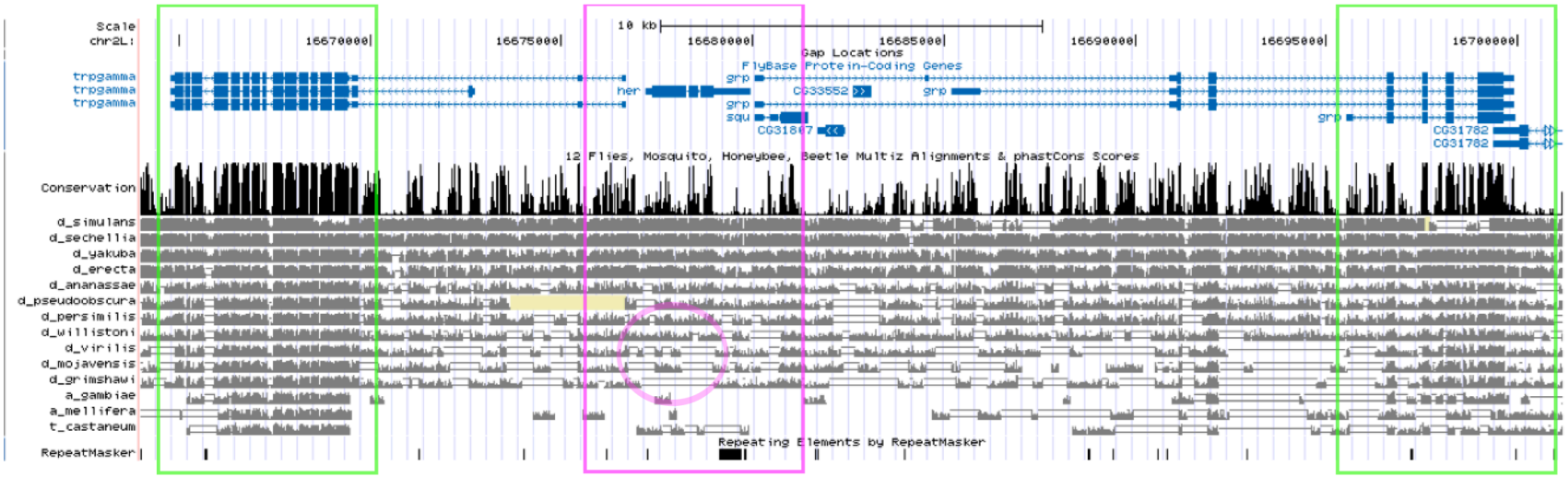


Figure 2.8: **Conservation of *her* Locus in *Drosophila*:** Browser Snapshot generated from the UCSC Genome Browser [288] showing conservation of genes flanking the *her* locus. Highly conserved flanking genes *trp-gamma* (left) and *grp* (right) are highlighted with green boxes. The *her* locus is shown with a magenta box; magenta circle highlights the low level of conservation in *Drosophila virilis*, *Drosophila mojavensis*, and *Drosophila grimshawi*. Bottom tracks are conservation score for each species. Note that synteny is conserved for all species except *Drosophila mojavensis*.

Chapter 3

Analysis of DNA Binding by the HER protein in *Drosophila melanogaster*

3.1 Abstract

Sex determination in *Drosophila melanogaster* takes place through a signaling cascade involving both transcriptional and post-transcriptional regulation. While most of the components of this pathway have been extensively characterized at both the genetic and molecular level, the molecular function of the *hermaphrodite* gene, a putative zinc-finger transcription factor proposed to play a role in specifying sexual fate in somatic tissues along with the transcription factor *Doublesex*, has remained mostly unknown. Using Chromatin Immunoprecipitation combined with high-throughput Illumina sequencing (ChIP-seq), I assayed the *in vivo* genome-wide binding of *hermaphrodite* in *Drosophila melanogaster* 4-6h embryos. These data show that the Hermaphrodite protein binds strongly to over 300 sites in the fly genome, most of which are located in intronic and intergenic regions close to annotated transcription starts, and are enriched with a previously unknown 15-bp sequence motif. In addition, I find that Hermaphrodite binds in the proximal promoter region of the sex determination gene *intersex* (*ix*). These data strongly support a role for *hermaphrodite* as a sequence-specific regulator of transcription, and indicate that its role in sex determination may be indirect.

3.2 Introduction

Many metazoan species use genetic signals to determine their phenotypic sex, but the nature of these genetic signals, and the mechanisms by which they are translated into primary and secondary sexual characteristics, show a remarkable degree of variability across even closely related taxa. In humans and other therian mammals, the genetic signal is the presence or absence of a specific locus normally carried on the Y chromosome (*Sry* [289]), which is sufficient to determine whether the embryo will adopt a male or a female fate respectively. In the fruitfly *Drosophila melanogaster* , on the other hand, the Y chromosome plays no role in sex determination. Instead, the embryo determines its sex early in development by counting the number of X chromosomes present in each cell; individuals with two X chromosomes are phenotypically female, while those with a single X become phenotypically male. Counting is accomplished through a set of genes on the X chromosome, the X-encoded signal element (XSE) genes [290–292]. These genes are expressed early in development, and when two copies of the X chromosome are present, XSE gene dosage is sufficient to activate a female-specific promoter region (P_e) of the master sex regulator *Sex lethal* (*Sxl*) [290, 291].

Transcription of *Sxl* from the P_e promoter produces a female-specific *Sxl* mRNA, encoding an RNA-binding protein which guides a set of regulated alternative splicing events [290, 291]. After the counting signal from the XSE genes is lost, a second *Sxl* promoter, P_M, is activated non-sex-specifically. The transcript from P_M requires autoregulatory *Sxl* splicing activity to remove a cassette exon containing a premature stop. Consequently, only females, with active SXL protein produced from P_e, are able to produce functional transcripts from P_M [290, 291]. SXL also directs female-specific splicing of *male-specific lethal - 2* (*msl-2*), which regulates dosage compensation, and *transformer*(*tra*) [290, 291]. The female isoform of *tra* encodes another RNA-binding splicing regulator, which directs

further female-specific splicing events in the transcription factors *fruitless*, which controls sex-specific behavior, and *Doublesex*(*Dsx*), which is responsible for most aspects of somatic sex differentiation [290, 291].

In genetically male individuals with a single X chromosome, dosage of the XSE genes is insufficient to drive transcription from P_e. When the XSE counting signal is lost, transcription from P_M begins, but because males do not have functional SXL, this transcript does not undergo the autoregulatory splicing event to remove the cassette exons containing the premature stop [290,291]. In the absence of regulated splicing from the SXL protein, *tra* mRNA is spliced in the default pattern, which includes a premature stop. The absence of functional *tra* in turn results in production of the default (male) isoform of *Dsx* [290, 291].

The male- and female- specific isoforms of *Dsx*, DSX^M and DSX^F, exert opposing transcriptional regulatory effects on target genes. DSX^M and DSX^F proteins contain identical DMRT DNA-binding domains, and show similar DNA binding preferences *in vitro* [293, 294]. Sex-specific differences in DSX transcriptional regulation are thought to take place through a C-terminal domain, which differs between the two sex-specific isoforms and likely participates in protein-protein interactions with transcriptional cofactors [295]. The canonical example of sex-specific regulation by DSX is the fat body enhancer (FBE) of the *Yp1* and *Yp2* genes [296]. In this enhancer, DSX^M and DSX^F both act through the same DNA binding site, but DSX^F is able to function as an activator by recruiting the cofactor *intersex* (*ix*), while DSX^M does not recruit *ix* and represses transcription through an unknown mechanism. Because *ix* mutants show a general intersexual phenotype in females similar to that seen in *Dsx* mutants, *ix* is likely to play a role in the activation of many other DSX^F targets. However, since some targets of DSX are activated in males and repressed in females (e.g. *Wnt2* and *bnl* in the genital imaginal disc [297, 298] it is likely that other cofactors are also involved in regulating transcription with *Dsx*.

The gene *hermaphrodite* (*her*) has been proposed to play two distinct roles in the sex determination pathway. The first of these is a maternal effect, upstream of *Sxl*, which results in a female specific reduction in viability in offspring of her^- mothers [266,268]. This *her* -mediated lethality is enhanced when combined with mutations in the XSE gene *sis-a* or *Sxl* loss-of-function mutants, and can be almost completely rescued by a constitutively expressed allele of *Sxl* [266]. In addition, female offspring of her^- mothers show a loss of expression from *Sxl* P_e promoter [266]. These results suggest that *hermaphrodite* may be necessary for initial transcriptional readout of the chromosomal sex signal that leads to *Sxl* expression in females, although more recent experiments have been unable to replicate these results observed in the original studies [299], indicating that these effects may be due to secondary effects from the mixed genetic background of the crosses.

In addition to its role at the top of the sex-determination signaling cascade, *hermaphrodite* has also been shown to act further downstream in the pathway, in the somatic sex determination branch mediated by *Dsx*. The most pronounced effects are seen in females, where the zygotic expression of certain combinations of hypomorphic *her* alleles results in a strong intersexual phenotype [266, 268]. This includes male-like pigmentation of the anterior portion of the fifth and sixth abdominal segments A5 and A6, partial masculinization of genital and anal structures, and partial rotation of bristles on the metatarsus of the foreleg analogous to the male sex combs [265, 266, 268]. Male *her* mutants show a

much weaker intersexual phenotype, which manifests in the presence of extra bristles on the sixth sternite (S6) [265, 266, 268]. Analysis of *dsx*/*dsx* ; *her*/*her* double mutants shows that some of these phenotypic effects are dependent on the presence of functional DSX (foreleg bristle rotation and A5/A6 pigmentation in females, S6 bristles in males). Other *her* mutant effects are not *dsx* dependent, because these effects are still visible in *dsx*/*dsx* intersex individuals (e.g. lateral anal plate fusion and vaginal tooth number) [265].

The sex-specific effects of *her* can only be observed in weaker alleles under semi-restrictive conditions; with stronger alleles and more restrictive conditions (higher rearing temperature), *her* mutants show non-sex-specific maternal and zygotic lethality, indicating a broader role for *her* outside of the sex determination pathway [266, 268]. *hermaphrodite* has also been shown to activate the *Yp* genes non-sex-specifically in a mostly *Dsx*-independent manner [267]. This effect, combined with the observation that *her* contains four C2H2 zinc finger domains, suggests that *her* functions as a sequence-specific transcription factor, although neither DNA binding nor transcriptional regulatory activity has been shown directly.

In this work I examine the DNA binding behavior of *her* using chromatin immunoprecipitation combined with high-throughput Illumina sequencing (ChIP-seq), with the goals of 1) determining whether the protein encoded by *her* functions as a sequence-specific transcriptional regulator and 2) gaining insight into the sex-specific and non-sex-specific molecular function of *her* by examining its set of target genes. This will provide insight into a previously understudied component of the *Drosophila* sex determination hierarchy, and will inform the inquiries into the evolutionary properties of *her* (see Ch. 2 and 4).

3.3 Materials and Methods

3.3.1 Protein Expression and Purification

her coding sequences for protein expression constructs were amplified by PCR from the full-length *her* cDNA. Full-length *her* coding sequence was amplified using primers ATGCTTAGTGCGGATCGGGA (her_fl_F) and TCAAGTTTGACTTAACGTGCC (her_fl_R); *her* N-terminal peptide (amino acids 1-250) was amplified using primers her_fl_F and TCGTATTTCATACTGCCTACC (her_Nt_R). C-terminal peptide (amino acids 251- 487) was amplified using primers GAGGAAAGTTTGGAATTTGTTT (her_Ct_F) and her_fl_R. BLAST [300] showed that these sequences contained no regions of homology > 8 to other proteins.

All sequences were cloned into the Invitrogen Gateway entry vector pDONR using BP-recombinase, then transferred to the Gateway pDEST17 vector, which contains an N-terminal 6xHis tag, for expression. Expression and lysis were carried out as in the QIAExpressionist Manual (QIAgen); BL21-AI cells (Invitrogen) containing expression constructs were grown to $OD_800 = 0.6$ and induced for 4h with 0.2% arabinose. Lysis was performed in 50mM HEPES buffer containing 8M Urea and 10mM Imidazole. Cleared lysate was bound to Ni-NT agarose (Qiagen), washed with 50mM HEPES with 4M Urea and 20mM Imidazole, and eluted in 50mM HEPES containing 4M Urea and 250mM imidazole.

3.3.2 Antibody Purification

Antisera were raised in rabbits against the full-length *Drosophila melanogaster* HER peptide at the Pocono Rabbit Farm and Laboratory. Antisera were affinity purified against either the N-terminal (herNt) or C-terminal (herCt) portions of the HER protein to generate two independent antibodies. Affinity purification was carried out using 1-3mg of herNt or herCt peptides bound to 333mg (dry weight) of Activated CNBr-sepharose 4B. Antisera were diluted in 10mM Tris, filtered, and passed over the affinity resin twice. Resin was washed with 20CV 10mM Tris, pH7.5, 20CV 10mM Tris pH7.5 with 1M NaCl. Antibodies were eluted with 3CV 4M $MgCl_2$ 0.25M glycine pH3.5 0.1% Tween-20, followed by 0.1M Glycine pH 2.5 0.1% Tween20. pH was adjusted to 7.5, and eluate was concentrated and resuspended into 10mM Tris pH7.5, and concentrated to approximately 3 mg/ml.

3.3.3 Chromatin Immunoprecipitation

Chromatin was prepared from 4-6h *Drosophila melanogaster* (Canton-S wild type) embryos raised at 25 °C. Approximately 5g staged embryos were fixed for 5min in 50ml hexane saturated with 10xPBS and 5% formaldehyde, then washed twice for $>$ 5min with 1xPBS + .5% Triton X-100 and flash frozen. Purified chromatin was prepared from fixed embryos as in [66]. Purified chromatin was fragmented by sonication using 8 cycles of 30 seconds each at setting 2 on a Sonifier 250 (Branson), yielding an average fragment size of 350bp.

Chromatin immunoprecipitation was carried out using 100μg of purified, fragmented chromatin. Chromatin was diluted into IP buffer (10mM Tris pH 8,1mM EDTA, 150mM NaCl, 0.5% Triton X-100, 0.1% Sodium Deoxycholate, 0.5% Sarkosyl) and cleared using 3μg of Normal Rabbit IgG (Santa Cruz Biotechnology) bound to Protein-A sephacryl beads prepared in-house (see [66] for protocol). 2% of total sample was saved for use as the input control. Pre-cleared chromatin was then incubated overnight at 4 °C with 3μg of herNt or herCt antibody or 3μg of normal rabbit IgG (Santa Cruz Biotechnology) and bound to Protein-A sephacryl prepared as above. After binding, the following washes were performed: 1) IP buffer for 5min 2) IP Buffer for 10min 3) two washes of 0.5x IP Buffer for 20min ea 4) LiCl Buffer (10mM Tris,1mM EDTA, 250mM LiCl, 1%NP-40, 1% Sodium Deoxycholate) for 20min 5) TE for 10min. Samples were treated with 1μL RNAse A (Fermentas) for 30 min, then with 1.8μL Proteinase K (Fermentas) at 55 °C overnight. Samples were then incubated at 65 °C for 6 hours to reverse formaldehyde crosslinks. De-crosslinked samples were eluted with TE, extracted with phenol-chloroform, and resuspended in TE.

3.3.4 Immunoprecipitation and Western Blots

Immunoprecipitation for antibody tests was performed on nuclear extracts prepared from 10g 0-12h *Drosophila melanogaster* embryos (extract kindly provided by Malik Francis). 250μL of extract was incubated with 3μg of her antibody (herNt or herCt) or 3μg normal rabbit IgG at 4 °C for 1hr. Samples were then bound to Protein A-sephacryl beads (prepared as above) and washed 3 times with IP was buffer (25mM HEPES pH7.6, 150mM KCl, 1mM EDTA, 1mM EGTA, 10%Glycerol, 1mM DTT, 0.4mM PMSF). After

washes, the supernatant was removed and beads were boiled for 7min with SDS-PAGE sample buffer and loaded directly onto 4-12% PAGE gel (Biorad, Inc).

Gels were transferred to nitrocellulose paper, blocked with Infrared Western Blocking Buffer (LiCor Biosciences), and incubated with 1:1000 diluted primary antibody. For IPs performed with herNt, herCt was used as the primary; for herNt IPs, herCt was used as the primary; for IgG IPs, herNt was used as the primary. All blots were incubated with IRDye-700 Protein-A secondary (Rockland Immunochemicals) and imaged using an Odyssey infrared scanner (LiCor Biosciences).

3.3.5 Illumina Library Preparation and Sequencing

Illumina Library preparation followed the protocol outlined in the Illumina ChIP-seq Sample Preparation handbook (Illumina). ChIP samples in TE were end-repaired using T4 DNA Polymerase, T4 Polynucleotide Kinase, and Klenow Polymerase (New England Biolabs), and A-tailed with (3'->5' exo$^-$) Klenow (New England Biolabs). Illumina single-end adaptors were ligated using Quick T4 DNA Ligase (New England Biolabs). Library PCR amplification was performed using cycle parameters outlined in Illumina ChIP-seq Sample Preparation handbook (Illumina) for 16 cycles, followed by size selection in 2% agarose gels for sequences 200-500bp in length. DNA purification in between steps was carried out using either QIAquick PCR Cleanup Kit (Qiagen) or QIAquick MinElute PCR Cleanup Kit (Qiagen). Purification following size selection was carried out using QIAquick MinElute Gel Extraction Kit (Qiagen). Libraries were quantitated using either a Bioanalyzer 2100 (Agilent Technologies) or quantitative PCR using a similarly sized library as a standard.

3.3.6 Data Analysis

Samples were sequenced using a Genome Analyzer 2 (Illumina, Inc). Image analysis and base calling were performed using the Illumina Analysis Pipeline version 1.3. Read mapping was performed using either Eland from Illumina Pipeline v1.3 or Bowtie [301] to the *Drosophila melanogaster* genome version 5.18. Peak calling was performed using MACS [302]. All other analysis performed using custom scripts written in Python.

3.4 Results

3.4.1 ChIP-seq analysis shows that *her* binds DNA

In order to determine the DNA-binding behavior of the HER protein, I performed ChIP-seq analysis using two antibodies prepared from *her* rabbit antiserum affinity-purified against either the N-terminal (herNt) or C-terminal (herCt) portion of the *her* protein. A control for nonspecific antibody binding (normal rabbit IgG) was included, as well as an input control. I chose to use 4-6h embryos for this analysis as previous work showed the peak of zygotic *her* expression occurred during this time period [267].

Illumina sequencing generated millions of reads for each sample as shown in Table 3.1. Reads were mapped to version 5.18 of the *Drosophila melanogaster* genome using Bowtie [301], with variable fractions of mapping reads shown in Table 3.1. Reads that

mapped to more than two positions in the genome were discarded to eliminate repetetive sequences from further analysis. Both herNt and herCt samples showed significant enrichment over the input control sample Fig. 3.2, Fig. 3.1. In contrast, IgG control samples showed no significant enrichment over background (data not shown). Excellent agreement was observed between herNt and herCt samples (Fig. 3.3) with normalized tag counts between the two samples for 100-bp windows giving $R^2 = 0.97$.

In order to confirm that the herNt and herCt antibodies were specifically recognizing the HER protein *in vivo*, I also performed immunoprecipitation followed by western blots on 0-12h embryo nuclear extracts. Immunoprecipitation with the herCt antibody followed by blotting with the herNt antibody showed enrichment of an approximately 60kD protein, close to the predicted size for the HER protein of 57kD (Fig. 3.4, center panel). Similar results were obtained for the converse experiment in which I immunoprecipitated using the herNt antibody and blotted using the herCt antibody (Fig. 3.4, left panel). Immunoprecipitation with Normal Rabbit IgG and probing with herNt as a control showed no enrichment (Fig. 3.4, right panel).

3.4.2 Distribution of HER-bound peaks is consistent with a role for *her* in transcriptional regulation

Peaks of *her* binding were identified from mapped ChIP-seq reads using MACS [302], using input reads as a negative control. MACS called 1484 total peaks for the herNt sample, and 1412 peaks for the herCt sample. Slight differences in the signal to noise ratio between the two samples meant that the false discovery rate (FDR) reported by MACS differed between samples for overlapping peaks. In order to choose a set of high-confidence peaks for further analysis, I identified overlapping peaks for the two IP samples at 1% , 2%, and 5% FDR and chose to use the set of overlapping peaks with FDR $< 5\%$ from the herCt set and $< 1\%$ FDR from the herNt set as it gave the best agreement between the two samples. This gave a total of 302 "high confidence" peaks.

I next examined distribution of these peaks relative to structural features of genes (Fig. 3.5 A). Most peaks fall either in intergenic regions (132) or within introns (87), with smaller numbers occuring in 5' untranslated regions (40) and coding sequences (41) and few (2) occuring in 3' untranslated regions. In addition, I calculated the distribution of distances from each peak to the nearest annotated mRNA transcription start site (Fig. 3.5 B). 67% of peaks occured within 1000bp of an annotated transcription start, and the peak of this distribution occurs near 0. These data are consistent with patterns observed for most transcriptional regulators examined by ChIP-Chip [66, 67], although the preference for the basal promoter region observed for HER is unusual. Taken together, I believe this pattern supports a role for *her* as transcriptional regulator.

3.4.3 HER peaks show enrichment of a novel 15-bp sequence motif

In order to determine if HER binding showed any sequence specificity, I used the motif-finding software MEME [303] to search for enriched sequences in 200bp windows surrounding the high-confidence peaks. This gave strong enrichment of 15-bp sequence motif (Fig. 3.6A) (MEME E-value = 1.6x10$^-$152). Searching peak regions with Patser [304] shows

that > 80% of bound regions contain a significant hit for this motif, and that this motif is most strongly enriched in the region immediately surrounding the peak (Fig. 3.6B).

The motif I identify does not appear to be similar to either any known core promoter motif [305] or known *Drosophila* transcription factor binding motif [224, 306] (data not shown). I also am able to recover a nearly identical motif when only peaks located > 1000bp from transcription starts are used for MEME analysis, indicating that the motif is likely a true *her* binding sequence and not an unknown core-promoter motif.

3.4.4 HER target genes show a wide range of functions and expression patterns

In order to gain further insight into the possible biological role of *her* , I examined putative *her* target genes. Targets for each bound region in my high-confidence set were defined as the gene whose transcription start was closest to the peak of binding. Putative *her* target genes are enriched for transcription factors and components of signaling pathways, including genes known to be involved in EGF(*csw*, *Src42A*, *sty*, *CrebB-17A*, *p38b*, *Csk*), FGF(*csw*, *Pk61C*, *SNF1-A*, *CG13398*, *Mio*), Notch(*H*, *Dl*, *dx*, *numb*), and TGF-β(*dpp*,*shn*) signaling [307]. I also find strong binding in the promoter region of the transcriptional cofactor *intersex* [308], a core component of the *Drosophila* sex-determination pathway (Fig. 3.7).

I also examined mRNA expression timecourse data (taken from the modENCODE project [309]) for putative target genes for the set of 302 high-confidence HER-bound peaks. These data are shown in Fig. 3.8. Several temporal patterns are apparent among the *her* targets: many show strong maternal expression (Fig. 3.8, red box), a small subset show strong expression throughout development (Fig. 3.8, grey box), and some show sex-specific expresion in adult flies (Fig. 3.8, magenta box). In addition, a subset of genes shows a pattern of embryonic expression that is similar to that of *her* , with strong maternal expression and a peak of expression at 4-6h of development (Fig. 3.8, orange box).

Finally, I examined spatial expression patterns of target genes using *in situ* hybridization data from the BDGP FlyExpress database [310] (http://www.flyexpress.org). Expression patterns for the 83 genes with images in the database are shown in Fig. 3.9. I again observed that many of the putative *her* targets are expressed maternally (Fig. 3.9, First column). Genes that are zygotically expressed show mostly ubiquitous expression, with only a few showing evidence for Anterior / Posterior (Fig. 3.9, magenta arrows) or Dorsal / Ventral (Fig. 3.9, green arrows) or other tissue-specific patterns (Fig. 3.9, orange arrows)

.

3.5 Discussion

3.5.1 *her* As a Seqeunce-Specific Transcription Factor

Because of its ability to bind DNA, its limited distribution in the genome, the presence of a strongly supported binding motif, and its localization to transcriptional start sites, I believe that *her* functions as a seqeunce-specific transcription factor. Although my

data cannot directly support a role for *her* as either an activator or repressor of transcription, the presence of a subset of target genes whose temporal expression pattern matches up well with *her* indicates that it may act as an activator in some contexts (also see discussion of the relationship between *her* and *ix* below). The close proximity of *her* binding peaks to core promoters is somewhat unusual in comparison to other transcription factors active in the early *Drosophila* embryo, which tend to bind in more distant enhancer regions (e.g. [66, 67, 311, 312]). However, cases of sequence-specific transcriptional regulators binding near the core promoter are known in *Drosophila melanogaster* , such as *ovo* [313]. Further experiments using transgenic reporters for *her* targets and modified *her* proteins should allow us to resolve the exact molecular role of *her* in transcription.

3.5.2 The Role of *her* in Sex Determination and Other Pathways

The direct binding of *her* to the promoter of the transcriptional cofactor *ix*, a key component of the somatic sex determination pathway [308], may provide an explanation for sex-specific phenotypes observed in many *her* mutants. If *her* functions as an activator of *ix*, then we would expect that loss of *her* would lead to loss of *ix* expression; as *ix* is required as a cofactor for DSX^F this would lead to a female-specific intersexual phenotype as is observed for both *ix* [295, 308] and *her* mutants [265–268]. However, a model of *her* function in which the only input of *her* into the sex determination pathway is through *ix* and DSX^F cannot explain all of the available genetic data, as some effects of *her* mutations are still visible in *dsx* null mutants [265]. This model also cannot explain the weak male-specific intersexual phenotypes observed in some experiments [265, 266], or the observation that at least some of the effect of *her* on *Yp* gene expression requires DNA sequences outside of the DSX-regulated Fat Body Enhancer [267]. It is therefore likely that *her* can also influence development of sex-specific structures via other mechanisms, though whether these take place through components of the canonical sex deterimination pathway is not clear.

Since I assayed embryos 4-6h old, I cannot draw any conclusions about the effects of *her* seen upstream of *Sxl* , which would be expected to take place prior to cellularization [290]. However, no binding to any of the XSE genes *sisA*, *sisB*, or *sisC*, or to *Sxl* itself was observed. Interestingly, binding by HER is seen near the promoter of the transcription factor *Stat92E* (data not shown), which acts as the transcriptional effector for the extracellular signaling initiated by *sisC* and may directly bind the *Sxl* $_{Pe}$ promoter [314, 315]. Although most of the effect on *Sxl* $_{Pe}$ expresion is expected to be through maternal *Stat92E* [314], if *her* is found to have a similar binding pattern in the early embryo or in the ovary, this model may explain the maternal *her* effects on sex determination. Experiments in progress to examine *her* binding at earlier stages in development may help to resolve this hypothesis.

Finally, the fact that the vast majority of *her* binding I observe is to genes outside of the core sex-determination pathway supports the hypothesis that *her* likely plays a broad pleiotropic role in many developmental pathways in the *Drosophila* embryo. Many *her* target genes have lethal mutant phenotypes (e.g. *dpp* [316], *ed* [317], *Dl* [318]), which may explain the non-sex-specific maternal and zygotic lethality observed for strong *her* alleles. The fact that *her* targets include components of a wide variety of signaling pathways active in many different tissues and embryonic stages is difficult to resolve with the apparent lack of tissue-specific expression for *her* . One possibility is that *her* provides a general-

ized activation of transcription for its targets, which are then acted on by tissue-specific repressors to restrict expression to a smaller anatomical region; alternatively, *her* may act combinatorially with other transcriptional activators to activate targets in a specific tissue. Given that there is not a unifying pattern to expression of *her* targets, it is also possible that *her* may act as a repressor in some cases, or that much of the binding we observe may be non-functional. Conditional expression and knockdown experiments may be able to shed some light on these issues in the future. In any case, I believe that *her* represents an important component of *Drosophila melanogaster* development, and this combined with its interesting evolutionary properties (see Ch. 2 and 4) makes it worthy of further study.

Table 3.1: *dmel* ChIP-seq Illumina Sequencing and Mapping

Sample	Total Reads	Mapped Reads	Percent Mapped
Input	22029632	7785127	35%
herNt	15298874	5849141	38%
herCt	16052607	4107609	25%

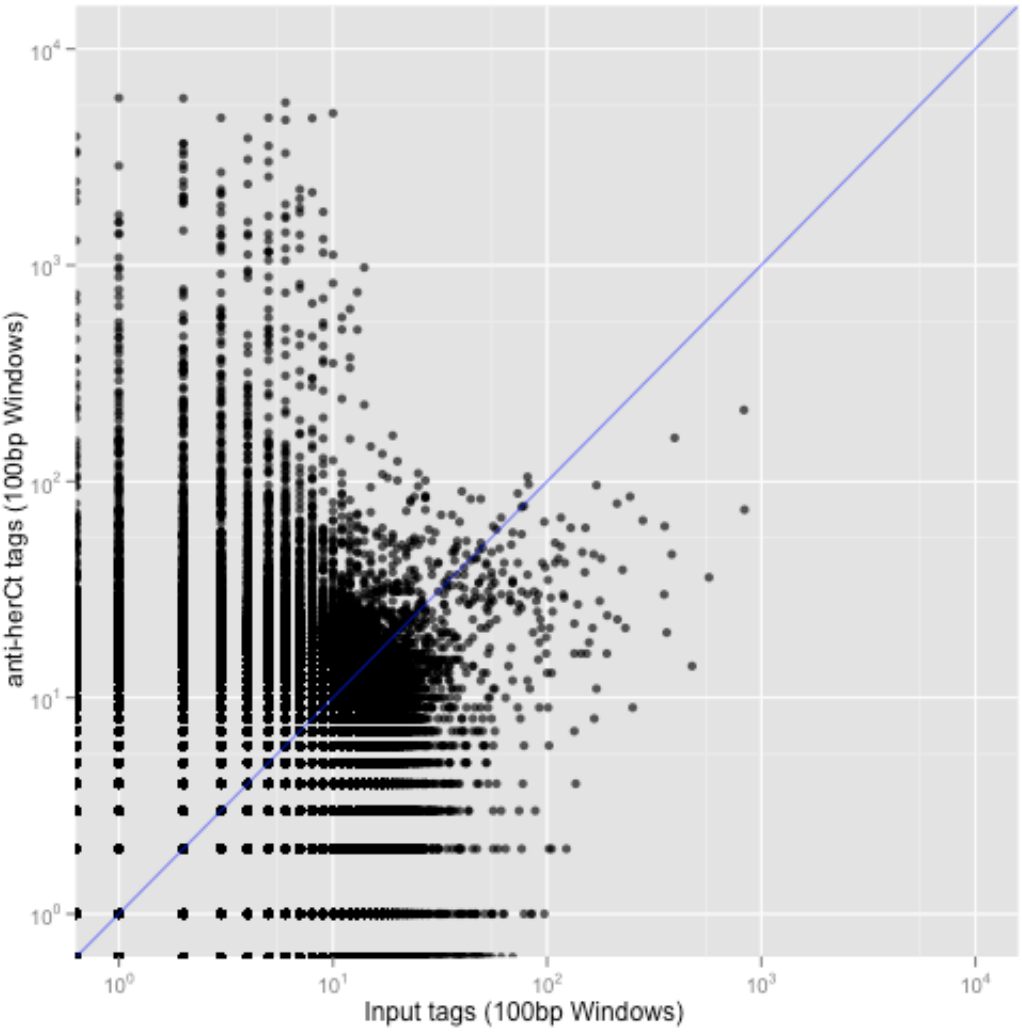


Figure 3.1: **Enrichment of herCt ChIP-seq tags relative to Input control:** *Drosophila melanogaster* herCt ChIP-seq tag counts (shifted to account for fragment size estimated by MACS [302]). Each point represents tag count in herCt (y-axis) and Input (x-axis) samples for a single 100bp window. Axes are scaled logarithmically. Blue line represents x=y. High point density above and left of the diagonal indicates that a high level of enrichment was achieved.

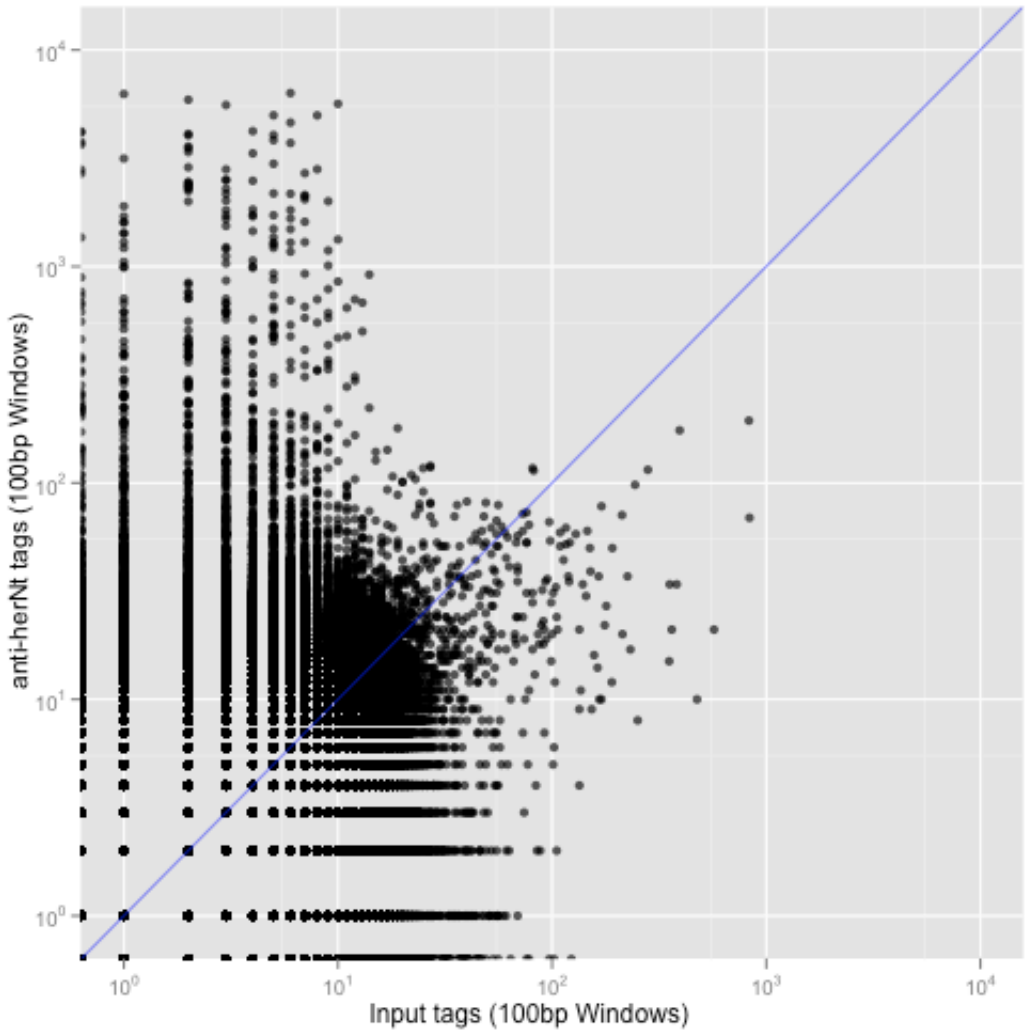


Figure 3.2: **Enrichment of herNt ChIP-seq tags relative to Input control:** *Drosophila melanogaster* herNt ChIP-seq tag counts (shifted to account for fragment size estimated by MACS [302]). Each point represents tag count in herNt (y-axis) and Input (x-axis) samples for a single 100bp window. Axes are scaled logarithmically. Blue line represents x=y. High point density above and left of the diagonal indicates that a high level of enrichment was achieved.

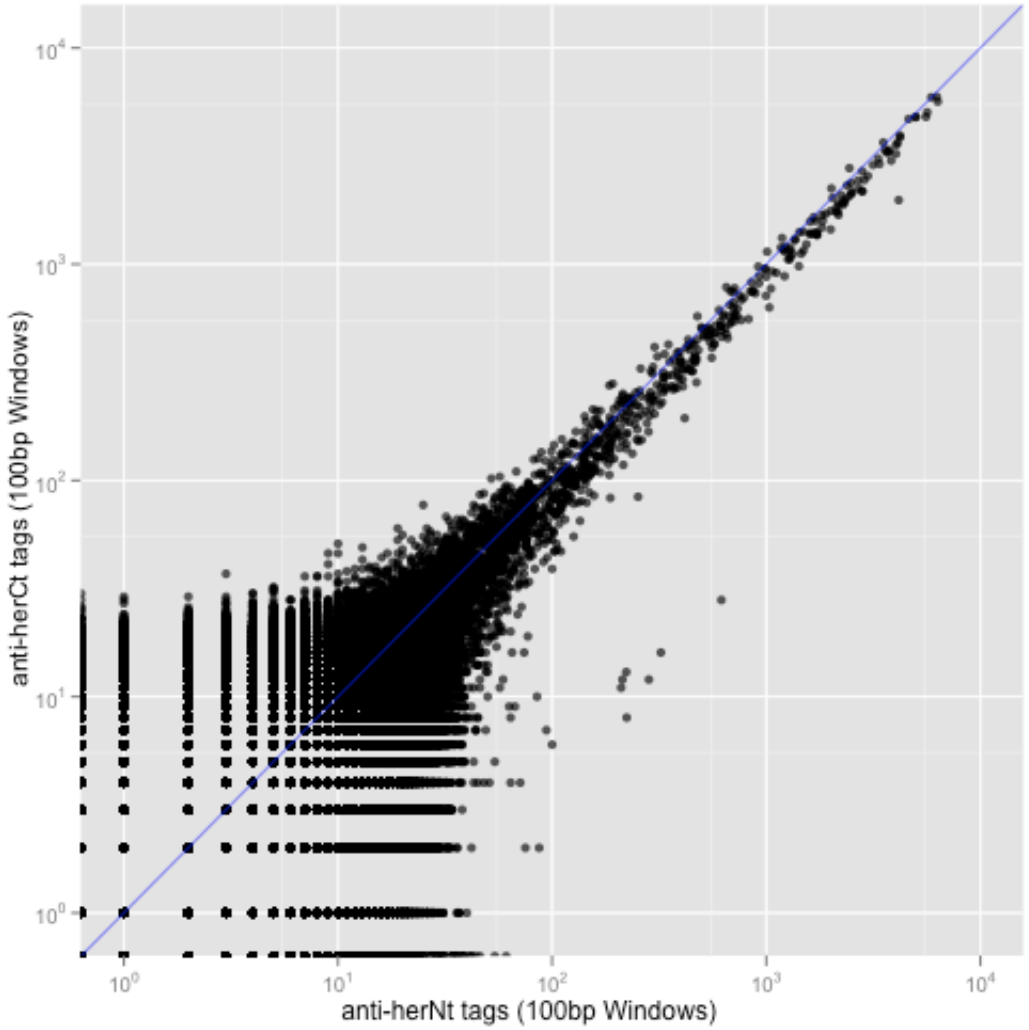


Figure 3.3: **Correlation of herCt and herNt Tag Counts:** Each point represents tag count in herNt (x-axis) and herCt (y-axis) samples for a single 100bp window. Axes are scaled logarithmically. Blue line represents x=y. Note that samples are highly correlated; $R^2 = 0.97$.

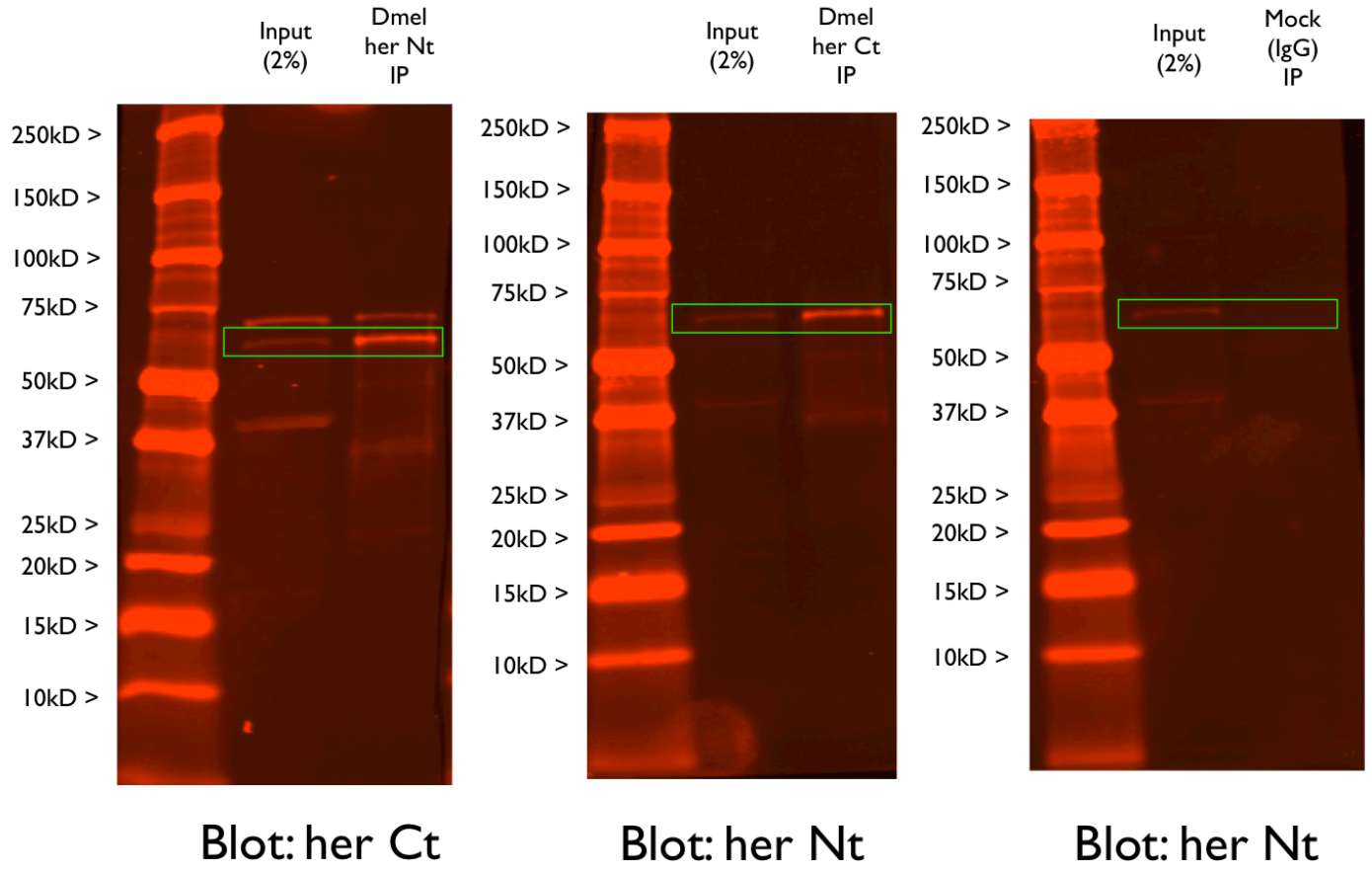


Figure 3.4: **Chromatin Immunoprecipitation and Western Blot of *Drosophila melanogaster* 0-12h Nuclear Extract With the dpherB Antibody:** Comparison of herNt IP to Input control is shown on the left; herCt IP compared to Input is shown at center; Input compared to Normal Rabbit IgG control is shown on the right.

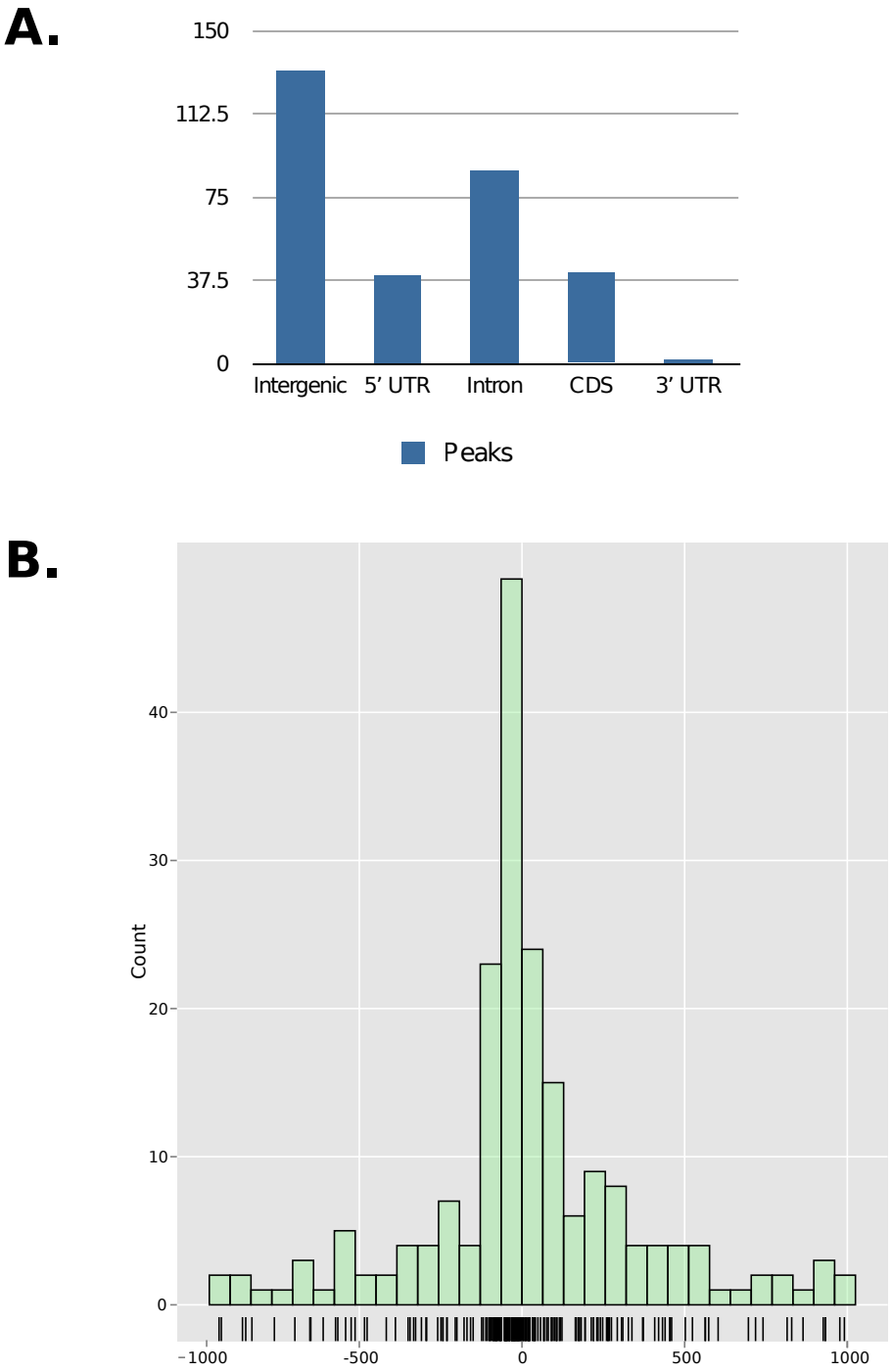


Figure 3.5: **Location of *her* Peaks Relative to Gene Structural Features:** (A) Shows the number of HER-boung peaks from the "high confidence" dataset lying within various gene features. (B) Shows a histogram of the distribution of the distance from each HER binding peak to the nearest annotated transcription start. Rug at bottom has a line indicating the position of each individual data point. Note that only peaks < 1000bp from an annotated promoter were included, representing 67% of peaks.

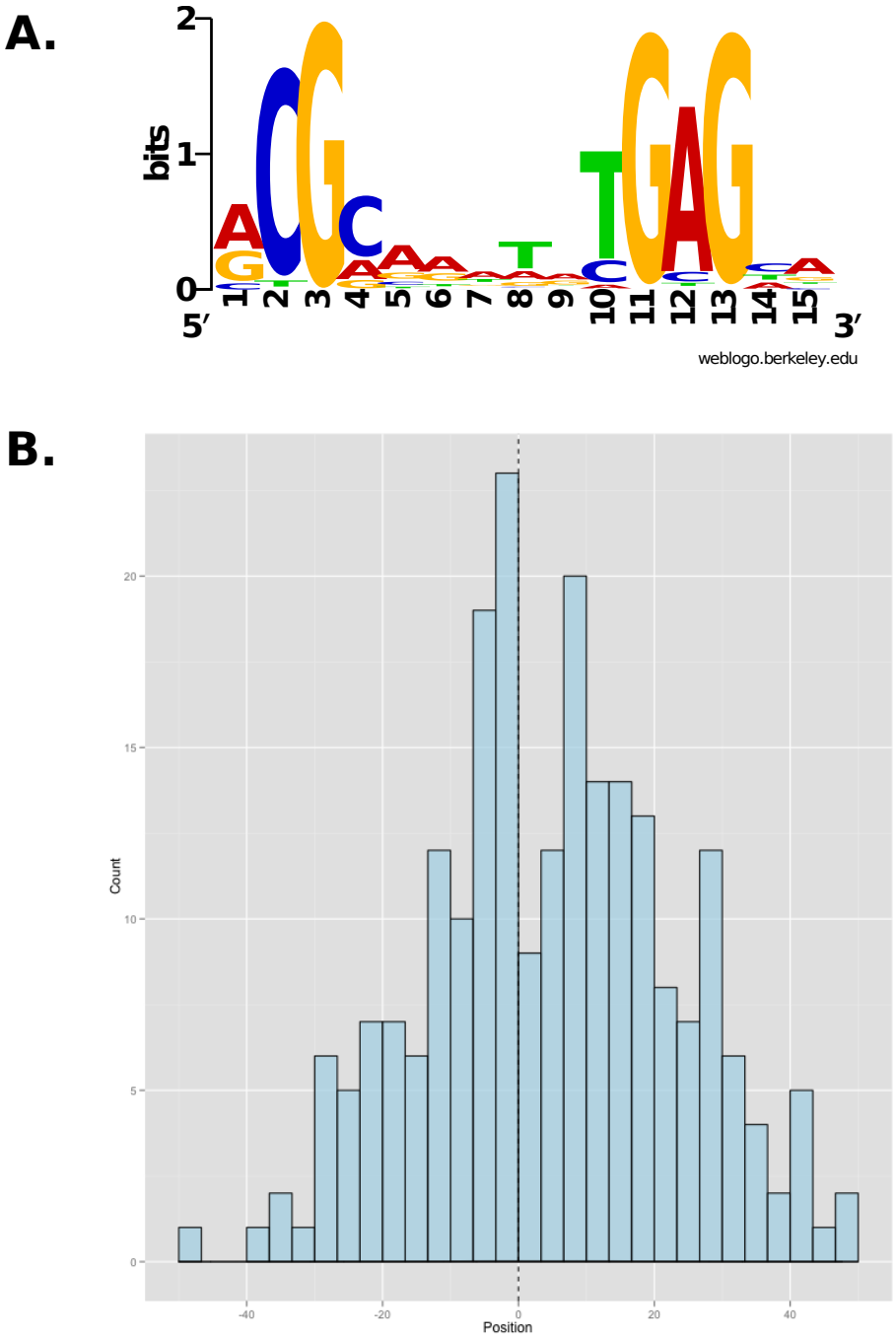


Figure 3.6: **Motif Enriched in *Drosophila melanogaster* HER-bound Regions and Distribution Relative to Binding Peaks:** (**A**) shows the motif identified by MEME [303] as being enriched in the *Drosophila melanogaster* HER-bound regions. Sequence logo generated using weblogo (http://weblogo.berkeley.edu [319]) (**B**) shows a histogram of hits for this motif called by Patser [304] in 100-bp regions centered around *Drosophila melanogaster* binding peaks. Note that the density is located near the center (peak) position.

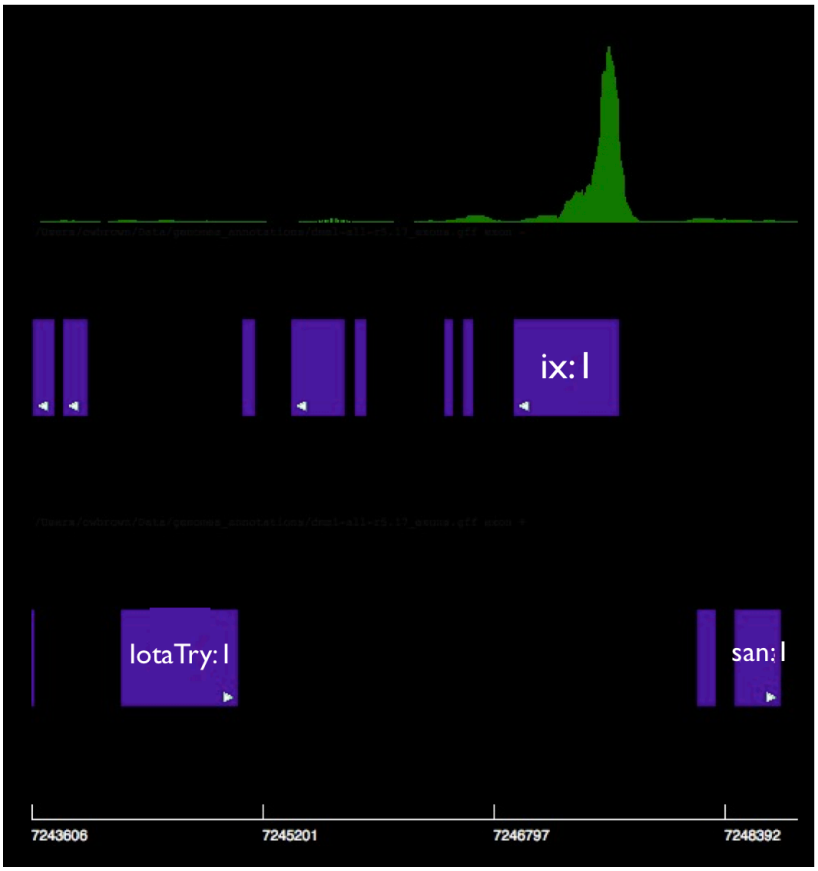


Figure 3.7: **HER Binding to the *ix* locus:** Tags from herNt for are shown in the top track. Middle track shows the location of *ix* exons; note that the transcription start lies almost directly below the peak. Plot generated using the Kelp Genome Browser (https://sourceforge.net/projects/kelpgenomebrows/).

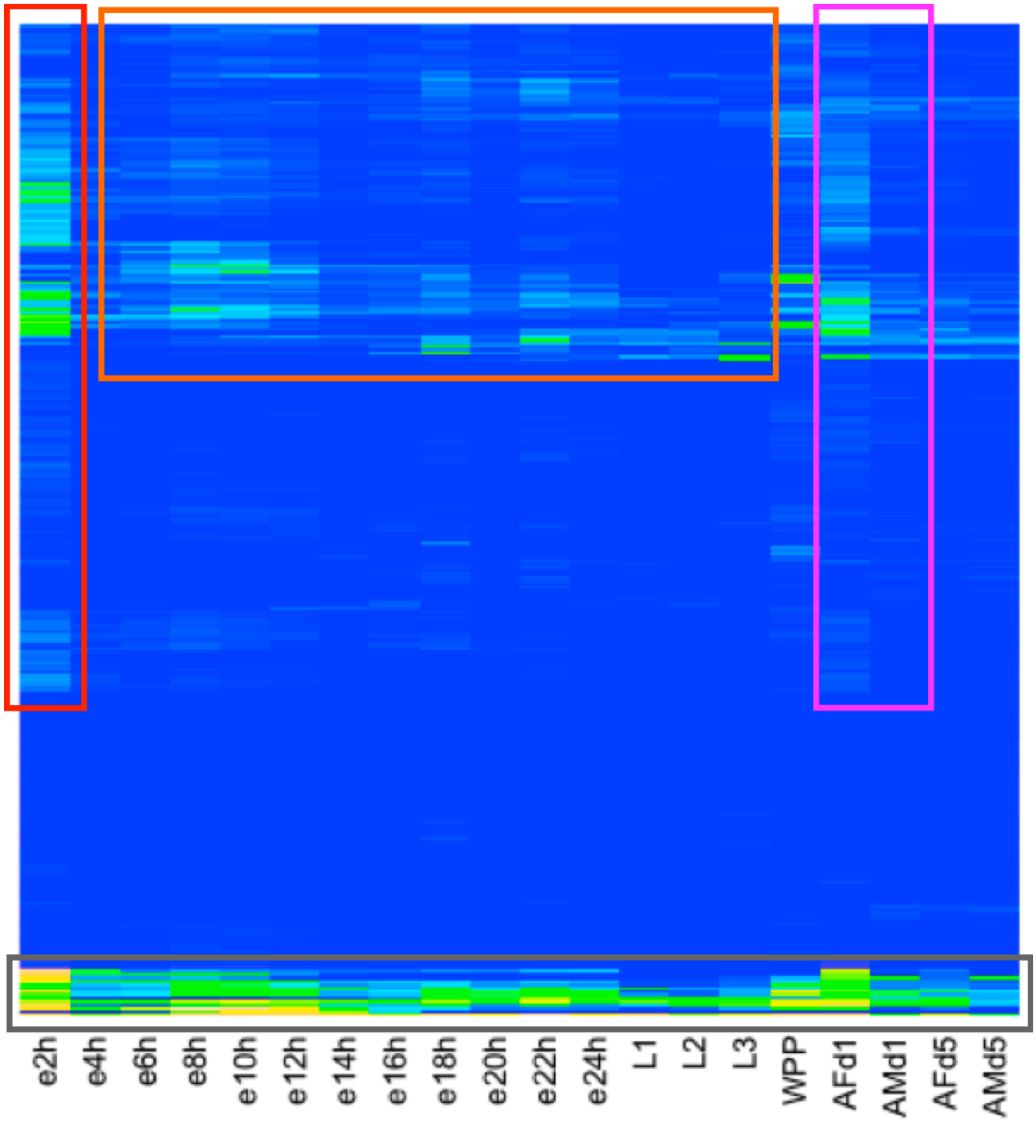


Figure 3.8: **Expression profiles of Putative *her* Targets during Development:** Heatmap shows expression level from modENCODE microarray data ([309]). Yellow represents is high expression, blue is low. Red box highlights genes with maternal expression; orange box highlights genes whose expression may track with *her* expression; magenta box highlights genes that show sexually dimorphic expression in adults; grey box highlights ubiquitously expressed genes. exh = $x$hours embryonic development; Lx = Larval instar x; WPP = Pupal; AFd1 = Adult Female, 1 day post eclosion; AMd1 = Adult Male 1 day post eclosion; AFd5 = Adult Female 5 days post eclosion; AMd5 = Adult Male, 5 days post eclosion

Figure 3.9: **Expression Patterns of *her* targets During Embryonic Development:** Expression patterns for the 83 *hermaphrodite* targets with image data available in the BDGP image database [310]. Orange arrows point to genes with tissue-specific patterns; Magenta Arrow point to genes with anteroposterior patterns; Green Arrows point to genes with dorsoventral patterns S1-3 = Stage 1-3, S4-6 = Stages 4-6, S7-8 = Stages 7 and 8, S9-10 = Stages 9 and 10, S11-12 = Stages 11-12, S13-16 = Stages 13-16

Chapter 4

Comparative Analysis of DNA binding by the Transcription Factor *hermaphrodite* in *Drosophila melanogaster* and *Drosophila pseudoobscura*

4.1 Abstract

In previous work I identified the *Drosophila* gene *hermaphrodite* (*her*), as a transcription factor that exhibits a high degree of divergence within its DNA-biding among several species of the genus *Drosophila*. The HER proteins present in *Drosophila melanogaster* and *Drosophila pseudoobscura* show a high degree of divergence within their 4 shared C2H2 zinc finger binding domains, and *Drosophila pseudoobscura* HER contains 3 extra copies of the C-terminal-most finger. Here I analyze *in vivo* binding of HER in both of these species using ChIP-seq, and find evidence of large-scale differences in binding targets between the species, as well as possible evidence for a shift in the DNA-binding specificity of HER.

4.2 Introduction

Although the idea that changes in transcription factor proteins can play an important role in regulatory evolution has gained some traction in recent years [7,13,14], there are still relatively few examples of this process in the literature, and most cases involve changes outside the DNA-binding domain ([135, 136, 198]). In addition, only a handful of cases have examined the relationship between changes in the transcription factor and changes in its target *cis*-acting sequences [147, 182], none of them in a metazoan system. The evolutionary importance of changes in transcription factor binding specificity, particularly in species with complex development and physiology, and the evolutionary mechanisms that allow such changes to take place therefore remain unknown.

The *Drosophila* gene *hermaphrodite* (*her*) displays several unique features which make it an attractive candidate for studying these questions in a well-established model system. A previous computational study (see Ch. 2) identified *her* as having an elevated rate of substitutions within the 4 C2H2 zinc fingers of its DNA binding domain. The changes are most dramatic in the species *Drosophila pseudoobscura* and *Drosophila persimilis* in the emphobscura subgroup, which show extensive changes at positions outside of the canonical specificity-determining positions (-1,2,3,6) of the zinc finger binding helices [26], and an A→G substitution at the 2 position of the fourth zinc finger. These species were also found to contain a large insertion downstream of the DNA-binding domain consisting of a 3-fold tandem duplication of the C-terminal-most zinc finger. Expansions of this type have been observed in zinc finger proteins from other species following gene duplication [132, 256], and may therefore represent a common mechanism through which zinc fingers can alter their binding specificity.

Genetic experiments have implicated a role for *hermaphrodite* in the *Drosophila melanogaster* sex determination pathway, both in the initial activation of the master sex regulator *Sxl* (although this result has been disputed - see [299]) and in the somatic sex pathway with the sex-specific transcription factor *Dsx* [265–268]. Interestingly, the sex determination pathway is known to be among the more rapidly evolving regulatory systems in insects. For example, the position of *Sxl* at the top of the pathway is specific to the family *Drosophilidae* [320]. In the fruitfly *Ceratitis capitata*, *Sxl* is non-sex-specifically expressed, with the gene *tra*, which acts downstream of *Sxl* in *Drosophila melanogaster* , under control of other mechanisms [282–284]. In addition, many sexually dimorphic traits

are divergent among *Drosophila* species, including male sex combs and genitalia [285] and body pigmentation [35].

Previous work in *Drosophila melanogaster* (see Ch. 3) confirmed that HER binds DNA, identified a set of potential HER target genes, and established a possible HER binding motif. In this work I compare and contrast these data with the genome wide DNA-binding behavior of the HER protein in *Drosophila pseudoobscura* , assessed using ChIP-seq. I find evidence for differences in both the genome-wide target sets and DNA binding preference between the HER proteins for these species.

4.3 Materials and Methods

4.3.1 RACE PCR and Cloning

I extracted total RNA from 0-12h *Drosophila pseudoobscura* embryos using Trizol (Invitrogen). $100\mu L$ of flash-frozen embryos were homogenized directly into Trizol buffer for RNA extraction. The aqueous fraction was removed and extracted with chloroform, precipitated in isopropanol, and resuspended in RNAse-free water. Six $100\mu L$ samples were pooled and concentrated using a Microcon centrifugal concentrator (Millipore). mRNA was isolated from total RNA using the Oligotex mRNA Midi Kit (Qiagen). Reverse Transcription and 5' and 3' RACE were carried out using the Marathon RACE kit (ClonTech) with 5' Gene-specific primer CGTCCCACCAAAAAGGAAAGTCATCC and 3' gene-specific primer CAACAATCCATTGAGAGGCGACAAGG. Final amplification of the transcript used primers CAATATTTGTTTTGGAAAATGCACAACA TATTTTAAGGACTCATCTTTATTGGTTT. Amplified transcript was cloned into the pGEM-T EZ vector (Promega) for sequencing.

4.3.2 Degenerate PCR

I prepared genomic DNA from the following species for use in degenerate PCR reactions: *Drosophila affinis* (stock #14012-0141.06), *Drosophila obscura* (stock #14011-0151.00), *Drosophila subobscura* (stock #14011-0131.08), *Drosophila sturtevanti* (stock #14043-0871.14), and *Drosophila miranda*(stock #14011-0101.08). Whole adult flies were homogenized in grinding buffer (5% sucrose, 80mM NaCl, 0.1M Tris-HCl pH8.5, 0.5% SDS, 50mM EDTA), centrifuged, and the supernatant precipitated in KOAc and Acetic Acid. PCR was carried out using the degenerate primers CCRTGYGCGTTGGCCAGCTG and GCGWTYYTCCTTGAAVGGYTTGTA which were designed to amplify regions corresponding to the insert region in *Drosophila pseudoobscura* . PCR reactions were prepared using 5mM MgCl and 20mM of each primer and amplified using Bio-X-Act short polymerase (Bioline). Amplified sequences were cloned into the pGEM-T EZ vector (Promega) and Sanger sequnced at the UC Berkeley Sequencing Core Facility.

4.3.3 Protein Expression and Purification

Sequences representing amino acid residues 1→149 (*D. pse* herA) and 541 to 731 (*D. pse* herB) were amplified from the full-length *Drosophila pseudoobscura her* cDNA and

cloned into the Gateway pDEST17 vector using BP and LR recombinase (Invitrogen), and into the pProEX-Htb vector (Invitrogen). Peptide expression in *E. Coli* and purification were carried out as described in Ch. 3.

4.3.4 Antibody Purification and Testing

Antiserum was raised in rabbits against both the *Drosophila pseudoobscura* herA and herB peptides. The independent antiserum samples were then affinity-purified against the herA and herB peptides using the methods described in Ch. 3. Antibodies were tested by western blot using either whole embryos lysed in SDS-Page sample buffer or nuclear extract prepared from 0-12h *Drosophila pseudoobscura* embryos. Blots were probed with herA or herB antibodies diluted 1:1000 in blocking buffer, then with 1:10,000 infrared fluorescent Goat anti-Rabbit secondary antibody (Rockland Immunochemicals). Scanning was performed using an Odyssey Infrared Scanner (LiCor Biosciences).

4.3.5 Chromatin Immunoprecipitation and Sequencing

Chromatin was prepared from *Drosophila pseudoobscura* embryos aged at 25 Cdeg to 4.5-6.5h post fertilization. This stage matches the 4-6h time window I used for experiments carried out with *Drosophila melanogaster* discussed in Chapter 3. Embryos were fixed using the procedure outlined in Chapter 3, and chromatin purification was carried out by CsCl gradient centrifugation as in [66, 67]. Purified chromatin was fragmented using 16 cycles of 10 min ea. (15s on, 45s off) in a Bioruptor bath-type sonicator (Diagenode), giving an average fragment size between 250 and 350 bp.

Two different methods were used for chromatin immunoprecipitation. In the first, I followed the procedure from Ch. 3 using 100μg of *Drosophila pseudoobscura* chromatin and 3ng of either the herA or herB antibody. In the second method, I pooled 50% each of 4-6h *Drosophila melanogaster* chromatin and 4.5-6.5h *Drosophila pseudoobscura* chromatin prior to sonication. 100μg of pooled, fragmented chromatin was then used for immunoprecipitation and sequencing.

Illumina library construction and sequencing were carried out as in Ch. 3. Tag mapping was performed using Bowtie [301], with filtering applied for sequencing quality scores and tags which map to more than two locations in the genome. Peak calling was performed using MACS [302] or a custom algorithm which fits a theoretically predicted peak shape to the peak distribution within the genome (Tommy Kaplan, unpublished). Analysis of peak enrichment was performed using custom Python scripts. Motif finding was perfomed using MEME [303], and motif finding within sequences performed using custom Python scripts and Patser [304].

4.4 Results

4.4.1 The *Drosophila pseudoobscura* and *Drosophila persimilis her* mRNA contains a 3-Fold tandem duplication of the C-terminal C2H2 zinc finger domain

Computational analysis had previously shown the DNA-binding domain of the *Drosophila pseudoobscura* HER protein to contain many amino acid substitutions relative to *Drosophila melanogaster* , including several at predicted DNA-contacting residues. In addition, this analysis showed the presence of a large insertion in the *hermaphrodite* gene in the region 3' of the putative DNA-binding domain, which was found to contain 3 duplications of the 3'-most C2H2 zinc finger domain. In order to confirm that *her* is expressed in *Drosophila pseudoobscura* and that the insertion is present in the final transcript, I used RACE PCR on total mRNA from 0-12h *Drosophila pseudoobscura* embryos to clone the *her* mRNA. A single transcript was recovered, which contained the insertion, and was missing both of the predicted intron sequences, indicating that it had been fully processed.

4.4.2 The zinc finger duplication seen in *Drosophila pseudoobscura* and *Drosophila persimilis* is conserved throughout the *Obscura* group

In order to determine the extent of conservation of the insertion observed in the *Drosophila pseudoobscura* and *Drosophila persimilis her* genes, I performed degenerate PCR using primers flanking a region of the *her* locus containing the 3'-most zinc finger domain shared with *Drosophila melanogaster* and all three of the duplicated C2H2 zinc fingers present in *Drosophila pseudoobscura* . I attempted to amplify this region from genomic DNA prepared from 4 other species with the emphObscura group (*Drosophila miranda*,*Drosophila obscura*,*Drosophila subobscura*,*Drosophila affinis*), and a closely related species outside of the *Obscura* group (*Drosophila sturtevanti*). I was able to clone and sequence the insertion region from all of these species with the exception of *D. sturtevanti*; alignments of these sequences with the *D. pseudoobscura* and *D. persimilis* sequences are shown in Fig. 4.1. The sequences showed that the duplicated zinc finger domains are present in all of these species, and that several of the species show further diversification of these binding domains, with multiple substitutions occuring at predicted DNA-contacting residues (Fig. 4.1, red boxes).

4.4.3 ChIP-seq analysis of *Drosophila pseudoobscura* HER Shows Significantly enriched Peaks

I examined the binding of *Drosophila pseudoobscura* HER by ChIP-seq, using antiserum raised against two independent peptides selected from the HER protein: *D.pse* herA (pA) (residues 1-149), which contains 4 of the 7 HER C_2H_2 -zf domains, and *D. pse* herB (pB) (residues 541 $\rightarrow$ 731) from the well-conserved C-terminal domain of *Drosophila pseudoobscura* HER Fig. 4.2. Antiserum from the first and second bleeds was purified against the corresponding peptides. Antibodies were tested by probing western blots prepared with either whole embryo or nuclear extract prepared from 0-14h *Drosophila pseudoobscura* embryos. Blotting results are shown in Fig. 4.3. The pA blot (Fig. 4.3, left panel) showed a

band between the 75kD and 100kD size markers in both the whole-embryo and nuclear extract lanes (Fig. 4.3, red box), near the expected size for the full-length HER protein (83kD). The pB antibody showed no strong signal in the nuclear extract lane (Fig. 4.3, right panel), and no bands that clearly match the 83kD band from the pA blot in the whole-embryo sample, although a band is present below the 75kD marker that may be shifted due to differences in how the two gels were run. The absence of signal was consistent across independent pB antibodies prepared using antiserum obtained from different animals and different bleeds. The absence of pB signal in these western blots is puzzling, because in subsequent ChIP analysis pB was able to specifically enrich a region of the *Drosophila pseudoobscura* genome which had been shown to be bound by HER in previous analysis, as measured by semi-quantitative PCR (data not shown); this may reflect differences in the epitopes recognized by pA and pB.

Because of the conflicting results from antibody tests, both pA and pB were used for subsequent ChIP experiments. Chromatin immunoprecipitation was performed using chromatin purified from 4.5-6.5h *Drosophila pseudoobscura* embryos, as previous analysis had indicated that this time period represents similar developmental stages to the 4-6h time period used for the *Drosophila melanogaster* ChIP experiments (see Ch. 3). A control ChIP reaction using normal rabbit IgG was also performed alongside the treatment samples. Results of Illumina high-throughput sequencing of ChIP samples are shown in Table 4.1. Mapping of reads was performed using BowTie [301], with settings to eliminate sequences which map to more than two sequences in the genome. The low percent of mapping reads (28%) in *D. pse* pB may due to higher than normal amounts of Illumina library adapter contamination.

Peak calling was performed using a custom algorithm, which scores the fit of enriched regions in the genome with a theoretical peak model (Tommy Kaplan, unpublished results). Results using this algorithm were similar to those obtained using MACS [302], with fewer low-significance peaks (Data not shown). This analysis resulted in 1109 bound regions called for *D.pse* pA and 1102 bound regions called for pB using standard cutoff values. Similar analysis of the Input sample alone yielded 103 regions. To account for differences in total tag counts between the samples, all subsequent analysis was performed on datasets normalized to the equivalent of 1000000 tags. In order to confirm that peaks are significantly enriched with respect to background signals, the normalized tag count (with normalized input tags subtracted) within each bound region was calculated for the IgG and treatment samples; these results are shown in figure Fig. 4.4**A**. Distributions of tag counts for pA and pB were found to be significantly different from the tag count distributions for matching regions in the IgG data (K-S p-value = 0.0 for both datasets). Median normalized tag counts were 13.3 for *D. pse* pA peaks and 19.0 for pB, compared to -0.677 and -0.12 for the respective IgG sets. Mean tag counts were 19.26 for pA and 37.4 for pB.

Examination of normalized tag counts for pA and pB antibody datasets across the pooled set of peaks showed that a subset of peaks from both antibodies show strong enrichment in one dataset, but low or nonexistent enrichment in the other Fig. 4.4**(B)**. These likely represent some cross-reaction by both antibodies with other DNA-binding factors. In order to correct for these discrepancies, all subsequent analysis was performed using the set of peaks that overlap between the two datasets; since the two antibodies were raised against

independent peptides that have low sequence similarity, this set most likely represents true HER binding. 396 bound regions (35% percent of peaks for each antibody) show overlap between the two datasets; when the set of overlapping regions are considered in isolation, the regions of differential enrichment disappear Fig. 4.4**(B)**. Correlation coefficient does not change significantly between the full and overlap-filtered sets (pearson r = 93.8 and 93.0, respectively), although the mean and median tag count shifts slightly towards higher values (median = 18.99 for pA, 26.8 for pB; mean = 26.4 for pA, 41.7 for pB). Most of the peaks that are eliminated by filtering are in the lower part of the distribution of tag counts (Fig. 4.4**(C)**, although there is a set of highly enriched peaks that were eliminated from the pB dataset. Median tag count values are similar to that obtained from a similar analysis of the *Drosophila melanogaster* HER peaks identified in Chapter 3 (23.12 tag equivalents), although this value masks a set of peaks with higher values in the *D. mel* data reflected in a higher mean (90.5 tag equivalents). This may indicate stronger DNA binding by the *Drosophila melanogaster* HER protein, or may be due to differences in affinity between the antibodies raised against the *D. mel* and *D. pse* HER proetins.

4.4.4 Most HER Binding is Divergent Between *Drosophila melanogaster* and *Drosophila pseudoobscura*

In order to assess conservation of binding, I used a *D. pse* - *D. mel* whole genome alignment constructed using the alignment software FSA [321]. For this analysis, I used the set of *D. mel* bound regions which overlap between peak calls made using the new peak-fitting algorithm and the set found by MACS used in Ch. 3 (n=256). Peaks from each species were mapped onto genomic coordinates from the other using the alignment as a guide, and divided into three categories: overlapping peaks (peak summit within 100bp of summit in other species), shifted peaks (bound region occurs within 1kb of bound region in second species) and species-specific peaks. Fig. 4.6 shows the distribution of these peak categories in both datasets. The majority of observed strong binding (93% of *D. pse* regions and 89% of *D. mel* regions) is species-specific; however, a small number of conserved cases were observed. Nine peaks (3.5% of *D. mel* and 3.0% of *D. pse*) are located at overlapping positions in the two species, and 15 peaks (6.6% of *D. mel* and 4.2% of *D. pse* regions) are shifted relative to one another.

Among the overlapping and shifted peaks sets, 17 pairs of *Drosophila melanogaster* -*Drosophila pseudoobscura* peaks were found to bind near a common orthologous gene between the two species (listed in Table 4.2). These genes show no obvious relationship in terms of function, but they do include a number of developmental transcription factors (*tribbles* (*trbl*), [322], *schnurri* (*shn*) [323], *spalt minor* (*salm*) [324], *no ocelli* (*noc*) [325], a Mediator subunit (*(*MED10) [326], and two [GAP]s, RhoGAP71E and RhoGAP93B [327]. Notably, one of the shared genes is *separation anxiety*(*san*) [328], which directly flanks the sex determination gene *intersex* (*ix* [295]); binding is conserved within the short intergenic region between the *san* and *ix* transcription start sites so the regulatory relationship between *her* and *ix* may also be conserved.

4.4.5 HER-bound regions in *Drosophila pseudoobscura* are enriched for a 15-bp motif which differs from the *Drosophila melanogaster* HER binding motif

I also used MEME [303] to analyze 100-bp windows around peaks within the 396 *Drosophila pseudoobscura* HER-bound peaks for evidence of motif enrichment. This analysis gave a well supported motif shown in Fig. 4.7**A** (MEME E-value = $1.4x10^{-}207$), which is enriched near the center of *Drosophila pseudoobscura* binding peaks (Fig. 4.7B). This motif closely resembles that obtained from a similar analysis in *Drosophila melanogaster* , with high information CGC and GAG positions separated by a 6-bp linker. In spite of this similarity, the motif obtained from the *Drosophila pseudoobscura* data is more specific, with a higher information content linker region with a stronger preference for Adenosine. Of the 396 bound regions, 217 (54%, 226 sites total) have an occurrence of this motif within 100 bp of a peak summit, representing 13% of the total predicted binding sites in the genome (1471). This is a remarkably high occurrence and occupancy compared to other fly transcription factors which have been examined using ChIP-seq; only 20% of the regions bound by the A-P morphogen TF *Bcd*, for example, show a strong match for the *in vitro* BCD binding site [66], and there are tens of thousands of unbound sites for this factor distuributed throughout the genome. Because the sequences matching the *Drosophila pseudoobscura* HER motif are a subset of those that match the *Drosophila melanogaster* HER motif, most of the observed *Drosophila pseudoobscura* HER sites are also strong *Drosophila melanogaster* HER sites; 51 sites within *Drosophila pseudoobscura* HER-bound regions are predicted matches for the *Drosophila melanogaster* HER motif but not the *Drosophila pseudoobscura* HER motif.

4.4.6 The relationship between HER binding conservation and conservation of the predicted HER binding site

In order to determine whether the large-scale divergence in binding observed for HER proteins from the two species was due to *cis* changes, I examined the bound peaks in each species for occurrences of either the *Drosophila pseudoobscura* or *Drosophila melanogaster hermaphrodite* motif. Sequences lying within each bound region from both species were obtained from the whole-genome alignment, and each sequence was searched for instances of either the *Drosophila melanogaster* HER motif or the *D. pse* HER motif. For regions bound in *D. mel* but not in *D. pse* , 8% of peaks contain a conserved *D. mel* motif, of which 28% (2.3% of total) also match the *D. pse* HER motif Fig. 4.8; in comparison, only 0.3% of peaks in *D. pse* contain a conserved instance of either motif. 50% of *D. mel* species-specific peaks contain a non-conserved instance of the *D. mel* HER motif, compared to 66.9% of *D. pse* species specific peaks. Of these species specific *D. pse* HER sites, 84% (56.6% of total) are *D. pse* HER sites which do not have a corresponding *D. mel* site; in contrast, only 13.3% of *D. mel* HER peaks contain a non-conserved site matching the *D. pse* HER motif. This indicates that *D. pse* HER may be binding to a large novel pool of strong binding sites in *D. pse* .

Of the nine shared peaks, more than half (5/9) contain a conserved instance of the *D. pse* HER motif (Fig. 4.9**A**), and all but one peak contains at least one instance of HER motif in both species. Fig. 4.9**B**) shows examples of conserved binding which appears to

be driven by conservation of a HER motif. For the 15 shifted binding regions (Fig. 4.10), the members of each pair were considered separately. For *D. pse* regions in a shifted pair, 7 showed a species-specific binding site, 5 a conserved site, and 3 no site in either species. For *D. mel* regions, 7 peaks contained a site in both species, 5 contained a species-specific site, 3 contained no motif, and one region contained a site in *D. pse* but not *D. mel* . Notably, of the seven species-specific sites in *D. pse* peaks, 5 are *D. pse* HER sites. The locus shown in Fig. 4.10 is an example of a peak shift which may be driven by reciprocal gain and loss of binding sites in each species.

4.4.7 Secondary Motifs With a Defined Linear Order are found in *Drosophila pseudoobscura* HER-bound regions

In addition to the primary *D. pse* HER motif identified by MEME, four other highly significant motifs were also enriched within 100bp of a HER binding peak (Fig. 4.11). These motifs do not appear to be related to either the *D. mel* or *D. pse* HER motif, or to one another. MEME E-values for these motifs range from $1e-253$ for dpse-motif-4 to $3.8e-296$ for dpse-motrif 3. Enrichment of these motifs is specific to *Drosophila pseudoobscura* HER-bound sequences; as shown in the right-hand column of figure 4, all 4 motifs are present in significant numbers in *D. pse* HER-bound regions (13.6% of regions for dpse-motif-2, 13.3% of regions for dpse-motif-3, 19.6% of peaks for dpse-motif-4, and 3.5% of peaks for dpse-motif-5), but are almost entirely absent from *D. mel* HER-bound regions (one occurrence of dpse-motif-3 (0.3%), ten occurrences of dpse-motif-4 (3.9%). All of the motifs with the exception of dpse-motif-4 (which most likely represents a simple sequence repeat) are highly specific and are found at a small number of locations within the genome — the 54 dpse-motif-2 sites represent 10% of the total predicted sites in the genome, and the 53 dpse-motif-3 sites more than a third of the genomic total (34.4%).

When the locations of these motifs were mapped back to the *D. pse* HER-bound sequences, a striking pattern was observed. 12.6% of *D. pse* bound regions contain at least 3 instances of these motifs, and they are found in a well-defined linear order within the bound region with respect to one another and the primary *D. pse* HER motif. Fig. 4.12 shows the position of these five motifs within the 50 *D. pse* HER-bound regions which contain at least 3 distinct motifs (sequences are sorted by position of the primary HER motif to emphasize order). Most of the sequences contain at least a partial match to the linear sequence dpse-her-mtx4+, dpse-her-mtx1+, dpse-her-mtx3+,dpse-her-mtx2+,dpse-her-mtx5. Spacing is also well conserved between sequences, particularly among the *D. pse* HER primary motif, dpse-mtx-3, and dpse-mtx4.

Because of the apparent high sequence similarity among the subset of *D. pse* HER targets containing these motifs, I hypothesized that the presence of a highly similar *D. pse* HER motif match in these sequences may be partially responsible for the differences in the motifs learned by MEME from the HER binding data for *Drosophila melanogaster* and *Drosophila pseudoobscura* . In order to test this, I split the sequences into three groups: one containing sequences which had at least three separate motif matches, one containing motifs with exactly two motif matches, and one containing only instances of the *D. pse* primary motif. *D. pse* primary motifs were then extracted from each set and used in separate MEME runs to reconstruct the *D. pse* HER primary PWM. The results of this analysis are shown

in Fig. 4.13. The PWM reconstructed from sequences containing two or more motifs in addition to the *D. pse* HER primary motif is highly specific, most notably within the 6-bp central linker region which differentiates the *D. mel* and *D. pse* HER motifs. The PWMs reconstructed from the 2-motif and single-motif sets show more degeneracy, with the single-motif PWM closely resembling that obtained from the whole dataset. The presence of the set of highly similar HER primary motifs alone therefore does not explain the difference in the motifs learned from the HER bound regions from the two species, although it is likely to be a contributing factor.

Finally, the sequence similarity among the bound regions in the 3-motif set extends beyond the regions containing the HER secondary motifs. A portion of the alignment of 200-bp regions containing the 50 3-motif sequences is shown in Fig. 4.14; although the highlighted motifs are clearly well conserved, many flanking positions also show a high degree of similarity among sequences; this may imply that the *D. pse* HER 3-motif regions may in fact be repetitive sequence elements (see Discussion).

4.5 Discussion

4.5.1 Evolution of the *her* DNA-binding domain

I showed that the three extra C2H2 zinc fingers unique to the *Drosophila pseudoobscura* and *Drosophila persimils her* gene are present in a spliced, A-tailed transcript that is expressed in *Drosophila pseudoobscura* embryos. I also showed that this insertion is conserved over 2 million years of evolution within the Obscura group [329]. This indicates that the insertion was present in the common ancestor of the Obscura group species. The timing of the duplication events is difficult to resolve since *Drosophila willistoni*, the most divergent species relative to *Drosophila melanogaster* , contains 5 zinc fingers. It is not clear whether the extra domain in *Drosophila willistoni* is orthologous to the Obscura group duplicated zinc fingers, so I cannot say with certainty whether the ancestral *her* gene contained 4 fingers (with a single duplication in *Drosophila willistoni* and a triple duplication in the lineage leading to the Obscura group) or 5 (with a double duplication in the Obscura group and a loss in the branch leading to the melanogaster group), as both possibilities are equally parsimonious.

4.5.2 Divergence of HER binding between *Drosophila melanogaster* and *Drosophila pseudoobscura*

ChIP-seq analysis of *Drosophila pseudoobscura* 4.5-6.5h embryos using two independent antibodies yielded a robust set of HER bound regions. Comparison with peaks derived from a similar analysis in *Drosophila melanogaster* showed that 90% of binding events are divergent between the two species, with a small set of clear examples of conserved binding events. This amount of divergence is much higher than what has been observed for other TFs in flies; for 6 A-P patterning factors examined in *Drosophila melanogaster* and *D. yakuba* [330], for example, only 0.3% - 14.2% of binding peaks were found to be species-specific. A similar pattern of extensive conservation has been observed for these factors in

Drosophila pseudoobscura as well (Mathilde Paris and Xiaoyong Li, personal communication). Lower levels of conservation have been observed for TFs in some yeast species; for example, ChIP analysis of the Tec1 and Ste12 proteins from *S. cerevisiae*, *S. mikatae*, and *S. bayanus* found conservation of only 20% of binding events among species [105], even though these species span a similar evolutionary distance to *D. mel* and *D. yak* (10 myr). Extensive divergence in TF binding has also been observed in the vertebrate TFs CEBPA and HNF4A among human, mouse, dog, possum and chick [106, 107], although these species are more divergent than the yeast and fly cases (80myr-300myr). No examples of TF binding divergence on the scale of that observed for HER over such a short evolutionary timescale (25myr between *D. pse* and *D. mel*) have yet been found for a metazooan, although given the patterns observed for more divergent species divergence is likely to be the rule rather than the exception [104].

My data indicate that the extensive divergence in HER binding is at least partially due to changes in the *cis*sequences that are predicted to be bound by HER. Examination of aligned sequences underlying divergent peaks shows that the majority of examples of differential binding between *D. pse* and *D. mel* are accompanied by the occurrence of a species-specific binding site near the binding peak. Conversely, 8 out of 9 shared peaks contain a conserved binding site or shifted binding site in both species. Although many shared peaks contain species-specific binding sites, a larger portion contain conserved sites than was observed for sites that have completely diverged. This indicates that factors other than gain and loss of binding sites could be responsible for the shift in binding in these species.

The effect of these massive changes in binding on HER function in each species is unclear; however, the clearest link between HER binding and the known mutant phenotypes of HER for somatic sex determination, the binding of HER to the intergenic region upstream of *ix*, does appear to be conserved. In other cases of large-scale binding divergence (e.g. [106, 107]), the conserved target genes appear to represent a core set of target genes that are known to be functionally regulated by the divergent factor. Other than *ix*, there are no obvious connections between the conserved target genes (Table 4.2) and known HER functions, although a number of them (e.g. *salm*) have lethal phenotypes and could therefore be involved in the functions of *her* outside of the sex-determination pathway.

4.5.3 Divergence of the HER binding motif between *Drosophila melanogaster* and *Drosophila pseudoobscura*

The original motivation for this study was to determine the effects of the sequence differences observed between the *Drosophila pseudoobscura* and *Drosophila melanogaster* HER proteins, specifically whether these changes had any effect on the DNA-binding preference of this protein in the two species. MEME motif enrichment analysis showed that peaks of HER binding in *Drosophila pseudoobscura* are enriched for a motif that closely resembles that observed from a similar analysis of *D. mel* HER bound regions. Both motifs show a strong preference for CGC and GAG, at the 5' and 3' ends respectively; however, the motif created from *D. pse* her-bound sequences shows a stronger preference for As within the six base pair linker region that connects these two regions. This significantly increases the specificity of the motif — although the HER position weight matrix

constructed using the *Drosophila melanogaster* data predicts 20,000 binding sites genome wide for both *Drosophila pseudoobscura* and *Drosophila melanogaster* , the *Drosophila pseudoobscura* motif predicts 1400 sites for *D. pse* and 1200 sites for *D. mel* . Although one appealing explanation for this difference is that the HER protein has diverged in its DNA-binding preference between the two species, without *in vitro* binding data for HER it is not possible to definitively say that this is the case. Since the size of the datasets (number of bound regions) examined in the two species was similar, it is unlikely that this difference could be due to a sampling effect; furthermore, the number of peaks containing a site matching the *D. pse* motif is more than 4 times in the *D. pse* dataset compared to the *D. mel* data, even though enrichment for *D. mel* data was generally higher. If the *Drosophila pseudoobscura* antibody had a lower affinity for HER compared to the *D. mel* antibody, a more specific motif could be obtained simply from the *D. pse* antibody only giving signal at the strongest-bound regions; if this were the case we would expect *D. pse* HER binding to be a subset of *D. mel* binding, and the number of *Drosophila pseudoobscura* motifs to be similar between the two species, neither of which is true in my data.

4.5.4 *D. pse* -specific Secondary HER motifs May explain the Divergence in Binding

The most striking feature of the *Drosophila pseudoobscura* HER data is the presence of a large subset of bound regions that are strongly enriched for a set of sequence motifs separate from the primary HER motif. Instances of these secondary HER motifs tend to co-occur within the same sequence, where they show remarkably tight linear ordering and spacing. These sequences are specific to the *D. pse* data; with the exception of her-mtx-4, which appears to be a simple repeat, only a single occurrence of any secondary motif is found within any of the *D. mel* HER bound regions. Although it is possible that these sequences could be a novel *cis*-regulatory element with strict requirements for site spacing, it is unlikely that such an element could have proliferated to such a great extent between *D. mel* and *D. pse* by point mutations alone.

Interestingly, alignments of the bound regions that contain hits for these motifs show extensive similarity even outside of the regions corresponding to motif hits. This indicates that the similarity between the bound regions is due not to common transcription factor binding sites, but rather to a common sequence origin for these regions, possibly from a transposon or other repetetive sequence element. The exaption of transposable elements for use in gene-regulatory roles is known to be a common mechanism for gene regulatory divergence [331, 332]; many of these elements have been shown to exhibit enhancer-like effects on gene expression and pre-existing TF binding sites [333], and a significant proportion of human *cis*-regulatory elements are known to be derived from transposons [334, 335]. However, this mechanism has not been observed in other cases of *Drosophila cis*-regulatory evolution, which generally focus on gain and loss of binding sites by point mutations. Interestingly, the sequences containing the *D. pse* HER secondary motifs appear to be specific to the obscura group — BLAST searches against other fly species using these sequences do not return any similar results for species other than *Drosophila pseudoobscura* and *D. persimilis*, although it is possible that such sequences could be too divergent to be recognized. Although the sequences are similar to one another, they do not appear to have the high level

of sequence similarity usually observed for active transposons [336]; it is therefore possible that these elements are the result of a burst of sequence duplication or transposition events rather than an ongoing transposition process.

The relationship between these repeated sequences and the divergence within the HER DNA binding domain is as yet unclear, but their presence does hint at some possible novel mechanisms for evolutionary change in transcription factor DNA-binding preference. One possibility is that HER has not changed its DNA binding preference, and that the observed difference in the motifs learned from MEME is due to the larger pool of identical binding sites contained within the repetitive elements. If HER can bind these sequences strongly, then even if no change in binding preference has occurred, any motif-enrichment analysis of bound regions would show a strong enrichment for the binding site contained in the repetetive element. Examination of the putative HER binding sites contained in the repetitive bound regions did show a strong enrichment for sequences matching the *D. pse* -specific motif; however, this alone does not explain all of the divergence in the *D. pse* HER motif because even binding sites not contained within repetitive regions are enriched for the *D. pse* -specific motif.

It is possible that some or all of these could be degenerate versions of the repetitive element. It is also possible that the proliferation of the repetitive element during *Drosophila pseudoobscura* evolution could have played a role in the divergence of HER in *dpse*. Because *her* appears to be a relatively young gene, with no close fly orthologs beyond *D. willistoni* (see Ch. 1), it is likely that *hermaphrodite* experienced a period of functional redundancy in the common ancestor of *D. wil*,*D. pse* , and *D. mel* ; if this redundancy led to subfunctionalization of target genes between *her* and its (unknown) paralog, it is possible that HER could have bound a relatively small set of targets in this ancestor. If the *D. pse* -specific repetetive element proliferated after the split of the *D. pse* and *D. mel* lineages, these sequences could have acquired regulatory function in the *D. pse* ancestor by inserting near promoters; HER could then bind to these novel regulatory regions, and if this binding were adaptive, there could be selective pressure to change binding preference to match the new targets. If the original set of HER targets were small, there would not be much selective pressure to maintain the wide range of binding specificity observed in *Drosophila melanogaster her* , and the *Drosophila pseudoobscura* HER could have acccumulated mutations that allowed it to strengthen its interaction with the binding site in the repetetive element by becoming more specific. Alternatively, this loss of selective pressure could have led to accumulation of neutral mutations that destroyed the ability of HER to bind a wide range of sequences while maintaining binding to the repetitive sequence and the original targets. Finally, it is possible that binding to these sequences is unrelated to the the change in binding — it is certainly conceivable that the *hermaphrodite* binding to these sequences is purely coincidental, and has no function consequence beyond acting as a binding "sink."

Although more experiments are needed to determine the exact basis of the differences in HER specificity and genome wide binding between the *D. pse* and *D. mel* HER proteins, it seems likely that HER represents an interesting and unusual case of regulatory divergence. In particular, newly-developed *in vitro* binding techniques such as MITOMI [337] should be able to definitively determine whether the differences in the sequences

HER binds in these two species is due to sequence changes in the protein or to a proliferation of *Drosophila pseudoobscura* HER binding sites within the genome, or both in combination. Better characterization of HER function in these species should also provide insight into possible selective mechanisms that may have influenced the changes in HER; in particular, it would be useful to know if the role of *her* in sex determination is in fact conserved in *Drosophila pseudoobscura* , and whether the extensive changes in target genes between these two species is reflected in novel functions for *her* in *Drosophila pseudoobscura* . Tests of the repetetive regions for regulatory function would also be help to determine whether the observed novel HER binding is in fact functional. Regardless of the exact mechanisms, the large scale divergence of *her* makes it unique, at least among fly TFs, and it therefore represents an excellent candidate for future studies of both *cis*and *trans*regulatory evolution.

Table 4.1: *dmel* ChIP-seq Illumina Sequencing and Mapping

Sample	**Total Reads**	**Mapped Reads**	**Percent Mapped**
Input	27376945	14068837	51.3%
D. pse herA	27811647	16029714	57.6%
D. pse herB	28692306	8134896	28.3%
IgG	25876520	10192362	39.3%

Table 4.2: Shared Nearby Genes (*D. mel* Orthologs)

Gene Name	**FBid**	**Distance to *D. mel* Peak**	**Distance to *D. pse* Peak**
trbl	FBgn0028978	3832	1022
RhoGAP71E	FBgn0036518	10	768
Mio	FBgn0032940	14	1884
shn	FBgn0003396	105	1488
CG1504	FBgn0031100	3880	56
RhoGAP93B	FBgn0038853	593	3138
noc	FBgn0005771	858	531
Csp	FBgn0004179	609	3210
mip40	FBgn0034430	12	908
CG7137	FBgn0034422	15	1341
CG12163	FBgn0037303	208	224
osa	FBgn0003013	41	2095
san	FBgn0024188	297	124
MED10	FBgn0036581	159	95
Neb-cGP	FBgn0083167	253	384
CG3529	FBgn0035995	165	287
salm	FBgn0004579	527	1705

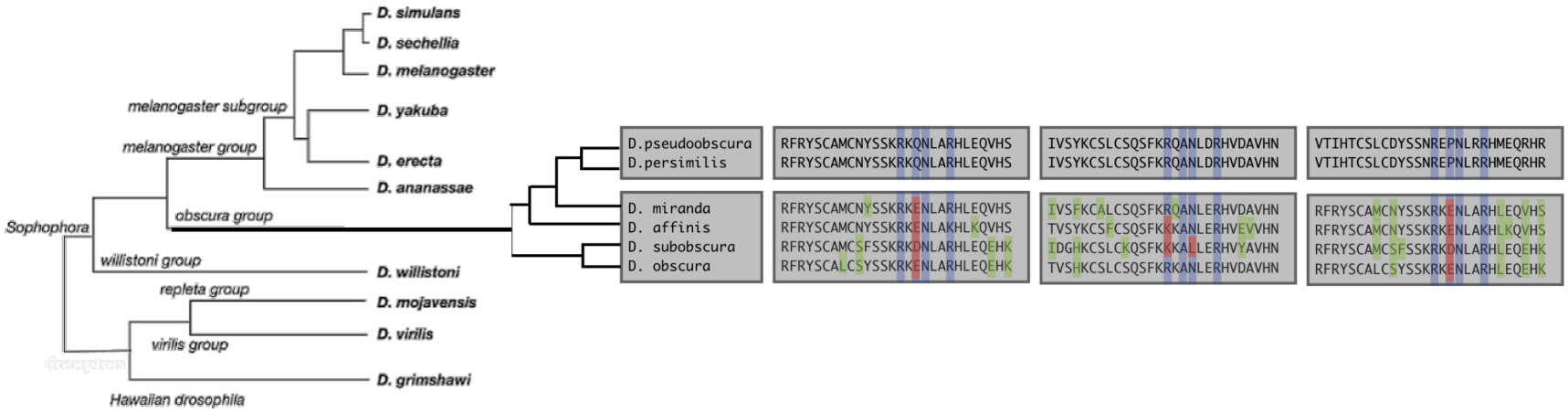


Figure 4.1: ***her* DNA binding Domains from Inserted Region in Species of the Obscura group:** Aligned C2H2 Zinc Finger sequences obtained by degenrate PCR from non-sequenced species of the Obscura group. Blue shaded residues are the -1,2,3,6 canonical DNA-contacting residues of the C2H2 zinc finger. Amino acid substitutions outside of DNA-binding domains are shaded green, substitutions at DNA-contacting residues red.

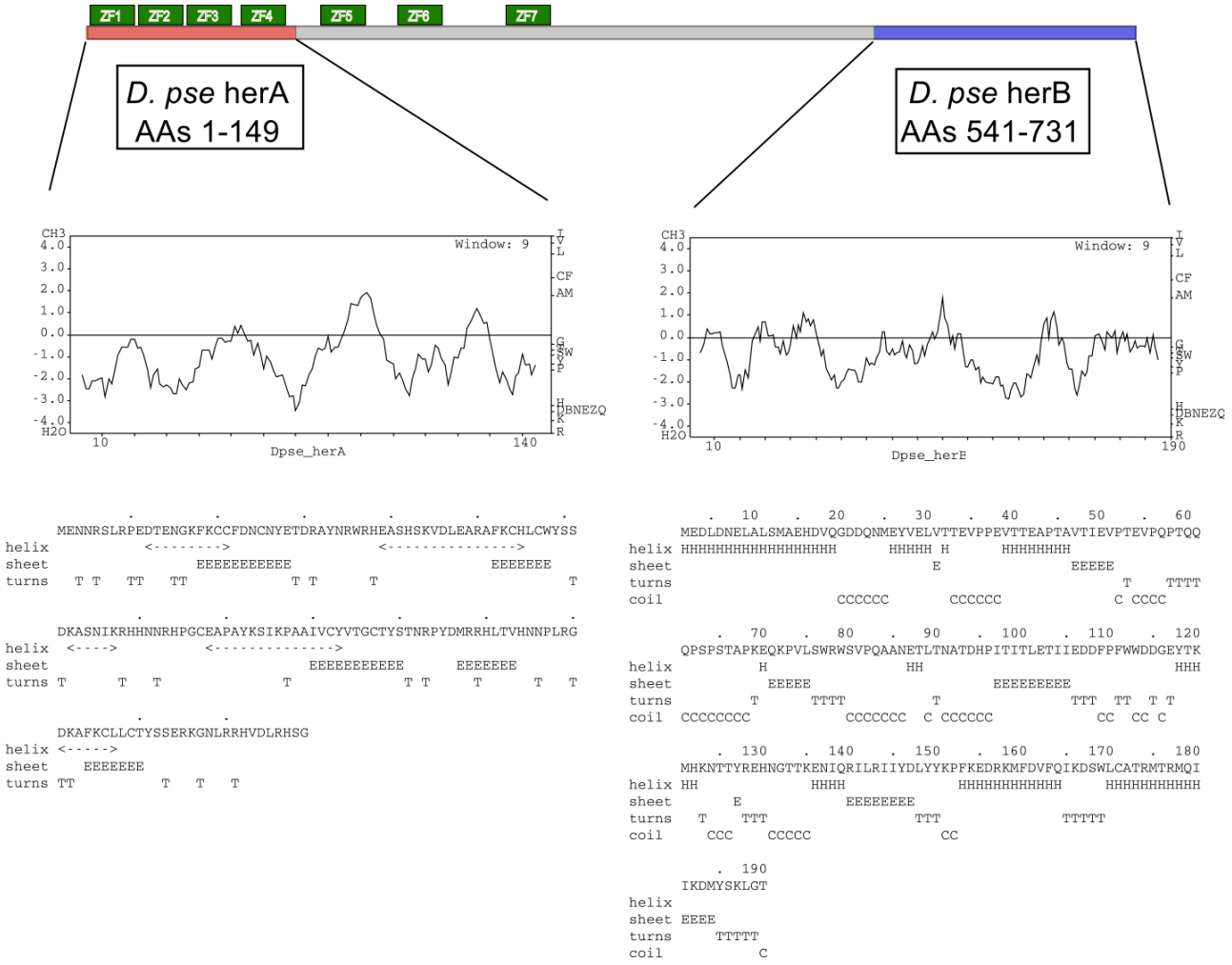


Figure 4.2: ***Drosophila pseudoobscura* anti-HER Antibody Construction:** Antibodies were raised against *Drosophila pseudoobscura* HER protein fragments pA and pB as shown. The Ponting:2002p7916 peptide included 4 of the C_2H_2 -zf binding domains; the pB antibody contained a conserved region at C-terminus of the protein. Hydropathy and predicted secondary structures for each protein are shown below the sequence map.

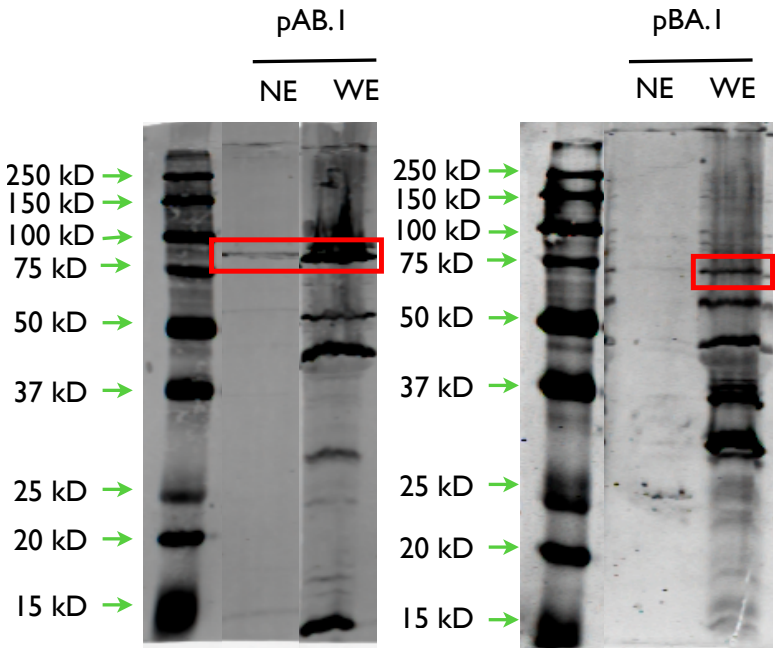


Figure 4.3: ***her Drosophila pseudoobscura* anti-HER Antibody Western Blots:** Blots using anti-(D. pse) HER antibodies to probe whole embryo (WE) and nuclear (NE) extracts prepared from 0-12h *Drosophila pseudoobscura* embryos. The Ponting:2002p7916 blot (left) shows a bands in both the for both nuclear extract and whole embryo extract between 75kD and 100kD, which matches the expected size of the *Drosophila pseudoobscura* HER protein (83kD). The pB blot shows no bands for nuclear extract, and a possible band near 75kD in the whole-embryo extract

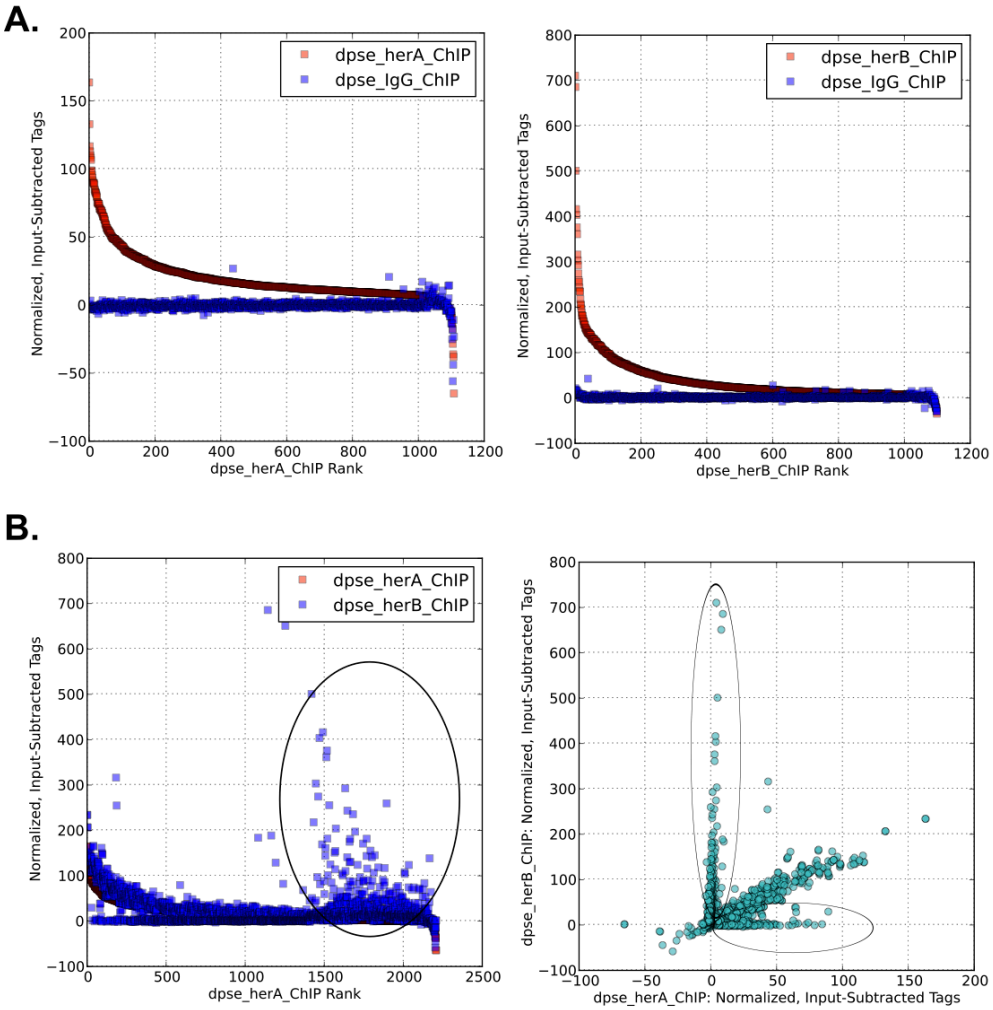


Figure 4.4: ***Drosophila pseudoobscura* HER ChIP-seq Peak Enrichment:** ChIP-seq was performed with chromatin prepared from 4.5-6.5h *Drosophila pseudoobscura* embryos , using the anti-*D.pse* HER antibodies pA and pB. Peaks were called using a custom algorithm which fits a modeled peak shape to the data (Tommy Kaplan, unpublished). Genome-wide tag counts were normalized to the equivalent of 1000000 total tags, and the normalized Input DNA tag count was subtracted from each bound region. Plots in **(A.)** compare the tag count from the treatment sample for each bound region (red points) with the tag count from the same region in the control ChIP reaction using normal rabbit IgG (blue points). Points are ranked left-to-right along the x-axis by the treatment signal. pA ChIP is shown in the left panel, and pB ChIP in the right. Both reactions show significant enrichment over control (K-S test p-value = 0.0 for pA, 0.0 for pB). **(B.)** shows a comparison of enrichment for pooled bound regions between the two antibody samples. Left panel shows all bound regions from pA and pB; red points represent the pA tag count, blue points the pB tag count for each region. Regions are sorted left-to-right by the strength of the pA signal. Right panel shows a scatterplot of values for the same regions; each datapoint represents the signal in a bound regions. A number of points show strong enrichment in one antibody, but not the other (circled). Pearson $r = 0.93$, Spearman $r = 0.99$

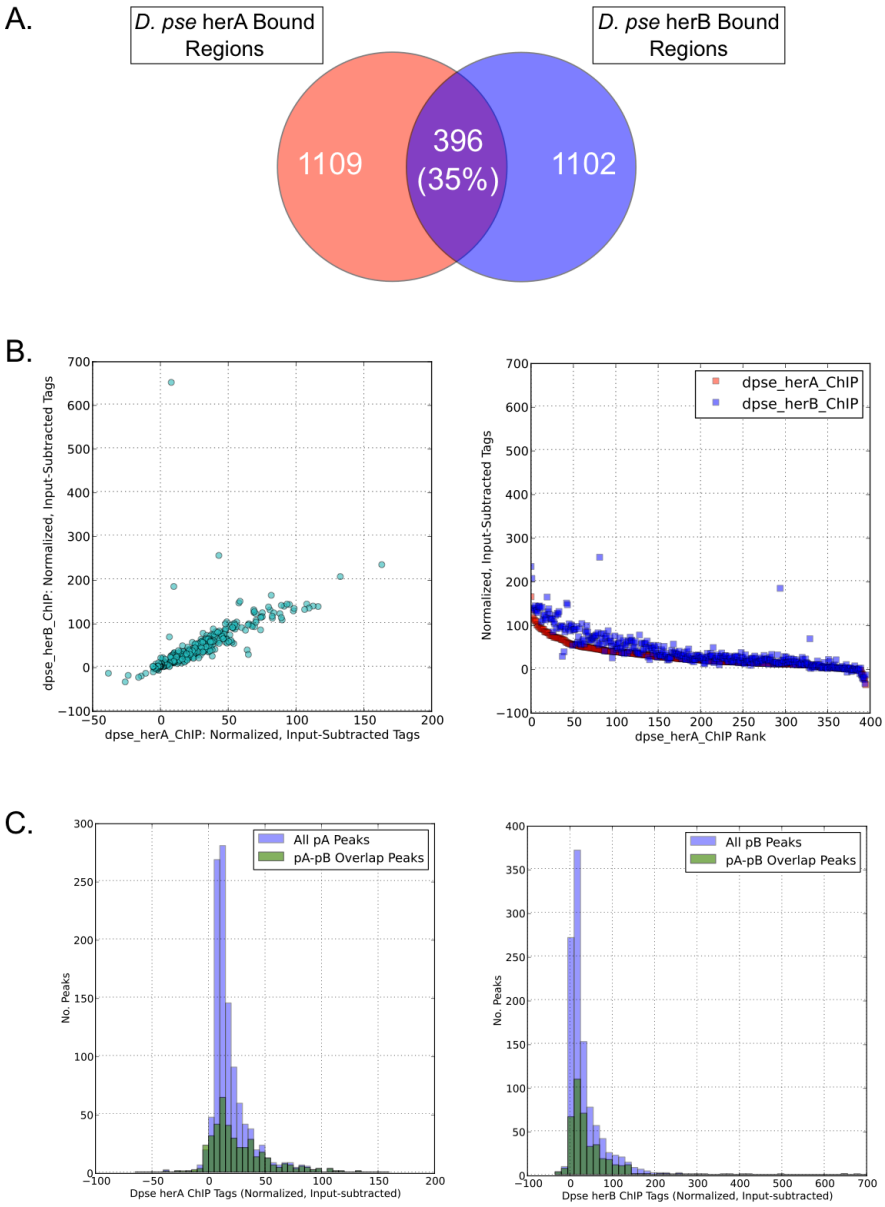


Figure 4.5: **Overlap in Peaks Between ChIP-seq Datasets for pA and pB Antibodies:** Overlap between the pA and pB datasets were taken for further analysis; approximately 35% of called peaks from each dataset overlap with a peak from the other dataset (**(A.)**). After filtering for overlaps, agreement between the two datasets is improved (Pearson r = 0.93, Spearman r = 1.0); **(B.)** shows a scatterplot of tag counts in each dataset following filtering for non-overlapping peaks (left) and a ranked-enrichment plot for the two datasets following filtering (right; points are ranked by pA value). **(C.)** shows histograms of the distribution of peak heights for each antibody before overlap filtering (blue bars) and after filtering (green bars). Filtering removed mostly low-enrichment peaks for pA, but a number of highly enriched peaks for pB.

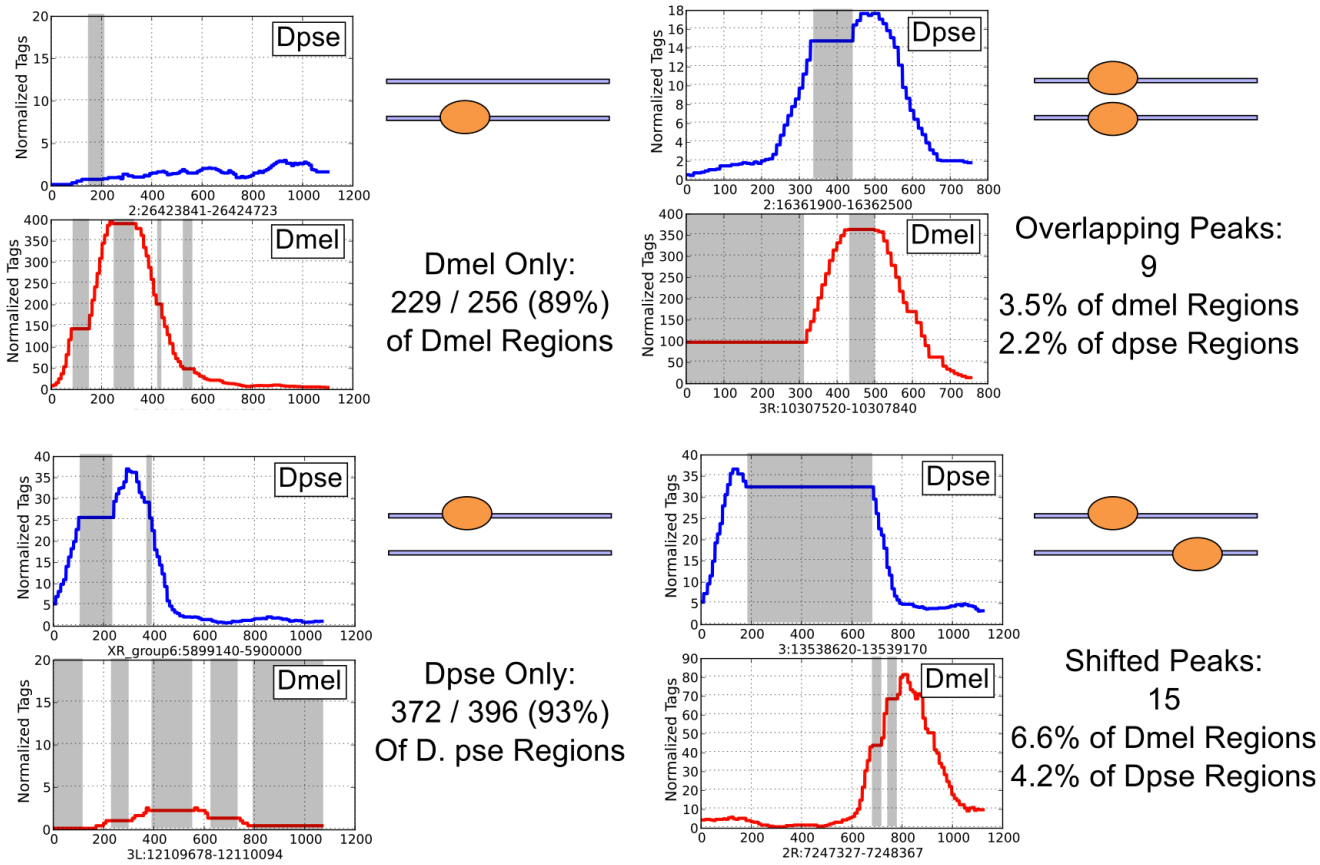


Figure 4.6: **Conservation and Divergence of HER Binding between *Drosophila melanogaster* and *Drosophila pseudoobscura* :** Overlap in bound regions between *Drosophila melanogaster* and *Drosophila pseudoobscura* HER datasets was examined using a whole-genome alignment between *Drosophila melanogaster* and *Drosophila pseudoobscura* using the alignment software FSA [321]. Shared peaks (upper right, 3% of total) were defined as peaks which were within 100bp of one another in transformed sequence coordinates in both genomes. Shifted peaks (lower right, 5-6% of total) were defined as peak regions which do not match the direct overlap criteria but have a peak in the aligned species within 1kb of the peak region endpoints. The remaining peaks (90%, upper left and lower left) are species-specific, with no nearby peaks in the aligned species. Plots show enrichment at aligned positions in the two species, grey areas represent gapped regions of the alignment in each species.

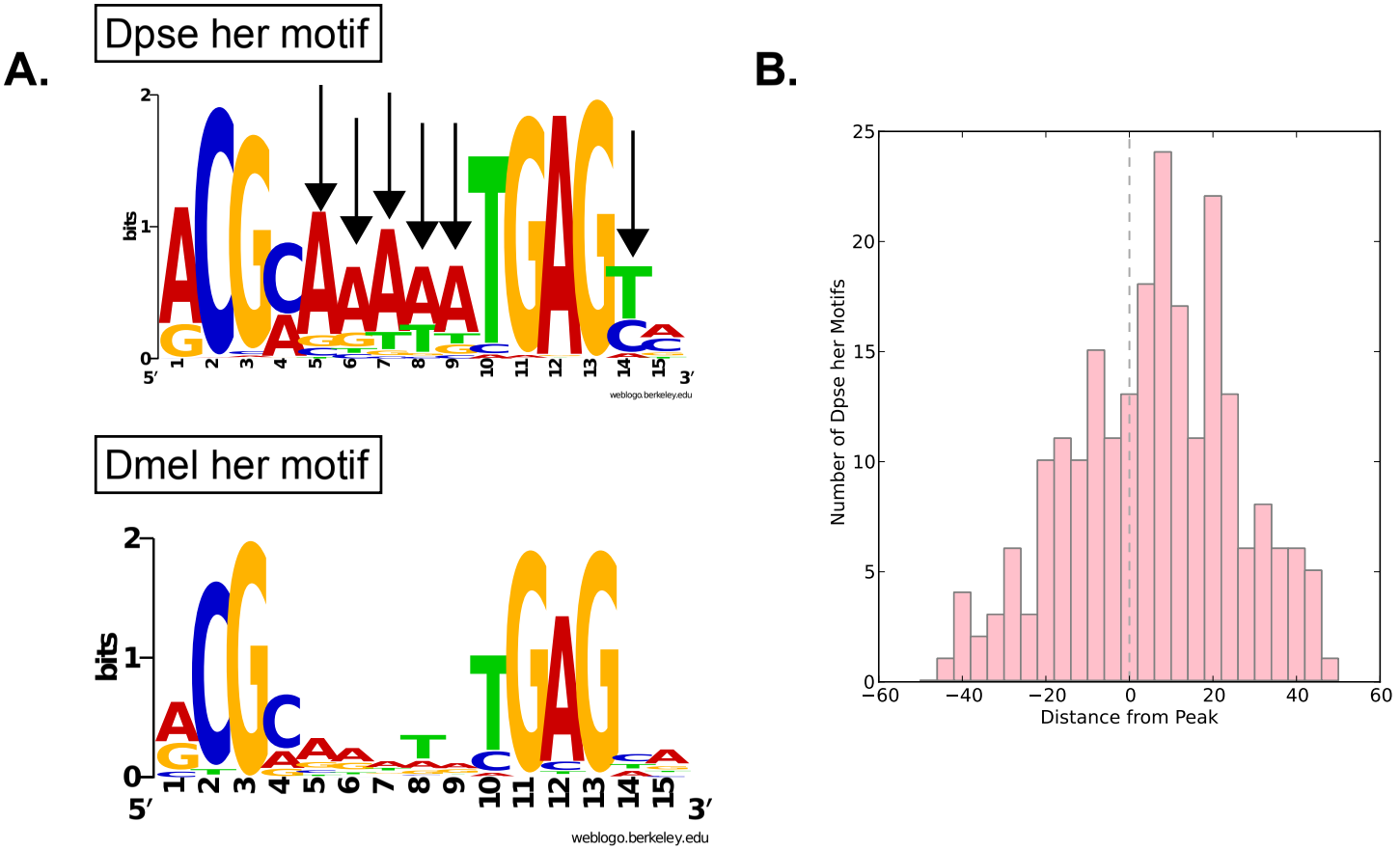


Figure 4.7: **HER-bound regions in *dpse* are Enriched for an Altered HER motif:(A.)Analysis of motif enrichment in *Drosophila pseudoobscura* HER peaks using the motif-finding software MEME [303] shows strong enrichment (E-value 1.4e-709) of a motif (top) which resembles the *Drosophila melanogaster* HER motif (bottom); positions with differences in *dpse* are marked with an arrow. (B.) shows distribution of motif hits called by Patser [304] within 100-bp windows surrounding HER peaks.**

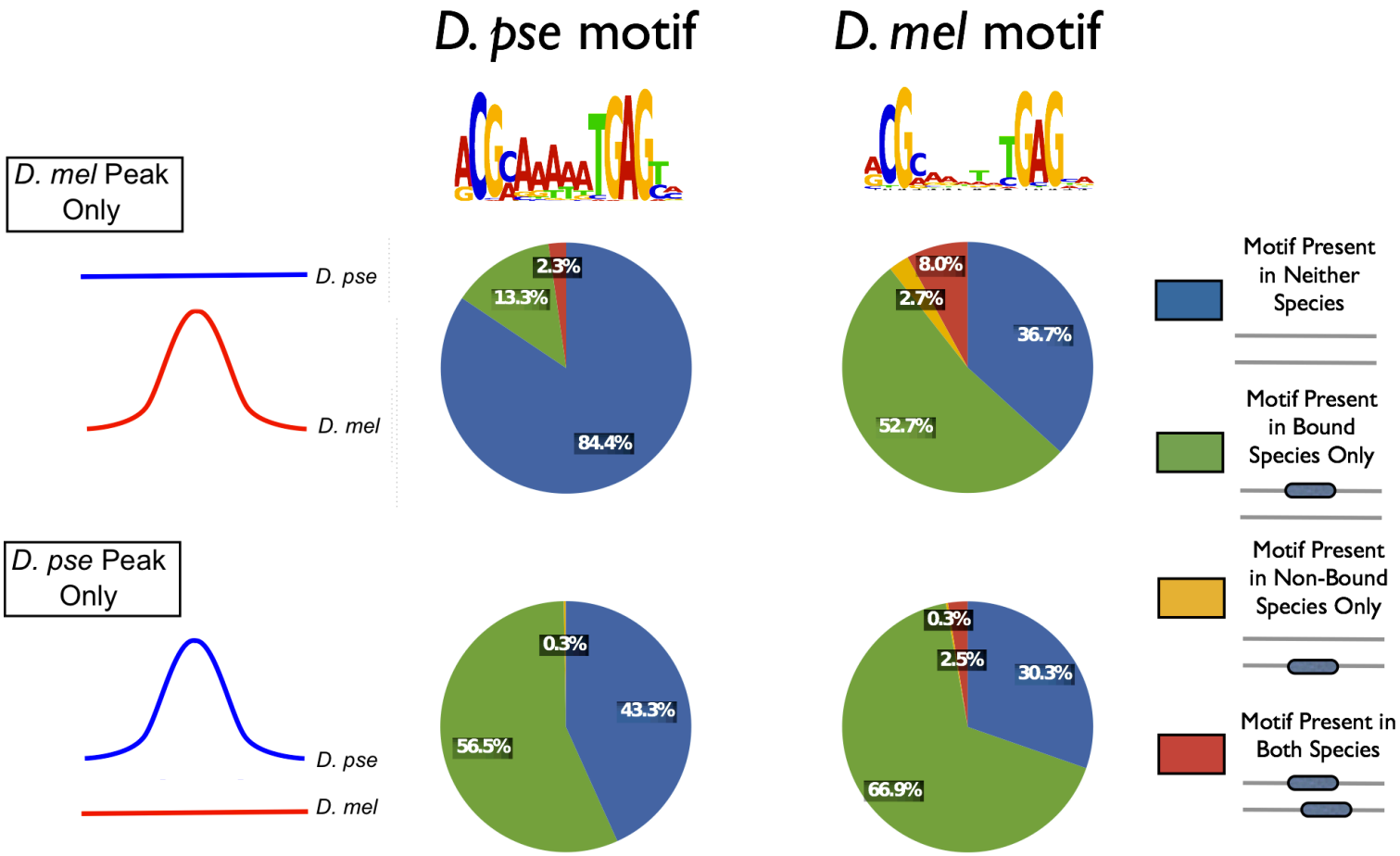


Figure 4.8: **Conservation and Divergence in *Drosophila pseudoobscura* and *Drosophila melanogaster* motifs in Non-conserved Bound Regions:** Charts show proportion of peaks which have an occurrence of the indicated motif within 100bp of the peak summit in aligned regions from each species. Peaks show either conservation of the motif in both species (red), an occurrence of the motif only in the bound species (green), an occurrence only in the unbound species (yellow) or absence of a motif in both species (blue). *Drosophila pseudoobscura* peaks show a higher occurrence of *Drosophila pseudoobscura* motifs which are not conserved at orthologous positions in *Drosophila melanogaster* (lower left)

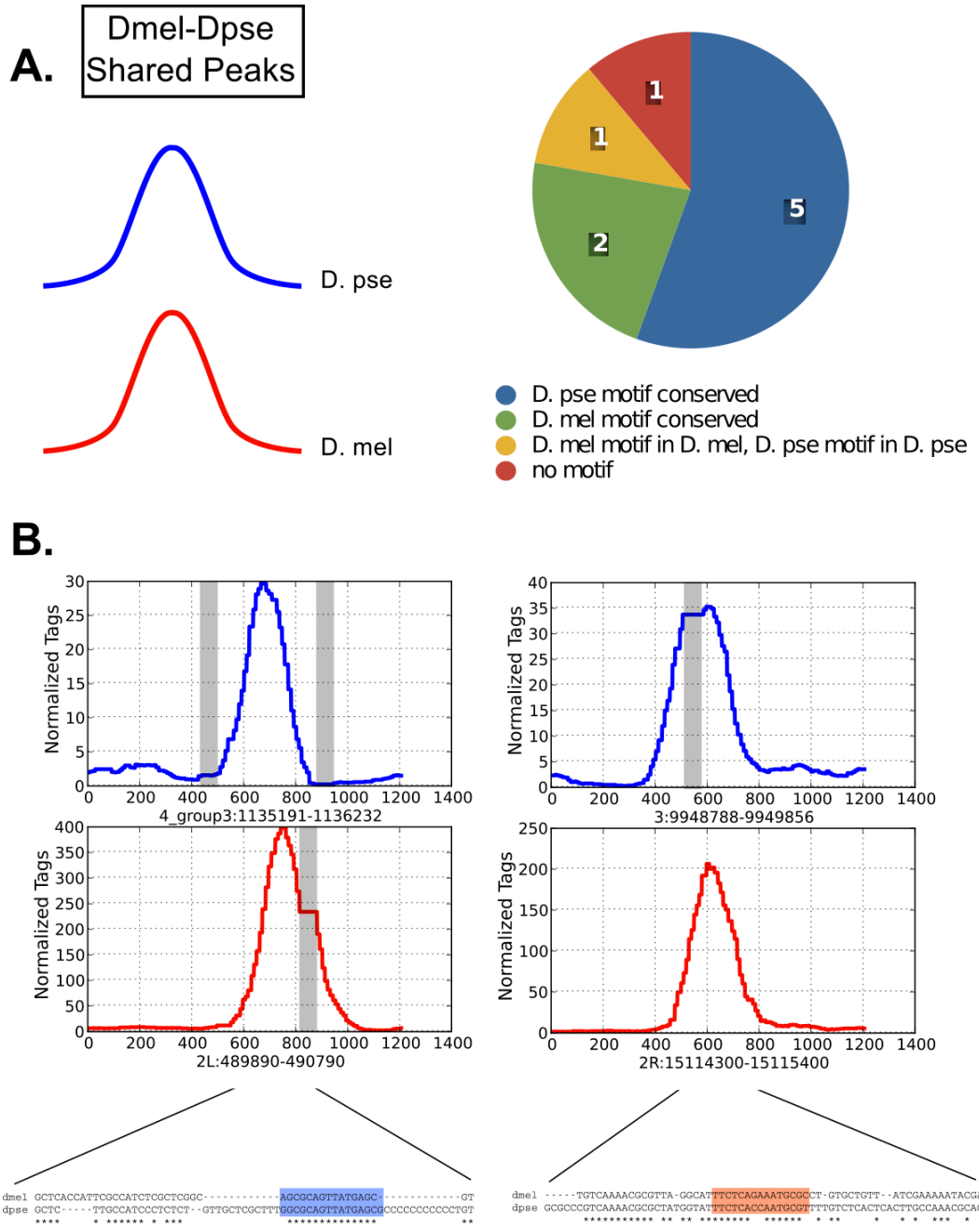


Figure 4.9: **Conservation and Divergence in *Drosophila pseudoobscura* and *Drosophila melanogaster* motifs in Shared Bound Regions: (A.)** Motif occurrence in 100-bp window flanking the shared peaks. 5/9 of the conserved peaks have a conserved *Drosophila pseudoobscura* HER motif **(B.)** Example plots showing motif conservation in shared peaks. Left plot shows conservation of the *Drosophila pseudoobscura* her motif in both species (blue box), right plot shows conservation of the *Drosophila melanogaster* motif in both species (red box).

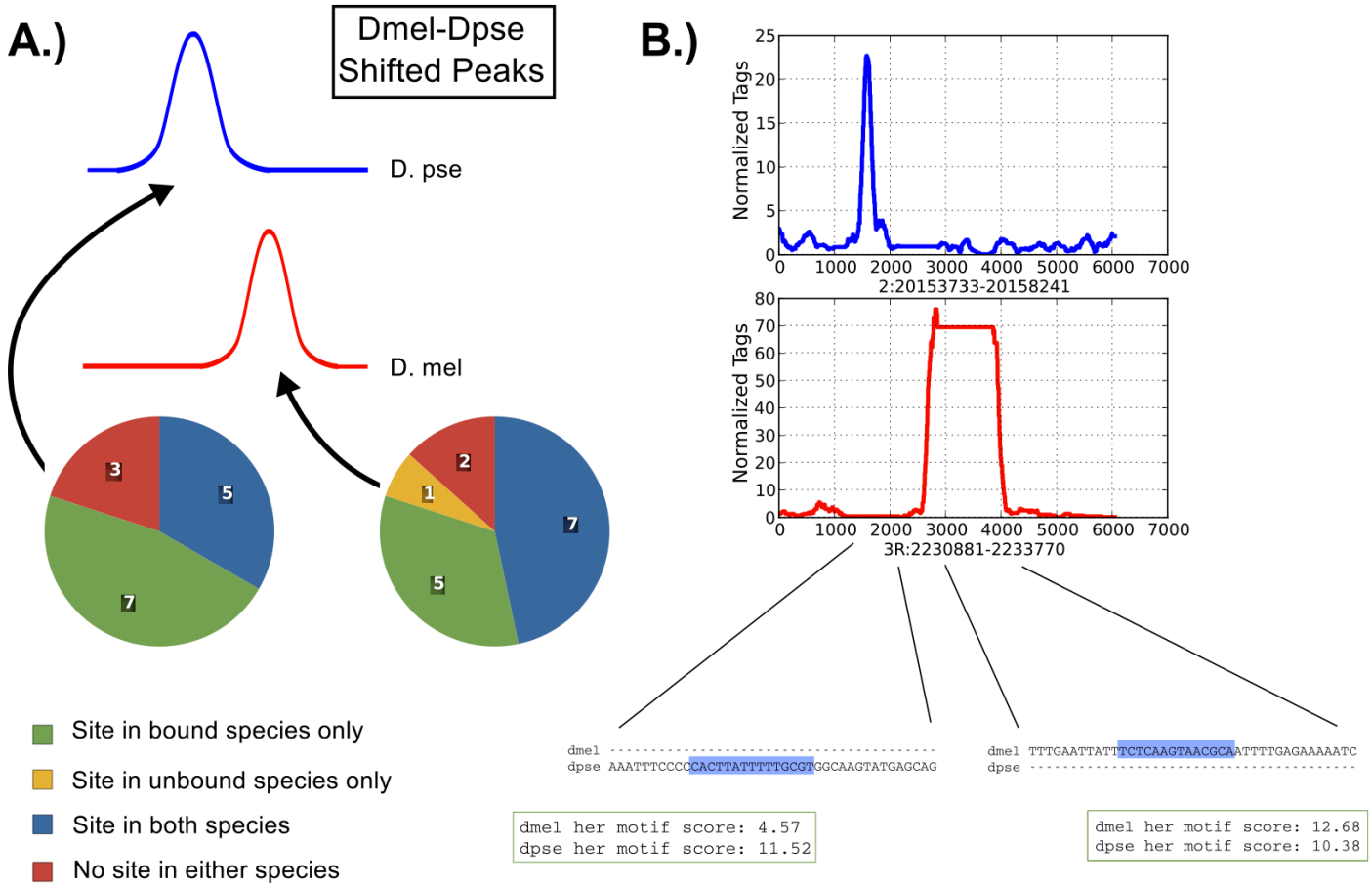


Figure 4.10: **Conservation and Divergence of *Drosophila pseudoobscura* and *Drosophila melanogaster* binding sites in Shifted HER-bound regions: (A.)** Conservation of a HER binding site beneath peaks which do not directly overlap, but are located within 1000bp of one another (shifted peaks, n=15); 'sites' here refer to binding sites matching either the *Drosophila melanogaster* or the *Drosophila pseudoobscura* HER binding motif. **(B)** shows an example plot of a pair of shifted peaks and the corresponding binding sites; each peak contains a species-specific binding site which lies within an apparent indel region; both of these sites are strong matches to the *D. pse* motif, but the *D. pse* -specific peak contains a binding site which strongly matches the *D. pse* but not the *D. mel* motif.

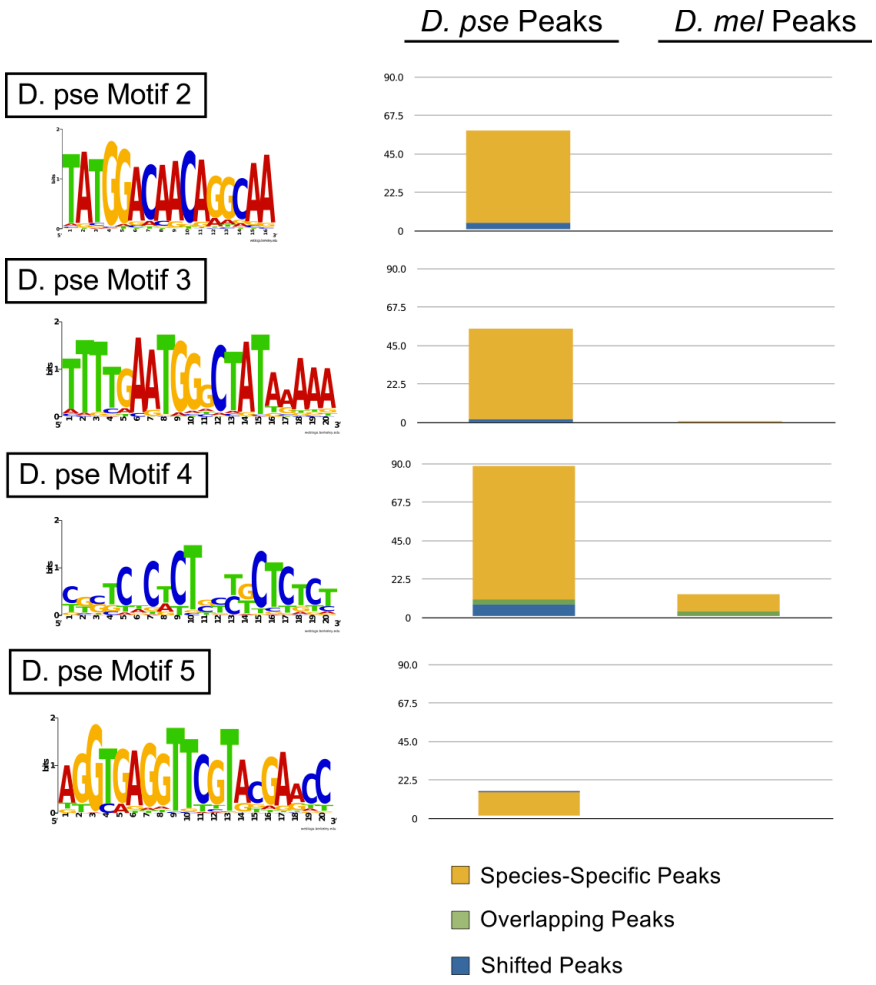


Figure 4.11: **Secondary Motifs Identified in *Drosophila pseudoobscura* HER-bound Regions:**MEME analysis of her-bound sequences identified four significant motifs in addition to the primary HER motif (Fig. 4.7) (left column). Plots on the right show the frequency of each motif in different peak categories for both *D. pse* and *D. mel* ; motifs 2, 3 and 5 show strong enrichment only in *D. pse* peaks; of these three, there is only a single occurance in a *D. mel* peak (of Motif 3)

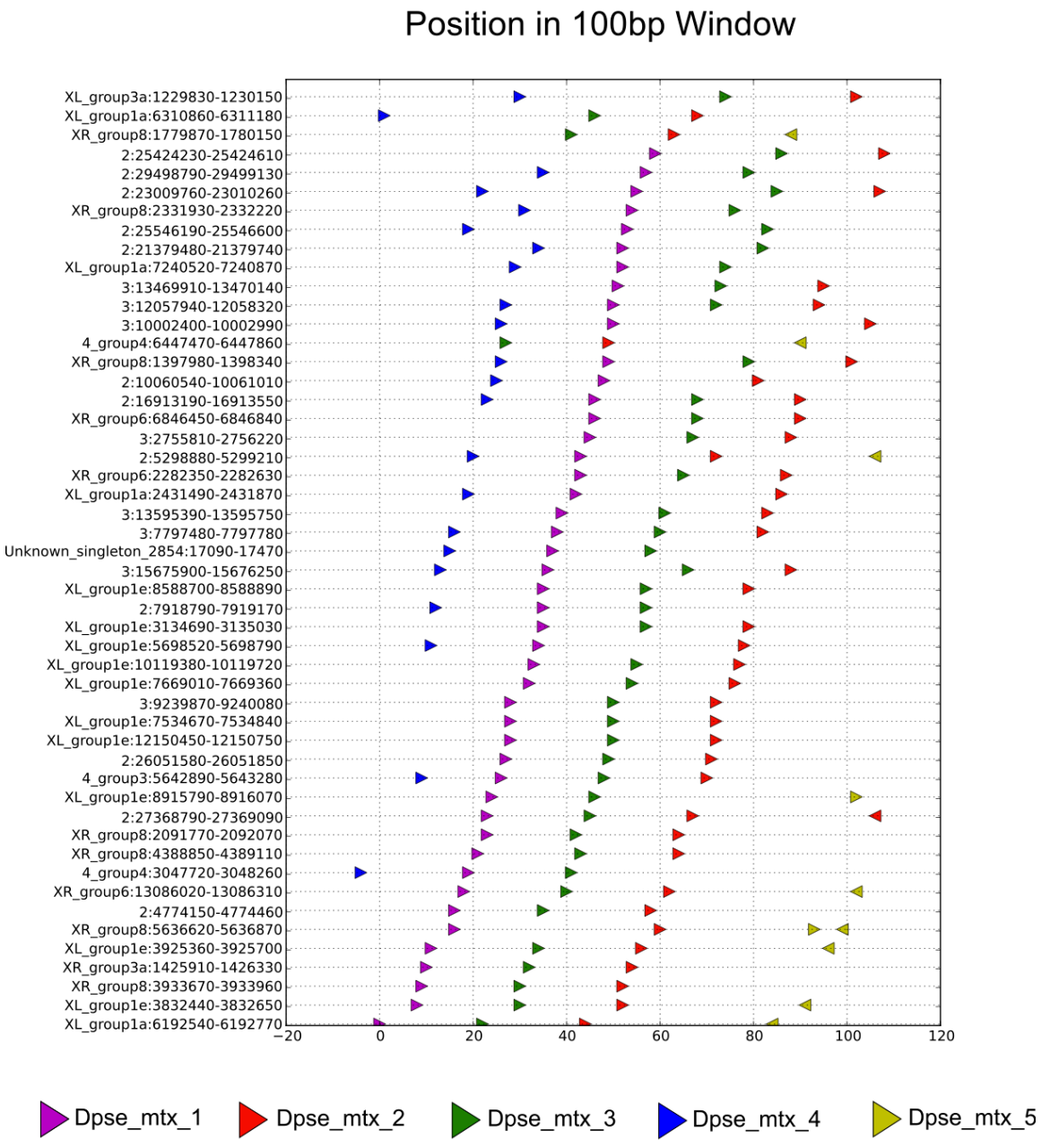


Figure 4.12: **Motif Distribution in HER-bound Regions Containing Multiple Occurrences of HER Secondary Motifs:** Plot shows the inear distribution of *D. pse* HER motifs in the 50 HER-bound regions which contain hits for three or more *D. pse* HER motifs. Sequences are oriented such that the dpse HER primary matrix (Dpse-mtx-1) is always in the positive orientation. Sequences are sorted by start position of the most central HER Dpse-mtx-1. Note that this does not change position or orientation of the other motifs.

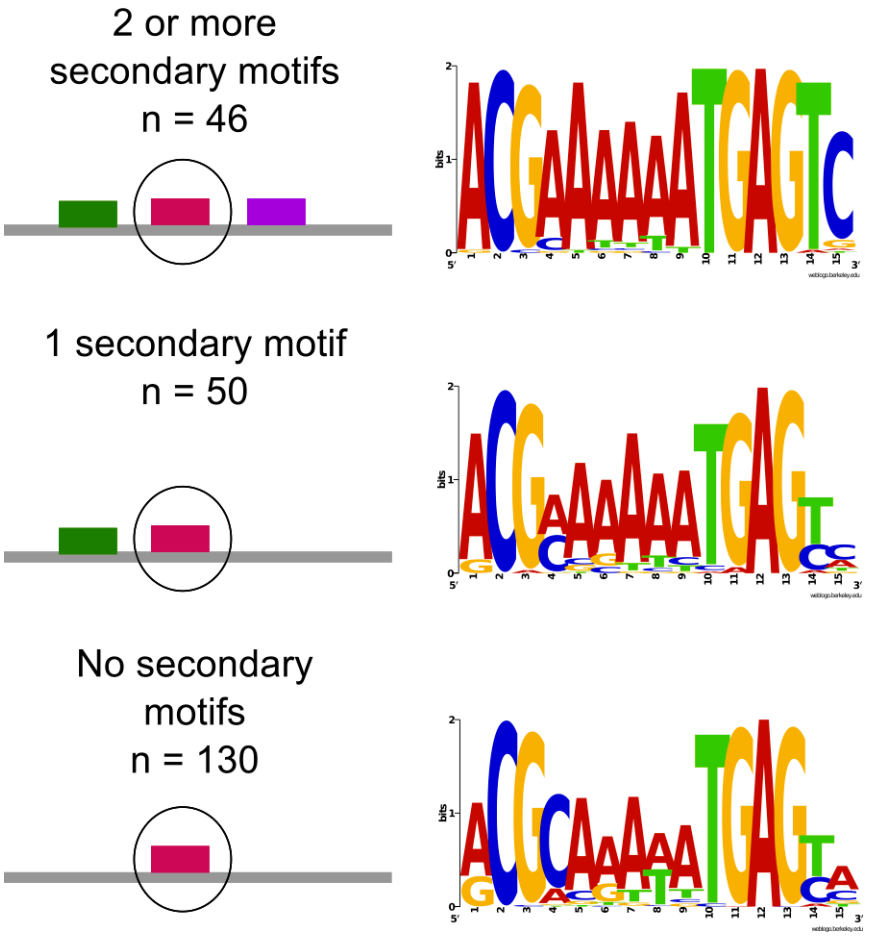


Figure 4.13: **HER Motifs Calculated from HER-bound Regions Containing *D. pse* HER Secondary Sites:** Motifs on right were re-calculated using predicted HER primary binding sites from HER bound regions containing two or more secondary motifs (Top), exactly one secondary motif (Middle) or only a primary HER site (Bottom)

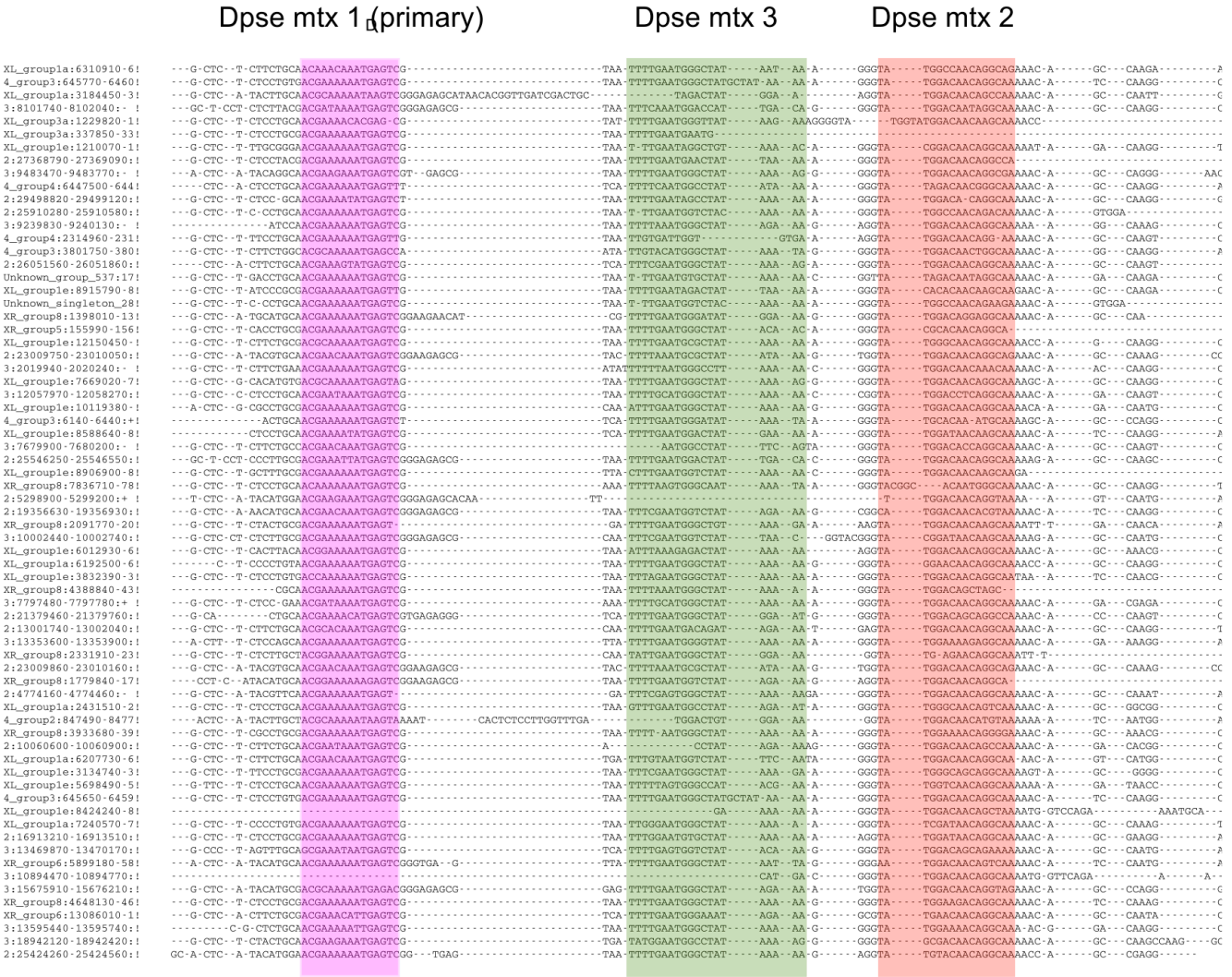


Figure 4.14: **Sequence Alignment of HER-bound regions containing 3 or more HER Motifs:** A portion of the global alignment of 50 HER-bound sequences containing three or more HER motifs, showing the region containing the HER primary motif (magenta), the dpse-mtx-3 motif (green) and the dpse-mtx-2 motif (red)

Bibliography

[1] Darwin C (2010) The Origin of Species: By Means of Natural Selection of the Preservation of Favoured Races in the Struggle for Life. Gale. URL http://books.google.com/books?id=ategSgAACAAJ.

[2] Morgan T (1916) A Critique of the Theory of Evolution. Lectures delivered at Princeton University, 1916. Princeton University Press.

[3] Jacob F, Monod J (1961) On the regulation of gene activity. Cold Spring Harb Symp Quant Biol 26:193–211.

[4] Zuckerkandl E, Pauling L (1965) Molecules as documents of evolutionary history. J Theor Biol 8:357–66. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WMD-4JJG0MS-C&_user=4420&_coverDate=03%252F31%252F1965&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=1d5fdb282a8d0eb7a8b08a4e566ab653.

[5] Ohno S (1972) An argument for the genetic simplicity of man and other mammals* 1. Journal of Human Evolution URL http://linkinghub.elsevier.com/retrieve/pii/0047248472900115.

[6] King MC, Wilson AC (1975) Evolution at two levels in humans and chimpanzees. Science 188:107–16.

[7] Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016. doi:10.1111/j.1558-5646.2007.00105.x. URL http://www3.interscience.wiley.com/journal/117958848/abstract?CRETRY=1&SRETRY=0.

[8] Carroll SB (2005) Evolution at two levels: on genes and form. PLoS Biol 3:e245. doi:10.1371/journal.pbio.0030245. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0030245.

[9] Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci USA 104 Suppl 1:8605–12. doi:10.1073/pnas.0700488104. URL http://www.pnas.org/content/104/suppl.1/8605.long.

[10] Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206–16. doi:10.1038/nrg2063. URL http://www.nature.com/nrg/journal/v8/n3/abs/nrg2063.html.

[11] Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, et al. (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20:1377–419. doi:10.1093/molbev/msg140. URL http://mbe.oxfordjournals.org/cgi/content/full/20/9/1377?view=long&pmid=12777501.

[12] Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134:25–36. doi:10.1016/j.cell.2008.06.030. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4SYCK28-9&_user=4420&_coverDate=07%252F11%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=341b39fad4c5c2fecba472e140d5aa66.

[13] Lynch VJ, Wagner GP (2008) Resurrecting the role of transcription factor change in developmental evolution. Evolution 62:2131–54. doi:10.1111/j.1558-5646.2008.00440.x. URL http://www3.interscience.wiley.com/journal/119880687/abstract?CRETRY=1&SRETRY=0.

[14] Hsia CC, McGinnis W (2003) Evolution of transcription factor function. Curr Opin Genet Dev 13:199–206.

[15] Carey M, Smale S (2000) Transcriptional Regulation in Eukaryotes : Concepts, Strategies, and Techniques. Cold Spring Harbor Laboratory Press.

[16] Smale ST, Kadonaga JT (2003) The rna polymerase ii core promoter. Annu Rev Biochem 72:449–79. doi:10.1146/annurev.biochem.72.121801.161520. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.biochem.72.121801.161520?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%253dncbi.nlm.nih.gov.

[17] Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. Crit Rev Biochem Mol Biol 41:105–78. doi:10.1080/10409230600648736.

[18] Lemon B, Tjian R (2000) Orchestrated response: a symphony of transcription factors for gene control. Genes Dev 14:2551–69. URL http://genesdev.cshlp.org/content/14/20/2551.long.

[19] Deato MDE, Tjian R (2008) An unexpected role of tafs and trfs in skeletal muscle differentiation: switching core promoter complexes. Cold Spring Harb Symp Quant Biol 73:217–25. doi:10.1101/sqb.2008.73.028. URL http://symposium.cshlp.org/content/73/217.long.

[20] Levine M, Tjian R (2003) Transcription regulation and animal diversity. Nature 424:147–51. doi:10.1038/nature01763. URL http://www.nature.com/nature/journal/v424/n6945/full/nature01763.html.

[21] Arnone MI, Davidson EH (1997) The hardwiring of development: organization and function of genomic regulatory systems. Development 124:1851–64. URL http://dev.biologists.org/content/124/10/1851.long.

[22] Moses AM, Pollard DA, Nix DA, Iyer VN, Li XY, et al. (2006) Large-scale turnover of functional transcription factor binding sites in drosophila. PLoS Comput Biol 2:e130. doi:10.1371/journal.pcbi.0020130. URL http://www.ploscompbiol.org/article/info%253Adoi%252F10.1371%252Fjournal.pcbi.0020130.

[23] Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-dna complexes. Genome Biol 1:REVIEWS001. URL http://genomebiology.com/2000/1/1/REVIEWS/001.

[24] Latchman D (2008) Eukaryotic Transcription Factors. Academic Press / Elsevier.

[25] Berg OG, von Hippel PH (1988) Selection of dna binding sites by regulatory proteins. Trends Biochem Sci 13:207–11.

[26] Wolfe SA, Nekludova L, Pabo CO (2000) Dna recognition by cys2his2 zinc finger proteins. Annu Rev Biophys Biomol Struct 29:183–212. doi:10.1146/annurev.biophys.29.1.183. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.biophys.29.1.183?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%253dncbi.nlm.nih.gov.

[27] Kaplan T, Friedman N, Margalit H (2005) Ab initio prediction of transcription factor targets using structural knowledge. PLoS Comput Biol 1:e1. doi:10.1371/journal.pcbi.0010001. URL http://www.ploscompbiol.org/article/info%253Adoi%252F10.1371%252Fjournal.pcbi.0010001.

[28] Persikov AV, Osada R, Singh M (2009) Predicting dna recognition by cys2his2 zinc finger proteins. Bioinformatics 25:22–9. doi:10.1093/bioinformatics/btn580. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/25/1/22?view=long&pmid=19008249.

[29] Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, et al. (2008) Variation in homeodomain dna-binding revealed by high-resolution analysis of sequence preferences. Cell 133:1266. doi:10.1016/j.cell.2008.05.024. URL http://www.ncbi.nlm.nih.gov/pubmed/18585359.

[30] Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14:283–91. doi:10.1016/j.sbi.2004.05.004.

[31] Stern DL (2000) Evolutionary developmental biology and the problem of variation. Evolution 54:1079–91.

[32] McGinnis N, Kuziora MA, McGinnis W (1990) Human hox-4.2 and drosophila deformed encode similar regulatory specificities in drosophila embryos and

larvae. Cell 63:969–76. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4C5946X-62&_user=4420&_coverDate=11%252F30%252F1990&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=39b326b6ca2486ec99a04b4e017387ed.

[33] Wittkopp PJ, True JR, Carroll SB (2002) Reciprocal functions of the drosophila yellow and ebony proteins in the development and evolution of pigment patterns. Development 129:1849–58. URL http://dev.biologists.org/content/129/8/1849.long.

[34] Jeong S, Rokas A, Carroll SB (2006) Regulation of body pigmentation by the abdominal-b hox protein and its gain and loss in drosophila evolution. Cell 125:1387–99. doi:10.1016/j.cell.2006.04.043. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4K95D04-S&_user=4420&_coverDate=06%252F30%252F2006&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=0ebf0d70a4e4a9b2704d8ac00f92f113.

[35] Williams TM, Selegue JE, Werner T, Gompel N, Kopp A, et al. (2008) The regulation and evolution of a genetic switch controlling sexually dimorphic traits in drosophila. Cell 134:610–23. doi:10.1016/j.cell.2008.06.052. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4T8B9FR-D&_user=4420&_coverDate=08%252F22%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=9de0b90d6ab0213daeae68b366cdf222.

[36] Rebeiz M, Pool JE, Kassner VA, Aquadro CF, Carroll SB (2009) Stepwise modification of a modular enhancer underlies adaptation in a drosophila population. Science 326:1663–7. doi:10.1126/science.1178357. URL http://www.sciencemag.org/cgi/content/full/326/5960/1663.

[37] Shirangi TR, Dufour HD, Williams TM, Carroll SB (2009) Rapid evolution of sex pheromone-producing enzyme expression in drosophila. PLoS Biol 7:e1000168. doi:10.1371/journal.pbio.1000168. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.1000168.

[38] Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in drosophila. Nature 433:481–7. doi:10.1038/nature03235. URL http://www.nature.com/nature/journal/v433/n7025/full/nature03235.html.

[39] Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, et al. (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440:1050–3. doi:10.1038/nature04597. URL http://www.nature.com/nature/journal/v440/n7087/full/nature04597.html.

[40] Sucena E, Delon I, Jones I, Payre F, Stern DL (2003) Regulatory evolution of shaven-baby/ovo underlies multiple cases of morphological parallelism. Nature 424:935–8. doi:10.1038/nature01768. URL http://www.nature.com/nature/journal/v424/n6951/full/nature01768.html.

[41] Skaer N, Simpson P (2000) Genetic analysis of bristle loss in hybrids between drosophila melanogaster and d. simulans provides evidence for divergence of cis-regulatory sequences in the achaete-scute gene complex. Dev Biol 221:148–67. doi:10.1006/dbio.1999.9661. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WDG-45F53JG-9B&_user=4420&_coverDate=05%252F01%252F2000&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=5d76cd83487028bf5359243fe1c91067.

[42] Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, et al. (2004) Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature 428:717–23. doi:10.1038/nature02415. URL http://status.nature.com/.

[43] Tournamille C, Colin Y, Cartron JP, Kim CLV (1995) Disruption of a gata motif in the duffy gene promoter abolishes erythroid gene expression in duffy-negative individuals. Nat Genet 10:224–8. doi:10.1038/ng0695-224. URL http://www.nature.com/ng/journal/v10/n2/abs/ng0695-224.html.

[44] Hamblin MT, Thompson EE, Rienzo AD (2002) Complex signatures of natural selection at the duffy blood group locus. Am J Hum Genet 70:369–83. doi:10.1086/338628. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B8JDD-4RDPT55-C&_user=4420&_coverDate=02%252F28%252F2002&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=60d53d066fa4b094ee834aa14c600510.

[45] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in africa and europe. Nat Genet 39:31–40. doi:10.1038/ng1946. URL http://www.nature.com/ng/journal/v39/n1/abs/ng1946.html.

[46] Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design. Blackwell Publishing. URL http://books.google.com/books?id=tayrCszYKdkC&printsec=frontcover.

[47] Davidson EH (2006) The Regulatory Genome: Gene Regulatory Networks in Development and Evolution. Academic Press / Elsevier.

[48] Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, et al. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94. doi:10.1126/science.1139158. URL http://www.sciencemag.org/content/317/5834/86.short.

[49] Srivastava M, Simakov O, Chapman J, Fahey B, Gauthier MEA, et al. (2010) The amphimedon queenslandica genome and the evolution of animal complexity. Nature 466:720–6. doi:10.1038/nature09201.

[50] King N, Westbrook MJ, Young SL, Kuo A, Abedin M, et al. (2008) The genome of the choanoflagellate monosiga brevicollis and the origin of metazoans. Nature 451:783. doi:doi:10.1038/nature06617. URL http://www.nature.com/nature/journal/v451/n7180/full/nature06617.html.

[51] Charoensawan V, Wilson D, Teichmann SA (2010) Genomic repertoires of dna-binding transcription factors across the tree of life. Nucleic acids research doi:10.1093/nar/gkq617.

[52] van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19:479–84.

[53] Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLoS Comput Biol 2:e48. doi:10.1371/journal.pcbi.0020048. URL http://www.ploscompbiol.org/article/info%253Adoi%252F10.1371%252Fjournal.pcbi.0020048.

[54] Wagner GP, Amemiya C, Ruddle F (2003) Hox cluster duplications and the opportunity for evolutionary novelties. Proc Natl Acad Sci USA 100:14603–6. doi:10.1073/pnas.2536656100. URL http://www.pnas.org/content/100/25/14603.long.

[55] Lynch VJ, Roth JJ, Wagner GP (2006) Adaptive evolution of hox-gene homeodomains after cluster duplications. BMC Evol Biol 6:86. doi:10.1186/1471-2148-6-86. URL http://www.biomedcentral.com/1471-2148/6/86.

[56] Tadepally HD, Burger G, Aubry M (2008) Evolution of c2h2-zinc finger genes and subfamilies in mammals: species-specific duplication and loss of clusters, genes and effector domains. BMC Evol Biol 8:176. doi:10.1186/1471-2148-8-176. URL http://www.biomedcentral.com/1471-2148/8/176.

[57] Crow KD, Stadler PF, Lynch VJ, Amemiya C, Wagner GP (2006) The "fish-specific" hox cluster duplication is coincident with the origin of teleosts. Mol Biol Evol 23:121–36. doi:10.1093/molbev/msj020. URL http://mbe.oxfordjournals.org/content/23/1/121.long.

[58] Edelstein LC, Collins T (2005) The scan domain family of zinc finger transcription factors. Gene 359:1–17. doi:10.1016/j.gene.2005.06.022. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T39-4H1001V-1&_user=4420&_coverDate=10%252F10%252F2005&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=6ee8914957f2279ff7102769ef564633&searchtype=a.

[59] Nowick K, Stubbs L (2010) Lineage-specific transcription factors and the evolution of gene regulatory networks. Brief Funct Genomics 9:65–78. doi:10.1093/bfgp/elp056.

[60] Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 12:1048–59. doi:10.1101/gr.174302. URL http://genome.cshlp.org/content/12/7/1048.long.

[61] Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM (2009) A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics 10:252–63. doi:10.1038/nrg2538. URL http://www.nature.com/nrg/journal/v10/n4/abs/nrg2538.html.

[62] Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, et al. (2005) Natural selection on protein-coding genes in the human genome. Nature 437:1153–7. doi:10.1038/nature04240. URL http://www.nature.com/nature/journal/v437/n7062/full/nature04240.html.

[63] De S, Lopez-Bigas N, Teichmann SA (2008) Patterns of evolutionary constraints on genes in humans. BMC Evol Biol 8:275. doi:10.1186/1471-2148-8-275. URL http://www.biomedcentral.com/1471-2148/8/275.

[64] Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP (2006) Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature 440:242–5. doi:10.1038/nature04559. URL http://www.nature.com/nature/journal/v440/n7081/full/nature04559.html.

[65] Babu MM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358:614–33. doi:10.1016/j.jmb.2006.02.019. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-4JCCHMB-1&_user=4420&_coverDate=04%252F28%252F2006&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=fea095a32d7b536df4ccd7c106c8c005.

[66] Li XY, MacArthur S, Bourgon R, Nix D, Pollard DA, et al. (2008) Transcription factors bind thousands of active and inactive regions in the drosophila blastoderm. PLoS Biol 6:e27. doi:10.1371/journal.pbio.0060027. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0060027.

[67] MacArthur S, Li XY, Li J, Brown JB, Chu HC, et al. (2009) Developmental roles of 21 drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol 10:R80. doi:10.1186/gb-2009-10-7-r80. URL http://genomebiology.com/2009/10/7/R80.

[68] Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. Science 311:796–800. doi:10.1126/science.1113832. URL http://www.sciencemag.org/cgi/content/full/311/5762/796.

[69] Jothi R, Balaji S, Wuster A, Grochow JA, Gsponer J, et al. (2009) Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network

architecture. Mol Syst Biol 5:294. doi:10.1038/msb.2009.52. URL http://www.nature.com/msb/journal/v5/n1/full/msb200952.html.

[70] Gehring WJ, Ikeo K (1999) Pax 6: mastering eye morphogenesis and eye evolution. Trends Genet 15:371–7. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TCY-3X8GYP5-G&_user=4420&_coverDate=09%252F01%252F1999&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=c399161555782d807f9ed90d89f243a0&searchtype=a.

[71] Ranganayakulu G, Elliott DA, Harvey RP, Olson EN (1998) Divergent roles for nk-2 class homeobox genes in cardiogenesis in flies and mice. Development 125:3037–48. URL http://dev.biologists.org/content/125/16/3037.long.

[72] Bhardwaj N, Kim PM, Gerstein MB (2010) Rewiring of transcriptional regulatory networks: hierarchy, rather than connectivity, better reflects the importance of regulators. Science Signaling 3:ra79. doi:10.1126/scisignal.2001014.

[73] Park M, Lewis C, Turbay D, Chung A, Chen JN, et al. (1998) Differential rescue of visceral and cardiac defects in drosophila by vertebrate tinman-related genes. Proc Natl Acad Sci USA 95:9366–71. URL http://www.pnas.org/content/95/16/9366.long.

[74] Jovelin R, Dunham JP, Sung FS, Phillips PC (2009) High nucleotide divergence in developmental regulatory genes contrasts with the structural elements of olfactory pathways in caenorhabditis. Genetics 181:1387–97. doi:10.1534/genetics.107.082651. URL http://www.genetics.org/cgi/content/full/181/4/1387.

[75] Tuch BB, Li H, Johnson AD (2008) Evolution of eukaryotic transcription circuits. Science 319:1797–9. doi:10.1126/science.1152398. URL http://www.sciencemag.org/cgi/content/full/319/5871/1797.

[76] Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. Nat Rev Genet 8:921–31. doi:10.1038/nrg2267. URL http://www.nature.com/nrg/journal/v8/n12/full/nrg2267.html.

[77] Crombach A, Hogeweg P (2008) Evolution of evolvability in gene regulatory networks. PLoS Comput Biol 4:e1000112. doi:10.1371/journal.pcbi.1000112. URL http://www.ploscompbiol.org/article/info%253Adoi%252F10.1371%252Fjournal.pcbi.1000112.

[78] Tsuda ME, Kawata M (2010) Evolution of gene regulatory networks by fluctuating selection and intrinsic constraints. PLoS Comput Biol 6. doi:10.1371/journal.pcbi.1000873.

[79] Stern DL, Orgogozo V (2009) Is genetic evolution predictable? Science 323:746–51. doi:10.1126/science.1158997. URL http://www.sciencemag.org/cgi/content/abstract/323/5915/746.

[80] Stern DL, Orgogozo V (2008) The loci of evolution: how predictable is genetic evolution? Evolution 62:2155–77. doi:10.1111/j.1558-5646.2008.00450.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.2008.00450.x/abstract.

[81] Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? Trends Ecol Evol (Amst) 23:26–32. doi:10.1016/j.tree.2007.09.011. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VJ1-4R5H1H5-2&_user=4420&_coverDate=01%252F31%252F2008&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=78451893cec0a2b5c5fdc271af5d766f&searchtype=a.

[82] Papa R, Martin A, Reed RD (2008) Genomic hotspots of adaptation in butterfly wing pattern evolution. Curr Opin Genet Dev 18:559–64. doi:10.1016/j.gde.2008.11.007.

[83] Steiner CC, Weber JN, Hoekstra HE (2007) Adaptive variation in beach mice produced by two interacting pigmentation genes. PLoS Biol 5:e219. doi:10.1371/journal.pbio.0050219.

[84] Hoekstra HE (2006) Genetics, development and evolution of adaptive pigmentation in vertebrates. Heredity 97:222–34. doi:10.1038/sj.hdy.6800861. URL http://www.nature.com/hdy/journal/v97/n3/full/6800861a.html.

[85] Colosimo PF, Hosemann KE, Balabhadra S, Villarreal G, Dickson M, et al. (2005) Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. Science 307:1928–33. doi:10.1126/science.1107239. URL http://www.sciencemag.org/content/307/5717/1928.long.

[86] Kitano J, Bolnick DI, Beauchamp DA, Mazur MM, Mori S, et al. (2008) Reverse evolution of armor plates in the threespine stickleback. Curr Biol 18:769–74. doi:10.1016/j.cub.2008.04.027. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VRT-4SHF3MD-1&_user=4420&_coverDate=05%252F20%252F2008&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=50a345ec4f2910659941b9def1238aef&searchtype=a.

[87] Cresko WA, Amores A, Wilson C, Murphy J, Currey M, et al. (2004) Parallel genetic basis for repeated evolution of armor loss in alaskan threespine stickleback populations. Proceedings of the National Academy of Sciences of the United States of America 101:6050–5. doi:10.1073/pnas.0308479101. URL http://www.pnas.org/content/101/16/6050.long.

[88] Fujimoto A, Kimura R, Ohashi J, Omi K, Yuliwulandari R, et al. (2008) A scan for genetic determinants of human hair morphology: Edar is associated with asian hair thickness. Hum Mol Genet 17:835–43. doi:10.1093/hmg/ddm355. URL http://hmg.oxfordjournals.org/content/17/6/835.long.

[89] McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, et al. (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. Nature 448:587–90. doi:10.1038/nature05988. URL http://www.nature.com/nature/journal/v448/n7153/full/nature05988.html.

[90] Marcellini S, Simpson P (2006) Two or four bristles: functional evolution of an enhancer of scute in drosophilidae. PLoS Biol 4:e386. doi:10.1371/journal.pbio.0040386.

[91] Shapiro MD, Bell MA, Kingsley DM (2006) Parallel genetic origins of pelvic reduction in vertebrates. Proc Natl Acad Sci USA 103:13753–8. doi:10.1073/pnas.0604706103. URL http://www.pnas.org/content/103/37/13753.long.

[92] Hobert O (2008) Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. Proc Natl Acad Sci USA 105:20067–71. doi:10.1073/pnas.0806070105. URL http://www.pnas.org/content/105/51/20067.long.

[93] Gibert JM, Marcellini S, David JR, Schlötterer C, Simpson P (2005) A major bristle qtl from a selected population of drosophila uncovers the zinc-finger transcription factor poils-au-dos, a repressor of achaete-scute. Dev Biol 288:194–205. doi:10.1016/j.ydbio.2005.09.032. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WDG-4H98T6Y-4&_user=4420&_coverDate=12%252F01%252F2005&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=12bfe18800cf6bb7aedba64d42e4a846&searchtype=a.

[94] Frankel N, Davis GK, Vargas D, Wang S, Payre F, et al. (2010) Phenotypic robustness conferred by apparently redundant transcriptional enhancers. Nature 466:490–3. doi:10.1038/nature09158. URL http://www.nature.com/nature/journal/v466/n7305/full/nature09158.html.

[95] Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. Proceedings of the National Academy of Sciences of the United States of America 104 Suppl 1:8597–604. doi:10.1073/pnas.0702207104. URL http://www.pnas.org/content/104/suppl.1/8597.long.

[96] Pigliucci M (2010) Genotype-phenotype mapping and the end of the 'genes as blueprint' metaphor. Philos Trans R Soc Lond, B, Biol Sci 365:557–66. doi:10.1098/rstb.2009.0241.

[97] Ludwig MZ, Bergman C, Patel NH, Kreitman M (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403:564–7. doi:10.1038/35000615. URL http://www.nature.com/nature/journal/v403/n6769/full/403564a0.html.

[98] Hare EE, Peterson BK, Eisen MB (2008) A careful look at binding site reorganization in the even-skipped enhancers of drosophila and sepsids. PLoS Genet 4:e1000268. doi:10.1371/journal.pgen.1000268. URL http://www.plosgenetics.org/article/info%253Adoi%252F10.1371%252Fjournal.pgen.1000268.

[99] Hare EE, Peterson BK, Iyer VN, Meier R, Eisen MB (2008) Sepsid even-skipped enhancers are functionally conserved in drosophila despite lack of sequence conservation. PLoS Genet 4:e1000106. doi:10.1371/journal.pgen.1000106. URL http://www.plosgenetics.org/article/info%253Adoi%252F10.1371%252Fjournal.pgen.1000106.

[100] MacArthur S, Brookfield JFY (2004) Expected rates and modes of evolution of enhancer sequences. Mol Biol Evol 21:1064–73. doi:10.1093/molbev/msh105. URL http://mbe.oxfordjournals.org/content/21/6/1064.long.

[101] Tsong AE, Tuch BB, Li H, Johnson AD (2006) Evolution of alternative transcriptional circuits with identical logic. Nature 443:415–20. doi:10.1038/nature05099. URL http://www.nature.com/nature/journal/v443/n7110/full/nature05099.html.

[102] Kuo D, Licon K, Bandyopadhyay S, Chuang R, Luo C, et al. (2010) Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome research doi:10.1101/gr.111765.110.

[103] Lynch M (2007) The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics 8:803–13. doi:10.1038/nrg2192. URL http://www.nature.com/nrg/journal/v8/n10/full/nrg2192.html.

[104] Li H, Johnson AD (2010) Evolution of transcription networks–lessons from yeasts. Curr Biol 20:R746–53. doi:10.1016/j.cub.2010.06.056.

[105] Borneman AR, Gianoulis TA, Zhang ZD, Yu H, Rozowsky J, et al. (2007) Divergence of transcription factor binding sites across related yeast species. Science 317:815–9. doi:10.1126/science.1140748. URL http://www.sciencemag.org/cgi/content/abstract/317/5839/815.

[106] Odom DT, Dowell RD, Jacobsen ES, Gordon W, Danford TW, et al. (2007) Tissue-specific transcriptional regulation has diverged significantly between human and mouse. Nat Genet 39:730–2. doi:10.1038/ng2047. URL http://www.nature.com/ng/journal/v39/n6/full/ng2047.html.

[107] Schmidt D, Wilson MD, Ballester B, Schwalie PC, Brown GD, et al. (2010) Five-vertebrate chip-seq reveals the evolutionary dynamics of transcription factor binding. Science 328:1036–40. doi:10.1126/science.1186176.

[108] Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, New York.

[109] Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–5. URL http://www.sciencemag.org/cgi/content/full/290/5494/1151.

[110] Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome research 18:137–47. doi:10.1101/gr.6341207. URL http://genome.cshlp.org/content/18/1/137.long.

[111] Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, et al. (2007) Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proc Natl Acad Sci USA 104:8397–402. doi:10.1073/pnas.0608218104. URL http://www.pnas.org/content/104/20/8397.long.

[112] Charoensawan V, Wilson D, Teichmann SA (2010) Lineage-specific expansion of dna-binding transcription factor families. Trends Genet 26:388–93. doi:10.1016/j.tig.2010.06.004.

[113] Ledent V, Vervoort M (2001) The basic helix-loop-helix protein family: comparative genomics and phylogenetic analysis. Genome Res 11:754–70. doi:10.1101/gr.177001. URL http://genome.cshlp.org/content/11/5/754.long.

[114] Sluder AE, Mathews SW, Hough D, Yin VP, Maina CV (1999) The nuclear receptor superfamily has undergone extensive proliferation and diversification in nematodes. Genome Res 9:103–20. URL http://genome.cshlp.org/content/9/2/103.long.

[115] Robinson-Rechavi M, Maina CV, Gissendanner CR, Laudet V, Sluder A (2005) Explosive lineage-specific expansion of the orphan nuclear receptor hnf4 in nematodes. J Mol Evol 60:577–86. doi:10.1007/s00239-004-0175-8. URL http://www.springerlink.com/content/kh577937250w0171/.

[116] Chung HR, Löhr U, Jäckle H (2007) Lineage-specific expansion of the zinc finger associated domain zad. Mol Biol Evol 24:1934–43. doi:10.1093/molbev/msm121.

[117] Chung HR, Schäfer U, Jäckle H, Böhm S (2002) Genomic expansion and clustering of zad-containing c2h2 zinc-finger genes in drosophila. EMBO Rep 3:1158–62. doi:10.1093/embo-reports/kvf243. URL http://www.nature.com/embor/journal/v3/n12/full/embor015.html.

[118] Huntley S, Baggott DM, Hamilton AT, Tran-Gyamfi M, Yang S, et al. (2006) A comprehensive catalog of human krab-associated zinc finger genes: insights into the evolutionary history of a large family of transcriptional repressors. Genome Res 16:669–77. doi:10.1101/gr.4842106. URL http://genome.cshlp.org/content/16/5/669.long.

[119] Looman C, Abrink M, Mark C, Hellman L (2002) Krab zinc finger proteins: an analysis of the molecular mechanisms governing their increase in numbers and complexity during evolution. Mol Biol Evol 19:2118–30. URL http://mbe.oxfordjournals.org/content/19/12/2118.long.

[120] Lorenz P, Dietmann S, Wilhelm T, Koczan D, Autran S, et al. (2010) The ancient mammalian krab zinc finger gene cluster on human chromosome 8q24.3 illustrates principles of c2h2 zinc finger evolution associated with unique expression profiles in human tissues. BMC Genomics 11:206. doi:10.1186/1471-2164-11-206.

[121] Jackson M, Watt AJ, Gautier P, Gilchrist D, Driehaus J, et al. (2006) A murine specific expansion of the rhox cluster involved in embryonic stem cell biology is under

natural selection. BMC Genomics 7:212. doi:10.1186/1471-2164-7-212. URL http://www.biomedcentral.com/1471-2164/7/212.

[122] Garcia-Fernàndez J (2005) The genesis and evolution of homeobox gene clusters. Nature Reviews Genetics 6:881–92. doi:10.1038/nrg1723.

[123] Abbasi AA, Grzeschik KH (2007) An insight into the phylogenetic history of hox linked gene families in vertebrates. BMC Evol Biol 7:239. doi:10.1186/1471-2148-7-239. URL http://www.biomedcentral.com/1471-2148/7/239.

[124] Scott MP (1992) Vertebrate homeobox gene nomenclature. Cell 71:551–3. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4C6BNRR-CK&_user=4420&_coverDate=11%252F13%252F1992&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=bb17fe3ce7c9b27b529c6ee5a92e5821&searchtype=a.

[125] Amoutzias GD, Veron AS, Weiner J, Robinson-Rechavi M, Bornberg-Bauer E, et al. (2007) One billion years of bzip transcription factor evolution: conservation and change in dimerization and dna-binding site specificity. Molecular Biology and Evolution 24:827–35. doi:10.1093/molbev/msl211.

[126] Bertrand S, Brunet FG, Escriva H, Parmentier G, Laudet V, et al. (2004) Evolutionary genomics of nuclear receptors: from twenty-five ancestral genes to derived endocrine systems. Molecular Biology and Evolution 21:1923–37. doi:10.1093/molbev/msh200.

[127] Moore AD, Björklund AK, Ekman D, Bornberg-Bauer E, Elofsson A (2008) Arrangements in the modular evolution of proteins. Trends Biochem Sci 33:444–51. doi:10.1016/j.tibs.2008.05.008.

[128] Hurles M (2004) Gene duplication: the genomic trade in spare parts. PLoS Biol 2:E206. doi:10.1371/journal.pbio.0020206. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0020206.

[129] Verstrepen KJ, Jansen A, Lewitter F, Fink GR (2005) Intragenic tandem repeats generate functional variability. Nat Genet 37:986–90. doi:10.1038/ng1618. URL http://www.nature.com/ng/journal/v37/n9/full/ng1618.html.

[130] Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev 15:621–7. doi:10.1016/j.gde.2005.09.010. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VS0-4H998XR-1&_user=4420&_coverDate=12%252F31%252F2005&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=7d84f4677670a86d2e57866662b4f2b4&searchtype=a.

[131] i Kubo K, Sakamoto A, Kobayashi A, Rybka Z, Kanno Y, et al. (1998) Cys2/his2 zinc-finger protein family of petunia: evolution and general mechanism of target-sequence recognition. Nucleic Acids Res 26:608–15.

[132] Emerson RO, Thomas JH (2009) Adaptive evolution in zinc finger transcription factors. PLoS Genet 5:e1000325. doi:10.1371/journal.pgen.1000325.

[133] Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ (2006) Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression. Proc Natl Acad Sci USA 103:9935–9. doi:10.1073/pnas.0509809103.

[134] Neduva V, Russell RB (2005) Linear motifs: evolutionary interaction switches. FEBS Lett 579:3342–5. doi:10.1016/j.febslet.2005.04.005.

[135] Galant R, Carroll SB (2002) Evolution of a transcriptional repression domain in an insect hox protein. Nature 415:910–3. doi:10.1038/nature717.

[136] Löhr U, Yussa M, Pick L (2001) Drosophila fushi tarazu. a gene on the border of homeotic function. Curr Biol 11:1403–12. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VRT-441NT53-H&_user=4420&_coverDate=09%252F18%252F2001&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=6c53372ff8b40866ad7329f5003a90b7.

[137] Morgenstern B, Atchley WR (1999) Evolution of bhlh transcription factors: modular evolution by domain shuffling? Mol Biol Evol 16:1654–63. URL http://mbe.oxfordjournals.org/content/16/12/1654.long.

[138] Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. Nature Reviews Genetics 11:572–82. doi:10.1038/nrg2808. URL http://www.nature.com/nrg/journal/v11/n8/abs/nrg2808.html.

[139] Luscombe NM, Thornton JM (2002) Protein-dna interactions: amino acid conservation and the effects of mutations on binding specificity. J Mol Biol 320:991–1009. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-46872M9-8&_user=4420&_coverDate=07%252F26%252F2002&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=ba15eecfa63b8e5f8671d795b805b6b8.

[140] Isalan M, Klug A, Choo Y (1998) Comprehensive dna recognition through concerted interactions from adjacent zinc fingers. Biochemistry 37:12026–33. doi:10.1021/bi981358z.

[141] Klug A (2010) The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Annu Rev Biochem 79:213–31. doi:10.1146/annurev-biochem-010909-095056.

[142] Maerkl SJ, Quake SR (2009) Experimental determination of the evolvability of a transcription factor. Proc Natl Acad Sci USA 106:18650–5. doi:10.1073/pnas.0907688106.

[143] Fong JH, Keating AE, Singh M (2004) Predicting specificity in bzip coiled-coil protein interactions. Genome Biol 5:R11. doi:10.1186/gb-2004-5-2-r11. URL http://genomebiology.com/2004/5/2/R11.

[144] Newman JRS, Keating AE (2003) Comprehensive identification of human bzip interactions with coiled-coil arrays. Science 300:2097–101. doi:10.1126/science.1084648. URL http://www.sciencemag.org/content/300/5628/2097.long.

[145] Grigoryan G, Reinke AW, Keating AE (2009) Design of protein-interaction specificity gives selective bzip-binding peptides. Nature 458:859–64. doi:10.1038/nature07885. URL http://www.nature.com/nature/journal/v458/n7240/full/nature07885.html.

[146] Grove CA, Masi FD, Barrasa MI, Newburger DE, Alkema MJ, et al. (2009) A multiparameter network reveals extensive divergence between c. elegans bhlh transcription factors. Cell 138:314–27. doi:10.1016/j.cell.2009.04.058. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4WV09S7-G&_user=4420&_coverDate=07%252F23%252F2009&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=f4dc29965a1578f35694bef5f166132d&searchtype=a.

[147] Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, et al. (2004) Conservation and evolution of cis-regulatory systems in ascomycete fungi. PLoS Biol 2:e398. doi:10.1371/journal.pbio.0020398. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0020398.

[148] Ting CT, Tsaur SC, Wu ML, Wu CI (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. Science 282:1501–4.

[149] Lynch VJ, Roth JJ, Takahashi K, Dunn CW, Nonaka DF, et al. (2004) Adaptive evolution of hoxa-11 and hoxa-13 at the origin of the uterus in mammals. Proc Biol Sci 271:2201–7. doi:10.1098/rspb.2004.2848.

[150] Lamb RS, Irish VF (2003) Functional divergence within the apetala3/pistillata floral homeotic gene lineages. Proc Natl Acad Sci USA 100:6558–63. doi:10.1073/pnas.0631708100.

[151] Martinez-Castilla LP, Alvarez-Buylla ER (2003) Adaptive evolution in the arabidopsis mads-box gene family inferred from its complete resolved phylogeny. Proc Natl Acad Sci USA 100:13407–12. doi:10.1073/pnas.1835864100. URL http://www.pnas.org/content/100/23/13407.long.

[152] Miller JC, Pabo CO (2001) Rearrangement of side-chains in a zif268 mutant highlights the complexities of zinc finger-dna recognition. Journal of Molecular Biology 313:309–15. doi:10.1006/jmbi.2001.4975. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-457CXY4-47&_user=4420&_coverDate=10%252F19%252F2001&_rdoc=1&_fmt=high&_orig=search&_origin=search&_

sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=612eef3ac9ee1e1e7cf2e115b1d22ea4&searchtype=a.

[153] Iwema T, Chaumot A, Studer RA, Robinson-Rechavi M, Billas IML, et al. (2009) Structural and evolutionary innovation of the heterodimerization interface between usp and the ecdysone receptor ecr in insects. Mol Biol Evol 26:753–68. doi:10.1093/molbev/msn302. URL http://mbe.oxfordjournals.org/content/26/4/753.long.

[154] Pabo CO, Nekludova L (2000) Geometric analysis and comparison of protein-dna interfaces: why is there no simple code for recognition? Journal of Molecular Biology 301:597–624. doi:10.1006/jmbi.2000.3918. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45F517N-8X&_user=4420&_coverDate=08%252F18%252F2000&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=e761cf361a065b4bdc0506fcc9304ec7&searchtype=a.

[155] Pufall MA, Lee GM, Nelson ML, Kang HS, Velyvis A, et al. (2005) Variable control of ets-1 dna binding by multiple phosphates in an unstructured region. Science 309:142–5. doi:10.1126/science.1111915.

[156] Mitchell PJ, Tjian R (1989) Transcriptional regulation in mammalian cells by sequence-specific dna binding proteins. Science 245:371–8. URL http://www.sciencemag.org/content/245/4916/371.long.

[157] Emili A, Greenblatt J, Ingles CJ (1994) Species-specific interaction of the glutamine-rich activation domains of sp1 with the tata box-binding protein. Mol Cell Biol 14:1582–93. URL http://mcb.asm.org/cgi/reprint/14/3/1582?view=long&pmid=8114696.

[158] Gerber HP, Seipel K, Georgiev O, Höfferer M, Hug M, et al. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–11. URL http://www.sciencemag.org/content/263/5148/808.long.

[159] Perutz M (1994) Polar zippers: their role in human disease. Protein Sci 3:1629–37. doi:10.1002/pro.5560031002. URL http://onlinelibrary.wiley.com/doi/10.1002/pro.5560031002/abstract;jsessionid=FB31EB422635C7D9AF748025D3532E7A.d01t01?systemMessage=There+will+be+a+release+of+Wiley+Online+Library+scheduled+for+Saturday+27th+November+2010.+Access+to+the+website+will+be+disrupted+as+follows%253A+New+York+0630+EDT+to+0830+EDT%253B+London+1130+GMT+to+1330+GMT%253B+Singapore+1930+SGT+to+2130+SGT.

[160] Xiao H, Jeang KT (1998) Glutamine-rich domains activate transcription in yeast saccharomyces cerevisiae. J Biol Chem 273:22873–6. URL http://www.jbc.org/content/273/36/22873.long.

[161] Fondon JW, Garner HR (2004) Molecular origins of rapid and continuous morphological evolution. Proc Natl Acad Sci USA 101:18058–63. doi:10.1073/pnas.0408118101. URL http://www.pnas.org/content/101/52/18058.long.

[162] Sears KE, Goswami A, Flynn JJ, Niswander LA (2007) The correlated evolution of runx2 tandem repeats, transcriptional activity, and facial length in carnivora. Evol Dev 9:555–65. doi:10.1111/j.1525-142X.2007.00196.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1525-142X.2007.00196.x/abstract?systemMessage=There+will+be+a+release+of+Wiley+Online+Library+scheduled+for+Saturday+27th+November+2010.+Access+to+the+website+will+be+disrupted+as+follows%253A+New+York+0630+EDT+to+0830+EDT%253B+London+1130+GMT+to+1330+GMT%253B+Singapore+1930+SGT+to+2130+SGT.

[163] Amoutzias G, Veron A, Weiner JI, Robinson-Rechavi M, Bornberg-Bauer E, et al. (2007) One billion years of bzip transcription factor evolution: Conservation and change in dimerization and dna-binding site specificity. Mol Biol Evol 24:827. doi:10.1093/molbev/msl211. URL http://mbe.oxfordjournals.org/cgi/content/full/24/3/827.

[164] Jones S (2004) An overview of the basic helix-loop-helix proteins. Genome Biology 5:226. doi:10.1186/gb-2004-5-6-226.

[165] Amoutzias GD, Robertson DL, Bornberg-Bauer E (2004) The evolution of protein interaction networks in regulatory proteins. Comp Funct Genomics 5:79–84. doi:10.1002/cfg.365.

[166] Carey M (1998) The enhanceosome and transcriptional synergy. Cell 92:5–8. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4194SB7-2&_user=4420&_coverDate=01%252F09%252F1998&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=72f8706a9ce866f82c466e060a6b7e4e&searchtype=a.

[167] Lebrecht D, Foehr M, Smith E, Lopes FJP, Vanario-Alonso CE, et al. (2005) Bicoid cooperative dna binding is critical for embryonic patterning in drosophila. Proc Natl Acad Sci USA 102:13176–81. doi:10.1073/pnas.0506462102.

[168] Amin J, Fernandez M, Ananthan J, Lis JT, Voellmy R (1994) Cooperative binding of heat shock transcription factor to the hsp70 promoter in vivo and in vitro. J Biol Chem 269:4804–11.

[169] Katsani KR, Hajibagheri MA, Verrijzer CP (1999) Co-operative dna binding by gaga transcription factor requires the conserved btb/poz domain and reorganizes promoter topology. The EMBO Journal 18:698–708. doi:10.1093/emboj/18.3.698.

[170] Jaynes JB, O'Farrell PH (1988) Activation and repression of transcription by homoeodomain-containing proteins that bind a common site. Nature 336:744–9. doi:10.1038/336744a0. URL http://www.nature.com/nature/journal/v336/n6201/abs/336744a0.html.

[171] Hiller MA, Lin TY, Wood C, Fuller MT (2001) Developmental regulation of transcription by a tissue-specific taf homolog. Genes & Development 15:1021–30. doi:10.1101/gad.869101. URL http://genesdev.cshlp.org/content/15/8/1021.long.

[172] Gaston K, Jayaraman PS (2003) Transcriptional repression in eukaryotes: repressors and repression mechanisms. Cell Mol Life Sci 60:721–41.

[173] Wagner GP, Lynch VJ (2008) The gene regulatory logic of transcription factor evolution. Trends Ecol Evol (Amst) 23:377–85. doi:10.1016/j.tree.2008.03.006. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VJ1-4SK2V2B-2&_user=4420&_coverDate=07%252F31%252F2008&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=af81a9f0338f3c537f37dfb95d6c7823&searchtype=a.

[174] Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. PLoS Biol 6:e38. doi:10.1371/journal.pbio.0060038.

[175] Keleher CA, Passmore S, Johnson AD (1989) Yeast repressor alpha 2 binds to its operator cooperatively with yeast protein mcm1. Mol Cell Biol 9:5228–30. URL http://mcb.asm.org/cgi/reprint/9/11/5228?view=long&pmid=2689875.

[176] Passmore S, Elble R, Tye BK (1989) A protein involved in minichromosome maintenance in yeast binds a transcriptional enhancer conserved in eukaryotes. Genes & Development 3:921–35. URL http://genesdev.cshlp.org/content/3/7/921.long.

[177] Hull CM, Raisner RM, Johnson AD (2000) Evidence for mating of the "asexual" yeast candida albicans in a mammalian host. Science 289:307–10. URL http://www.sciencemag.org/content/289/5477/307.long.

[178] Tsong AE, Miller MG, Raisner RM, Johnson AD (2003) Evolution of a combinatorial transcriptional circuit: a case study in yeasts. Cell 115:389–99. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4B1X0XX-6&_user=4420&_coverDate=11%252F14%252F2003&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=5d7cd07143865cf02d5c774f3fc69f12&searchtype=a.

[179] Messenguy F, Dubois E (2003) Role of mads box proteins and their cofactors in combinatorial control of gene expression and cell development. Gene 316:1–21. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T39-49KH2XM-3&_user=4420&_coverDate=10%252F16%252F2003&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=487e0cea3fc39ab7b1cc28c910afa8f7&searchtype=a.

[180] Tanay A, Regev A, Shamir R (2005) Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. Proc Natl Acad Sci USA 102:7203–8. doi:10.1073/pnas.0502521102. URL http://www.pnas.org/content/102/20/7203.long.

[181] Hogues H, Lavoie H, Sellam A, Mangos M, Roemer T, et al. (2008) Transcription factor substitution during the evolution of fungal ribosome regulation. Mol Cell 29:552–62. doi:10.1016/j.molcel.2008.02.006. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSR-4S21JBK-6&_user=4420&_coverDate=03%252F14%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=466450fd1c94feb2f51f281937182925.

[182] Lavoie H, Hogues H, Mallick J, Sellam A, Nantel A, et al. (2010) Evolutionary tinkering with conserved components of a transcriptional regulatory network. PLoS Biol 8:e1000329. doi:10.1371/journal.pbio.1000329.

[183] Powers T (2004) Ribosome biogenesis: giant steps for a giant problem. Cell 119:901–2. doi:10.1016/j.cell.2004.12.011. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4F493RP-1&_user=4420&_coverDate=12%252F29%252F2004&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=92c38f3e3c4d88d7999e6bf3f5a30405&searchtype=a.

[184] Wade JT, Hall DB, Struhl K (2004) The transcription factor ifh1 is a key regulator of yeast ribosomal protein genes. Nature 432:1054–8. doi:10.1038/nature03175. URL http://www.nature.com/nature/journal/v432/n7020/full/nature03175.html.

[185] Martin DE, Soulard A, Hall MN (2004) Tor regulates ribosomal protein gene expression via pka and the forkhead transcription factor fhl1. Cell 119:969–79. doi:10.1016/j.cell.2004.11.047. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4F493RP-9&_user=4420&_coverDate=12%252F29%252F2004&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=1fb440eb0c8b31b1bee897fbe9779ac9&searchtype=a.

[186] Schawalder SB, Kabani M, Howald I, Choudhury U, Werner M, et al. (2004) Growth-regulated recruitment of the essential yeast ribosomal protein gene activator ifh1. Nature 432:1058–61. doi:10.1038/nature03200. URL http://www.nature.com/nature/journal/v432/n7020/full/nature03200.html.

[187] Rudra D, Zhao Y, Warner JR (2005) Central role of ifh1p-fhl1p interaction in the synthesis of yeast ribosomal proteins. The EMBO Journal 24:533–42. doi:10.1038/sj.emboj.7600553. URL http://www.nature.com/emboj/journal/v24/n3/full/7600553a.html.

[188] Hall DB, Wade JT, Struhl K (2006) An hmg protein, hmo1, associates with promoters of many ribosomal protein genes and throughout the rrna gene locus in saccharomyces

cerevisiae. Mol Cell Biol 26:3672–9. doi:10.1128/MCB.26.9.3672-3679.2006. URL http://mcb.asm.org/cgi/content/full/26/9/3672?view=long&pmid=16612005.

[189] Zhao Y, McIntosh KB, Rudra D, Schawalder S, Shore D, et al. (2006) Fine-structure analysis of ribosomal protein gene transcription. Mol Cell Biol 26:4853–62. doi:10.1128/MCB.02367-05. URL http://mcb.asm.org/cgi/content/full/26/13/4853?view=long&pmid=16782874.

[190] Kasahara K, Ohtsuki K, Ki S, Aoyama K, Takahashi H, et al. (2007) Assembly of regulatory factors on rrna and ribosomal protein genes in saccharomyces cerevisiae. Mol Cell Biol 27:6686–705. doi:10.1128/MCB.00876-07. URL http://mcb.asm.org/cgi/content/full/27/19/6686?view=long&pmid=17646381.

[191] Merz K, Hondele M, Goetze H, Gmelch K, Stoeckl U, et al. (2008) Actively transcribed rrna genes in s. cerevisiae are organized in a specialized chromatin associated with the high-mobility group protein hmo1 and are largely devoid of histone molecules. Genes & Development 22:1190–204. doi:10.1101/gad.466908. URL http://genesdev.cshlp.org/content/22/9/1190.long.

[192] Martchenko M, Levitin A, Hogues H, Nantel A, Whiteway M (2007) Transcriptional rewiring of fungal galactose-metabolism circuitry. Curr Biol 17:1007–13. doi:10.1016/j.cub.2007.05.017. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VRT-4NVTH3P-4&_user=4420&_coverDate=06%252F19%252F2007&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=63b7b03fc6b02844df1f19e131b9472e&searchtype=a.

[193] Johnston M (1987) A model fungal gene regulatory mechanism: the gal genes of saccharomyces cerevisiae. Microbiol Rev 51:458–76. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2830478&dopt=abstractplus.

[194] Lohr D, Venkov P, Zlatanova J (1995) Transcriptional regulation in the yeast gal gene family: a complex genetic network. FASEB J 9:777–87. URL http://www.fasebj.org/content/9/9/777.long.

[195] Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, et al. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169:1915–25. doi:10.1534/genetics.104.036871. URL http://www.genetics.org/cgi/content/full/169/4/1915.

[196] Mann RS, Hogness DS (1990) Functional dissection of ultrabithorax proteins in d. melanogaster. Cell 60:597–610. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4D0YBCM-54&_user=4420&_coverDate=02%252F23%252F1990&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=d6e137617e9d7419b4d59c9b5a967724&searchtype=a.

[197] Casares F, Calleja M, Sánchez-Herrero E (1996) Functional similarity in appendage specification by the ultrabithorax and abdominal-a drosophila hox genes. The EMBO Journal 15:3934–42. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8670898&dopt=abstractplus.

[198] Grenier JK, Carroll SB (2000) Functional evolution of the ultrabithorax protein. Proc Natl Acad Sci USA 97:704–9.

[199] Johnson FB, Parker E, Krasnow MA (1995) Extradenticle protein is a selective cofactor for the drosophila homeotics: role of the homeodomain and ypwm amino acid motif in the interaction. Proc Natl Acad Sci USA 92:739–43. URL http://www.pnas.org/content/92/3/739.long.

[200] Chang CP, Shen WF, Rozenfeld S, Lawrence HJ, Largman C, et al. (1995) Pbx proteins display hexapeptide-dependent cooperative dna binding with a subset of hox proteins. Genes Dev 9:663–74. URL http://genesdev.cshlp.org/content/9/6/663.long.

[201] van Dijk MA, Murre C (1994) extradenticle raises the dna binding specificity of homeotic selector gene products. Cell 78:617–24. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4CXMRNG-B&_user=4420&_coverDate=08%252F26%252F1994&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=d107f3e5ae13380fa729a468b6881382&searchtype=a.

[202] Yussa M, Löhr U, Su K, Pick L (2001) The nuclear receptor ftz-f1 and homeodomain protein ftz interact through evolutionarily conserved protein domains. Mech Dev 107:39–53. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T9H-43S5T18-4&_user=4420&_coverDate=09%252F30%252F2001&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=fe01bf684b7a64765735009ec9efb0b7&searchtype=a.

[203] Schwartz CJ, Sampson HM, Hlousek D, Percival-Smith A, Copeland JW, et al. (2001) Ftz-factor1 and fushi tarazu interact via conserved nuclear receptor and coactivator motifs. The EMBO Journal 20:510–9. doi:10.1093/emboj/20.3.510. URL http://www.nature.com/emboj/journal/v20/n3/full/7593557a.html.

[204] Heery DM, Kalkhoven E, Hoare S, Parker MG (1997) A signature motif in transcriptional co-activators mediates binding to nuclear receptors. Nature 387:733–6. doi:10.1038/42750. URL http://www.nature.com/nature/journal/v387/n6634/full/387733a0.html.

[205] Lynch VJ, Tanzer A, Wang Y, Leung FC, Gellersen B, et al. (2008) Adaptive changes in the transcription factor hoxa-11 are essential for the evolution of pregnancy in mammals. Proc Natl Acad Sci USA 105:14928–33. doi:10.1073/pnas.0802355105. URL http://www.pnas.org/content/105/39/14928.long.

[206] Lynch VJ, Brayer K, Gellersen B, Wagner GP (2009) Hoxa-11 and foxo1a cooperate to regulate decidual prolactin expression: towards inferring the core transcriptional regulators of decidual genes. PLoS ONE 4:e6845. doi:10.1371/journal.pone.0006845. URL http://www.plosone.org/article/info%253Adoi%252F10.1371%252Fjournal.pone.0006845.

[207] Amoutzias GD, Weiner J, Bornberg-Bauer E (2005) Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene 347:247–53. doi:10.1016/j.gene.2004.12.031. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T39-4FJTP2F-3&_user=4420&_coverDate=03%252F14%252F2005&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=ff04cde3609f1da9a47dbdd7a7524086&searchtype=a.

[208] Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, et al. (2002) bzip transcription factors in arabidopsis. Trends Plant Sci 7:106–11.

[209] Pinney JW, Amoutzias GD, Rattray M, Robertson DL (2007) Reconstruction of ancestral protein interaction networks for the bzip transcription factors. Proc Natl Acad Sci USA 104:20449–53. doi:10.1073/pnas.0706339104.

[210] Taubert S, Ward JD, Yamamoto KR (2010) Nuclear hormone receptors in nematodes: Evolution and function. Molecular and cellular endocrinology doi:10.1016/j.mce.2010.04.021. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T3G-5004R48-1&_user=4420&_coverDate=05%252F01%252F2010&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=81a909e6aa6e07ef6151e063e798e834&searchtype=a.

[211] Urrutia R (2003) Krab-containing zinc-finger repressor proteins. Genome Biol 4:231. doi:10.1186/gb-2003-4-10-231. URL http://genomebiology.com/content/4/10/231.

[212] Hamilton AT, Huntley S, Kim J, Branscomb E, Stubbs L (2003) Lineage-specific expansion of krab zinc-finger transcription factor genes: implications for the evolution of vertebrate regulatory networks. Cold Spring Harb Symp Quant Biol 68:131–40.

[213] Witzgall R, O'Leary E, Leaf A, Onaldi D, Bonventre JV (1994) The krüppel-associated box-a (krab-a) domain of zinc finger proteins mediates transcriptional repression. Proc Natl Acad Sci USA 91:4514–8.

[214] Abrink M, Ortiz JA, Mark C, Sanchez C, Looman C, et al. (2001) Conserved interaction between distinct krüppel-associated box domains and the transcriptional intermediary factor 1 beta. Proc Natl Acad Sci USA 98:1422–6. doi:10.1073/pnas.041616998. URL http://www.pnas.org/content/98/4/1422.long.

[215] Ayyanathan K, Lechner MS, Bell P, Maul GG, Schultz DC, et al. (2003) Regulated recruitment of hp1 to a euchromatic gene induces mitotically heritable, epigenetic gene silencing: a mammalian cell culture model of gene variegation. Genes Dev 17:1855–69. doi:10.1101/gad.1102803. URL http://genesdev.cshlp.org/content/17/15/1855.long.

[216] Williams AJ, Blacklow SC, Collins T (1999) The zinc finger-associated scan box is a conserved oligomerization domain. Mol Cell Biol 19:8526–35. URL http://mcb.asm.org/cgi/content/full/19/12/8526?view=long&pmid=10567577.

[217] Jauch R, Bourenkov GP, Chung HR, Urlaub H, Reidt U, et al. (2003) The zinc finger-associated domain of the drosophila transcription factor grauzone is a novel zinc-coordinating protein-protein interaction module. Structure 11:1393–402. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VSR-49YDS35-G&_user=4420&_coverDate=11%252F30%252F2003&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=49ac75d07e94fedab9de804f68c4c01d&searchtype=a.

[218] Wingender E, Dietze P, Karas H, Knüppel R (1996) Transfac: a database on transcription factors and their dna binding sites. Nucleic acids research 24:238–41. URL http://nar.oxfordjournals.org/lookup/pmid?view=long&pmid=8594589.

[219] Portales-Casamar E, Thongjuea S, Kwon AT, Arenillas D, Zhao X, et al. (2010) Jaspar 2010: the greatly expanded open-access database of transcription factor binding profiles. Nucleic acids research 38:D105–10. doi:10.1093/nar/gkp950. URL http://nar.oxfordjournals.org/content/38/suppl_1/D105.long.

[220] Itzkovitz S, Tlusty T, Alon U (2006) Coding limits on the number of transcription factors. BMC Genomics 7:239. doi:10.1186/1471-2164-7-239. URL http://www.biomedcentral.com/1471-2164/7/239.

[221] Massari M, Murre C (2000) Helix-loop-helix proteins: Regulators of transcription in eucaryotic organisms. Mol Cell Biol 20:429. doi:10.1128/MCB.20.2.429-440.2000. URL http://mcb.asm.org/cgi/content/full/20/2/429.

[222] Beltran AC, Dawson PE, Gottesfeld JM (2005) Role of dna sequence in the binding specificity of synthetic basic-helix-loop-helix domains. Chembiochem 6:104–13. doi:10.1002/cbic.200400184.

[223] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860. doi:doi:10.1038/35057062. URL http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html.

[224] Noyes MB, Christensen RG, Wakabayashi A, Stormo GD, Brodsky MH, et al. (2008) Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133:1277–89. doi:10.1016/j.cell.2008.05.023. URL http:

//www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4SVCPF4-P&_user=4420&_coverDate=06%252F27%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=8cde2b8b70bf11cbc2000ef9ac5670b4.

[225] Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, et al. (2009) Diversity and complexity in dna recognition by transcription factors. Science 324:1720–3. doi:10.1126/science.1162327. URL http://www.sciencemag.org/cgi/content/abstract/324/5935/1720.

[226] Biggin MD, McGinnis W (1997) Regulation of segmentation and segmental identity by drosophila homeoproteins: the role of dna binding in functional activity and specificity. Development 124:4425–33.

[227] Gebelein B, McKay DJ, Mann RS (2004) Direct integration of hox and segmentation gene inputs during drosophila development. Nature 431:653–9. doi:10.1038/nature02946. URL http://www.nature.com/nature/journal/v431/n7009/full/nature02946.html.

[228] Carr A, Biggin MD (1999) A comparison of in vivo and in vitro dna-binding specificities suggests a new model for homeoprotein dna binding in drosophila embryos. The EMBO Journal 18:1598. doi:doi:10.1093/emboj/18.6.1598. URL http://www.nature.com/emboj/journal/v18/n6/abs/7591589a.html.

[229] Guertin MJ, Lis JT (2010) Chromatin landscape dictates hsf binding to target dna elements. PLoS Genet 6. doi:10.1371/journal.pgen.1001114.

[230] Ponting CP, Russell RR (2002) The natural history of protein domains. Annu Rev Biophys Biomol Struct 31:45–71. doi:10.1146/annurev.biophys.31.082901.134314.

[231] Gramzow L, Ritz MS, Theissen G (2010) On the origin of mads-domain transcription factors. Trends Genet 26:149–53. doi:10.1016/j.tig.2010.01.004.

[232] Falciani F, Hausdorf B, Schröder R, Akam M, Tautz D, et al. (1996) Class 3 hox genes in insects and the origin of zen. Proc Natl Acad Sci USA 93:8479–84. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=8710895&dopt=abstractplus.

[233] Stauber M, Jäckle H, Schmidt-Ott U (1999) The anterior determinant bicoid of drosophila is a derived hox class 3 gene. Proceedings of the National Academy of Sciences of the United States of America 96:3786–9. URL http://www.pnas.org/content/96/7/3786.long.

[234] Stauber M, Prell A, Schmidt-Ott U (2002) A single hox3 gene with composite bicoid and zerknullt expression characteristics in non-cyclorrhaphan flies. Proc Natl Acad Sci USA 99:274–9. doi:10.1073/pnas.012292899. URL http://www.pnas.org/content/99/1/274.long.

[235] Hughes CL, Liu PZ, Kaufman TC (2004) Expression patterns of the rogue hox genes hox3/zen and fushi tarazu in the apterygote insect thermobia domestica. Evol Dev 6:393–401. doi:10.1111/j.1525-142X.2004.04048.x. URL http://www3.interscience.wiley.com/journal/118815517/abstract?CRETRY=1&SRETRY=0.

[236] Panfilio KA, Akam M (2007) A comparison of hox3 and zen protein coding sequences in taxa that span the hox3/zen divergence. Dev Genes Evol 217:323–9. doi:10.1007/s00427-007-0133-8. URL http://www.springerlink.com/content/r2688057r2061572/.

[237] Papillon D, Telford MJ (2007) Evolution of hox3 and ftz in arthropods: insights from the crustacean daphnia pulex. Dev Genes Evol 217:315–22. doi:10.1007/s00427-007-0141-8. URL http://www.springerlink.com/content/0492536281161578/.

[238] Ephrussi A, Johnston DS (2004) Seeing is believing: the bicoid morphogen gradient matures. Cell 116:143–52. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4BJK582-2&_user=4420&_coverDate=01%252F23%252F2004&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=1e40068c25d3fe85406e3c995a04f5e6&searchtype=a.

[239] Rushlow C, Doyle H, Hoey T, Levine M (1987) Molecular characterization of the zerknüllt region of the antennapedia gene complex in drosophila. Genes & Development 1:1268–79. URL http://genesdev.cshlp.org/content/1/10/1268.long.

[240] Wakimoto BT, Turner FR, Kaufman TC (1984) Defects in embryogenesis in mutants associated with the antennapedia gene complex of drosophila melanogaster. Dev Biol 102:147–72. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WDG-4DV058X-DB&_user=4420&_coverDate=03%252F31%252F1984&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=56fb7513087dedf278bb24c30555a53e&searchtype=a.

[241] Haerty W, Artieri C, Khezri N, Singh RS, Gupta BP (2008) Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. BMC Genomics 9:399. doi:10.1186/1471-2164-9-399. URL http://www.biomedcentral.com/1471-2164/9/399.

[242] Hamilton AT, Huntley S, Tran-Gyamfi M, Baggott DM, Gordon L, et al. (2006) Evolutionary expansion and divergence in the znf91 subfamily of primate-specific zinc finger genes. Genome Res 16:584–94. doi:10.1101/gr.4843906. URL http://genome.cshlp.org/content/16/5/584.long.

[243] Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L (2003) Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome Res 13:1097–110. doi:10.1101/gr.963903.

[244] Ruez C, Payre F, Vincent A (1998) Transcriptional control of drosophila bicoid by serendipity delta: cooperative binding sites, promoter context, and co-evolution. Mech Dev 78:125–34.

[245] Ferrer P, Crozatier M, Salles C, Vincent A (1994) Interspecific comparison of drosophila serendipity delta and beta: multimodular structure of these c2h2 zinc finger proteins. J Mol Evol 38:263–73.

[246] Shultzaberger R (2009) Functional Variability in Transcription Initiation Complexes. Ph.D. thesis, University of California, Berkeley.

[247] Mannhaupt G, Schnall R, Karpov V, Vetter I, Feldmann H (1999) Rpn4p acts as a transcription factor by binding to pace, a nonamer box found upstream of 26s proteasomal and other genes in yeast. FEBS Lett 450:27–34. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6T36-3WHK65P-6&_user=4420&_coverDate=04%252F30%252F1999&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=152b9e0c191287d3bbc783dd05edd492&searchtype=a.

[248] Xie Y, Varshavsky A (2001) Rpn4 is a ligand, substrate, and transcriptional regulator of the 26s proteasome: a negative feedback circuit. Proceedings of the National Academy of Sciences of the United States of America 98:3056–61. doi:10.1073/pnas.071022298. URL http://www.pnas.org/content/98/6/3056.long.

[249] Fernandes L, Rodrigues-Pousada C, Struhl K (1997) Yap, a novel family of eight bzip proteins in saccharomyces cerevisiae with distinct biological functions. Mol Cell Biol 17:6982–93. URL http://mcb.asm.org/cgi/reprint/17/12/6982?view=long&pmid=9372930.

[250] Rodrigues-Pousada C, Menezes RA, Pimentel C (2010) The yap family and its role in stress response. Yeast 27:245–58. doi:10.1002/yea.1752. URL http://onlinelibrary.wiley.com/doi/10.1002/yea.1752/abstract;jsessionid=3A3A6C09DF376D52BE861EC8C0011FEA.d02t01.

[251] Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, et al. (2004) Transcriptional regulatory code of a eukaryotic genome. Nature 431:99–104. doi:10.1038/nature02800. URL http://www.nature.com/nature/journal/v431/n7004/full/nature02800.html.

[252] Tan K, Feizi H, Luo C, Fan SH, Ravasi T, et al. (2008) A systems approach to delineate functions of paralogous transcription factors: role of the yap family in the dna damage response. Proceedings of the National Academy of Sciences of the United States of America 105:2934–9. doi:10.1073/pnas.0708670105. URL http://www.pnas.org/content/105/8/2934.long.

[253] Suckow M, Kisters-Woike B, Hollenberg CP (1999) A novel feature of dna recognition: a mutant gcn4p bzip peptide with dual dna binding

specificities dependent of half-site spacing. J Mol Biol 286:983–7. doi:10.1006/jmbi.1999.2537. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WK7-45V8044-K&_user=4420&_coverDate=03%252F05%252F1999&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=1dacaf1d8e206bc5b8556992e46834cf&searchtype=a.

[254] Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T (2000) Structural basis for the diversity of dna recognition by bzip transcription factors. Nat Struct Biol 7:889–93. doi:10.1038/82822. URL http://www.nature.com/doifinder/10.1038/82822.

[255] Macisaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, et al. (2006) An improved map of conserved regulatory sites for saccharomyces cerevisiae. BMC Bioinformatics 7:113. doi:10.1186/1471-2105-7-113. URL http://www.biomedcentral.com/1471-2105/7/113.

[256] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, et al. (2009) Accelerated evolution of the prdm9 speciation gene across diverse metazoan taxa. PLoS Genet 5:e1000753. doi:10.1371/journal.pgen.1000753. URL http://www.plosgenetics.org/article/info%253Adoi%252F10.1371%252Fjournal.pgen.1000753.

[257] Good JM, Dean MD, Nachman MW (2008) A complex genetic basis to x-linked hybrid male sterility between two species of house mice. Genetics 179:2213–28. doi:10.1534/genetics.107.085340. URL http://www.genetics.org/cgi/content/full/179/4/2213.

[258] Good JM, Handel MA, Nachman MW (2008) Asymmetry and polymorphism of hybrid male sterility during the early stages of speciation in house mice. Evolution 62:50–65. doi:10.1111/j.1558-5646.2007.00257.x. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1558-5646.2007.00257.x/abstract.

[259] Storchová R, Gregorová S, Buckiová D, Kyselová V, Divina P, et al. (2004) Genetic analysis of x-linked hybrid sterility in the house mouse. Mamm Genome 15:515–24. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=15366371&dopt=abstractplus.

[260] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J (2009) A mouse speciation gene encodes a meiotic histone h3 methyltransferase. Science 323:373–5. doi:10.1126/science.1163601. URL http://www.sciencemag.org/content/323/5912/373.short.

[261] Sun S, Ting CT, Wu CI (2004) The normal function of a speciation gene, odysseus, and its hybrid sterility effect. Science 305:81–3. doi:10.1126/science.1093904.

[262] Ting CT, Tsaur SC, Sun S, Browne WE, Chen YC, et al. (2004) Gene duplication and speciation in drosophila: evidence from the odysseus locus. Proc Natl Acad Sci USA 101:12232–5. doi:10.1073/pnas.0401975101.

[263] Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in drosophila hybrids. Science 323:376–9. doi:10.1126/science.1163934. URL http://www.sciencemag.org/content/323/5912/376.short.

[264] Consortium DG, Clark AG, Eisen MB, Smith DR, Bergman CM, et al. (2007) Evolution of genes and genomes on the drosophila phylogeny. Nature 450:203–18. doi:10.1038/nature06341. URL http://www.nature.com/nature/journal/v450/n7167/full/nature06341.html.

[265] Li H, Baker BS (1998) hermaphrodite and doublesex function both dependently and independently to control various aspects of sexual differentiation in drosophila. Development 125:2641–51. URL http://dev.biologists.org/content/125/14/2641.long.

[266] Pultz MA, Baker BS (1995) The dual role of hermaphrodite in the drosophila sex determination regulatory hierarchy. Development 121:99–111. URL http://dev.biologists.org/content/121/1/99.long.

[267] Li H, Baker BS (1998) Her, a gene required for sexual differentiation in drosophila, encodes a zinc finger protein with characteristics of zfy-like proteins and is expressed independently of the sex determination hierarchy. Development 125:225–35. URL http://dev.biologists.org/content/125/2/225.long.

[268] Pultz MA, Carson GS, Baker BS (1994) A genetic analysis of hermaphrodite, a pleiotropic sex determination gene in drosophila melanogaster. Genetics 136:195–207. URL http://www.genetics.org/cgi/reprint/136/1/195.

[269] Markow TA, O'Grady PM (2007) Drosophila biology in the genomic age. Genetics 177:1269–76. doi:10.1534/genetics.107.074112. URL http://www.genetics.org/cgi/content/full/177/3/1269.

[270] Adryan B, Teichmann SA (2006) Flytf: a systematic review of site-specific transcription factors in the fruit fly drosophila melanogaster. Bioinformatics 22:1532–3. doi:10.1093/bioinformatics/btl143. URL http://bioinformatics.oxfordjournals.org/cgi/content/full/22/12/1532?view=long&pmid=16613907.

[271] Finn R, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The pfam protein families database. Nucleic Acids Research 38:D211. doi:10.1093/nar/gkp985. URL http://nar.oxfordjournals.org/cgi/content/full/38/suppl_1/D211.

[272] Eddy SR (1998) Profile hidden markov models. Bioinformatics 14:755–63. URL http://bioinformatics.oxfordjournals.org/cgi/reprint/14/9/755?view=long&pmid=9918945.

[273] Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, et al. (2007) Creating a honey bee consensus gene set. Genome Biol 8:R13. doi:10.1186/gb-2007-8-1-r13. URL http://genomebiology.com/2007/8/1/R13.

[274] Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113. doi:10.1186/1471-2105-5-113.

[275] Yang Z (1997) Paml: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–6. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=9367129&dopt=abstractplus.

[276] Perkins KK, Dailey GM, Tjian R (1988) Novel jun- and fos-related proteins in drosophila are functionally homologous to enhancer factor ap-1. The EMBO Journal 7:4265. URL http://www.ncbi.nlm.nih.gov/pubmed/3149584?dopt=AbstractPlus&tool=FlyBase.

[277] Ellenberger TE, Brandl CJ, Struhl K, Harrison SC (1992) The gcn4 basic region leucine zipper binds dna as a dimer of uninterrupted alpha helices: crystal structure of the protein-dna complex. Cell 71:1223–37. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4HHXWYJ-J&_user=4420&_coverDate=12%252F24%252F1992&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=9b9102ee74fc8fc0b34b22d0bfcba28b.

[278] Ollmann M, Young LM, Como CJD, Karim F, Belvin M, et al. (2000) Drosophila p53 is a structural and functional homolog of the tumor suppressor p53. Cell 101:91–101. doi:10.1016/S0092-8674(00)80626-1.

[279] Cho Y, Gorina S, Jeffrey PD, Pavletich NP (1994) Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations. Science 265:346–55.

[280] Reuter R (1994) The gene serpent has homeotic properties and specifies endoderm versus ectoderm within the drosophila gut. Development 120:1123–35.

[281] Omichinski JG, Clore GM, Schaad O, Felsenfeld G, Trainor C, et al. (1993) Nmr structure of a specific dna complex of zn-containing dna binding domain of gata-1. Science 261:438–46. URL http://www.sciencemag.org/cgi/reprint/261/5120/438.

[282] Gabrieli P, Falaguerra A, Siciliano P, Gomulski LM, Scolari F, et al. (2010) Sex and the single embryo: early deveiopment in the mediterranean fruit fly, ceratitis capitata. BMC Dev Biol 10:12. doi:10.1186/1471-213X-10-12. URL http://www.biomedcentral.com/1471-213X/10/12.

[283] Haag ES, Doty AV (2005) Sex determination across evolution: connecting the dots. PLoS Biol 3:e21. doi:10.1371/journal.pbio.0030021. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0030021.

[284] Schutt C, Nothiger R (2000) Structure, function and evolution of sex-determining systems in dipteran insects. Development URL http://dev.biologists.org/cgi/content/abstract/127/4/667.

[285] Kopp A, True JR (2002) Evolution of male sexual characters in the oriental drosophila melanogaster species group. Evol Dev 4:278–91. URL http://www3.interscience.wiley.com/journal/118932864/abstract.

[286] Birney E, Clamp M, Durbin R (2004) Genewise and genomewise. Genome Res 14:988. doi:10.1101/gr.1865504. URL http://www.ncbi.nlm.nih.gov/pubmed/15123596.

[287] Sonnhammer EL, Durbin R (1995) A dot-matrix program with dynamic threshold control suited for genomic dna and protein sequence analysis. Gene 167:GC1–10.

[288] Karolchik D, Bejerano G, Hinrichs AS, Kuhn RM, Miller W, et al. (2007) Comparative genomic analysis using the ucsc genome browser. Methods Mol Biol 395:17–34. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=17993665&dopt=abstractplus.

[289] Wallis MC, Waters PD, Graves JAM (2008) Sex determination in mammals–before and after the evolution of sry. Cell Mol Life Sci 65:3182–95. doi:10.1007/s00018-008-8109-z.

[290] Cline TW, Meyer BJ (1996) Vive la différence: males vs females in flies vs worms. Annu Rev Genet 30:637–702. doi:10.1146/annurev.genet.30.1.637. URL http://arjournals.annualreviews.org/doi/abs/10.1146/annurev.genet.30.1.637?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub%253dncbi.nlm.nih.gov.

[291] Camara N, Whitworth C, Doren MV (2008) The creation of sexual dimorphism in the drosophila soma. Curr Top Dev Biol 83:65–107. doi:10.1016/S0070-2153(08)00403-1. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B7CTT-4V87HSB-7&_user=4420&_coverDate=12%252F31%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=7027c6fa736331b7d72cdc6d21b6e93e.

[292] Erickson JW, Quintero JJ (2007) Indirect effects of ploidy suggest x chromosome dose, not the x:a ratio, signals sex in drosophila. PLoS Biol 5:e332. doi:10.1371/journal.pbio.0050332. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.0050332.

[293] Burtis KC, Coschigano KT, Baker BS, Wensink PC (1991) The doublesex proteins of drosophila melanogaster bind directly to a sex-specific yolk protein gene enhancer. The EMBO Journal 10:2577–82. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=1907913&dopt=abstractplus.

[294] Cho S, Wensink PC (1997) Dna binding by the male and female doublesex proteins of drosophila melanogaster. J Biol Chem 272:3185–9. URL http://www.jbc.org/content/272/6/3185.long.

[295] Garrett-Engele CM, Siegal ML, Manoli DS, Williams BC, Li H, et al. (2002) intersex, a gene required for female sexual development in drosophila, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation. Development 129:4661–75. URL http://www.google.com/search?client=safari&rls=en-us&q=intersex,+a+gene+required+for+female+sexual+development+in+Drosophila,+is+expressed+in+both+sexes+and+functions+together+with+doublesex+to+regulate+terminal+differentiation&ie=UTF-8&oe=UTF-8.

[296] Coschigano KT, Wensink PC (1993) Sex-specific transcriptional regulation by the male and female doublesex proteins of drosophila. Genes Dev 7:42–54. URL http://genesdev.cshlp.org/content/7/1/42.long.

[297] DeFalco T, Camara N, Bras SL, Doren MV (2008) Nonautonomous sex determination controls sexually dimorphic development of the drosophila gonad. Dev Cell 14:275–86. doi:10.1016/j.devcel.2007.12.005. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WW3-4RTD1MT-J&_user=4420&_coverDate=02%252F12%252F2008&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=ec6c87ffe81c0df5abc1db2b73f1f775.

[298] Ahmad SM, Baker BS (2002) Sex-specific deployment of fgf signaling in drosophila recruits mesodermal cells into the male genital imaginal disc. Cell 109:651–61. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-463G2Y1-F&_user=4420&_coverDate=05%252F31%252F2002&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=27237fd1e1209254e3ce2e74f8f2d8eb.

[299] ten Bosch JR, Benavides JA, Cline TW (2006) The tagteam dna motif controls the timing of drosophila pre-blastoderm transcription. Development 133:1967–77. doi:10.1242/dev.02373. URL http://dev.biologists.org/content/133/10/1967.long.

[300] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–10. doi:10.1006/jmbi.1990.9999. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2231712&dopt=abstractplus.

[301] Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 10:R25. doi:10.1186/gb-2009-10-3-r25.

[302] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, et al. (2008) Model-based analysis of chip-seq (macs). Genome Biol 9:R137. doi:10.1186/gb-2008-9-9-r137.

[303] Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=7584402&dopt=abstractplus.

[304] Hertz GZ, Hartzell GW, Stormo GD (1990) Identification of consensus patterns in unaligned dna sequences known to be functionally related. Comput Appl Biosci 6:81–92. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2193692&dopt=abstractplus.

[305] Rach EA, Yuan HY, Majoros WH, Tomancak P, Ohler U (2009) Motif composition, conservation and condition-specificity of single and alternative transcription start sites in the drosophila genome. Genome Biol 10:R73. doi:10.1186/gb-2009-10-7-r73. URL http://genomebiology.com/2009/10/7/R73.

[306] Noyes MB, Meng X, Wakabayashi A, Sinha S, Brodsky MH, et al. (2008) A systematic characterization of factors that regulate drosophila segmentation via a bacterial one-hybrid system. Nucleic Acids Research 36:2547–60. doi:10.1093/nar/gkn048. URL http://nar.oxfordjournals.org/cgi/content/full/36/8/2547?view=long&pmid=18332042.

[307] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Research 38:D355. doi:10.1093/nar/gkp896. URL http://nar.oxfordjournals.org/cgi/content/full/38/suppl_1/D355.

[308] Chase BA, Baker BS (1995) A genetic analysis of intersex, a gene regulating sexual differentiation in drosophila melanogaster females. Genetics 139:1649. URL http://www.ncbi.nlm.nih.gov/pubmed/7789766?dopt=AbstractPlus&tool=FlyBase.

[309] Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, et al. (2009) Unlocking the secrets of the genome. Nature 459:927–30. doi:10.1038/459927a. URL http://www.nature.com/nature/journal/v459/n7249/full/459927a.html.

[310] Tomancak P, Beaton A, Weiszmann R, Kwan E, Shu S, et al. (2002) Systematic determination of patterns of gene expression during drosophila embryogenesis. Genome Biol 3:RESEARCH0088.

[311] Zeitlinger J, Zinzen RP, Stark A, Kellis M, Zhang H, et al. (2007) Whole-genome chip-chip analysis of dorsal, twist, and snail suggests integration of diverse patterning processes in the drosophila embryo. Genes Dev 21:385–90. doi:10.1101/gad.1509607.

[312] Sandmann T, Girardot C, Brehme M, Tongprasit W, Stolc V, et al. (2007) A core transcriptional network for early mesoderm development in drosophila melanogaster. Genes Dev 21:436–49. doi:10.1101/gad.1509007. URL http://genesdev.cshlp.org/content/21/4/436.long.

[313] Lü J, Oliver B (2001) Drosophila ovo regulates ovarian tumor transcription by binding unusually near the transcription start site. Development 128:1671–86.

[314] Sefton L, Timmer JR, Zhang Y, Béranger F, Cline TW (2000) An extracellular activator of the drosophila jak/stat pathway is a sex-determination signal element. Nature 405:970–3. doi:10.1038/35016119. URL http://www.nature.com/nature/journal/v405/n6789/full/405970a0.html.

[315] Jinks TM, Polydorides AD, Calhoun G, Schedl P (2000) The jak/stat signaling pathway is required for the initial choice of sexual identity in drosophila melanogaster. Mol Cell 5:581–7. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSR-4194KG4-J&_user=4420&_coverDate=03%252F31%252F2000&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=63319809248340baaf94f5404d46d57d.

[316] Irish VF, Gelbart WM (1987) The decapentaplegic gene is required for dorsal-ventral patterning of the drosophila embryo. Genes Dev 1:868–79.

[317] Bai J, Chiu W, Wang J, Tzeng T, Perrimon N, et al. (2001) The cell adhesion molecule echinoid defines a new pathway that antagonizes the drosophila egf receptor signaling pathway. Development 128:591–601.

[318] Vässin H, Campos-Ortega JA (1987) Genetic analysis of delta, a neurogenic gene of drosophila melanogaster. Genetics 116:433–445.

[319] Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) Weblogo: a sequence logo generator. Genome Res 14:1188–90. doi:10.1101/gr.849004. URL http://genome.cshlp.org/content/14/6/1188.long.

[320] Siera SG, Cline TW (2008) Sexual back talk with evolutionary implications: stimulation of the drosophila sex-determination gene sex-lethal by its target transformer. Genetics 180:1963–81. doi:10.1534/genetics.108.093898. URL http://www.genetics.org/cgi/content/full/180/4/1963.

[321] Bradley RK, Roberts A, Smoot M, Juvekar S, Do J, et al. (2009) Fast statistical alignment. PLoS Comput Biol 5:e1000392. doi:10.1371/journal.pcbi.1000392.

[322] Seher TC, Leptin M (2000) Tribbles, a cell-cycle brake that coordinates proliferation and morphogenesis during drosophila gastrulation. Curr Biol 10:623–9.

[323] Grieder NC, Nellen D, Burke R, Basler K, Affolter M (1995) Schnurri is required for drosophila dpp signaling and encodes a zinc finger protein similar to the mammalian transcription factor prdii-bf1. Cell 81:791–800. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6WSN-4D16PVD-M&_user=4420&_coverDate=06%252F02%252F1995&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=a214410b6d4c8161484b250e78ebff33&searchtype=a.

[324] Tsonis PA, Goetinck PF (1988) The drosophila homoeotic gene spalt is structurally related to collagen alpha 1(iv) chain. Coll Relat Res 8:451–2. URL http://www.ncbi.nlm.nih.gov/sites/entrez?Db=pubmed&Cmd=Retrieve&list_uids=2906282&dopt=abstractplus.

[325] Cheah PY, Meng YB, Yang X, Kimbrell D, Ashburner M, et al. (1994) The drosophila l(2)35ba/noca gene encodes a putative zn finger protein involved in the development

of the embryonic brain and the adult ocellar structures. Mol Cell Biol 14:1487–99. URL http://mcb.asm.org/cgi/reprint/14/2/1487?view=long&pmid=8289824.

[326] Boube M, Faucher C, Joulia L, Cribbs DL, Bourbon HM (2000) Drosophila homologs of transcriptional mediator complex subunits are required for adult cell and segment identity specification. Genes & Development 14:2906–17.

[327] Lundström A, Gallio M, Englund C, Steneberg P, Hemphälä J, et al. (2004) Vilse, a conserved rac/cdc42 gap mediating robo repulsion in tracheal cells and axons. Genes & Development 18:2161–71. doi:10.1101/gad.310204.

[328] Williams BC, Garrett-Engele CM, Li Z, Williams EV, Rosenman ED, et al. (2003) Two putative acetyltransferases, san and deco, are required for establishing sister chromatid cohesion in drosophila. Curr Biol 13:2025–36. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VRT-4B5333W-N&_user=4420&_coverDate=12%252F02%252F2003&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=31b9fc6b09f950243bf2502a875191a4&searchtype=a.

[329] Powell J (1997) Progress and Prospects in Evolutionary Biology : the Drosophila Model. Oxford University Press, New York.

[330] Bradley RK, Li XY, Trapnell C, Davidson S, Pachter L, et al. (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related drosophila species. PLoS Biol 8:e1000343. doi:10.1371/journal.pbio.1000343. URL http://www.plosbiology.org/article/info%253Adoi%252F10.1371%252Fjournal.pbio.1000343.

[331] Bourque G (2009) Transposable elements in gene regulation and in the evolution of vertebrate genomes. Curr Opin Genet Dev 19:607–12. doi:10.1016/j.gde.2009.10.013. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6VS0-4XP2D3P-1&_user=4420&_coverDate=12%252F31%252F2009&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=e6a4a96ec47a5b959409c63e780b2d6a&searchtype=a.

[332] Bourque G, Leong B, Vega VB, Chen X, Lee YL, et al. (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. Genome Res 18:1752–62. doi:10.1101/gr.080663.108. URL http://genome.cshlp.org/content/18/11/1752.long.

[333] Sasaki T, Nishihara H, Hirakawa M, Fujimura K, Tanaka M, et al. (2008) Possible involvement of sines in mammalian-specific brain formation. Proceedings of the National Academy of Sciences of the United States of America 105:4220–5. doi:10.1073/pnas.0709398105. URL http://www.pnas.org/content/105/11/4220.long.

[334] Jordan IK, Rogozin IB, Glazko GV, Koonin EV (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. Trends Genet 19:68–72. URL http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6TCY-47DT7P2-1&_user=4420&_coverDate=02%252F28%252F2003&_rdoc=1&_fmt=high&_orig=search&_origin=search&_sort=d&_docanchor=&view=c&_acct=C000059607&_version=1&_urlVersion=0&_userid=4420&md5=862a06475df950ff39946b0925426128&searchtype=a.

[335] Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, et al. (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42:631–4. doi:10.1038/ng.600. URL http://www.nature.com/ng/journal/v42/n7/full/ng.600.html.

[336] Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, et al. (2002) The transposable elements of the drosophila melanogaster euchromatin: a genomics perspective. Genome Biology 3:RESEARCH0084.

[337] Maerkl SJ, Quake SR (2007) A systems approach to measuring the binding energy landscapes of transcription factors. Science 315:233–7. doi:10.1126/science.1131007. URL http://www.sciencemag.org/cgi/content/full/315/5809/233.