UNIVERSITY OF CAMBRIDGE

Evolutionary patterns of group B Sox binding and function in *Drosophila*

Sarah Hamilton Carl

Darwin College

A thesis submitted for the Degree of Doctor of Philosophy

October 2014

Abstract

Genome-wide binding and expression studies in *Drosophila melanogaster* have revealed widespread roles for Dichaete and SoxNeuro, two group B Sox proteins, during fly development. Although they have distinct target genes, these two transcription factors bind in very similar patterns across the genome and can partially compensate for each other's loss, both phenotypically and at the level of DNA binding. However, the inherent noise in genome-wide binding studies as well as the high affinity of transcription factors for DNA and the potential for non-specific binding makes it difficult to identify true functional binding events. Additionally, external factors such as chromatin accessibility are known to play a role in determining binding patterns in *Drosophila*. A comparative approach to transcription factor binding facilitates the use of evolutionary conservation to identify functional features of binding patterns. In order to discover highly conserved features of group B Sox binding, I performed DamID-seq for SoxNeuro and Dichaete in four species of *Drosophila, D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. I also performed FAIRE-seq in *D. pseudoobscura* embryos to compare the chromatin accessibility landscape between two fly species and to examine the relationship between open chromatin and group B Sox binding.

I found that, although the sequences, expression patterns and overall transcriptional regulatory targets of Dichaete and SoxNeuro are highly conserved across the drosophilids, both binding site turnover and rates of quantitative binding divergence between species increase with phylogenetic distance. Elevated rates of binding conservation can be found at bound genomic intervals overlapping functional sites, including known enhancers, direct targets of Dichaete and SoxNeuro, and core binding intervals identified in previous genome-wide studies. Sox motifs identified in intervals that show binding conservation are also more highly

conserved than those in intervals that are only bound in one species. Notably, regions that are bound in common by SoxNeuro and Dichaete are more likely to be conserved between species than those bound by one protein alone. However, by examining binding intervals that are uniquely bound by one protein and conserved, I was able to identify distinctive features of the targets of each transcription factor that point to unique aspects of their functions.

My comparative analysis of group B Sox binding suggests that sites that are commonly bound by Dichaete and SoxNeuro, primarily at targets in the developing nervous system, are highly constrained by natural selection. Uniquely bound targets have different tissue expression profiles, leading me to propose a model whereby the unique functions of Dichaete and SoxNeuro may arise from a combination of differences in their own expression patterns and the broader nuclear environment, including tissue-specific cofactors and patterns of accessible chromatin. These results shed light on the evolutionary forces that have maintained deep conservation of the complex functional relationships between group B Sox proteins from insects to mammals.

Declaration

This dissertation:

- is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text.
- is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution except as specified in the text; and
- does not exceed the prescribed limit of 60,000 words for the Degree Committee of the Faculty of Biology.

Sarah Hamilton Carl
October 2014

Acknowledgements

First and foremost, I would like to thank my boss, Steve. From the first time I met with him, I knew this was a lab I wanted to work in. Thank you for our in-depth conversations, for sharing your passion for science, for challenging me and encouraging me when it seemed like everything was going wrong. Mostly, thank you for believing in me as a scientist.

To the other members of the lab, thank you for all the support you have provided, in big ways and small. Enrico and Katerina, you were two of the first people I met in the lab, and it's been great fun sharing this ride with you. Thanks for your advice and commiseration! Josh and Alex, I've really enjoyed getting to know you guys over the last year and sharing your enthusiasm as you've started your projects. I know there will be frustrating times, but I hope you find your Ph.D.s to be as rewarding of a journey as I have. A big thanks is due to Sang for invaluable cloning advice and for injecting my flies (and injecting them again ... and again!).

Bettina, you've provided amazing support (and wonderful rhubarb tarts!) along the way. Thank you for answering every question, no matter how small or silly, and helping me overcome my fear of R! Jelena, I credit you with teaching me to program for the first time and opening my eyes to the world of bioinformatics. I don't think it's an exaggeration to say that your influence has changed the path of my career. I've been lucky to have you as a friend and mentor.

Many other members of the Genetics Department have shared their friendship, advice and encouragement over the last four years. I won't try to name everyone because I will surely leave someone out, but thanks to you all for fond memories of Happy Hour shenanigans, Tuesday evenings in the Flying Pig and many other

adventures!

A special thanks to my boyfriend Sam (<3), for broadening my world, encouraging me to learn more about programming, lively scientific and philosophical discussions, and your unwavering support.

I owe a big thanks to my parents, Norm and Theresa. I know it's been hard for you to have me so far away, and I appreciate your putting up with my wanderlust and continuing to be so supportive! On a more pragmatic note, thank you for providing me with food, an internet connection, and a place to work and sleep during the last 6 weeks or so of the thesis-writing process. I love you both. Dad, I know you've been busy conducting your own research. I hope this thesis is up to your standards, and I eagerly anticipate the chance to finally compare our results.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

Although a large part of modern biology is devoted to uncovering the functions of the vast array of DNA, RNA and protein molecules that make up an organism, the concept of function remains surprisingly slippery. This can be best illustrated by the recent uproar surrounding the publication of the largest collection of datasets related to non-coding DNA to date by the ENCODE and modENCODE projects (The modENCODE Consortium *et al.*, 2010; Dunham *et al.*, 2012). Famously, through integrating all of its datasets, the ENCODE consortium was able to grant 80.4% of the nucleotides in the human genome a function; this figure, however, was quickly and hotly disputed (Dunham *et al.*, 2012; Graur *et al.*, 2013). It can be said that the function of a transcription factor (TF) is to bind DNA and regulate the expression of target genes; however, the complexity of combinatorial binding patterns and the sheer quantity of binding events, even in the model organism *Drosophila*, which has a smaller and more compact genome than humans, suggest that TF function is complex and context-dependent (Biggin, 2011; Kaplan *et al.*, 2011; Neph *et al.*, 2012; Zinzen *et al.*, 2009). One possible measure of biological function comes from the effect of natural selection, which, given a large enough population and free flow of alleles, should remove mutations that are detrimental to an organism and preserve those that allow for correct molecular function. Therefore, sequences or, by extension, TF binding events that are functional should be conserved by selection during evolution. In this thesis, I have

applied the preceding hypothesis to the binding and function of two group B Sox proteins, a family of TFs that is both deeply conserved in animal evolution and shows complex interplays in binding patterns. Here I present an introduction to group B Sox proteins in vertebrates and insects, a review of previous studies that have used evolutionary comparisons to elucidate TF function and an overview of the experiments that I performed.

1.1 Glossary

- **Transcription factor (TF):** A protein whose primary function is to bind to DNA at specific recognition sites, either alone or in a complex with itself (as a homodimer) or other cofactors (as a heterodimer), in order to induce a positive or negative change in the level of transcription of a nearby gene.
- **Regulatory DNA:** Non-coding sequences of DNA that, when bound by the appropriate transcription factors, are necessary and sufficient to direct spatially and temporally specific expression patterns of nearby genes. Regulatory sequences may be located in intergenic DNA (upstream or downstream of genes) as well as in introns. Individual units of regulatory DNA are often referred to as enhancers or cis-regulatory elements (CRMs).
- **Transcription factor binding site (TFBS):** A small stretch of DNA, typically ranging from 6-12 nucleotides, that is recognized and bound by a transcription factor, often resulting in upregulation or downregulation of a nearby target gene. The preferred DNA sequence recognized by a particular TF is often referred to as a sequence motif; however, the sequences of individual TFBS instances can vary, a phenomenon known as degeneracy. Not all binding events of a TF to a TFBS result in a change in gene expression.
- **Target gene:** A gene whose regulatory DNA is bound by a particular TF. Genes whose expression has been demonstrated to change in response to TF binding are typically referred to as direct targets of that TF; however, TF binding at a target gene can also play an indirect role in gene regulation, for example through recruiting and stabilizing cofactors or changing the local chromatin environment.

1.2 Group B Sox Proteins

Sox genes encode a deeply-conserved family of transcription factors (TFs) that serve as broad developmental regulators in metazoa. They are thought to have evolved in conjunction with the origin of multicellular animal life, as they are present in all animal genomes in which they have been searched for, including basal members such as sponges and placozoa (Jager *et al.*, 2006, 2008; Larroux *et al.*, 2006; Phochanukul and Russell, 2010; Srivastava *et al.*, 2008). Members of the Sox (Sry-related high-mobility-group box) family contain one highly conserved HMG (high-mobility group) DNA-binding domain, which typically shares greater than 50% sequence homology to that of the mammalian testis-determining factor SRY (Bowles *et al.*, 2000; Guth and Wegner, 2008; Phochanukul and Russell, 2010; Sinclair *et al.*, 1990). They bind to DNA in the minor groove, recognizing variants of the motif A/TA/TCAAAG, and are known to induce DNA bending (Bowles *et al.*, 2000; Ferrari *et al.*, 1992; Giese *et al.*, 1992). Sox genes are classified into ten groups, A through J, based on HMG sequence and full-length protein structure (Schepers *et al.*, 2002). Members of each subgroup are often expressed in overlapping patterns in particular subsets of tissues during development and play important roles in directing the correct differentiation of cells in those tissues; for example, in vertebrates, group B genes are expressed in the developing central nervous system and eye (Bergsland *et al.*, 2011; Kamachi *et al.*, 1998; Uwanogho *et al.*, 1995; Wood and Episkopou, 1999), while group C genes are expressed in the kidney and pancreas (Huang *et al.*, 2013; Sock *et al.*, 2004; Wilson *et al.*, 2005), groups C, D and E are expressed in the skeleton and cartilage (Akiyama *et al.*, 2002; Smits *et al.*, 2001), and group F genes are expressed in the developing vascular and lymphatic systems (Downes and Koopman, 2001; Matsui, 2006). Based on these observations and genomic studies that have identified many targets of various Sox proteins, it appears that the Sox family has evolved to regulate cell fate decisions in diverse tissue types across the animal phylogeny (Lefebvre *et al.*, 2007; Whyte *et al.*, 2013). While mammalian genomes contain multiple paralogues for most of these groups, invertebrates typically have far fewer Sox genes. Sequenced insect genomes, including that of Drosophila, typically contain one gene in each of groups C, D, E, and F, and four genes in group B, although occasional extra genes have originated in

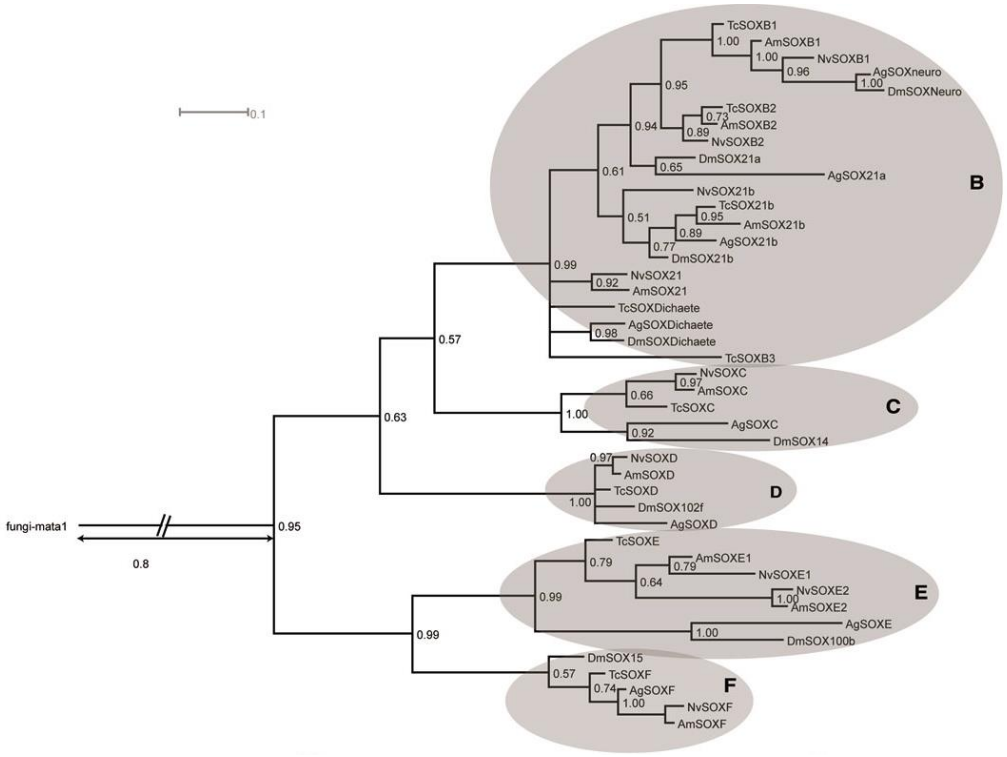


Figure 1.1: Rooted Bayesian phylogeny of representative insect Sox proteins. All species have four group B proteins except for *T. castaneum*, which has an extra group B member (SoxB3), and all species have one member of each other subgroup except for the hymenopterans *N. vitripennis* and *A. mellifera*, which have undergone a gene duplication in group E. Figure reproduced from Wilson and Dearden (2008). Abbreviations: Tc, *Tribolium castaneum*; Am, *Apis mellifera*; Nv, *Nasonia vitripennis* Ag, *Anopheles gambiae*; Dm, *Drosophila melanogaster*.

particular lineages (Figure 1.1) (Bowles *et al.*, 2000; Phochanukul and Russell, 2010).

Group B *Sox* genes are some of the best characterized members of the *Sox* family. In addition to being the most closely related *Sox* genes to *Sry*, they appear to have highly conserved functions throughout evolution (Collignon *et al.*, 1996; McKimmie *et al.*, 2005). In mammals, group B *Sox* genes have been implicated in stem cell pluripotency and self-renewal, ectoderm formation, neural induction, central nervous system (CNS) development, placode formation, and gametogenesis (Guth and Wegner, 2008). A role for group B *Sox* genes in neural development appears to be conserved throughout the higher metazoa, making *Drosophila* an attractive system in which to study group B *Sox* function and evolution more closely

(Uwanogho *et al.*, 1995; Wood and Episkopou, 1999; Wegner and Stolt, 2005). Group B *Sox* genes have also been analyzed at both sequence and expression levels in several species of invertebrates, showing strong evidence for functional conservation but also revealing a complex evolutionary history whose details are not fully resolved (Wilson and Dearden, 2008; McKimmie *et al.*, 2005; Wei *et al.*, 2010; Pioro and Stollewerk, 2006; Zhong *et al.*, 2011). There are four group B *Sox* genes in the *Drosophila melanogaster* genome: *SoxNeuro (SoxN)*, *Dichaete*, *Sox21a*, and *Sox21b* (McKimmie *et al.*, 2005). Of these, the most extensively studied to date are *SoxN* and *Dichaete*.

In vertebrates, group B *Sox* genes are divided into two subgroups: group B1, which includes *Sox1*, *Sox2* and *Sox3* (Collignon *et al.*, 1996), and group B2, which includes *Sox14* and *Sox21* (Malas *et al.*, 1999; McKimmie *et al.*, 2005). In the chicken, group B1 proteins act as transcriptional activators during development, while group B2 proteins act as transcriptional repressors (Uchikawa *et al.*, 1999, 2011). Group B1 and B2 genes play opposing roles in the developing vertebrate CNS, with group B1 proteins conveying early neuroectodermal competence and maintaining neural precursors while group B2 proteins promote neuronal differentiation (Wegner and Stolt, 2005; Wegner, 2011). Although it has been argued based on sequence orthology that *SoxN* is a group B1 gene while *Dichaete* is more closely related to the B2 subgroup (Bowles *et al.*, 2000; Guth and Wegner, 2008; Wegner and Stolt, 2005; Zhong *et al.*, 2011), functional arguments place *Dichaete* with the group B1 genes (McKimmie *et al.*, 2005). For example, *Dichaete* specific mutant phenotypes in the *Drosophila* CNS midline are rescued by expression of the mouse Sox2 protein, supporting the idea that both *Dichaete* and *SoxN* may be orthologous to vertebrate group B1 genes (Sánchez-Soriano and Russell, 1998). Additionally, Dichaete is known to interact molecularly with the POU-domain protein Ventral veins lacking (Vvl), while mammalian Sox2 interacts with the POU protein Oct4 and can also interact with Vvl when expressed in the fly (Ambrosetti *et al.*, 1997; Archer *et al.*, 2011; Bery *et al.*, 2013; Ma *et al.*, 2000; Masui *et al.*, 2007; Sánchez-Soriano and Russell, 1998; Tanaka *et al.*, 2004). Further functional data suggests that the B1-B2 division may not be functionally relevant in insects, as both *Dichaete* and *SoxN* play a number of complex roles during development that correspond to those played by vertebrate group B1 and B2 *Sox* genes and that cannot be neatly divided into activator

and repressor functions (Ferrero *et al.*, 2014). Although it is difficult to assign orthology between vertebrate and insect group B *Sox* genes due to their divergent evolutionary histories (McKimmie *et al.*, 2005; Wilson and Dearden, 2008; Zhong *et al.*, 2011), the similarities in the expression patterns and functions of *Sox1*, *Sox2* and *Sox3* in vertebrates and *SoxN* and *Dichaete* in insects suggest that a combination of descent from a common group B *Sox* ancestor and functional convergent evolution have shaped a deeply conserved yet complex relationship between these two sets of *Sox* genes (Crémazy *et al.*, 2000; Sánchez-Soriano and Russell, 1998; Uwanogho *et al.*, 1995; Wood and Episkopou, 1999; Zhong *et al.*, 2011).

Studies of *in vivo* binding patterns of Sox proteins in mammals and flies have identified a large number of conserved orthologous targets, reinforcing the observation that the division of functions between group B paralogues cannot be simply translated from vertebrates to invertebrates. In the mouse, the group B1 genes *Sox2* and *Sox3* as well as the group C gene *Sox11* are expressed in a successive fashion in the developing CNS; a recent ChIP-seq study examined binding patterns of Sox2, Sox3 and Sox11 in neural precursor cells (NPCs) and differentiated neurons. Although Sox2 and Sox3 are primarily responsible for maintaining NPCs, while Sox11 plays an opposite role by promoting the differentiation of neurons, all three proteins share a large proportion of their bound intervals and target genes. In addition to showing extensive common binding patterns, it appears that group B1 proteins expressed at earlier developmental timepoints can pre-bind target genes of later Sox proteins, priming them for later regulation by establishing bivalent chromatin marks without actually activating transcription (Bergsland *et al.*, 2011). In the case of *Drosophila*, Dichaete and SoxN share large numbers of targets with both Sox2 and Sox11, demonstrating that they can play roles carried out by both group B and group C proteins in mammals and that their function cannot be easily split between the roles of maintaining neural precursors and promoting neural differentiation. Dichaete in particular shares a high number of orthologous targets with mouse Sox2, which is consistent with the functional rescue of *Dichaete* mutant phenotypes achieved by expressing Sox2 protein (Sánchez-Soriano and Russell, 1998). These shared targets are highly associated with transcriptional regulation and the generation of neurons, including genes involved in the neuroblast regulatory network, Notch signalling and neu-

roblast cell fate (Aleksic *et al.*, 2013). Slightly fewer Sox2 targets are shared with core SoxN target genes; however, these genes are also strongly associated with CNS development. Interestingly, a much higher overlap in targets is observed between SoxN and Sox11, suggesting that SoxN in particular has a conserved role in neuronal differentiation and that some of its functions may have been co-opted by group C *Sox* genes in mammals (Ferrero *et al.*, 2014).

As with *Sox1*, *Sox2* and *Sox3* in vertebrates, both *Dichaete* and *SoxN* are expressed in overlapping patterns in the *Drosophila* CNS and are necessary for its normal development, although they do not show sequential expression as do *Sox2* and *Sox3* (Bergsland *et al.*, 2011; Buescher *et al.*, 2002; Crémazy *et al.*, 2000; Girard *et al.*, 2006; Sánchez-Soriano and Russell, 2000; Shen *et al.*, 2013). *Dichaete* mutant embryos show axonal and midline defects, which can be rescued by expressing Dichaete (or mammalian Sox2) in the midline (Sánchez-Soriano and Russell, 2000). *SoxN* mutant embryos also show axonal defects and loss of lateral neurons (Buescher *et al.*, 2002; Overton *et al.*, 2002). In *Drosophila*, neuroblasts delaminate from the neuroectoderm in three columns on either side of the midline: the medial, intermediate, and lateral columns. *Dichaete* and *SoxN* expression patterns partially overlap in these columns; *Dichaete* is expressed from the midline outwards to the intermediate column, while *SoxN* is excluded from the midline but is expressed from the medial column to the lateral column (Overton *et al.*, 2002) (Figure 1.2). *SoxN/Dichaete* double mutants have more severe CNS defects than either single mutant; in particular, they show an increased loss of neuroblasts in the medial column in comparison to single mutants, which is where *SoxN* and *Dichaete* expression overlaps most strongly (Figure 1.3) (Buescher *et al.*, 2002; Overton *et al.*, 2002).A similar effect is observed among mutants for the three vertebrate group B1 *Sox* genes, where mice lacking *Sox1* or *Sox3* show only mild brain and spinal cord phenotypes, and neuroectoderm development is normal in *Sox2* hypomorphs (Ferri, 2004; Guth and Wegner, 2008; Nishiguchi *et al.*, 1998; Rizzoti *et al.*, 2004; Wegner and Stolt, 2005). In zebrafish, in which six group B1 genes are present, severe embryonic and CNS defects are only present in quadruple *sox2/sox3/sox19a/sox19b* knockdowns (Okuda *et al.*, 2010). Such apparent redundancy is also observed with paralogous vertebrate *Sox* genes in other subgroups, including the group C genes *Sox4*, *Sox11* and *Sox12* and the group F genes *Sox17* and *Sox18* (Bhattaram *et al.*, 2010; Matsui, 2006).

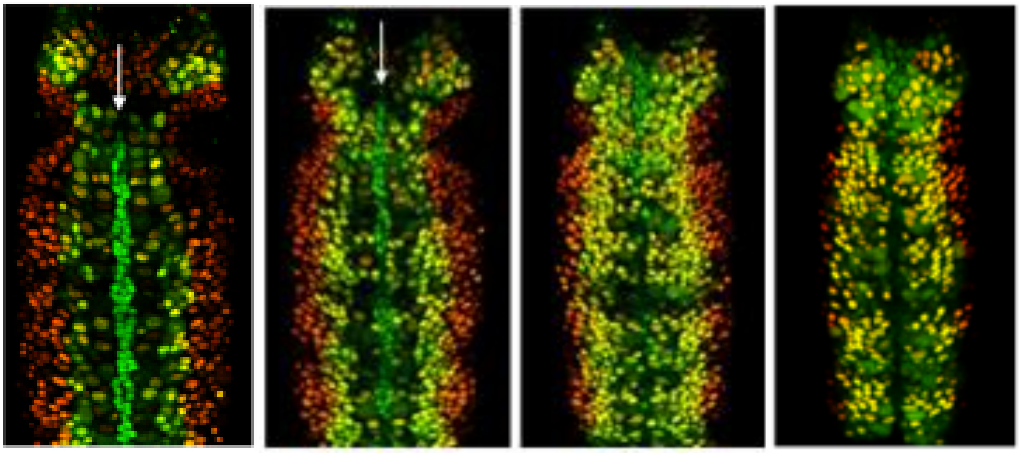


Figure 1.2: Dichaete and SoxN expression in the neuroectoderm of stage 10 *D. melanogaster* embryos. Several planes of focus are shown. Dichaete is expressed in the ventral midline (green cells, indicated by white arrows) as well as the medial and intermediate columns of neuroblasts. SoxN is expressed with Dichaete in the medial and intermediate columns (yellow) and alone in the lateral column of neuroblasts (red). Figure reproduced from Overton (2003).

These results strongly suggest functional compensation between Sox family members is widespread; however, the evolutionary driver for this phenomenon is not fully understood.

In addition to functional compensation at the level of neural phenotypes, *in vivo* binding and expression studies of Dichaete and SoxN in *D. melanogaster* show that they have highly similar genome-wide binding patterns and share a large number of gene targets (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). Commonly bound gene targets cover many of the core functionalities of both Dichaete and SoxN, including over a hundred other TFs active in the CNS, the proneural genes of the *achaete-scute* complex, *Dr* and *vnd*, which encode TFs involved in dorsoventral patterning in the CNS (Zhao *et al.*, 2007), and the neuroblast temporal identity genes *svp*, *hb*, *Kr* and *pdm2* (Ferrero *et al.*, 2014; Isshiki *et al.*, 2001; Maurange and Gould, 2005). Previous *in vivo* binding studies of Dichaete have provided evidence that it can bind to highly occupied target (HOT) regions, which are areas of the genome that are bound commonly by many TFs and are associated with open chromatin (Aleksic *et al.*, 2013; Kvon *et al.*, 2012). A role for Dichaete as a modulator of DNA architecture that supports the binding of

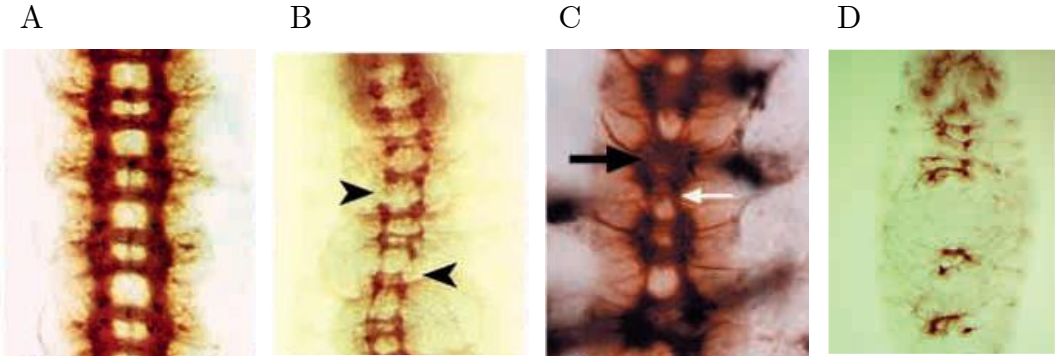


Figure 1.3: *Dichaete/SoxN* double mutants show a more severe CNS phenotype than either single mutant. Flat preparation of stage 16 *D. melanogaster* embryos stained for BP102 to show the axonal structure of the CNS. A.) Wild type embryo. B.) *SoxN*-mutant ($SoxNeuro^{U6\text{-}35}$) embryo. Arrowheads show lack of longitudinal staining in hemisegments. C.) *Dichaete*-mutant (D^{r72}/Df) embryo. The white arrow shows thinning of longitudinal connectives, and the black arrow shows fusion of commissural connectives. D.) $SoxNeuro^{U6\text{-}35}/D^{r72}$ double-mutant embryo. Longitudinal axons are almost completely absent and the neuropil shows frequent gaps. Figures reproduced from Overton *et al.* (2002) and Sánchez-Soriano and Russell (1998).

other TFs has also been proposed (Russell *et al.*, 1996). Together, these suggest that the binding patterns of group B Sox proteins, like many other developmental TFs that have been studied in the fly, may be strongly influenced by patterns of chromatin accessibility in addition to recognition of specific sequence motifs (Ferrero *et al.*, 2014; MacArthur *et al.*, 2009). However, it is unknown to what extent the chromatin environment drives Dichaete and SoxN binding or if all binding events in open chromatin are associated with gene regulation.

Further complicating the picture, not only do Dichaete and SoxN share many targets, they also display a complex pattern of compensatory binding in each others absence. DamID experiments examining SoxN binding in *Dichaete* mutants and vice versa have identified loci where one TF can compensate for the others absence by increasing its own binding. In addition, there are loci where the loss of one of these two Sox proteins appears to result in a loss of binding by the other (Figure 1.4). These observations suggest that Dichaete and SoxN can compensate for one another in some instances, but that they are also dependent on one another in order to function correctly in others. Furthermore, in some genomic locations the loss of one TF does not affect the binding of the other, indicating that their functions at certain loci are independent (Ferrero *et al.*, 2014). Consid-

ering the deep conservation of *Dichaete* and *SoxN* as paralogues throughout the insects (McKimmie *et al.*, 2005; Wilson and Dearden, 2008), it remains unclear why evolution has maintained these two partially redundant proteins.

The generation of new paralogues through gene duplications events has occurred frequently during metazoan evolution and is a major driver of increased complexity in genetic regulatory networks (Larroux *et al.*, 2008). The theoretical expectation after gene duplication has occurred is that the new paralogue experiences reduced selective pressure, as it is essentially a redundant copy of the original gene. This opens the door for the accumulation of mutations, which can lead to loss of function and transformation of one of the new paralogues into a pseudogene. Alternatively, if favorable mutations occur, then subfunctionalization, in which the role of the original gene is divided amongst the new paralogues either by functional domain or by spatial/temporal expression pattern, or neofunctionalization, in which the new copy acquires functions that did not belong to the original gene, can occur (Force *et al.*, 1999; Lynch, 2000). One well-studied example of subfunctionalization and neofunctionalization is the evolution of *Hox* genes, which code for a highly-conserved family of transcription factors that are primarily involved in establishing segmental identity along the anterior-posterior (AP) axis (Kappen and Ruddle, 1993).

Paralogous *Hox* genes have specific, though sometimes overlapping, expression domains along the AP axis and provide spatial information to downstream genes in order to direct the development of appropriate segmental morphology. In vertebrates, *Hox* genes have undergone tandem duplications followed by multiple whole-group duplications to result in in four *trans*-paralogous clusters, located on four different chromosomes (Foronda *et al.*, 2009; Maconochie *et al.*, 1996). Interestingly, *trans*-paralogous genes in the same relative positions (e.g. *Hoxa1* and *Hoxb1*) have retained greater similarities in sequence and expression patterns than *cis*-paralogous genes in each cluster (e.g. *Hoxa1* and *Hoxa2*). Although *Hox* single mutants typically do show specific phenotypes, there is some evidence for partial redundancy between *trans*-paralogues such as *Hoxa3* and *Hoxd3* (Greer *et al.*, 2000). In contrast, in both flies and vertebrates, *Hox* paralogues that arose through linear gene duplications have acquired largely unique expression domains and functions. Mutant phenotypes associated with each paralogous member of a single *Hox* cluster appear in specific domains along the AP axis that correspond

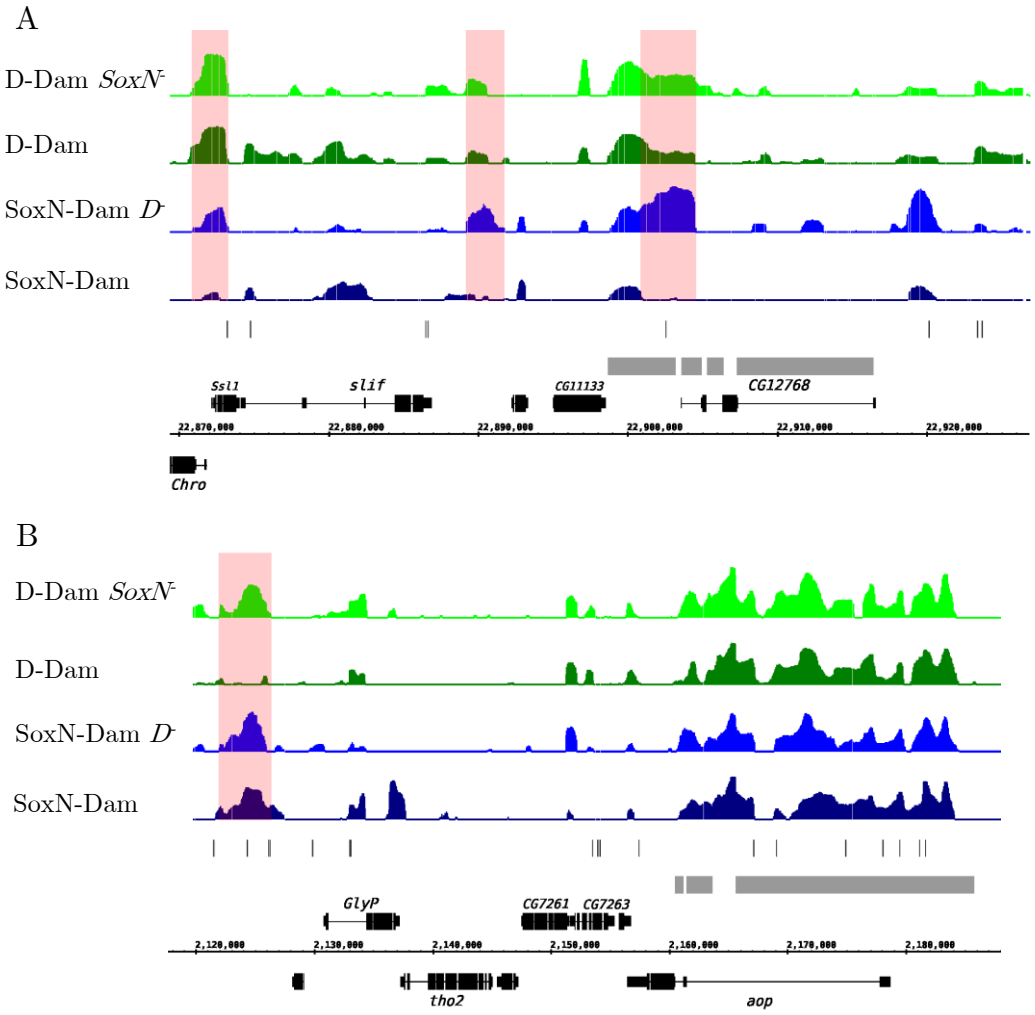


Figure 1.4: Reciprocal binding compensation by Dichaete and SoxN. Tracks show, from the bottom, gene models (black), known enhancers (gray), Sox motifs (gray), SoxN DamID in wild type (dark blue), SoxN DamID in *Dichaete* mutant background (blue), Dichaete DamID in wild type (dark green), Dichaete DamID in *SoxN* mutant background (light green). A.) Pink boxes show SoxN compensating for Dichaete binding. Dichaete is normally bound at these loci while SoxN is not; however, in a *Dichaete* mutant background, SoxN binds here (blue). B.) The pink box shows Dichaete compensating for SoxN binding. SoxN is normally bound at this locus while Dichaete is not; however, in a *SoxN* mutant background, Dichaete binds here (light green). Figures reproduced from Ferrero *et al.* (2014).

to the expression patterns of that member (Maconochie *et al.*, 1996). ChIP-chip experiments in *Drosophila* have confirmed that Hox proteins show a high level of *in vivo* specificity in their binding targets, although this specificity is likely to arise from a combination of specific DNA recognition sequences and the presence of unique combinations of cofactors (Hueber *et al.*, 2007; Hueber and Lohmann, 2008; Mann *et al.*, 2009).

Such specialization of paralogous genes after duplication has been suggested to drive the evolution of new gene regulatory modules, which can, in turn, facilitate adaptability and evolutionary innovation (Espinosa-Soto and Wagner, 2010). However, cases of genetic redundancy appear be conserved as a stable evolutionary state more often than theoretically predicted and in many different taxa (Lynch *et al.*, 2001; Vavouri *et al.*, 2008). Redundancy between a pair of forkhead transcription factors, *pes-1* and *fkh-2*, has been shown to be conserved between two species of nematode, *C. elegans* and *C. briggsae* (Molin *et al.*, 2000). In yeast, persistent functional redundancy among pairs of duplicated genes, measured in terms of overall fitness, appears to be widespread (Dean *et al.*, 2008). In contrast to the phenotypic and regulatory target specificity seen for *cis*-paralogous *Hox* genes in insect and vertebrates, functional redundancy in *Sox* genes from the same subgroup seems to be a common theme across evolution, with paralogues in multiple subgroups and in many different taxa showing overlapping patterns of expression and a lack of strong single-mutant phenotypes (Bhattaram *et al.*, 2010; Buescher *et al.*, 2002; Ferri, 2004; Guth and Wegner, 2008; Matsui, 2006; Nishiguchi *et al.*, 1998; Okuda *et al.*, 2010; Overton *et al.*, 2002; Rizzoti *et al.*, 2004; Uchikawa *et al.*, 2011; Uwanogho *et al.*, 1995; Wegner and Stolt, 2005; Wood and Episkopou, 1999).

One possible explanation for the compensation shown by *Dichaete* and *SoxN* is to provide greater regulatory robustness to the developing CNS; it has been argued that functional redundancy may be a general mechanism for promoting robustness in genetic regulatory networks (Nowak *et al.*, 1997; Tautz, 1992; Wagner, 2005, 2008). If regulation of the developing neuroectoderm represents the ancestral group B *Sox* function, then the unique, and sometimes opposing, roles of *Dichaete* and *SoxN* may be examples of partial neofunctionalization in the insects (Ferrero *et al.*, 2014). Both genes have independent functions; for example, *Dichaete* is expressed in unique domains, including the embryonic brain

and hindgut, where it has important regulatory functions (Sánchez-Soriano and Russell, 2000). Similarly, *SoxN* is prominently expressed in the ectoderm of the late embryo, where it has roles in cuticle patterning that are only partially compensated for by *Dichaete* (Overton *et al.*, 2007). If both the unique and common functions of the two proteins are conserved by natural selection, one would expect to find evidence of similar functionality and binding patterns throughout the insect phylogeny. In order address this question, I set out to examine the genome-wide *in vivo* binding patterns of both Dichaete and SoxN in four species of *Drosophila*. My goal was both to understand the evolutionary dynamics of group B Sox binding, including the rates of gain and loss of binding sites, as well as to test whether Dichaete and SoxN binding at common gene targets and specific binding at unique targets are equally conserved. In order to do so, I used a strategy of comparative binding analysis, drawn from several previous evolutionary studies of transcription factor binding in both *Drosophila* and vertebrates.

1.3 Comparative studies of transcription factor binding

The importance of regulatory DNA in development, disease and evolution is widely accepted and becoming a key focus for genomics as large-scale studies such as the ENCODE project attempt to map diverse elements of the non-coding genome (Dunham *et al.*, 2012; Gordon and Ruvinsky, 2012; Neph *et al.*, 2012; Wray, 2007). One of the major roles of regulatory DNA is to bind transcription factors and, together with other genomic elements such as promoters, to direct gene expression in a temporally and spatially specific manner. In the model organism *Drosophila melanogaster*, significant strides have been made towards understanding how multiple inputs are integrated to determine transcription factor occupancy in the nucleus, and how, in turn, combinatorial rules of transcription factor binding describe functional regulatory elements (Kaplan *et al.*, 2011; Li *et al.*, 2011; Zinzen *et al.*, 2009). However, the primary methods for determining transcription factor binding, both *in vivo* and *in silico*, suffer from difficulties in distinguishing between true functional events and biological noise, resulting in high numbers of potential false positives and making it difficult to tease apart un-

derlying regulatory networks (Biggin, 2011; Fisher *et al.*, 2012; MacArthur *et al.*, 2009). One potential way to circumvent this problem is via comparative studies of transcription factor binding in multiple *Drosophila* species, which facilitate the use of patterns of conservation to identify functional features of the regulatory genome as well as an analysis of the evolutionary dynamics of transcriptional regulation.

A number of different techniques for directly or indirectly studying genome-wide transcription factor binding patterns in *Drosophila* are available. Two of the primary *in vivo* techniques are ChIP (chromatin immunoprecipitation) and DamID, the latter based on DNA methylation by a tethered DNA adenine methyltransferase (dam) (Greil *et al.*, 2006) (Figure 1.5). Both of these techniques can be combined with either hybridization to a microarray or high-throughput sequencing in order to identify preferentially-bound regions genome-wide (Aleksic and Russell, 2009; van Steensel *et al.*, 2001); however, because arrays are generally not commercially available for non-model species and the cost of sequencing has dropped significantly in the last decade, sequencing has become the method of choice for most comparative studies. With the publication of the modENCODE data in 2010 (The modENCODE Consortium *et al.*, 2010), a large number of ChIP-chip and ChIP-seq datasets from *Drosophila melanogaster* were made publicly available; at the time of writing, the modMine database, which houses the modENCODE datasets, contains 279 entries for ChIP-chip and ChIP-seq datasets for transcription factor binding as well as chromosomal proteins and histone modifications in *D. melanogaster* (Contrino *et al.*, 2011). In addition, a more focused study on the binding of 31 transcription factors involved in early embryonic patterning, along with matching chromatin accessibility data, are available from the Berkeley Drosophila Transcriptional Network Project (MacArthur *et al.*, 2009). The availability of these datasets, as well as data-processing tools, quality control guidelines and experimental best practices from the modENCODE consortium (Landt *et al.*, 2012; Trinh *et al.*, 2013), provides a valuable resource for researchers wishing to undertake comparative studies in other *Drosophila* species. ChIP-seq experiments have been successfully performed with transcription factors in *D. simulans, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura* and *D. virilis* (Bradley *et al.*, 2010; He *et al.*, 2011b; Paris *et al.*, 2013; Villar *et al.*, 2014), representing an evolutionary span of approximately 40 million years.

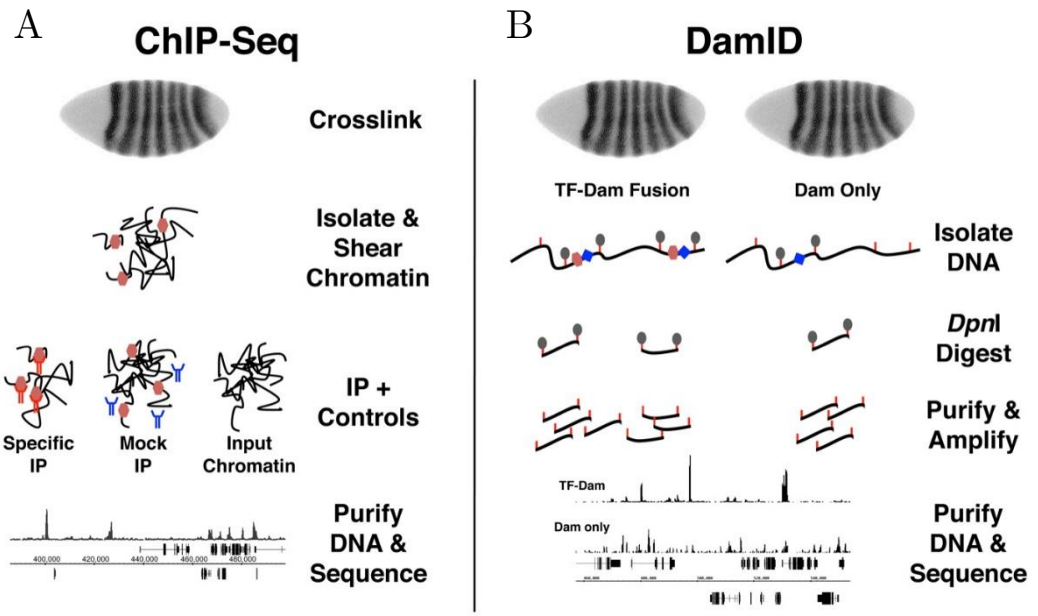


Figure 1.5: Overview of ChIP-seq and DamID-seq pipelines. A.) In ChIP-seq, embryos are first crosslinked with formaldehyde, and then chromatin is isolated and fragmented. An antibody specific to the TF of interest is used to enrich the sample for bound DNA fragments; mock IP and input chromatin controls are prepared in parallel. After reversal of crosslinks, the DNA is purified and sequenced, and reads are mapped to the reference genome. TFs are represented by orange hexagons, specific antibodies in red and control antibodies for mock IPs in blue. B.) In DamID-seq, two transgenic lines are created: one expressing a TF-Dam fusion and one expressing a Dam-only control. DNA is isolated from embryos of each line and digested with *Dpn*I, which cuts GATC sequences when the A is methylated. The resulting DNA is purified and sequenced, and rads are mapped to the reference genome. The top trace represents that Dam-TF profile and the bottom trace the Dam-only control profile. TFs are represented by orange hexagons, the Dam enzyme by a blue diamond, GATC motifs by red lines, and methylated adenine residues by gray ovals. Figure reproduced from Carl and Russell (in press).

One of the most fundamental questions that comparative transcription factor binding studies can ask is whether, and to what extent, individual binding events are conserved between different species. Several studies, focusing on different transcription factors and using different sets of species, have independently attempted to estimate binding conservation as well as the rate of binding site turnover in *Drosophila*. One of the first of these used ChIP-chip to measure genome-wide binding of the transcription factor Zeste. ChIP-chip was performed only in *D. melanogaster*, and the resulting binding intervals were aligned against the genomes of *D. simulans, D. erecta* and *D. yakuba* (Moses *et al.*, 2006). Since *in vivo* binding data was only available for one species, an analysis of quantitative differences in binding between species was not possible; instead, the authors considered binding as a binary state based on called peaks. Using a conservative approach, only binding intervals identified in *D. melanogaster* that could be unambiguously aligned to orthologous sequences in each of the other species were included, and the analysis was further restricted to those intervals containing matches to a Zeste binding motif positional weight matrix (PWM). Nonetheless, the authors found that at least 5% of Zeste binding sites identified in *D. melanogaster* were not conserved in the other species they examined, implying that those sites were either gained in the *D. melanogaster* lineage or lost in the other lineages since the divergence of the *melanogaster* sub-group (Moses *et al.*, 2006).

Several more recent studies employing ChIP-seq to measure transcription factor binding in multiple species of *Drosophila* generated broadly similar estimates of binding site conservation. Bradley and colleagues examined binding of 6 transcription factors involved in anterior-posterior (AP) patterning in the early embryo (Bicoid (Bcd), Hunchback (Hb), Kruppel (Kr), Giant (Gt), Knirps (Kni) and Caudal (Cad)) in the closely-related species *D. melanogaster* and *D. yakuba* (Bradley *et al.*, 2010). A subsequent experiment by the same group expanded the phylogenetic distance by measuring the binding of four of these factors (Bcd, Gt, Hb and Kr) in the same two species along with *D. pseudoobscura* and *D. virilis* (Paris *et al.*, 2013). A third study focused on the mesodermal regulator Twist in six species: *D. melanogaster, D. simulans, D. yakuba, D. erecta, D. ananassae* and *D. pseudoobscura*, which span approximately 25 million years of evolutionary time (He *et al.*, 2011b). Each of these studies considered both presence/absence of

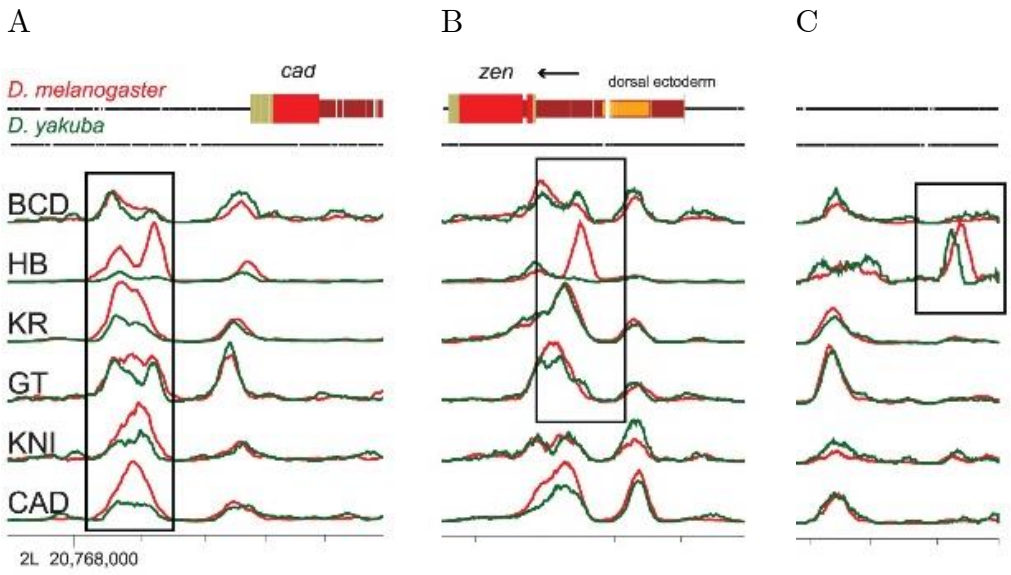


Figure 1.6: Examples of quantitative and qualitative changes in binding between five AP factors in *D. melanogaster* (red) and *D. yakuba* (green). A.) Examples of quantitative changes in binding strength between species with peak location conserved for several factors. B.) An example of the complete gain and loss of a peak between species for the factor Hb. C.) An example of a shift in binding site location between species with peak strength conserved for the factor Hb. Figure reproduced with modifications from Bradley *et al.* (2010).

peaks in each species as well as quantitative changes in binding strength (Figure 1.6).

Bradley *et al.* found that, for each of the 6 factors studied, between 1% and 15% of peaks that were identified in one species were absent in the other. They measured quantitative binding divergence by calculating the genome-wide correlations between binding strength at all peaks for each factor in *D. melanogaster* and *D. yakuba*; these values ranged from 0.57 0.75 for peaks at genes not known to be regulated by the AP patterning factors and were higher at known target genes (Bradley *et al.*, 2010). In similar pairwise comparisons between binding strengths of peaks in *D. melanogaster* and *D. pseudoobscura*, the correlations ranged from 0.37 for Gt to 0.64 for Kr, reflecting the greater phylogenetic distance between the two species (Paris *et al.*, 2013). In the case of Twist, around 80% of peaks identified in *D. melanogaster* were found to be conserved in *D. simulans* and *D. yakuba*, with the percentage decreasing to around 60% for *D.*

pseudoobscura. The authors measured quantitative divergence by computing the number of peaks whose binding strength changed between *D. melanogaster* and each other species; this ranged from around 10% to 35% of total peaks (He *et al.*, 2011b). One common finding among these studies, as well as two others that focused on the insulator proteins CTCF and BEAF-32 (Ni *et al.*, 2012; Yang *et al.*, 2012), is that differences in binding between species, measured either qualitatively or quantitatively, increase with the phylogenetic distance of the species being compared, prompting the hypothesis that binding divergence may follow a molecular clock mechanism (He *et al.*, 2011b).

Besides simply estimating rates of binding conservation and divergence, comparative studies of transcription factor binding can identify new features of transcription factor function by considering differences in binding conservation relative to genomic annotations or patterns of binding by other factors. This type of analysis builds on the hypothesis that functional sites will be subject to purifying selection and thus will be preferentially conserved. One way to test this hypothesis is to evaluate conservation at a set of well-characterized functional regulatory elements. For example, peaks for AP patterning regulators are more conserved at known AP target genes compared to all genes, and peaks for Twist binding are highly conserved at regulatory elements that are known Twist targets (Bradley *et al.*, 2010; He *et al.*, 2011b; Paris *et al.*, 2013). Additionally, the most highly conserved Twist peaks show an enrichment near genes that are down-regulated in *twist* mutants as well as genes that are annotated with Gene Ontology (GO) functions related to Twist's developmental role, both of which are also indicators of function. Clustered Twist sites assigned to the same gene are significantly more likely to be conserved than singleton sites assigned uniquely to a gene. This effect was observed up to an inter-peak distance of 5 kb, leading the authors to suggest that Twist binding to shadow enhancers might also have an effect on ensuring robustness of gene expression patterns (He *et al.*, 2011b). In the case of AP transcription factors, Paris *et al.* found that peaks in regions that were commonly bound by more than one factor were better conserved than those where only one factor bound, suggestive of a role for combinatorial binding between AP factors (Paris *et al.*, 2013).

It is also possible to examine the effect of sequence level conservation on transcription factor binding. Both the two AP factor studies and the Twist study

described above show that, while overall sequence conservation in bound regions does not correlate strongly with binding divergence, conservation of short sequence motifs within binding intervals does show some correlation with binding divergence (Bradley *et al.*, 2010; He *et al.*, 2011b; Paris *et al.*, 2013). He *et al.* found that Twist peaks present in all four species studied had significantly more fully-conserved Twist motifs than peaks that were only present in *D. melanogaster*. Similarly, the quality of Twist motifs present in peaks was also correlated with quantitative changes in binding strength between species. However, changes in motif quality alone do not explain all of the observed binding divergence in any of the cases studied, suggesting that other factors are at play in shaping binding patterns. After observing that not all losses of Twist binding could be attributed to a corresponding loss of a Twist motif, the authors decided to investigate whether other factors acting as binding partners for Twist had an effect on the conservation of its binding. A search for motifs that were significantly more conserved in highly-conserved Twist peaks compared to divergent Twist peaks or the background genome yielded two transcription factors known to act together with Twist: Snail and Dorsal. For Twist peaks in one species containing a Snail or Dorsal motif in addition to a conserved Twist motif, loss of the partner motif was sufficient to explain loss of Twist binding in another species in 19% of cases. Furthermore, the top ten motifs identified in Twist binding intervals explained 49% of losses of Twist binding despite conservation of a Twist motif. These findings go one step beyond a simple search for enriched motifs, identifying those that have a functional effect on binding patterns. Integration of an evolutionary analysis of gains and losses of Twist binding with a search for conserved co-occurring motifs led to both the validation of known Twist co-regulators such as Dorsal and Snail as well as the identification of new factors that could potentially bind to enhancers with Twist in a combinatorial manner to direct specific patterns of gene expression during development (He *et al.*, 2011b).

By studying 6 different transcription factors, Bradley *et al.* were in a unique position to examine the relationships between quantitative binding divergence for different factors across the genome. By performing principal component analysis (PCA) on regions bound by any factor, they found both a strong correlation between quantitative changes in binding strength across all factors (explaining 38% of all binding divergence between *D. melanogaster* and *D. yakuba*) as well as

both positive and negative correlations between changes in the binding of specific pairs of factors. For example, increases in binding of Giant, a repressor, were correlated with decreases in binding of Hunchback, an activator. A search for sequence motifs that were associated with the correlated binding divergence of all the AP factors revealed a CAGGTAG binding motif for the zygotic transcriptional activator Zelda (Bradley *et al.*, 2010). This strong association between AP factors and Zelda was later confirmed and extended into the more distant species *D. pseudoobscura* and *D. virilis* (Paris *et al.*, 2013). Zelda has since been shown to be a key factor in establishing regulatory regions in the early embryo that will be active later in development, and it has been suggested that it plays an important role in shaping the chromatin landscape during zygotic genome activation (Harrison *et al.*, 2011; Satija and Bradley, 2012). This example highlights a case where patterns of binding conservation for one set of transcription factors illuminated a new functional role for a different protein as well as a general feature of *Drosophila* embryonic development.

In contrast to *Drosophila*, comparative studies of transcription factor binding in vertebrate species show that binding patterns appear to have diverged much more over equivalent phylogenetic distances. The majority of binding sites of tissue-specific TFs in human, mouse, dog, opossum and chicken are species-specific, despite the highly-conserved DNA binding preferences of the orthologous proteins (Odom *et al.*, 2007; Schmidt *et al.*, 2010). Even among closely-related mouse and rat species, TF binding patterns show less similarities than among *Drosophila* species separated by similar periods of evolutionary time (Stefflova *et al.*, 2013). Potential explanations for these discrepancies include the vast differences in genome size and density of functional elements between vertebrates and *Drosophila* and the larger effective population size of insects in comparison to vertebrates, which tends to make natural selection more effective (Villar *et al.*, 2014). Remarkably, mice carrying a copy of human chromosome 21 show TF binding patterns on that chromosome that recapitulate those seen in humans, rather than on the orthologous mouse chromosome 16, demonstrating that species-specific differences largely stem from the *cis*-regulatory code itself, rather than other factors in the nuclear environment (Wilson *et al.*, 2008).

The degree of conservation of binding events in *Drosophila* makes it a particularly suitable model system in which to study the evolution of regulatory DNA

and to deduce information about TF function from evolutionary comparisons. In addition, the amenability of *Drosophila* to molecular techniques and genetic manipulation, as well as the publication of the sequenced genomes and phylogenetic relationships of twelve *Drosophila* species (Clark *et al.*, 2007) and the ongoing community efforts to sequence more species make the fruit fly a compelling model in which to conduct comparative studies of transcription factor binding. With this in mind, I chose to study the binding patterns of the two group B Sox proteins Dichaete and SoxN in four species of *Drosophila*: *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. These four species span divergence times from approximately two million years to 25 million years, allowing for a range of evolutionary comparisons, yet their genomes are close enough for accurate alignment, which is critical for a comparative binding analysis (Russo *et al.*, 1995) (Figure 1.7). I aimed to use such an analysis to shed new light on the functional and evolutionary dynamics of group B Sox binding in *Drosophila*.

1.4 Overview of experiments

The main questions that I set out to answer during my Ph.D. can be summarized as follows:

1. Where do Dichaete and SoxN bind in the genomes of *D. simulans, D. yakuba* and *D. pseudoobscura*, and what proportion of those binding sites are conserved with *D. melanogaster*?

2. Are there certain categories of binding sites that are more highly conserved across the drosophilids than others, and what can this tell us about Dichaete and SoxN function in invertebrates? Specifically, are sites that are commonly bound by both TFs equally conserved as those that are only bound by one?

3. To what extent do patterns of chromatin accessibility differ between *D. melanogaster* and *D. pseudoobscura*, and what is the relationship between open chromatin and group B Sox binding?

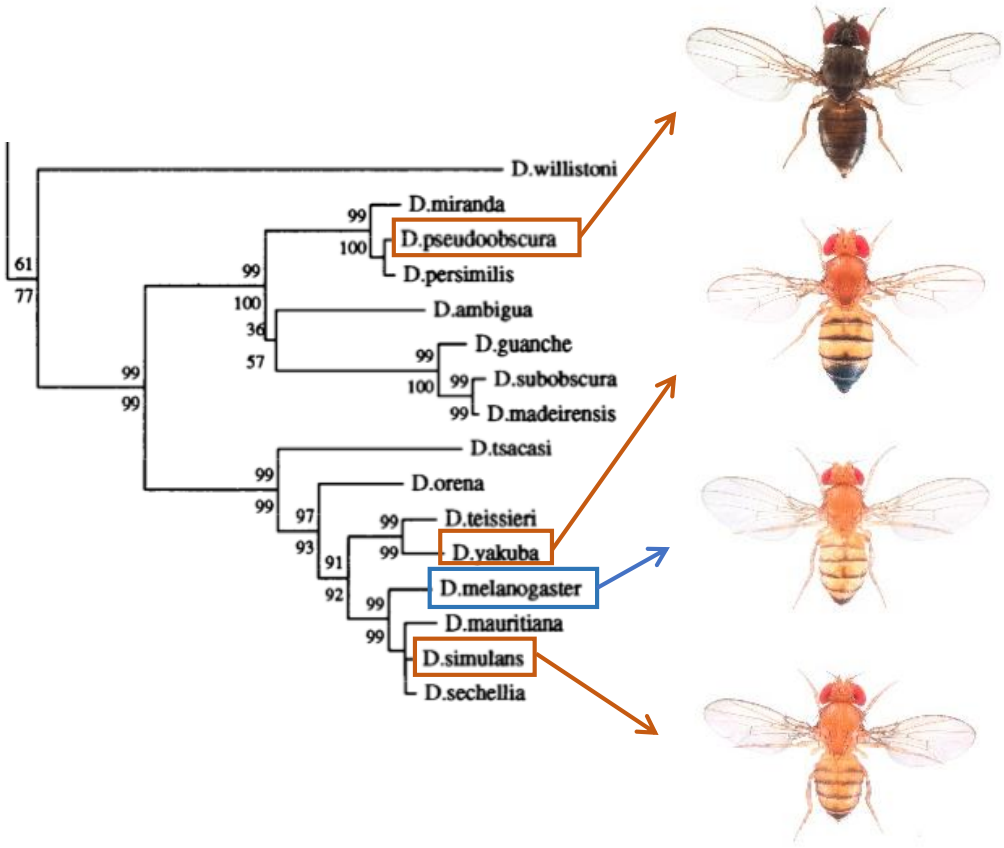


Figure 1.7: Phylogenetic relationship of *Drosophila* species used in this thesis. The reference species, *D. melanogaster*, is highlighted in the blue box. All non-model species are highlighted in red boxes. *D. melanogaster, D. simulans* and *D. yakuba* are located in the *melanogaster* subgroup, while *D. pseudoobscura* falls into the *obscura* subgroup. The phylogenetic tree is a neighbor-joining tree based on *Adh* nucleotide sequences from each species and is reproduced from Russo *et al.* (1995). The confidence probability is shown above each branch, and the bootstrap confidence level from 1000 replications is shown below each branch. *Drosophila* images are by Nicolas Gompel (`http://www.ibdml.univ-mrs.fr/equipes/BP_NG/Illustrations/melanogaster%20subgroup.html`).

In order to address the first question, I initially set out to perform ChIP-seq for Dichaete and SoxN in all four species of interest. After verifying the similarities between Dichaete and SoxN expression patterns in each species via immunohistochemistry, I performed ChIP-PCR in each species and ChIP-chip in *D. melanogaster* to test the performance of the antibodies against the two TFs in immunoprecipitations. Although the initial results were promising, two attempts at ChIP-seq for Dichaete failed to produce biological replicates with any significant, reproducible enrichment. The data from these preliminary experiments are presented in Chapter 3. After deciding that the ChIP-seq data was too noisy for any useful further analysis, I changed my experimental strategy and focused on performing DamID-seq for both Dichaete and SoxN in all four species. My first task was to create transgenic lines carrying Dichaete-Dam, SoxN-Dam and Dam-only constructs in each species; the details of this work are described in the methods section (Chapter 2). I then successfully carried out DamID-seq for Dichaete in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*, and for SoxN in *D. melanogaster* and *D. simulans*. In *D. pseudoobscura*, I was unable to generate a SoxN-Dam line, while in *D. yakuba* the DamID experiment failed, possibly due to a mutation in the transgenic SoxN sequence. A presentation of the DamID-seq datasets and a functional analysis of the binding patterns of the two TFs in each species can be found in Chapter 4.

Next, I compared the binding patterns of Dichaete-Dam and SoxN-Dam on both qualitative and quantitative levels in pairwise comparisons, and, in the case of Dichaete, in a three-way comparison between species. This allowed me to identify binding intervals that are unique to one species or conserved between two, three or four species. The detailed analysis of group B Sox binding conservation is presented in Chapter 5. In this section, I also address the second major question of my thesis. I examined differences in the rate of binding conservation between binding intervals associated with certain functional categories, such as those overlapping known enhancers or previously-identified Dichaete and SoxN target genes and core intervals. I also integrated the *in vivo* binding data with the genome sequences available in all four species to search for Sox motifs within bound intervals and analyzed the relationship between the number, quality and sequence conservation of Sox motifs and binding conservation. Finally, I considered the rates of conservation of common binding by Dichaete and SoxN versus

unique binding by either TF. In order to do so, I first performed a quantitative differential analysis of Dichaete and SoxN binding in both *D. melanogaster* and *D. simulans*, resulting in the detection of intervals that are commonly bound or uniquely bound in either one or both species. This allowed me to identify a strong relationship between common binding by both TFs and binding conservation, supporting the prior evidence for common regulation of many targets, as well as to examine the functions of potential targets that are uniquely bound by each TF across multiple species.

In order to address the third question, the role of chromatin accessibility in directing group B Sox binding and its differences between species, I performed FAIRE-seq in *D. pseudoobscura* embryos collected at five developmental stages. A detailed description of the *D. pseudoobscura* staging process as well as the FAIRE-seq protocol can be found in Chapter 2. These datasets, as well as a functional analysis of the accessible regions that I identified, are presented in Chapter 6. I used publicly-available ChIP-seq datasets for several TFs in *D. pseudoobscura* to investigate the relationship between accessible chromatin identified by FAIRE and TF binding, as well as examining the correlation between FAIRE accessibility and Dichaete binding as identified by DamID in *D. pseudoobscura*. A comparison of my FAIRE datasets with several chromatin accessibility datasets in *D. melanogaster* embryos revealed that the *D. pseudoobscura* FAIRE data may suffer from a lack of sensitivity, which could be due to technical problems during the chromatin preparation stage. Nonetheless, I was able to use these data to find significant associations between conserved Dichaete binding and open chromatin, supporting a role for chromatin accessibility not only in determining TF binding patterns but also in maintaining them during evolution.

As reviewed here, the importance of regulatory DNA during evolution has been increasingly recognized and studied over the last decade. However, conservation or divergence of regulatory regions can occur on several levels, and it is important to consider all of them in order to build a comprehensive picture of the function and evolution of transcriptional regulation. The central dogma of molecular biology often describes DNA as a language that must be read in order to produce RNA and proteins (Gerstein *et al.*, 2007), and this linguistic metaphor has been extended to create more complex models of molecular grammar (Searls, 1997, 2001, 2002). Although regulatory DNA is not typically transcribed or translated

itself, it can also be considered to have a type of grammar. If we consider an enhancer as a sentence, the most fundamental level, that of DNA sequence, can be compared to orthography or spelling; changes in a single letter may render the sequence unintelligible. Clearly this can be conserved during evolution, as most classical tests for selection rely on nucleotide sequence. The next level, which consists of binding sites for specific TFs, may be represented by the lexicon or set of words in a language. The primary goal of techniques such as ChIP-seq and DamID is to determine which words are present in which sentences. Conservation can also be studied at this level, as each TF may or may not bind to orthologous enhancers in multiple species. Just as words have different meaning depending on their positions relative to one another, TF binding can have different functions depending on the presence of cofactors or clustered binding sites. This regulatory syntax is perhaps the least well understood in terms of evolution, although TF combinatorial binding has been addressed in several studies in *Drosophila* (He *et al.*, 2011b; Zinzen *et al.*, 2009). Finally, the regulatory output of an enhancer, measured either by changes in gene expression or network-wide perturbations, corresponds to the semantics of a sentence. Studies integrating RNA-seq data with ChIP-seq binding data in multiple species attempt to address conservation at this level (Paris *et al.*, 2013). Clearly all of these functional levels are related, yet they also have a certain amount of independence. In this thesis, I attempt to address the conservation of group B Sox binding sites on all four levels, by examining expression patterns, genome-wide binding, potential cofactors and sequence motifs. My goal is to create an integrated view of Dichaete and SoxN regulatory function in *Drosophila*.

CHAPTER 2

Materials and Methods

2.1 Fly husbandry and stock keeping

The wild-type strains of the following *Drosophila* species were used in all experiments: *D. melanogaster* Oregon-R and w^{1118}; *D. simulans* $w^{[501]}$ (reference strain - `http://www.ncbi.nlm.nih.gov/genome/200?genome_assembly_id=28534`); *D. yakuba* Cam-115 (Coyne *et al.*, 2004); *D. pseudoobscura pseudoobscura* (reference strain - `http://www.ncbi.nlm.nih.gov/genome/219?genome_assembly_id=28567`). *D. melanogaster, D. simulans* and *D. yakuba* flies were kept at 25° C on standard cornmeal medium. *D. pseudoobscura* flies were kept at 22.5° C in low humidity, on banana-opuntia-malt medium (1000 ml water, 30 g yeast, 10 g agar, 20 ml Nipagin, 150 g mashed banana, 50 g molasses, 30 g malt, 2.5 g opuntia powder). All embryo collections were performed at 25° C with the exception of the *D. pseudoobscura* staged collections for FAIRE-seq, which were performed at 22.5° C. Flies were allowed to lay for varying periods of time on agar plates supplemented with grape juice and streaked with fresh yeast paste.

All microinjections to generate transgenic lines were performed by Sang Chan in the Department of Genetics injection facility. Before injections, flies were kept in cages for 2 days at 25° C, with a fresh grape juice-agar plate with yeast paste provided twice a day. After 2 days, the plates were changed every 30 minutes for 2 hours, and then embryos were collected after a 30-minute lay. In an 18° C

injection room, embryos were washed and dechorionated in 50% bleach for 3 minutes. They were then rinsed with cold water, blotted dry on a paper towel and transferred with a paintbrush to a coverslip on which a stripe of heptane-glue had been painted (made by dissolving sellotape in heptane). The embryos were aligned on the heptane-glue with forceps and covered with 10 S Voltalef oil (VWR). The posterior end of each embryo was injected using a glass needle loaded on a Leitz micromanipulator. The injection mix consisted of a piggyBac helper plasmid at 0.4 μg/μl and a piggyBac plasmid containing the construct of interest at 0.6 μg/μl.

Injected embryos were transferred on the coverslip to a grape juice-agar plate with a small dot of yeast paste and left to develop for 24 hours at 25° C. *D. pseudoobscura* embryos were allowed to develop for up to 48 hours to account for slower developmental times. Any hatched larvae were then transferred with the yeast paste into a fresh tube containing cornmeal medium. For *D. simulans, D. yakuba* and *D. pseudoobscura*, surviving adults were backcrossed to males or virgin females from the parental, wild-type strain. F1 progeny were then scored for eye-specific GFP expression, and transgenic lines were set up by crossing GFP-positive siblings. Because I was unable to identify flies carrying two copies of the transgene, these lines consisted of a mixed population of homozygous and heterozygous flies, meaning that the populations had to be periodically checked and GFP-positive flies selected in order to prevent loss of the transgene through genetic drift. For *D. melanogaster*, surviving adults were backcrossed to *w; Sco/SM6a* males or virgin females. F1 males were scored for eye-specific GFP expression and crossed singly to *w; Sco/SM6a* virgins, then the same males were crossed to *w; TM2/TM6c* virgins. F2 progeny of the *Sco/SM6a* cross were scored for eye-specific GFP expression and a curly wing phenotype, while F2 progeny of the *TM2/TM6c* cross were scored for eye-specific GFP expression and a Stubble phenotype. Siblings of each class were mated together. Balanced transgenic lines were identified in the F3 generation as stocks where all flies showed eye-specific GFP expression.

2.2 Immunohistochemistry

Embryos were collected from each species after an overnight lay following the protocol described above. They were then dechorionated in 50% bleach for 3 minutes, rinsed in cold water, and fixed by shaking for 20 minutes in 1.8 ml fixation solution (0.1 M PIPES, 1 mM MgSO4, 2 mM EGTA, pH 6.9) with 0.5 ml formaldehyde and 4 ml heptane. The aqueous phase was removed and 6 ml of methanol was added, followed by vortexing for 30 seconds. Any embryos that sank to the bottom of the tube were collected, rinsed with methanol, and stored at -20° C until needed for staining. Staining was performed as described (Patel, 1994) with primary antibodies at the following concentrations: rabbit anti-Dichaete, 1:100; rabbit anti-SoxN, 1:100 or 1:50. Primary antibodies were detected with biotin-conjugated secondary antibodies (goat anti-rabbit) at 1:200 using the ABC Elite kit (Vectastain). Stained embryos were mounted in 70% glycerol and photographed using Openlab v.4.0.2 imaging software on a Zeiss Axioplan microscope with a 20x objective.

2.3 Chromatin immunoprecipitation

For chromatin immunoprecipitations (ChIP), embryos were collected after an overnight or 12-hour lay and dechorionated as described above. They were fixed by shaking for 20 minutes in 670 μl crosslinking solution (50 mM HEPES, 1mM EDTA, 0.5 mM EGTA, 100 mM NaCl, pH 8.0) with 33 μl 37% formaldehyde and 3 ml heptane added. The crosslinking reaction was stopped by centrifuging for 2 minutes at 1000g to pellet the embryos, removing the supernatant and adding 2 ml PBT with 125 mM glycine. Embryos were then weighed in an Eppendorf tube, flash-frozen in liquid nitrogen and stored at -80° C. Approximately 200 mg of embryos were used per biological replicate. ChIPs were performed as described with some modifications for a small amount of starting material (Ghavi-Helm and Furlong, 2012; Sandmann *et al.*, 2007). Embryos were homogenized in Eppendorf tubes using a plastic pestle rather than in a Dounce homogenizer. Each sample was homogenized for 30 seconds in 1 ml cold PBT supplemented with protease inhibitors (Complete Mini Protease Inhibitor cocktail tablets, Roche),

then allowed to rest on ice for 30 seconds, then homogenized again for 30 seconds. The lysate was spun at 400g for 1 minute at 4° C, and the supernatant decanted into a fresh Eppendorf tube. After centrifugation at 1100g for 10 minutes at 4° C, the supernatant was discarded and the pellet resuspended in 1 ml cold cell lysis buffer supplemented with protease inhibitors. The sample was homogenized again for 30 seconds with a plastic pestle and the lysate spun at 2000g for 4 minutes at 4° C to pellet the nuclei. The pellet was resuspended in 1 ml cold nuclear lysis buffer and incubated for 20 minutes at room temperature to lyse the nuclei.

A Diagenode Bioruptor was used for sonication, with the energy settings on high. Chromatin was sonicated in 100 μl aliquots for 16 cycles of 30 seconds on, 30 seconds off. A 50 μl input aliquot was removed from each sample and treated with RNaseA for 30 minutes at 37° C, then with proteinase K overnight at 37° C. Crosslinks were reversed by incubating at 65° C for 6 hours. The distribution of DNA fragment sizes was assessed by performing a phenol-chloroform extraction and running the resulting DNA on a 3% agarose gel. Fragment sizes ranged from approximately 100 bp to 1000 bp, with the majority of the fragments falling between 300 and 700 bp. Immunoprecipitation was carried out with protein A-agarose beads (Millipore), using the buffers and wash protocol described in Sandmann *et al.* (2006). Anti-Dichaete antibody was pre-cleared by incubating for 3 hours at 4° C with methanol-fixed embryos, then added to a final concentration of 1:300. Affinity-purified anti-SoxN antibody was added without pre-clearing to a final concentration of 1:100. For mock IP controls, a rabbit anti-beta galactosidase antibody (AbCam ab616) was added to a final concentration of 1:1000. After performing the immunoprecipitation, crosslinks were reversed and the DNA purified using the same protocol described above for input samples.

2.3.1 ChIP-PCR

Targets were chosen for PCR amplification to test the specific enrichment of each ChIP by examining previous ChIP-chip and DamID experiments carried out in our lab (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). For each gene, a highly bound interval was identified in *D. melanogaster* and its sequence was used as a query

to search the genome of each other species using BlastN (Altschul *et al.*, 1990), with the goal of identifying an orthologous region of 500-800 bp. A negative control region was also identified for each factor in each species where binding was not observed in previous experiments in *D. melanogaster*. Primers were designed to amplify each region using Primer3 Plus (Untergasser *et al.*, 2007). Oligonucleotide sequences are shown in Table 2.1. PCR conditions were identical for each set of samples and were as follows: 95° C for 2 min.; 45 cycles of 95° C for 30 sec., 58° C for 30 sec., 72° C for 30 sec.; 72° C for 5 min. 1 μl of ChIP, mock IP, or input DNA was used as a template for each reaction. PCR products were run out on a 1% agarose gel, and the specificity of each antibody was assayed by comparing the presence and brightness of bands for the ChIP samples versus the mock IP and input samples.

TF	**Target gene**	**Species**	**Forward primer (5'-3')**	**Reverse primer (5'-3')**
Dichaete	*slit*	*D. melanogaster*	GATGCGAACC CAACTGAACT	AAACTCAAAC GTGCCGTAGA
	achaete	*D. melanogaster*	TGATGTCTGG ACCTTGTTGC	CCATTAAAGG CCGAAGATGA
	comm	*D. melanogaster*	AGAACCGGTT TTCGAGTGG	ATAAGCCTGA GCGCGAAGTT
	klingon (neg.)	*D. melanogaster*	ATCCGAATTC AAATCCACCA	GCAATCGAAA AAGTGGCAAT
	slit	*D. simulans*	GATGCGAACC CAACTGAACT	GCCACAGACA ATGCGACTTA
	achaete	*D. simulans*	TGATGTCTGG ACCTTGTTGC	TTAACGGCCG AAGATGATTC
	comm	*D. simulans*	GAACGCAAAA TCTCGACCAT	AGTGACATTC CATGGGGAGA
	klingon (neg.)	*D. simulans*	CAAAATCAGG AGCAGCACAA	GGATGTTGGA TTTGGATTCG
	slit	*D. yakuba*	AGTGACATTC CATGGGGAGA	ATACGTGCCA CAGACAATGC
	achaete	*D. yakuba*	ATACAAATTG CATGGCCACA	GAGACGATGG TCCTTGCTTC

	comm	*D. yakuba*	AGGGAAATGG GAAAATCCAC	AAAGTGGCCA AGAGCTGAAA
	klingon (neg.)	*D. yakuba*	CAAAATCAGG AGCAGCACAA	GAATGTTGCA TTTGCCTCCT
	slit	*D. pseudoobscura*	GCTGTGGACA CACACTCACC	GCGAGACCCG TAAAACAGTC
	achaete	*D. pseudoobscura*	CCACCCCTGA TTTATTGTGG	CAGCATCAAT GTGGCTCACT
	comm	*D. pseudoobscura*	CTCTCGGGCT GTACTCAAGG	TTCCGTTCCT TGTTTGTTCC
	klingon (neg.)	*D. pseudoobscura*	ATAGCCACGT AAGCCAATCG	GGGGGAGCAA AGTATTAGCC
SoxN	*nerfin-1*	*D. melanogaster*	GAGCCCATTG AAAAGCTCAG	GCTCGTCGTC ATAGCTCTCC
	gcm-2	*D. melanogaster*	GCCGTATGTG GAGGACAACT	GTGATGGTGA TGGTGGTACG
	castor	*D. melanogaster*	ACCTCTATCC GGGAATGACC	TTGGTTTTTG TGGAGGGAAG
	ppd6 (neg.)	*D. melanogaster*	AATTCGGTGG AAACGATCAC	ACCTCGATCA CTCGATGTCC
	nerfin-1	*D. simulans*	CTGAAAACCA GGTGCGAAAT	GAGTGGCTTT ATTGCGGAAG
	gcm-2	*D. simulans*	GCCGTATGTG GAGGACAACT	GGTGGTGATG GTGGTAGGTC
	castor	*D. simulans*	GCCACCCAAG AAAATCGTAA	GGTCATTCCC GGATAGAGGT
	ppd6 (neg.)	*D. simulans*	AACTCGGTGG AAACGATCAC	GGTAGCTAAC ACCCCGACA
	nerfin-1	*D. yakuba*	CTGAAAACCA GGTGCGAAAT	TGGTTTTAGG CGCTGTATCC
	gcm-2	*D. yakuba*	AACAGTACGG CGGAAATCAG	TGAGTAATCC TCCGGTGTCC
	castor	*D. yakuba*	CTCTTCCAGC TGCAAAATCC	TCAAAGTGTG GCTGAGTTGG
	ppd6 (neg.)	*D. yakuba*	AATTCGGTGG AAACGATCAC	ACCTCGATCA CTCGATGTCC

	nerfin-1	*D. pseudoobscura*	ACCGCAGTCG CTATCTGAAT	TCCTCCTCTT CGTCGATGTT
	gcm-2	*D. pseudoobscura*	TACGAGTCGA GTCCCCAGTT	GCGCTCTCGT AGAAGTGTCC
	castor	*D. pseudoobscura*	CCACCCCTCT CTCCTCTCTC	TGGTACAAGA GGGGGTTCTG
	ppd6 (neg.)	*D. pseudoobscura*	TGGAGGAGAG CAAGAGGAAA	AGTTGACCAA TGGCGGATAG

Table 2.1: Primers used to amplify target sequences of Dichaete and SoxN in each species for ChIP-PCR.

2.3.2 ChIP-chip

Dichaete ChIP samples from *D. melanogaster* were hybridized to a dual-color Nimblegen HD2 (2.1M probe) whole-genome tiling array in order to validate the specificity of the immunoprecipitation reactions. Probe libraries were constructed as described (Sandmann *et al.*, 2007). ChIP samples and their respective mock IP controls were labelled with either Cy3 or Cy5 dyes, with a dye swap in one of three biological replicates. Each ChIP sample and its matched control were hybridized to the same microarray. Hybridization was performed according to the manufacturers specifications. Spot-finding was carried out using NimbleScan, a proprietary software package developed by Roche. The raw data were quantile-normalized in R and analyzed with two different peak-calling algorithms, TiMAT (http://bdtnp.lbl.gov/TiMAT/) and Ringo (Toedling *et al.*, 2007) at false discovery rate (FDR) values of 1%, 5%, 10% and 25%.

2.3.3 ChIP-seq

ChIP reactions for ChIP-sequencing were performed as described above, with the exception that the protein A-agarose beads were changed to protein A/G PLUS-

agarose beads (Santa Cruz Biotechnology), as these do not contain salmon sperm DNA, a potential sequencing contaminant. Before library construction, sample concentrations were measured on a Qubit using the DNA High Sensitivity Assay (Life Technologies). Initially, libraries were constructed for sequencing in-house on an Ion Torrent PGM using the Ion Plus Fragment Library Kit (Life Technologies), quantified via qPCR, and templated using the Ion OneTouch Template Kit (Life Technologies). Libraries were sequenced on 316 chips; however, the quality and coverage of the resulting reads was insufficient for identifying binding peaks.

For the first attempt at Illumina sequencing, libraries were prepared using 10 ng of ChIP DNA or mock IP DNA, or the entire sample if less than 10 ng were available, with the NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina (NEB). Samples were barcoded using the NEBNext Multiplex Oligos for Illumina (Index Primers 1-12) (NEB). For the second attempt, libraries were constructed using 10 ng of ChIP or input DNA, or the entire sample if less than 10 ng were available, with the TruSeq DNA LT Sample Prep kit (Illumina). Samples were barcoded using the indexed adapters included in the kit, which were diluted 1:250 to account for the low amount of starting material. Size selection was performed using Agencourt AMPure XP beads (Beckman Coulter), with the aim of recovering fragments between 250 and 400 bp. In all cases, the library concentrations were measured using a Qubit with the DNA High Sensitivity Assay (Life Technologies), and the size distributions of DNA fragments were measured using a 2100 Bioanalyzer with the High Sensitivity DNA kit and chips (Agilent). 10 nM libraries were sent to the EMBL GeneCore sequencing facility in Heidelberg, Germany (http://genecore3.genecore.embl.de/genecore3/index.cfm) for sequencing on an Illumina HiSeq 2000 with v3 chemistry, with 12 samples multiplexed per lane. All libraries were sequenced as 50-bp single-end reads.

2.4 DamID

2.4.1 Cloning

Three constructs were created for DamID: Dichaete-Dam, SoxN-Dam and Dam-only. The SoxN-Dam fusion protein coding sequence was initially cloned from

an existing pUAST vectors (from Enrico Ferrero). Primers were designed to amplify the coding regions of the SoxN-Dam fusion protein as well as upstream UAS sites, an *HSP70* promoter, and the *SV40* 5' UTR. An *Spe*I site was introduced upstream of the cloned region. The Dichaete-Dam fusion protein coding sequence was cloned from genomic DNA extracted from a *D. melanogaster* line carrying this construct, which was created by Faysal Riaz (Riaz, 2009). Primers were designed to amplify the fusion protein coding region, upstream UAS sites, an *HSP70* promoter, and the *kayak* 5' UTR. A forward primer was used to introduce an *Spe*I site upstream of the cloned region, and a reverse primer introduced an *Avr*II site downstream of the Dichaete-Dam cloned region. The Dam coding region, as well as upstream UAS sites, an *HSP70* promoter and the *SV40* 5' UTR, was cut directly out of an existing pUAST vector (from Tony Southall). All primer sequences are available in Table 2.2.

Target	**Forward Primer**	**Reverse Primer**
Dichaete-Dam	GGGACTAGTCGAGTAC GCAAAGCTTCTGCAT	GGGCCTAGGAGTAAG GTTCCTTCACAAAGAT
SoxN-Dam	GGGACTAGTCGAGTAC GCAAAGCTTCTGCAT	GCGCTGACTTTGAGT GGAAT

Table 2.2: Primers used to amplify Dichaete-Dam and SoxN-Dam coding sequences and flanking regions in *D. melanogaster*.

A two-step cloning process was employed, first cloning inserts into the pSLfa1180fa shuttle vector, and then from the shuttle vector into a *piggyBac* vector marked with 3xP3-EGFP, pBac3xP3-EGFPafm (Horn and Wimmer, 2000). To clone all inserts into the shuttle vector, PCR amplicons and the pSLfa1180fa vector were cut with the following restriction enzymes (NEB):

- SoxN-Dam, pSLfa1180fa: *Spe*I/*Stu*I
- Dichaete-Dam, pSLfa1180fa: *Spe*I/*Avr*II
- Dam, pSLfa1180fa: *Sph*I/*Stu*I

Digestions were performed at 37° C for 1.5 hours and were stopped by incubation at 65° C for 20 minutes. All samples of cut pSLfa1180fa were dephosphorylated after restriction digestion by incubation with Antarctic phosphatase (NEB) at 37° C for 1 hour. The dephosphorylation reaction was stopped by incubation at

70° C for 10 minutes. Digestion products were run out on a 0.8% agarose gel and bands of the desired size were cut out and purified using a QIAQuick Gel Extraction Kit (Qiagen). DNA concentrations were measured using a Nanodrop (Thermo Scientific), and an aliquot of purified DNA was again run on 0.8% agarose gel to check for bands of the appropriate size.

Digested inserts and shuttle vectors were ligated at 16° C overnight using T4 DNA ligase (Roche). The ligation reaction was stopped by incubation at 65° C for 10 minutes. An aliquot of each product was run on a 0.8% agarose gel to check for ligation, and DNA concentrations were measured using a Qubit with the DNA High Sensitivity Assay (Life Technologies). Ligated plasmids were used to transform chemically competent *E. coli* (BIOBlue from Bioline or One Shot TOP10 from Invitrogen) and plated on LA+ampicillin plates, which were incubated overnight at 37° C . 24 colonies were picked for each plasmid the following day and were grown in 3 ml of LB in an orbital shaker at 37° C for 24 hours. 1.5 ml of the resulting cultures were used in minipreps to isolate the plasmid DNA using the Merlin system (Ravi Iyer). 12 plasmid preparations for each construct were chosen, and an aliquot of each was digested with a restriction enzyme that was expected to cut the plasmid only once (*Stu*I, *Spe*I, *Avr*I or *EcoR*I). Digested DNA was run on a 0.8% agarose gel to check for a band of the expected size. Of the clones that showed bands of the correct size, 4-5 for each construct were verified by Sanger sequencing.

One shuttle vector containing the desired insert for each construct was chosen for cloning into the final pBac3xP3-EGFP vector. These shuttle vectors as well as the *piggyBac* vector were cut with the octo-cutter restriction enzymes *Fse*I and *Asc*I (NEB) as follows:

- pSLfa1180fa-Dichaete-Dam, pBac3xP3-EGFP: *Fse*I/*Asc*I
- pSLfa1180fa-SoxN-Dam, pBac3xP3-EGFP: *Fse*I
- pSLfa1180fa-Dam, pBac3xP3-EGFP: *Fse*I

Digestions were performed at 37° C for 1.5 hours and were stopped by incubation at 65° C for 20 minutes. All samples of cut pBac3xP3-EGFP were dephosphorylated after digestion with Antarctic phosphatase. The dephosphorylation reaction was stopped by incubation at 70° C for 10 minutes. Digestion products were run

out on a 0.8% agarose gel and bands of the desired size were cut out and purified with a QIAQuick Gel Purification Kit (Qiagen). Purified DNA concentrations were measured with a Nanodrop (Thermo Scientific), and an aliquot was again run on a 0.8% agarose gel to check for bands of the appropriate size. Digested inserts and pBac3xP3-EGFP vectors were ligated at 16° C overnight using T4 DNA ligase (Roche). The ligation reaction was stopped by incubation at 65° C for 10 minutes. An aliquot of each product was run on a 0.8% agarose gel to check for ligation, and DNA concentrations were measured using a Qubit with the DNA High Sensitivity Assay (Life Technologies). Although some unligated pBac3xP3-EGFP vector was still visible in all samples, ligated vectors and inserts were also present, so transformation was attempted without further purification.

The ligated pBac3xP3-EGFP-Dichaete-Dam (pBac-Dichaete-Dam), pBac3xP3-EGFP-SoxN-Dam (pBac-SoxN-Dam) and pBac3xP3-Dam (pBac-Dam) constructs were introduced into chemically competent *E. coli* (BIOBlue from Bioline or One Shot TOP10 from Invitrogen) and plated on LA+ampicillin plates, which were incubated overnight at 37° C . 24 colonies were picked for each plasmid the following day and were grown in 3 ml of LB in an orbital shaker at 37° C for 24 hours. 1.5 ml of the resulting cultures were used in minipreps to isolate the plasmid DNA using the Merlin system. 12 plasmid preparations for each construct were chosen, and an aliquot of each was digested with *EcoR*I, which was expected to cut each plasmid in three places. Digested DNA was run on a 0.8% agarose gel to check for three bands of the expected sizes. Additionally, three clones of pBac-Dam with the correct band sizes were chosen for verification via Sanger sequencing.

1-2 clones that showed the correct pattern of bands and, for pBac-Dam, had the correct insert sequence, were chosen for each construct, and the corresponding *E. coli* cultures containing each plasmid were diluted in 50 ml LB and grown on an orbital shaker overnight at 37° C. Plasmid DNA was purified from each culture using a Qiagen HiSpeed Plasmid Midi Kit. The concentration of purified DNA was measured using a Nanodrop (Thermo Scientific) and an aliquot was run on a 0.8% agarose gel to check that the size was still correct. Purified plasmids were concentrated in a speedvac to a final concentration of 1 μg/μl. Plasmids were injected into embryos from each species of *Drosophila* along with a plasmid containing a helper *piggyBac* transposase (phsp-pBac or 1409 *D. mel hsp70* hyperactive *piggyBac*, supplied by Ernst Wimmer). Transformants were

obtained for each construct in all species except for pBac-SoxNDam in *D. pseudoobscura*; although this injection was repeated several times, no transformants were recovered.

2.4.2 Isolation of DamID DNA fragments

For each transgenic line in each species, embryos were collected after overnight lays and dechorionated in 50% bleach. They were then rinsed in homogenization buffer (10 mM Tris-HCl pH 7.6, 60 mM NaCl, 10 mM EDTA, 0.15 mM spermine, 0.15 mM spermidine, 0.5% Triton X-100), flash-frozen in liquid nitrogen and stored at -80° C. Three biological replicates were collected from each line, with each replicate consisting of approximately 50-150 μl of dry embryos. To extract high-molecular weight genomic DNA, each aliquot of embryos was homogenized in a Dounce 15-ml homogenizer in 10 ml of homogenization buffer. 10 strokes were applied with pestle B, followed by 10 strokes with pestle A. The lysate was then spun for 10 minutes at 6000g. The supernatant was discarded, and the pellet was resuspended in 10 ml homogenization buffer, then spun again for 10 minutes at 6000g. The supernatant was again discarded, and the pellet was resuspended in 3 ml homogenization buffer. 300 μl of 20% n-lauroyl sarcosine were added, and the samples were inverted several times to lyse the nuclei. The samples were treated with RNaseA followed by proteinase K at 37° C. They were then purified by two phenol-chloroform extractions and one chloroform extraction. Genomic DNA was precipitated by adding 2X EtOH and 0.1X NaOAc, dried, and resuspended in 50-150 μl TE buffer, depending on the starting amount of embryos. DNA was run on a 1% agarose gel to check for the presence of a single clean, high-molecular weight band, and the concentration was measured on a Nanodrop (Thermo Scientific).

Molecular biology for DamID was performed essentially as described, with some modifications (Vogel *et al.*, 2007). 30 μl of each gDNA sample was used in *Dpn*I digestions, which were performed at 37° C for two hours. Initially, distinct bands were detected after the PCR step which displayed a characteristic pattern for each species, indicating that they were likely due to the DamID primers binding non-specifically to non-digested genomic DNA. To prevent this, a size-selection step was added between the *Dpn*I digestion step and the ligation step. 0.7X

Agencourt AMPure XP beads (Beckman Coulter) were added to each sample to remove high-molecular weight DNA, leaving behind digested fragments. The supernatant was retained, and 1.1X Agencourt AMPure XP beads were added to recover all remaining DNA. The DNA was eluted in 30 μl TE buffer, then used in the ligation step. This size-selection effectively eliminated the non-specific genomic bands. For some of the non-model species of *Drosophila*, faint bands were still observed at regular size intervals; it was determined that these were due to oligomerization of the DamID adapters. These were eliminated by titrating the adapters at the ligation step down to the minimum concentration that still resulted in amplification of expected products, either 1:2 or 1:4. This also eliminated the faint amplification that was sometimes visible in the no-*Dpn*I control. Some adapter oligomers were still sequenced, but these could be filtered out computationally.

2.4.3 Preparation of DamID libraries for sequencing

After PCR amplification, the DamID DNA samples were purified using a phenol-chloroform extraction followed by a chloroform extraction. The DNA was precipitated and resuspended in 50 μl TE buffer. They were then sonicated in order to reduce the average fragment size using a Covaris S2 sonicator with the following settings: Intensity 5, duty cycle 10%, 200 cycles/burst, 300 seconds. After sonication, the samples were purified using a QIAquick PCR Purification kit to remove small fragments. The sample concentrations were measured using a Qubit with the DNA High Sensitivity Assay (Life Technologies), and the size distributions of DNA fragments were measured using a 2100 Bioanalyzer with the High Sensitivity DNA kit and chips (Agilent). Samples were sent to BGI Tech Solutions (HongKong) Co., Ltd., for library construction and sequencing on an Illumina MiSeq or HiSeq. Libraries were multiplexed with 2 samples per run for the MiSeq and 9-12 samples per lane for the HiSeq. MiSeq libraries were run as 150-bp single-end reads, while HiSeq libraries were run as 50-bp single-end reads.

2.5 FAIRE-seq

2.5.1 Isolation of FAIRE DNA fragments

The timing of developmental stages for *Drosophila pseudoobscura* embryos was calculated using the species-specific function from Kuntz and Eisen (2013), with the temperature set to 25° C. Adults were kept in cages at 22.5° C and were given fresh grape juice-agar plates streaked with fresh yeast paste every hour for at least 2 hours before collections began. For each collection, the flies were allowed to lay for 1 hour at 22.5° C, then the agar plates were removed, replaced with fresh plates, and placed at 25° C for the embryos to age to the correct stage. To verify the developmental stages, an aliquot of embryos at each stage was dechorionated with 50% bleach, devitellinized with heptane, and examined on a Zeiss Axioplan microscope with a 10x and a 20x objective. Calculated developmental times were added to the observed time that it took for embryos to reach cellularization at Stage 5, as this was considered the zero timepoint by Kuntz and Eisen (2013). The final times used for embryo staging were as follows and indicate the time that each agar plate was allowed to age after a 1-hour lay:

- Stage 5: 4 hours, 35 minutes
- Stage 9: 6 hours
- Stage 10: 6 hours, 45 minutes
- Stage 11: 8 hours, 45 minutes
- Stage 14: 13 hours, 45 minutes

Staged embryos were collected, dechorionated in 50% bleach, and fixed as described for ChIP, except that volumes were halved and fixation was for 15 minutes. Following fixation, embryos were flash-frozen in liquid nitrogen and stored at -80° C. Three biological replicates were collected for each stage, with approximately 60 mg of embryos per replicate. Embryos were homogenized in a 1-ml Dounce homogenizer in 1 ml of PBT supplemented with protease inhibitors (Complete Mini Protease Inhibitor cocktail tablets, Roche). 20 strokes were applied with the loose pestle, then the lysate was allowed to rest on ice for 30 seconds,

and then 20 more strokes were applied with the loose pestle. The lysate was centrifuged at 1100g for 10 minutes at 4° C and the supernatant was discarded. The pellet was resuspended in 1 ml cold cell lysis buffer supplemented with protease inhibitors and further homogenized by applying 20 strokes with the tight pestle. The lysate was centrifuged at 2000g for 4 minutes at 4° C to pellet the nuclei, the supernatant discarded, and the pellet resuspended in 1 ml cold nuclear lysis buffer supplemented with protease inhibitors. The mix was incubated at room temperature for 20 minutes to lyse the nuclei, split into 160 μl aliquots and sonicated in a Diagenode Bioruptor with the energy settings on high.

All aliquots were initially sonicated for 16 cycles of 30 seconds on, 30 seconds off. A 100 μl input aliquot was removed from each sample, treated with RNaseA for 30 minutes at 37° C and then with proteinase K for 1 hour at 55° C, and then incubated overnight at 65° C to reverse crosslinks. The following day, phenol-chloroform extractions were performed on the input samples as described (Simon *et al.*, 2012). The input samples were then run on a 3% agarose gel to estimate the size distribution of DNA fragments. If the average fragment size was greater than 500 bp, the entire FAIRE sample was sonicated again for up to 4 additional cycles of 30 seconds on, 30 seconds off, and another input aliquot was taken to determine the fragment size distribution. The final fragment sizes ranged from 100 bp to 1000 bp, with the majority of fragments falling between 250 and 500 bp. Phenol-chloroform extractions were performed on the FAIRE samples as described (Simon *et al.*, 2012).

2.5.2 Preparation of FAIRE libraries for sequencing

FAIRE DNA was purified using a QIAquick PCR Purification Kit (Qiagen).The concentration of each FAIRE sample was measured using a Qubit with the High Sensitivity DNA Assay (Life Technologies). The size distribution of fragments in each sample was measured using a 2100 Bioanalyzer with the High Sensitivity DNA kit and chips (Agilent). Input samples were not sequenced, but were used to gauge the percentage of the genome recovered by FAIRE (Simon *et al.*, 2012). Samples with at least 10 ng of DNA present were sent to BGI Tech Solutions (HongKong) Co., Ltd., for library construction and sequencing on an Illumina

HiSeq. Libraries were multiplexed with 12 samples per lane and were run as 50-bp single-end reads.

2.6 Sequencing data analysis

2.6.1 Quality control and mapping

All high-throughput sequencing data from Illumina Hiseq and Miseq platforms was received in Fastq format. The program FastQC was used to evaluate the overall quality and attributes of each dataset (`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`). FastQC produces a report for each sample containing the following sections: Basic Statistics, Per Base Sequence Quality, Per Sequence Quality Scores, Per Base Sequence Content, Per Base GC Content, Per Sequence GC Content, Per Base N Content, Sequence Length Distribution, Duplicate Sequences, Overrepresented Sequences and Overrepresented Kmers. This information was used to identify potential sequencing contaminants, such as adapters. In the case of the DamID data, cutadapt was used to trim adapter sequences from both ends of reads by running the command:

```
cutadapt -a GATCCTCGGCCGCGACC -g ^GGTCGCGGCCGAGGATC
        -o output.fastq.gz input.fastq.gz
```

FastQC was then run again to verify that all 5' and 3' adapter sequences had been removed.

Reads were mapped against each reference genome using bowtie2 with the default settings (Langmead and Salzberg, 2012). The reference genomes used were: *D. melanogaster* April 2006 (UCSC dm3, BDGP) (Adams, 2000; Celniker *et al.*, 2002), *D. simulans* April 2005 (UCSC droSim1, The Genome Institute at Washington University (WUSTL)) (Clark *et al.*, 2007), *D. yakuba* November 2005 (UCSC droYak2, The Genome Institute at Washington University (WUSTL)) (Clark *et al.*, 2007) and *D. pseudoobscura* November 2004 (UCSC dp3, Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC)) (Clark *et al.*, 2007; Richards, 2005). The most recent UCSC version was used for each

species except *D. pseudoobscura*; I decided to use the dp3 (Nov. 2004) release instead of the dp4 (Feb. 2006) release because annotation tables were not available for the dp4 release and because the unassembled chromosomes were broken into separate scaffolds, making it difficult to run certain downstream analysis tools. All reference genomes were downloaded from the UCSC Genome Browser (`http://hgdownload.soe.ucsc.edu/downloads.html`). Mapped reads were sorted and indexed using SAMtools (Li *et al.*, 2009).

2.6.2 ChIP-seq processing and peak calling

All ChIP and input libraries were normalized to a total library size of 1,000,000 reads. Peaks were called on each matched pair of ChIP and input replicates independently using both MACS and Peakzilla (Bardet *et al.*, 2013; Zhang *et al.*, 2008).

2.6.3 DamID processing, peak calling and annotation

The position of every GATC site in each genome was determined using the HOMER utility scanMotifGenomeWide.pl (`http://homer.salk.edu/homer/index.html`) (Heinz *et al.*, 2010). For each sample, reads were extended to the average fragment length (200 bp) using the BEDTools slop utility. The number of extended reads overlapping each GATC fragment was then calculated using the BEDTools coverage utility (Quinlan and Hall, 2010). The resulting counts for each sample were collated to form a count table for each species, consisting of one column for each fusion protein or Dam-only sample and one row for each GATC fragment in the genome. These count tables served as inputs to run DESeq2 (run in R version 3.1.0 using RStudio version 0.98), which was used to test for differential enrichment in the fusion protein samples versus the Dam-only samples in each GATC fragment (Love *et al.*, 2014). Fragments flagged as differentially enriched (log2 fold change >0 and adjusted p-value <0.05 or <0.01) were extracted, and neighboring GATC fragments with less than 100 bp separating them were merged to form binding intervals using a perl script. Binding intervals were scanned for both *de novo* and known motifs using HOMER

findMotifsGenome.pl (Heinz *et al.*, 2010).

For the *D. melanogaster* data, as well as translated data from each other species, each binding interval was assigned to the closest gene using a perl script, bed2closestGene_v2.pl, written by Bettina Fischer. Genomic feature annotations were performed using the Bioconductor package ChIPSeeker (Yu, 2014). The distances from binding intervals to TSSs were calculated and plotted using both the Bioconductor package ChIPpeakAnno (Zhu *et al.*, 2010), which considers the distance between every interval and the closest TSS, and an unpublished suite of R scripts written by Bettina Fischer, CHIPPAVI, which considers the distance between every TSS and the surrounding intervals. All calculations of overlaps between interval datasets were performed using the BEDTools intersect utility (Quinlan and Hall, 2010).

2.6.4 FAIRE-seq processing and peak calling

Each FAIRE-seq replicate was processed separately. For each sample, reads were extended to the average fragment length (300 bp) using the BEDTools slop utility (Quinlan and Hall, 2010). MOSAiCS (run in R version 3.1.0 using RStudio version 0.98) was used to call peaks, using a one-sample analysis with 50 bp-binned mappability, GC-content and N-content scores as covariates (Chung *et al.*, 2012). Mappability files were generated for the *D. pseudoobscura* reference genome using code available as part of the PeakSeq package (`http://archive.gersteinlab.org/proj/PeakSeq/Mappability_Map/Code/`) (Rozowsky *et al.*, 2009) and supplemental code from the MOSAiCS website (`http://www.stat.wisc.edu/~keles/Software/mosaics/`). GC-content and N-content files were generated using the *D. pseudoobscura* dp3.2bit binary file from the UCSC Genome Browser (`http://hgdownload.soe.ucsc.edu/downloads.html`) and supplemental code from the MOSAiCS website. After model-fitting, the Bayesian Information Criterion (BIC) and Goodness of Fit plot (GOF) for each replicate was examined, and the best-fitting model from either the one-signal-component or two-signal-component model was chosen for peak calling. Peak calling was performed at both FDR 5 and FDR 10. In order to establish a high-confidence list of peaks for each developmental stage, the FDR 10 peaks from each replicate in the

same stage were intersected using BEDTools (Quinlan and Hall, 2010), and any peaks that were present in at least 2 out of the 3 replicates were kept. DiffBind (run in R version 3.1.0 using RStudio version 0.98) was used to cluster replicate samples using both peak datasets and raw reads, as well as to perform principal component analysis (PCA) of replicate read density profiles and to identify differentially enriched peaks between each developmental stage. Peaks were scanned for both *de novo* and known motifs using HOMER findMotifsGenome.pl (Heinz *et al.*, 2010).

For the analysis of FAIRE tag count in transcription factor binding intervals, peaks for the factors Pipsqueak (Psq) and Trithorax-like (Trl) in 0-4 hour *D. pseudoobscura* embryos, and for the factors Hunchback (Hb), Giant (Gt), Bicoid (Bcd) and Kruppel (Kr) in blastoderm-stage *D. pseudoobscura* embryos, were downloaded from GEO (accession numbers GSE25666, GSE25667 and GSE50771). Peak coordinates were translated from the dp4 assembly to the dp3 assembly of the *D. pseudoobscura* genome using the UCSC LiftOver tool (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`). For Dichaete, DamID-seq peaks called by DESeq2 with an adjusted p-value <0.05 were used. Perl scripts were used to find the midpoint of each peak, then extend it 2500 bp in either direction, resulting in 5-kb intervals around each peak. These intervals were then split into 50-bp bins, and the BEDTools coverage utility was used to calculate the number of FAIRE-seq tags overlapping each bin for each biological replicate from Stage 5 (Quinlan and Hall, 2010). For each bin, the average score was taken over all peaks. FAIRE score versus position surrounding peaks was plotted in R version 3.1.0 using RStudio version 0.98.

Both Genscan and GeneID gene predictions for the UCSC dp3 version of the *D. pseudoobscura* genome were downloaded in BED format from the UCSC Table Browser (`http://genome.ucsc.edu/cgi-bin/hgTables?db=dp3`) (Burge and Karlin, 1997; Karolchik, 2004; Karolchik *et al.*, 2014; Parra, 2000). The BEDTools intersect utility (Quinlan and Hall, 2010) was used to classify each FAIRE peak from all stages as either exonic (falling entirely in an exon), exon boundary (partially overlapping an exon), intronic (falling entirely in an intron), gene boundary (partially overlapping a gene at either the 5' or 3' end) or intergenic (having no overlap with any gene).

2.6.5 Cross-species comparison

Both peaks and reads from non-*D. melanogaster* species were translated to the *D. melanogaster* UCSC dm3 reference genome using the LiftOver utility from the UCSC genome browser (`http://genome.ucsc.edu/cgi-bin/hgLiftOver`) (Bardet *et al.*, 2011). For *D. simulans* and *D. yakuba*, the minMatch parameter was set to 0.7, while for *D. pseudoobscura* it was set to 0.5; for all species, multiple outputs were not permitted. To enable a quantitative comparison between DamID datasets from all species, both translated peaks and reads were analyzed with DiffBind, which was run in R v3.1.0 using RStudio v0.98 (Ross-Innes *et al.*, 2012). Translated reads in BED format were converted into SAM format using a custom perl script, and then into BAM format for use with DiffBind using SAMtools (Li *et al.*, 2009). For each analysis, the translated reads from each sample were normalized together in DiffBind using the DESeq2 normalization method.

2.6.6 Data visualization

All visualization of sequence data was done with the Integrated Genome Browser (IGB) (Nicol *et al.*, 2009). Sequencing coverage was visualized in WIG or BIGWIG format, while peaks were visualized in BED format. The UCSC wigToBigWig utility (`http://hgdownload.cse.ucsc.edu/admin/exe/`) was used to convert WIG files to the BIGWIG format for easier storage and loading.

2.6.7 Code availability

All custom perl scripts can be found at `www.github.com/sarahhcarl/flychip`. An overview of the DamID data processing and peak calling pipeline can be found in the wiki at:
`https://github.com/sarahhcarl/flychip/wiki/Basic-DamID-analysis-pipeline`.

2.7 Molecular evolutionary analyses

2.7.1 Sequence analysis of group B Sox proteins

All orthologous group B Sox sequences were retrieved by using BLASTX (translated BLAST: `http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastx&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome`) against the genomes of *D. simulans, D. yakuba* and *D. pseudoobscura* with the DNA sequences of each group B *Sox* gene in *D. melanogaster* as queries (Altschul *et al.*, 1990). Sequences for group B Sox proteins in other species were downloaded from either NCBI (`http://www.ncbi.nlm.nih.gov/`) or, in the case of *Aedes aegypti*, VectorBase (`https://www.vectorbase.org/`). Amino acid sequences were aligned and a neighbor-joining tree was constructed using ClustalW (Chenna, 2003). Multiple alignments were visualized using BoxShade (`http://ch.embnet.org/software/BOX_form.html`), and phylogenetic trees were visualized using FigTree v1.3.1 (*http://tree.bio.ed.ac.uk/software/figtree/*).

2.7.2 Multiple alignment of conserved and unique binding regions

For the evolutionary analysis of Dichaete-Dam binding intervals, I used DiffBind to identify a set of binding intervals in *D. melanogaster* that were conserved in all four species and a set that were only present in *D. melanogaster* (Ross-Innes *et al.*, 2012). I extracted the *D. melanogaster* genome coordinates of these intervals and used the UCSC LiftOver utility to translate them to the *D. simulans* droSim1, *D. yakuba* droYak2 and *D. pseudoobscura* dp3 reference genome assemblies. For all species, the minMatch parameter was set to 0.7. The sequences of each orthologous binding interval in each species were obtained using the fetch-UCSC sequences tool from RSAT, preserving strand information (Thomas-Chollier *et al.*, 2011). For each interval for which one unambiguous orthologous sequence could be identified in all four species, I performed a multiple alignment of the sequences using the phylogeny-aware multiple aligner PRANK (Löytynoja and Goldman, 2005, 2008). In each case, I estimated the guide tree from the data directly,

resulting in a calculation of the substitution rate on each branch. I decided to estimate the guide tree from the data rather than using an independent estimate of branch lengths because selection is expected to act differently on different classes of DNA; therefore, branch lengths determined using coding sequences or averaging over the genome might over- or under-estimate the expected differences in regulatory sequences between species. Note, however, that the branch lengths estimated by prank are quite close to those determined from 1000 random 10-kb noncoding regions in *D. melanogaster, D. simulans, D. yakuba* and *D. erecta* by Moses *et al.* (2006). I calculated the percentage of perfectly conserved nucleotides in each interval from these multiple alignments using a custom perl script available at `www.github.com/sarahhcarl/Flychip/DamID_analysis`.

2.7.3 Predicting transcription factor binding sites

Two different strategies were used to predict Sox binding sites within the DamID binding intervals. All binding interval sequences were scanned independently for matches to Sox motifs using FIMO (Grant *et al.*, 2011), and subsets of aligned sequences were scanned using the RSAT tool matrix-scan (Sand *et al.*, 2008; Turatsinze *et al.*, 2008). In each case, I first downloaded the positional weight matrix (PWM) representing the top-scoring *de novo* Sox motif identified via HOMER in each DamID binding interval dataset (Heinz *et al.*, 2010). I then used FIMO to search for matches to each of these PWMs in all DamID binding datasets, using the original sequences from each species genome. The resulting hits were used to calculate the average number of motifs per binding interval overall, as well as in binding intervals that are conserved in all species versus those that are unique to one.

I used the same PWMs to scan the multiple alignments of both 4-way conserved and unique Dichaete-Dam binding intervals for potential Sox binding sites using matrix-scan. I chose matrix-scan for this analysis because, unlike FIMO, it can accept multiple alignments directly as input, greatly facilitating the assignment of positional orthology to putative binding sites. Matrix-scan was run using the pre-compiled *Drosophila* background file provided by RSAT as the background for scanning and with the cutoff for reporting matches set to a PWM weight-score

of >=4 and a p-value of <0.0001. If a binding site was identified at the same aligned position in an orthologous binding interval in more than one species, it was considered to be an orthologous site between those species. Sites that partially overlapped in position between orthologous enhancers were not considered to be orthologous; however, these were not common. Sites identified as matching multiple, overlapping PWMs in the same species and dataset were considered as separate hits.

2.7.4 Tests of conservation

The percentages of conserved nucleotides present in binding intervals and motifs were calculated using custom perl scripts available at `https://github.com/sarahhcarl/Flychip`. Randomly shuffled control motifs were generated using the RSAT tool permute-matrix (Thomas-Chollier *et al.*, 2011). All statistical tests were performed in R v3.1.0 using RStudio v0.98.

CHAPTER 3

Exploratory Analysis of Dichaete and SoxNeuro in Four Species of *Drosophila*

3.1 Overview and motivation

Before performing genome-wide binding experiments for Dichaete and SoxN in non-model species of *Drosophila*, I set out to characterize the orthologous proteins in each species of interest at the levels of sequence, expression pattern and specific target binding. In addition to verifying that the orthologous transcription factors are similar enough to make direct inter-species comparisons of binding valid, this process also allowed me to test the specificity of the Dichaete and SoxN antibodies that I planned to use for ChIP-seq in each species. Although the HMG DNA-binding domains of group B Sox proteins are highly conserved over the evolutionary distances that I examined, other domains of Dichaete and SoxN have diverged somewhat, making it possible that antibodies raised against these proteins in one species might not display the same reactivity with targets in another species (McKimmie *et al.*, 2005). I analyzed the expression pattern

of each protein by collecting embryos from each species and performing immunohistochemistry using both Dichaete and SoxN antibodies. I then compared the resulting staining patterns at a variety of developmental stages.

Because antibodies can often work well in immunohistochemistry but be ineffective in ChIP experiments, it was also necessary to test for enrichment of specific target sequences using ChIP-PCR before proceeding to ChIP-seq. This was an imprecise process in non-model species because PCR targets had to be designed in each species based on regions of known binding, which were only available in *D. melanogaster*. Although orthologous sequences can be approximated using tools such as BLAST (Altschul *et al.*, 1990) or by simply examining gene models and identifying, for example, introns at orthologous positions in each species, it is unknown *a priori* whether the transcription factors of interest will bind to those exact sequences in each species. Nonetheless, I designed primers for three potential target regions for each transcription factor in each species, along with one negative control region for each transcription factor, which were chosen by examining the available *in vivo* binding data in *D. melanogaster* and identifying regions where no binding was observed. I performed each ChIP reaction on three biological replicates of chromatin derived from fixed embryos collected from each species, using both input chromatin and a mock IP as controls. I then tested for target enrichment by performing PCR using each of the primer sets that I designed, with the ChIP DNA, the mock IP DNA, and the input as templates. Although qPCR would have given a more quantitative view of the enrichment of target sequences in the ChIP DNA compared to the control, classical PCR yielded a qualitative overview of the presence or absence of target sequences in each sample.

In order to compare Dichaete and SoxNeuro binding patterns on a genome-wide scale between four *Drosophila* species, I initially intended to use ChIP-seq. As discussed in the introduction, this technique has both advantages and disadvantages; ChIP allows for the detection of binding by the endogenous protein in the native spatial and temporal context, provided a suitable antibody is available. Genome-wide binding can be measured from ChIP experiments either by hybridizing the resulting DNA libraries to a microarray (ChIP-chip) or sequencing them on a high-throughput sequencer (ChIP-seq). I chose to primarily use sequencing rather than microarrays because of the lack of availability of tiling

arrays for non-model *Drosophila* species and also because of the increase in resolution and dynamic range possible with sequencing in comparison to microarray technology (Aleksic and Russell, 2009). However, I also performed one ChIP-chip experiment in *D. melanogaster* in order to further validate the Dichaete antibody before investing in sequencing.

For both the ChIP-chip and ChIP-seq experiments, I planned to sequence 3 replicates of each experimental condition (Dichaete and SoxN ChIP) and 3 control replicates in each species studied. I planned to compare the same control replicates against each experimental condition; however, due to the lack of target enrichment in ChIP-PCR experiments with the SoxN antibody, I only performed ChIP-chip and ChIP-seq for Dichaete. Although high-quality ChIP-seq biological replicates often show high correlation with each other, leading the modENCODE Consortium to accept a minimum of 2 biological replicates per experiment (Landt *et al.*, 2012), I decided to use 3 replicates to provide greater statistical confidence and in order to reduce the effect of any potential technical problems that could lead to one replicate being an outlier.

Theoretically, the most appropriate control for a ChIP experiment is a mock IP, in which preimmune serum or an antibody that is not expected to bind to anything in the sample is used in parallel to the real IP; this type of control is commonly used in ChIP-chip experiments (Ghavi-Helm and Furlong, 2012; Park *et al.*, 2013). I used a mock IP control with an antibody against beta-Galactosidase for the Dichaete ChIP-chip experiment and for one Dichaete ChIP-seq experiment. However, mock IPs often yield very low amounts of DNA, making it difficult to construct sequencing libraries and leading to the introduction of bias from PCR overamplification. For this reason, input chromatin, which is a sample of the original chromatin that is set aside after sonication and then purified without performing any type of immunoprecipitation, is often used instead of a mock IP for ChIP-seq (Ghavi-Helm and Furlong, 2012). An input control helps to correct for some of the biases associated with high-throughput sequencing, which can stem from factors such as GC content, differences in mappability and local chromatin accessibility, because it is subject to the same sources of bias as the experimental sample (Dohm *et al.*, 2008). After observing high amounts of PCR overamplification in my mock IP control samples for ChIP-seq, I switched to using input controls for all subsequent ChIP-seq experiments. In total, I gen-

erated the following genome-wide ChIP datasets: three replicates of Dichaete ChIP-chip with three replicates of beta-Galactosidase mock IP controls in *D. melanogaster*; four, two and three replicates respectively of Dichaete ChIP-seq with beta-Galactosidase mock IP controls in *D. melanogaster, D. simulans* and *D. yakuba*; and three replicates of Dichaete ChIP-seq with three replicates of input chromatin controls in *D. melanogaster, D. simulans, D. yakuba and D. pseudoobscura*. A detailed description of methods used for both immunohistochemistry and all ChIP-based techniques can be found in Chapter 2.

3.2 Sequence and phylogenetic analysis

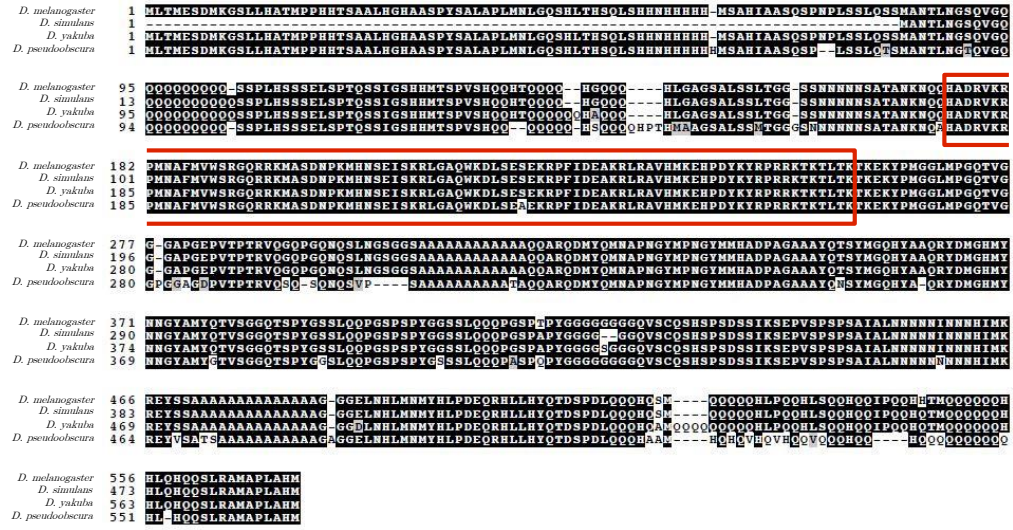
Previous to the work of this thesis, the amino acid sequences of the group B Sox proteins Dichaete, SoxN, Sox21a and Sox21b had been aligned in the insects *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Anopheles gambiae* and *Apis mellifera*, along with the corresponding orthologous sequences in mouse, revealing a deep conservation of the HMG box DNA binding domains in each protein alongside considerably higher divergence in other domains of the proteins. Additionally, the amino acid sequences of just the HMG domains of all known group B Sox proteins at the time were aligned together, illustrating the high levels of sequence conservation among all group B Sox proteins as well as specific amino acid substitutions that are common among orthologs but differ between paralogous proteins in each species analyzed (McKimmie *et al.*, 2005). On a larger scale, the genomic organization of group B *Sox* genes, with *Dichaete, Sox21a* and *Sox21b* located nearby on the same chromosome in the Dichaete cluster and *SoxN* located on a separate chromosome, has been shown to be conserved in insects ranging from *D. melanogaster* to the hymenopteran *Nasonia vitripennis* and the coleopteran *Tribolium castaneum*. An independent duplication in *Tribolium* has resulted in a fifth group B *Sox* gene, also located in the *Dichaete* cluster (Phochanukul and Russell, 2010). Previous phylogenetic analysis of group B Sox proteins in insects has shown that orthologs for each family member cluster into clades together, supporting the hypothesis that the four common group B Sox proteins diverged before the radiation of the major insect phyla (Wilson and Dearden, 2008; Zhong *et al.*, 2011).

I extended this analysis and focused it on the species relevant to my study by using BLAST to identify the orthologous sequences of the group B Sox proteins in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura* (Altschul *et al.*, 1990). I used ClustalW2 to align the entire amino acid sequences of each orthologous protein in the four species and found that they are very highly conserved, with near perfect conservation in the HMG domains and a relatively low number of substitutions and indels in other areas of the proteins (Figure 3.1A-B) (Chenna, 2003). The most divergent sequences belong to *D. pseudoobscura*, which is the farthest from the other three species phylogenetically. In *D. simulans*, I observed a large deletion at the N-terminal end of SoxNeuro; however, it was uncertain whether this was a true deletion or a fragment of missing sequence due to the lower quality of the *D. simulans* genome assembly. I also used the HMG domains from these species along with the amino acid sequences from the HMG domains of all identified group B Sox proteins in the mosquito *Aedes aegypti*, the beetle *Tribolium castaneum*, the honeybee *Apis mellifera*, the nematode *C. elegans* and the vertebrates mouse (*Mus musculus*) and human (*Homo sapiens*) to construct a neighbor-joining tree (Figure 3.1C). I used an established outgroup, the fungal protein MATA-1, to root the tree (Laudet *et al.*, 1993). This analysis shows that, for each group B Sox protein, the orthologous sequences from the four *Drosophila* species form a monophyletic clade, with the nearest sister group in each case except for that of Dichaete being the orthologous protein in *A. aegypti*. The results of the phylogenetic analysis of group B Sox proteins support the idea that orthologous proteins in *Drosophila* are highly conserved and are suitable for an inter-species comparison of binding. Nonetheless, I still needed to test whether antibodies against Dichaete and SoxN in *D. melanogaster* would react specifically with the orthologous proteins in each other species; to do so, I proceeded to use immunohistochemistry and ChIP-PCR.

A


D. melanogaster
D. simulans
D. yakuba
D. pseudoobscura

B


D. melanogaster
D. simulans
D. yakuba
D. pseudoobscura

C

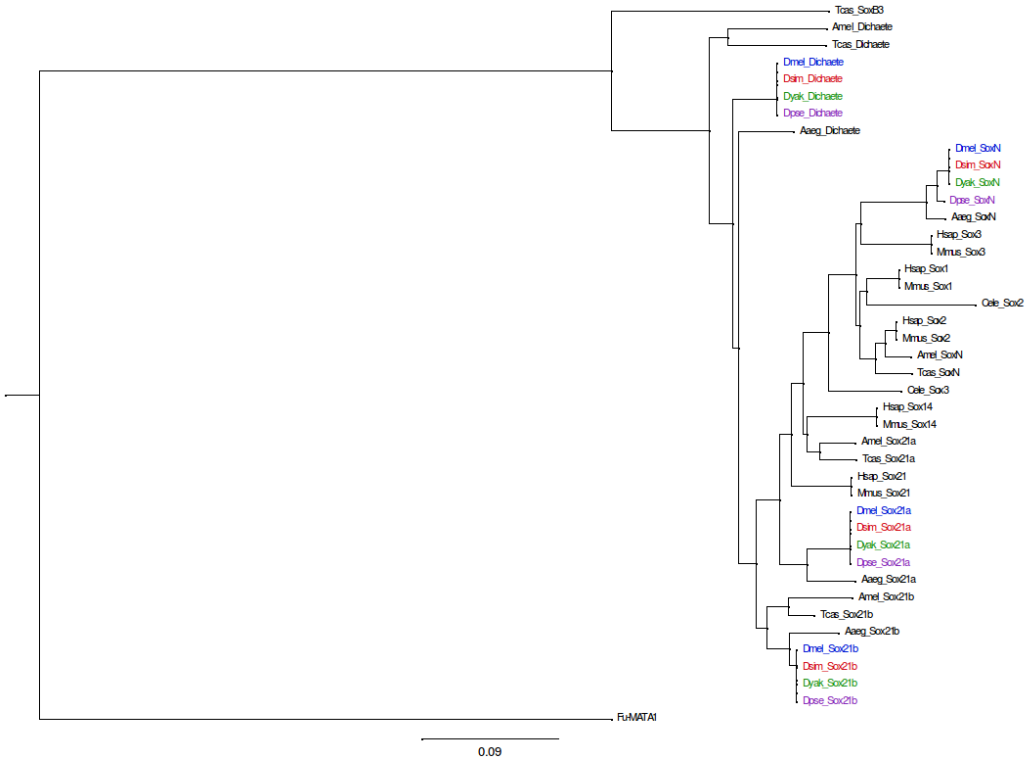


Figure 3.1: Phylogenetic analysis of group B Sox amino acid sequences. A.) Multiple alignment of entire amino acid sequence of Dichaete in, starting from the top, *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. The HMG domains of each orthologous protein are highlighted in the red box and are nearly identical. B.) Multiple alignments of entire amino acid sequence of SoxNeuro in, starting from the top, *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. The HMG domains of each orthologous protein are highlighted in the red box and are nearly identical. C.) Multiple alignment of entire neighbor-joining tree constructed from multiple alignment of the amino acid sequences of group B Sox HMG domains from the four species of *Drosophila* of interest as well as several other invertebrates and vertebrates. Species used in this study are highlighted in blue (*D. melanogaster*), red (*D. simulans*), green (*D. yakuba*) and purple (*D. pseudoobscura*). The tree was rooted using the fungal protein MATA-1, an established outgroup (Laudet *et al.*, 1993). The sequences from orthologous proteins in each species of *Drosophila* form monophyletic clades, with the nearest outgroup in each case being *A. aegypti*. Abbreviations: *Drosophila melanogaster (Dmel), Drosophila simulans (Dsim), Drosophila yakuba (Dyak), Drosophila pseudoobscura (Dpse), Aedes aegypti (Aaeg), Tribolium castaneum (Tcas), Apis mellifera (Amel), Caenorhabditis elegans (Cele), Homo sapiens (Hsap), Mus musculus (Mmus).*

3.3 Assessing expression patterns

The specific gene regulatory activity of transcription factors is tightly coupled to their spatial expression patterns, which often change throughout development. In *Drosophila*, even orthologous *cis*-regulatory regions with divergent sequences and positioning of TF binding sites have been shown to drive equivalent patterns of expression in transgenic assays (Hare *et al.*, 2008); it has been speculated that this phenotypic conservation is due to the evolution of compensatory binding events. Binding of TFs to shadow enhancers, which are secondary regions of regulatory DNA often located farther away from their target genes than primary enhancers and which can drive nearly identical expression patterns, can also confer robustness on expression (Ludwig *et al.*, 2011; Perry *et al.*, 2010). However, there are also well-documented cases in which evolutionary changes in the *cis*-regulatory region of a transcription factor have resulted in both a change in its expression pattern and the regulation of its downstream target genes, yielding novel phenotypes (Arnoult *et al.*, 2013; Frankel *et al.*, 2012). Therefore, before examining the genome-wide binding patterns of Dichaete and SoxN in different species of *Drosophila*, I first wanted to examine the expression patterns of each orthologous protein in each species of interest in order to verify that they were expressed in grossly equivalent domains during each stage of development. In order to do so, I performed immunohistochemistry on embryos collected from each species using antibodies for Dichaete and SoxN raised against the *D. melanogaster* proteins (Ferrero *et al.*, 2014; Sánchez-Soriano and Russell, 1998); this also served the purpose of determining whether the antibodies would react specifically with their respective orthologous proteins in each species.

The expression patterns of Dichaete and SoxN in the *D. melanogaster* embryo have been previously characterized using immunohistochemistry, fluorescent immunohistochemistry and *in situ* hybridizations, as well as in the *D. pseudoobscura* embryo using *in situ* hybridizations (Crémazy *et al.*, 2000; McKimmie *et al.*, 2005; Overton *et al.*, 2002; Sánchez-Soriano and Russell, 1998). Using these as references for comparison, I stained whole embryos from *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura* for Dichaete and SoxN and examined the expression patterns of Dichaete and SoxN at different stages of embryonic development

(Figure 3.2, Figure 3.3). I observed stronger staining with the Dichaete antibody than the SoxN antibody in all species; nonetheless, a clear and specific pattern of staining could be observed for both proteins in embryos of all species. Qualitatively, the spatial and temporal expression patterns of both proteins in all species were extremely similar. For Dichaete, a broad domain of staining was clearly visible in the blastoderm at stage 5, with a smaller anterior stripe. Characteristic strong staining in the central nervous system appeared at stage 9, with expression detectable in the neuroectoderm and ventral midline in stages 9 and 11. At stage 13, staining became visible in the hindgut, and ectodermal stripes appeared at stage 16. For SoxN, staining was visible throughout the neuroectoderm at stage 8, but was excluded from the ventral midline. The neuroectodermal expression began to take on a segmental pattern at stage 10. The segmental stripes were extended laterally at stage 12, and at stage 16 ectodermal stripes were apparent. These observations provide evidence that both Dichaete and SoxN are expressed in equivalent spatial and temporal patterns during embryonic development in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*.

3.4 Targeted binding analysis

After determining that Dichaete and SoxN have comparable patterns of expression in each species of interest and that the antibodies against each protein were capable of reacting specifically with the orthologous protein in each species, I decided to test the efficacy of the antibodies in chromatin immunoprecipitation (ChIP), as not all antibodies that work well for immunohistochemistry also work well in ChIP reactions (Landt *et al.*, 2012). To do so, I performed ChIP-PCR using chromatin derived from embryos of each species. This also served to determine whether enrichment for specific known targets of Dichaete and SoxN could be detected in non-*melanogaster* species. Although both the Dichaete and SoxN antibodies have been previously used in ChIP-chip experiments, in both cases the data was of variable quality, making it worthwhile to test the antibodies with a targeted analysis before proceeding to perform ChIP-seq (Aleksic, 2011; Ferrero, 2014a).

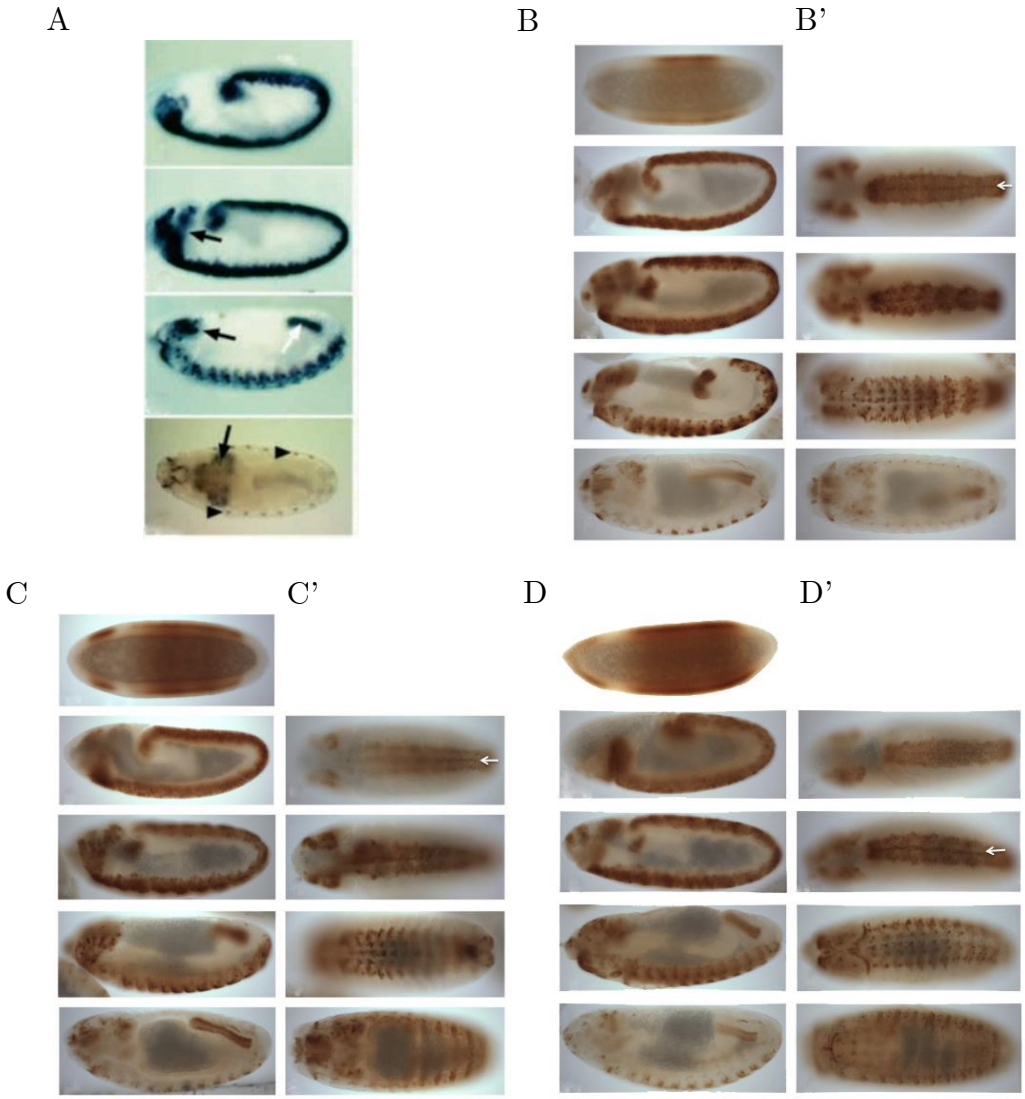


Figure 3.2: Dichaete expression patterns in developing embryos from *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. *D. melanogaster* stainings are reproduced from Sánchez-Soriano and Russell (1998) and were taken at stages 9, 11, 13 and 16. For all other species, images were taken at stages 5, 9, 11, 13 and 16. A.) Lateral views of Dichaete expression in *D. melanogaster* embryos. Black arrows indicate the brain and the white arrow indicates the hindgut. Black arrowheads indicate the chordotonal organs. B-D, lateral views; B'-D', dorsal or ventral views. B.) and B'.) Dichaete expression in *D. simulans* embryos. C.) and C'.) Dichaete expression in *D. yakuba* embryos. D.) and D'.) Dichaete expression in *D. pseudoobscura* embryos. Expression patterns are qualitatively the same at the equivalent stages in each species. White arrows in B', C' and D' indicate staining in the ventral midline.

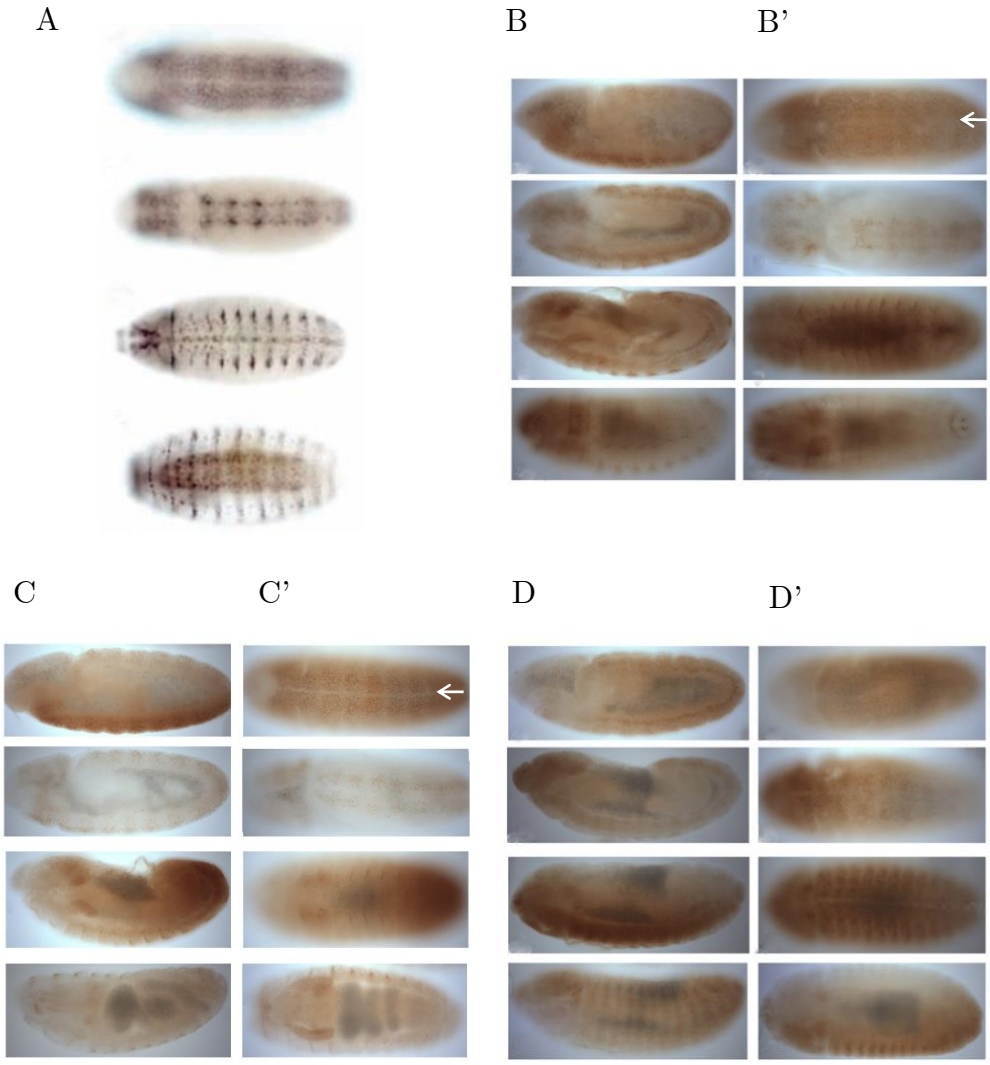


Figure 3.3: SoxNeuro expression patterns in developing embryos from *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. *D. melanogaster* stainings are reproduced from Buescher *et al.* (2002). Images were taken at stages 8, 10, 12 and 16. A.) Ventral views of SoxN expression in *D. melanogaster* embryos. B-D, lateral views; B'-D', dorsal or ventral views. B.) and B'.) SoxN expression in *D. simulans* embryos. C.) and C'.) SoxN expression in *D. yakuba* embryos. D.) and D'.) SoxN expression in *D. pseudoobscura* embryos. Although staining is weaker than for Dichaete, expression patterns are qualitatively the same at the equivalent stages in each species. White arrows in B' and C' indicate lack of expression in the ventral midline at early stages.

I performed ChIP on three biological replicates of chromatin from each species with each antibody. For each replicate, I also performed a mock IP control and set aside an aliquot of input chromatin, resulting in nine samples per TF per species. For each TF, I identified three high-confidence target intervals in *D. melanogaster* as well as one control region where binding was not detectable (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). I then used PCR to amplify the target regions as well as the control region in each sample, testing the enrichment of the ChIP samples in comparison to both the input and the mock IP control samples. The target regions chosen for Dichaete were in regulatory regions of the genes *slit* (*sli*), *achaete* (*ac*) and *commissureless* (*comm*), and the negative control region was near the gene *klingon* (*klg*). For the *D. melanogaster* Dichaete ChIP samples, strong amplification was visible at each target locus, whereas little or no amplification was visible for the mock IP control samples (Figure 3.4). As expected, the input samples also showed amplification at each target region; however, this was not as bright as the amplification present in the ChIP samples. Unexpectedly, amplification was also present at the negative control region for both the ChIP samples and the input samples. As ChIP is an inherently noisy technique, it is difficult to know whether this was due to true, previously undetected binding of Dichaete in this region or to contamination of the ChIP samples by non-specifically bound chromatin (Aleksic and Russell, 2009; Buck and Lieb, 2004).

A similar pattern of enrichment was observed in each other species, with some variation in the strength of amplification at different target regions. In the *D. simulans* ChIP samples, the *sli* and *comm* regions were strongly amplified, while the *ac* region was much weaker, as was the *klg* negative control region. In the *D. pseudoobscura* ChIP samples, the *sli* and *ac* regions were strongly amplified, while the *comm* region showed no amplification. In the *D. yakuba* ChIP samples, amplification was present in all target regions, but at a lower level than for the other species and also at a lower level than the corresponding input samples. However, given that the PCR probes were designed based solely on sequence orthology, without any direct evidence of *in vivo* binding in species other than *D. melanogaster*, this variation was not surprising.

For SoxN, the targets chosen were in regulatory regions of the genes *nervous fingers 1* (*nerfin-1*), *glial cells missing-2* (*gcm-2*) and *castor* (*cas*), and the neg-

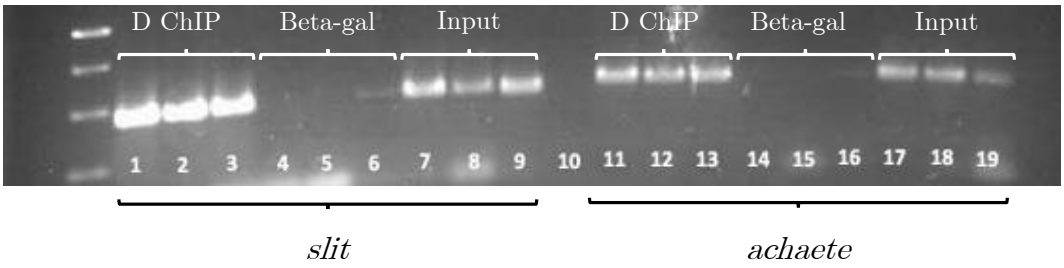


Figure 3.4: ChIP-PCR for Dichaete targets in *D. melanogaster*. Lanes 1-9, PCR for *slit* target region. Lanes 11-19, PCR for *achaete* target region. Lane 10, negative control. Lanes 1-3 and 11-13 are three replicates from Dichaete ChIP DNA, lanes 4-6 and 14-16 are three replicates from mock IP control DNA, and lanes 7-9 and 17-19 are three replicates from input chromatin, which was set aside from each sample before immunoprecipitation. Amplification is strongly visible in all Dichaete ChIP samples for both targets, indicating enrichment of target sequences, while little or no amplification is visible for any controls. As expected, amplification is also visible for input chromatin.

ative control region was near the gene *Protein phosphatase D6* (*PpD6*). However, unlike the Dichaete samples, the SoxN ChIP samples did not show a pattern of significant target enrichment in PCR assays. While amplification of each target region was generally detectable for the input DNA in all species, it was weak and variable in both the ChIP samples and mock IP negative controls. To determine whether this lack of enrichment was due to an inappropriate selection of target loci, I performed PCRs on the same samples with a new set of primers designed to detect regions that had been identified as SoxN targets in an earlier study (Girard *et al.*, 2006) (primers from E. Ferrero). Again, no significant enrichment was observed in experimental samples versus negative controls. For all replicates, less DNA was recovered from the IP reaction for SoxN than for Dichaete, raising the question of whether the lack of target amplification in the SoxN ChIP samples was due to lack of specificity of the antibody or simply an insufficient quantity of template DNA. However, given the results of all of the ChIP-PCR experiments, I decided to focus primarily on Dichaete for performing ChIP-seq.

3.5 Genome-wide binding analysis of Dichaete via ChIP-chip and ChIP-seq

3.5.1 ChIP-chip for Dichaete in *D. melanogaster*

Initially, I used three biological replicates to perform a ChIP-chip experiment for Dichaete in *D. melanogaster* in order to further validate the ChIP-PCR results. Of these three, one produced highly noisy data, and the remaining two suffered from problems during loading of the microarrays. These two replicate ChIP/control pairs were hybridized to new microarrays; however, one of the new arrays leaked during hybridization and had to be discarded. The best resulting ChIP/-control pair was therefore analyzed on its own, using the software tools TiMAT (`http://bdtnp.lbl.gov/TiMAT/`) and Ringo (Toedling *et al.*, 2007). TiMAT found 5444 peaks at FDR1, 9807 peaks at FDR5, 12822 peaks at FDR10, and 19044 peaks at FDR25 for this dataset, while Ringo found 10322 peaks at FDR1, 18724 peaks at FDR5, 23189 peaks at FDR10, and 31915 peaks at FDR25. The TiMAT FDR1 results were chosen for further analysis, as they represent the most stringent and high-confidence dataset, and are also in line with the number of peaks predicted by previous ChIP-chip and DamID experiments for Dichaete (Aleksic *et al.*, 2013).

Each interval in the TiMAT FDR1 dataset was assigned to the closest gene within 10kb upstream or downstream, with intervals that fell an equal distance between two genes being assigned to both. This resulted in 3807 gene assignments. The list of genes was uploaded onto FlyMine (`www.flymine.org`) (Lyne *et al.*, 2007), and the Gene Ontology Enrichment widget was used with a Holm-Bonferroni correction for multiple hypothesis testing to determine the what terms from the biological process ontology were enriched in the putative target genes. Terms with the most significant p-values included organ development (p = 1.19E-43), anatomical structure morphogenesis (p = 5.81E-42), biological regulation (p = 6.14E-37), generation of neurons (p = 9.06E-34) and neuron differentiation (p = 1.82E-30). A graphical overview of biological process enrichments was created using the Ontologizer and the PANTHER GOSlim ontology, which is a subset of the Gene Ontology containing high-level terms (Figure 3.5) (Bauer *et al.*, 2008).

When the TiMAT FDR1 binding intervals were visualized against the genome using IGB (Nicol *et al.*, 2009) binding was observed at known Dichaete targets such as *slit* and *commissureless* (Figure 3.6).

A crude comparison to previous Dichaete binding datasets was performed by taking the intersection of the list of FDR1 gene hits with a list of core Dichaete target genes compiled by J. Aleksic from three ChIP-chip datasets and one DamID dataset, using the intersect tool on FlyMine (Aleksic *et al.*, 2013; Lyne *et al.*, 2007). This intersection contains 1626 genes, representing 43% of the FDR1 gene list and 34% of the core target gene list. An intersection of the list of FDR5 gene hits with the same core target gene list contains 2330 genes, representing 40% of the FDR5 gene list and 48% of the core target gene list. These percentages are within the range of expected values based on pairwise comparisons of previously generated Dichaete binding datasets. Although the data were somewhat noisy and I was only able to use one biological replicate, overall the ChIP-chip results supported the hypothesis that the Dichaete antibody was specifically binding to Dichaete protein and pulling down true target sequences.

3.5.2 ChIP-seq for Dichaete in four species of *Drosophila*

3.5.2.1 Sequencing on the Ion Torrent PGM

In a first attempt at performing ChIP-seq, three biological replicates of Dichaete ChIP samples from *D. melanogaster*, along with three matched replicate mock IP control samples, were sequenced in-house on an Ion Torrent PGM. Difficulties in generating the correct enrichment of templated DNA molecules from each library on the Ion Sphere Particles (ISPs) and in loading the ISPs onto the sequencing chips led to highly variable and low numbers of reads from each run. A summary of the reads generated for each sample can be found in Table 3.1.

Given the insufficient numbers of reads, as well as the high amount of variation in coverage between replicate samples and generally low quality scores obtained, these datasets were not analyzed further. In the hope of getting more reads per sample, I decided to move to Illumina sequencing.

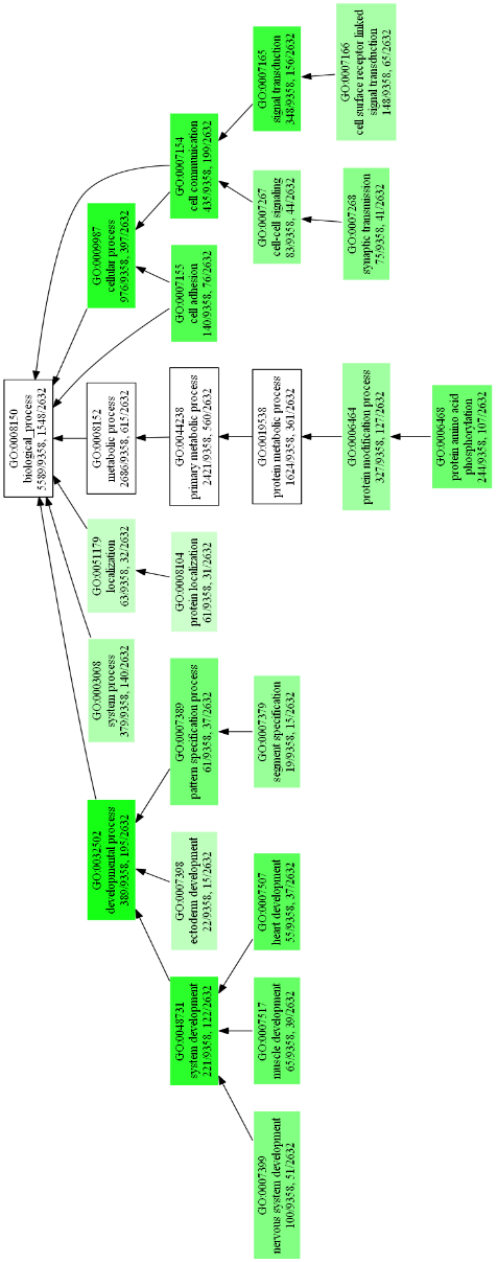


Figure 3.5: Gene Ontology Biological Process GOSlim terms enriched in annotated targets of Dichaete ChIP-chip binding intervals in *D. melanogaster*. All terms highlighted in green are statistically significant ($p < 0.05$) after correction for multiple hypothesis testing, with the intensity of the green correlating to lower p-values. Arrows go from child terms in the ontology to parent terms, which are related by either an is_a or a part_of relationship.

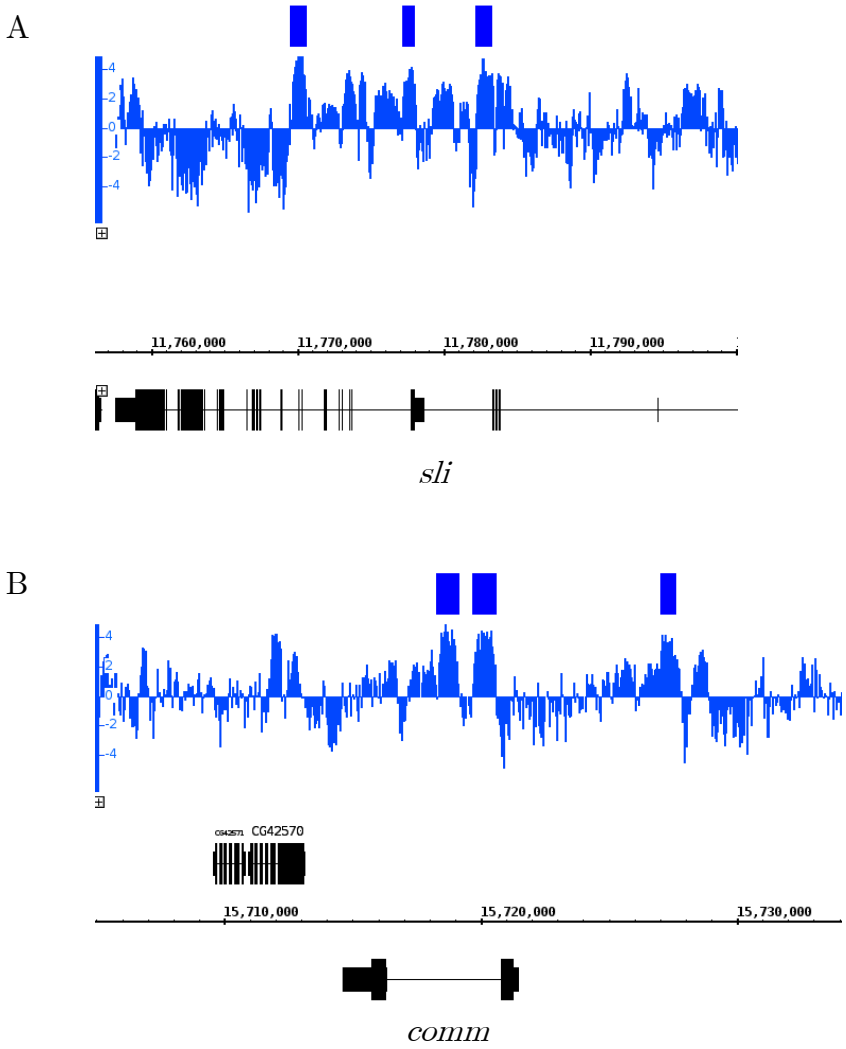


Figure 3.6: Dichaete ChIP-chip binding at known Dichaete targets in *D. melanogaster*. Gene models are in black, the Dichaete ChIP-chip binding profile is in blue and FDR1 Dichaete binding intervals are represented by blue bars above the binding profile. The Dichaete binding profile represents log2 ratio scores of Dichaete ChIP intensity versus mock IP intensity at each probe on the microarray. A.) Dichaete binding in several introns of the gene *slit*, on chromosome 2R. B.) Dichaete binding in the intron and downstream of the gene *comm*, on chromosome 3L.

Sample	Raw reads
Dichaete ChIP 1	64,204
Dichaete ChIP 2	601,960
Dichaete ChIP 3	2,354,296
Mock IP 1	318,257
Mock IP 2	1,210,355
Mock IP 3	85563

Table 3.1: Summary of reads obtained for ChIP-seq libraries on the Ion Torrent PGM

3.5.2.2 Sequencing on the Illumina HiSeq

I attempted to perform ChIP-seq for Dichaete in embryos from all four species of *Drosophila* on the Illumina HiSeq 2000 platform using two different controls: a mock IP and input chromatin. In the first instance, I generated and sequenced matched ChIP-seq and mock IP libraries for one biological replicate of *D. melanogaster* chromatin, two biological replicates of *D. simulans* chromatin and three biological replicates of *D. yakuba* chromatin. All of these samples were multiplexed in one lane and sequenced as 50-bp single end reads. On average, Illumina sequencing resulted in many more reads and higher quality scores for each sample than Ion Torrent sequencing. However, I also observed a very high level of duplication in all samples (Table 3.2). Although some duplication may be expected in a ChIP-seq dataset with very strong enrichment of target sequences, for this dataset the majority of the duplicates were 10-fold or more, indicating that the most likely cause of duplication was PCR overamplification during the library preparation step (Bardet *et al.*, 2011). High duplication was also present in the mock IP control samples; theoretically, these samples should not display significant enrichment, meaning that the duplication was again likely due to PCR overamplification stemming from a very low amount of starting material. After mapping the reads to their respective genomes and visualizing the read densities, it was clear that most of the sequenced reads for the mock IP samples were PCR artefacts, as they formed discrete, high peaks, rather than a random background distribution (Figure 3.7). There was also evidence for contamination by adapter sequences, as a relatively low fraction of reads from each sample mapped uniquely to the genome. I attempted to call peaks for each ChIP-control pair using MACS; while MACS was able to identify some peaks in each sample, the

numbers of peaks ranged widely (from 16 in the *D. melanogaster* sample to 4006 in the *D. yakuba* replicate 1), and in most cases more negative peaks were called than positive peaks.

Sample	**Clean reads**	**Mapped reads**	**% Duplicate reads**
D. mel Dichaete 1	19,504,942	14,022,763	96.6
D. mel mock IP 1	19,334,291	9,887,936	97.0
D. sim Dichaete 1	19,699,015	8,601,165	96.8
D. sim Dichaete 2	15,322,447	6,482,716	96.7
D. sim mock IP 1	19,567,960	9,130,168	96.4
D. sim mock IP 2	22,398,847	7,244,943	96.5
D. yak Dichaete 1	17,031,519	5,227,683	96.1
D. yak Dichaete 2	17,577,632	4,742,744	95.8
D. yak Dichaete 3	19,569,826	5,224,909	96.7
D. yak mock IP 1	8,562,845	5,391,420	96.1
D. yak mock IP 2	17,792,354	8,539,244	95.8
D. yak mock IP 3	20,003,005	13,004,641	97.2

Table 3.2: Summary of reads obtained for ChIP-seq libraries with mock IP controls on the Illumina HiSeq 2000. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba.*

Given these results, I decided to repeat the ChIP experiments and to switch to using input chromatin as controls, as they should be less vulnerable to PCR overamplification. I also changed the strategy for constructing libraries, switching from the NEBNext library kit to the Illumina TruSeq kit, which comes with pre-barcoded adapters, in order to try to decrease adapter contamination. I generated and sequenced matched ChIP and input libraries for three biological replicates each of *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura* chromatin. The samples were multiplexed in two lanes and sequenced as 50-bp single end reads. For some samples, considerably less reads were generated than in the previous sequencing attempt, due to technical variation of the sequencer (Table 3.3). Overall, the rates of duplication improved; however, they were still relatively high, indicating that both adapter contamination and PCR overamplification continued to be problematic.

In order to assess the level of reproducibility between biological replicates, I calculated the Pearson's correlation coefficient (PCC) between the read densities for

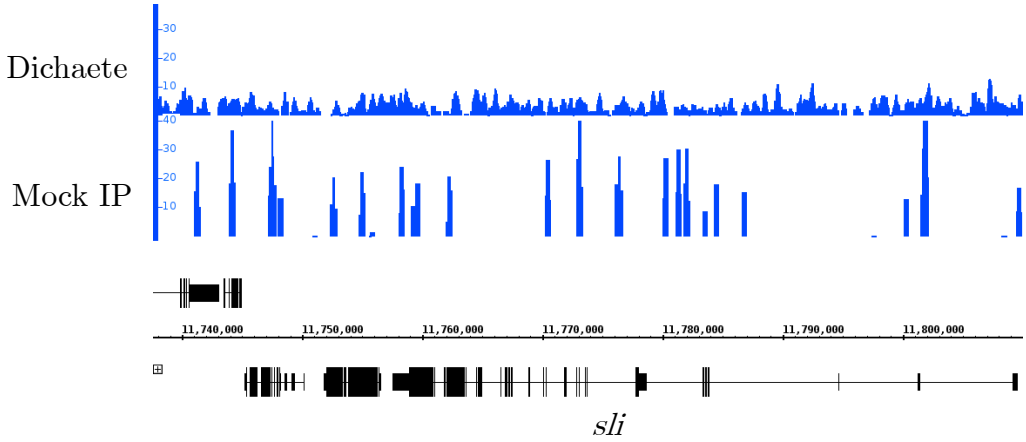


Figure 3.7: Dichaete ChIP-seq reads in *D. melanogaster* with mock IP control. A 70-kb region around the gene *slit* is shown. All samples are scaled to 1,000,000 reads for visualization purposes; the y-axis ranges from 0 to 40. Top track, Dichaete ChIP-seq mapped reads. Bottom track, mock IP control mapped reads. The mock IP reads show a sparser distribution with high, narrow peaks, indicative of PCR overamplification.

each set of replicates for the same condition, using a script provided by Bardet *et al.* (2011). The highest PCCs between pairs of replicates ranged from 0.71-0.85; however, these values are still below the anticipated PCC of >0.9 for high-quality replicate ChIP samples. For many pairs of replicates, particularly in *D. simulans* and *D. yakuba*, the PCCs were considerably lower and more variable, with some close to 0. Nonetheless, I attempted to call peaks on each matched ChIP-input replicate set using MACS as an exploratory analysis. However, in part due to the different amounts of mapped reads between replicates, the numbers of peaks called differed widely between replicates (2-605 for *D. melanogaster*, 66-632 for *D. simulans*, 2113-6636 for *D. yakuba* and 105-4458 for *D. pseudoobscura*). Even between the best-matched replicates in terms of numbers of peaks called, which were *D. pseudoobscura* replicates 1 and 3, only 114 peaks overlapped. Visualizing the read densities for these samples confirms the presence of PCR overamplification artefacts and a low degree of reproducibility between biological replicates (Figure 3.8, top three tracks). Interestingly, the input replicates show a greater apparent concordance than the ChIP samples, indicating that the lack of reproducibility in the ChIP samples is not due to the sequencing process alone (Figure 3.8, bottom three tracks). Replicate 3 of the *D. pseudoobscura* ChIP samples also appears to match the input samples better than it matches the other two ChIP replicates, suggesting that the ChIP itself might have failed in this sample.

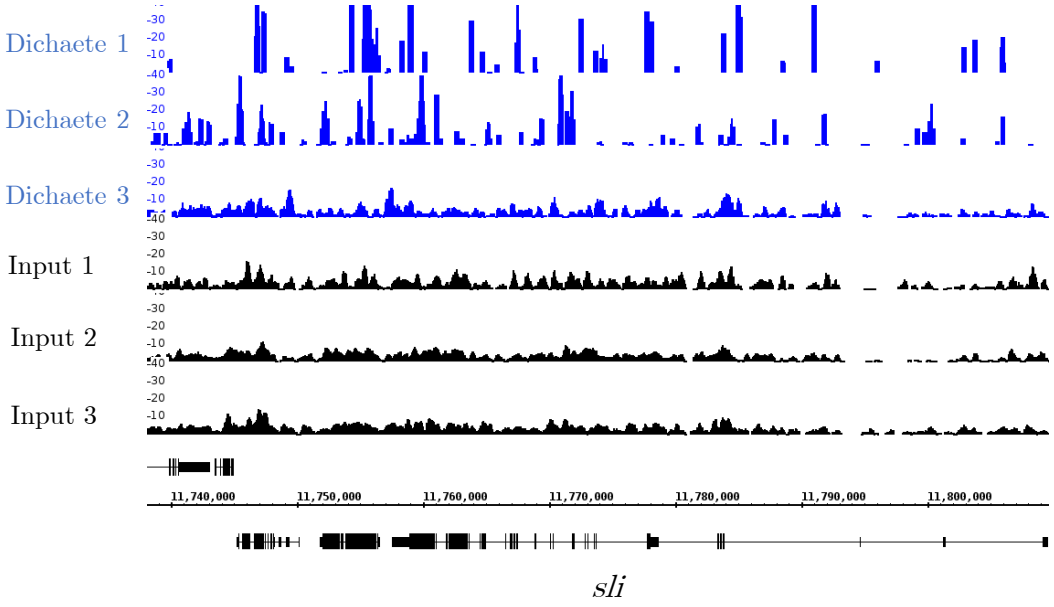


Figure 3.8: Dichaete ChIP-seq reads and input reads from three biological replicates in *D. pseudoobscura*. Top three tracks (blue), Dichaete ChIP-seq mapped reads. Bottom three tracks (black), input mapped reads. The same 70-kb region around the gene *slit* is shown as in Figure 3.7. All samples are scaled to 1,000,000 reads for visualization purposes; the y-axis ranges from 0 to 40. Although some similarities are visible between the three ChIP-seq replicates, the input replicates are clearly more reproducible. ChIP-seq replicates 1 and 2 suffer from PCR overamplification in places, as evidenced by sparse coverage and tall, narrow peaks representing highly duplicated reads. ChIP-seq replicate 3 shows more similarity to the input replicates, suggesting that the ChIP reaction might have failed in this replicate.

Sample	Clean reads	Mapped reads	% Duplicate reads
D. mel Dichaete 1	16,476,257	3,039,886	92.2
D. mel Dichaete 2	3,884,815	869,958	89.9
D. mel Dichaete 3	15,694,790	5,680,372	92.4
D. mel input 1	565,635	178,253	94.4
D. mel input 2	6,873,349	3,687,833	73.7
D. mel input 3	12,277,355	6,619,627	71.2
D. sim Dichaete 1	6,053,616	1,796,028	88.6
D. sim Dichaete 2	6,391,812	1,074,217	86.7
D. sim Dichaete 3	542,281	252,277	71.4
D. sim input 1	1,778,512	798,846	91.0
D. sim input 2	9,951,360	4,607,383	87.5
D. sim input 3	2,277,242	937,074	87.1
D. yak Dichaete 1	3,641,069	1,957,927	70.7
D. yak Dichaete 2	5,227,275	845,896	69.8
D. yak Dichaete 3	12,480,232	1,861,494	93.7
D. yak input 1	9,886,306	5,354,002	84.5
D. yak input 2	3,791,180	1,948,138	42.8
D. yak input 3	7,091,621	3,966,281	82.4
D. pse Dichaete 1	7,552,737	4,519,842	69.9
D. pse Dichaete 2	9,716,950	2,473,247	91.9
D. pse Dichaete 3	3,922,525	1,939,907	83.9
D. pse input 1	19,351,331	11,090,369	80.6
D. pse input 2	18,854,725	8,908,703	69.5
D. pse input 3	13,679,916	8,788,161	89.5

Table 3.3: Summary of reads obtained for ChIP-seq libraries with input controls on the Illumina HiSeq 2000. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura*

3.6 Discussion of results and conclusions

Confirming previous knowledge about group B Sox proteins in *Drosophila*, my exploration of Dichaete and SoxN in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura* showed strong evidence for both sequence conservation and functional conservation at the level of spatial and temporal expression patterns. These basic explorations of the orthologous proteins gave me confidence that I could make meaningful comparisons between the binding patterns of each protein in each species without being concerned that their overall functions had diverged

too widely. It was also interesting to note that in terms of both sequence and expression pattern, each set of orthologs displayed greater similarity amongst themselves than that displayed by the paralogs Dichaete and SoxN within any one species. In the light of previous data showing partial functional compensation and highly similar binding profiles between Dichaete and SoxN in *D. melanogaster* (Ferrero *et al.*, 2014; Overton *et al.*, 2007), these data suggest that certain specific, differentiating functions of the two transcription factors have been conserved throughout the evolution of the *obscura* and *melanogaster* groups of drosophilids and are likely ancestral to their divergence (Russo *et al.*, 1995).

One practical purpose of the immunohistochemistry and ChIP-PCR experiments was to determine the suitability of the antibodies raised against the *D. melanogaster* Dichaete and SoxN proteins for performing ChIP-seq against the orthologous proteins in each other species of *Drosophila* studied. In this respect, I needed to determine whether each antibody reacted specifically with each orthologous protein as well as whether it performed well in ChIP reactions with chromatin extracted from each species, as antibodies that work well for immunohistochemistry do not necessarily work well for ChIP. I was able to show that the first question was the case for both Dichaete and SoxN via immunohistochemistry. Embryos from each species stained with each antibody showed highly similar patterns of expression, and background staining was not substantially higher in the non-*melanogaster* species compared to in *D. melanogaster*, indicating that both antibodies react specifically with orthologous proteins from all the species studied.

The ChIP-PCR experiments gave more mixed results. In the case of Dichaete, the antibody performed well in *D. melanogaster*, yielding greater enrichment for target sequences in ChIP samples than either mock IP or input samples. Its performance was more variable in other species, although it was unknown whether this was due to the antibody or, as discussed above, the fact that the target sequences were chosen without direct evidence for binding in these species. The SoxN antibody was less successful and did not appear to give significant enrichment of target sequences in any species. Despite the fact that this antibody has been previously used in a ChIP-chip experiment in *D. melanogaster* (Ferrero *et al.*, 2014), I decided not to pursue its use in ChIP-seq in the first instance. Instead, I focused on Dichaete for my initial genome-wide binding experiments. As I was still not completely convinced of the Dichaete antibody's specificity in ChIP

experiments, I decided to verify it by first performing a ChIP-chip experiment in *D. melanogaster* and then proceeding to perform ChIP-seq in each species of interest.

Unfortunately, due to technical problems I was only able to analyze one biological replicate of ChIP-chip data. Although this dataset showed promising enrichment for known Dichaete targets, the results did not have statistical confidence, as I was unable to measure biological or technical variability between samples. My ChIP-seq experiments also suffered from a number of technical problems, most notably contamination by adapter sequences and low library complexity due to PCR overamplification. The results that I was able to generate suggested a higher level of variability between replicate ChIP samples than between replicate input samples. ChIP-chip experiments for Dichaete with the same antibody by a previous lab member were quite noisy; it is my hypothesis that this noise was exacerbated by the greater resolution of ChIP-seq (Aleksic, 2011). Having made these observations, and motivated by decreasing time and budget, I decided that the best course of action was not to pursue further ChIP-seq experiments, but rather to focus entirely on DamID-seq as an alternative method of assaying the genome-wide binding patterns of Dichaete and SoxN.

CHAPTER 4

Functional Analysis of *in vivo* Genome-Wide Binding of Dichaete and SoxNeuro

4.1 Experimental motivation and design

After experiencing a number of difficulties in attempting to measure Dichaete and SoxNeuro binding patterns on a genome-wide scale using ChIP-chip and ChIP-seq, I switched the focus of my work to performing comparative DamID experiments for each TF in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*. Unlike ChIP, DamID measures the binding of a transgenic protein expressed alongside the endogenous protein, which does have certain disadvantages. Although a technique for targeted DamID now exists (Southall *et al.*, 2013), most DamID experiments use constructs expressed globally at a low level, removing temporal and spatial specificity. On the other hand, DamID does not depend on the availability of a validated antibody and can therefore be used for any DNA-binding protein. As with ChIP, genome-wide DamID binding patterns can be assayed using either microarrays or high-throughput sequencing. Since

genome tiling arrays for non-model species are not readily available, I elected to use sequencing for my DamID experiments.

The use of appropriate controls and biological replicates is important in both DamID and ChIP experiments to account for the noise inherent in each technique. In DamID, the high affinity of the Dam protein for DNA must be controlled for to prevent the identification of non-specific enrichment. This is usually achieved by expressing a Dam-only construct and comparing the resulting binding patterns with those of TF-Dam fusions (Greil *et al.*, 2006; Vogel *et al.*, 2007). Since biological replicates for the Dam-only control tend to show reproducible peaks in specific genomic regions, rather than a flat distribution of background reads, a differential enrichment analysis strategy can be used to identify true binding peaks in each experimental condition in comparison to the control. As with the ChIP experiments, for DamID I planned to sequence three replicates for each experimental condition (Dichaete-Dam and SoxN-Dam) and three control Dam-only replicates in each species. I was able to do so for all conditions except for SoxN-Dam in *D. pseudoobscura*, as I was unable to generate a transgenic line using the SoxN-Dam construct in this species. For the species in which I generated both Dichaete-Dam and SoxN-Dam samples, I compared the same Dam-only controls against each experimental condition.

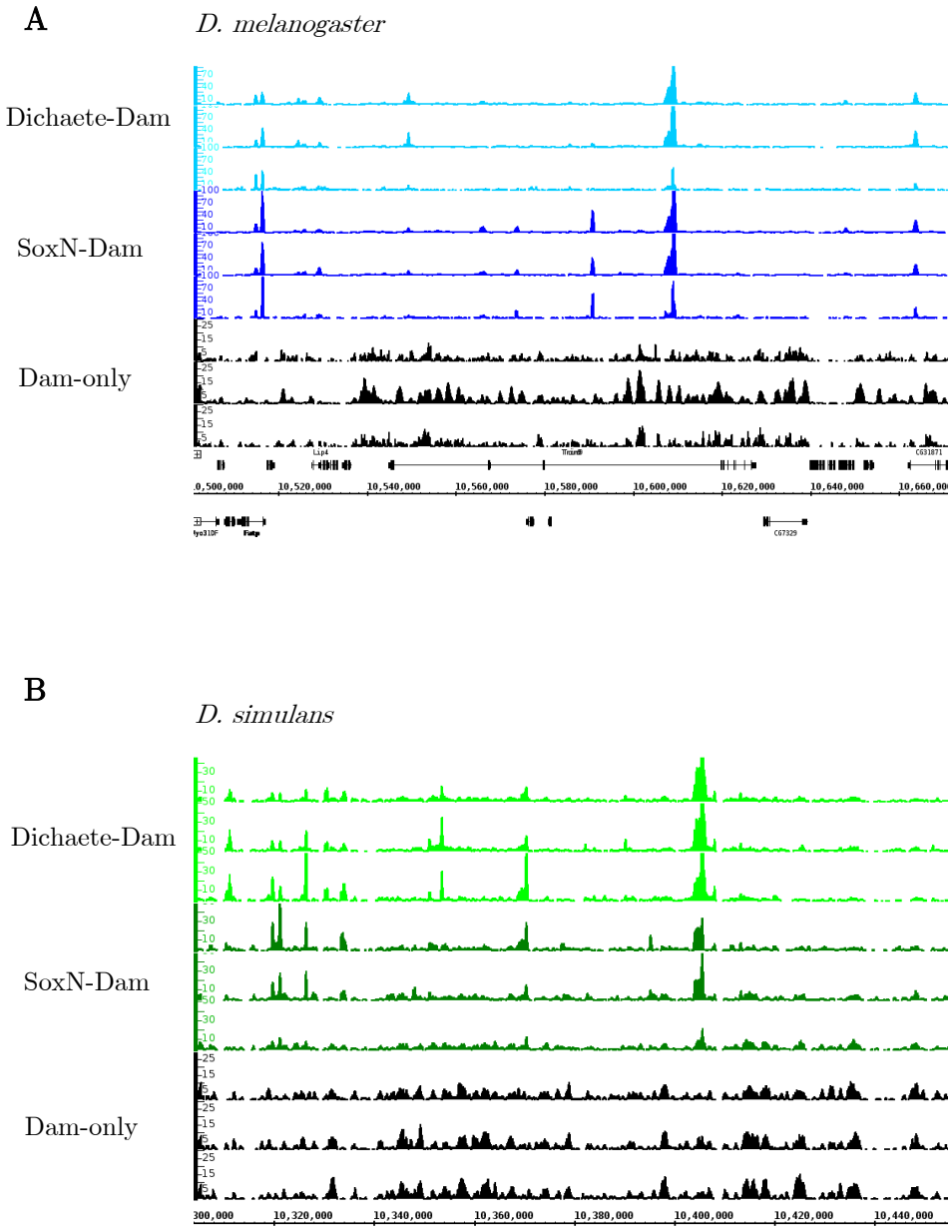
In total, I generated the following genome-wide DamID binding datasets: three replicates of Dichaete DamID-seq (Dichaete-Dam) with three replicates of Dam-only controls in *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*; and three replicates of SoxNeuro DamID-seq (SoxN-Dam) in *D. melanogaster, D. simulans* and *D. yakuba*. Detailed descriptions of the methods used to produce each dataset, including the generation of the transgenic fly lines, can be found in Chapter 2.

4.2 Overview of DamID results

4.2.1 Dichaete and SoxN binding datasets produced in each species

I performed DamID and sequenced the resulting libraries for three biological replicates of Dichaete-Dam, SoxN-Dam and Dam-only in *D. melanogaster, D. simulans* and *D. yakuba* embryos, and three biological replicates of Dichaete-Dam and Dam-only in *D. pseudoobscura* embryos. All embryos were collected after overnight lays, resulting in a mix of embryos from approximately 0-14 hours of age. One replicate each of *D. melanogaster* Dichaete-Dam and SoxN-Dam and two replicates of *D. melanogaster* Dam-only were sequenced as 150-bp single-end reads on an Illumina MiSeq, with two samples multiplexed per run. All other samples were multiplexed with 9-12 samples per lane and sequenced as 50-bp single-end reads on an Illumina HiSeq 2000. All samples showed some duplication, but the rates of duplication were within the expected range for highly enriched samples, with the control samples showing similar or lower rates of duplication than the fusion protein samples (Table 4.1).

In general, I observed a lower rate of unique mapping with the Dam-only samples than with the Sox fusions in all of the species. Upon inspection of unmapped reads, it was apparent that the majority were due to contamination by the DamID adapters. It is possible that the Dam-only controls were more affected by adapter contamination because the Dam-only protein binds to chromatin less frequently than the Sox fusions, resulting in fewer unique *Dpn*I-cut fragments. At the ligation step, this would then result in a greater molarity of adapter molecules relative to DNA fragments, meaning that more adapters could self-ligate and form concatemers that would then be amplified during the PCR step. Nonetheless, the large numbers of reads generated for each sample yielded sufficient depth of coverage for genome-wide binding analysis. Both the Sox fusion samples and the Dam-only samples showed high reproducibility between biological replicates, although the *D. pseudoobscura* samples were noisier and therefore less reproducible than those of the other species (Figure 4.1).


A
D. melanogaster
Dichaete-Dam
SoxN-Dam
Dam-only
10,500,000
10,520,000
10,540,000
10,560,000
10,580,000
10,600,000
10,620,000
10,640,000
B
D. simulans
Dichaete-Dam
SoxN-Dam
Dam-only
10,300,000
10,320,000
10,340,000
10,360,000
10,380,000
10,400,000
10,420,000
10,440,000

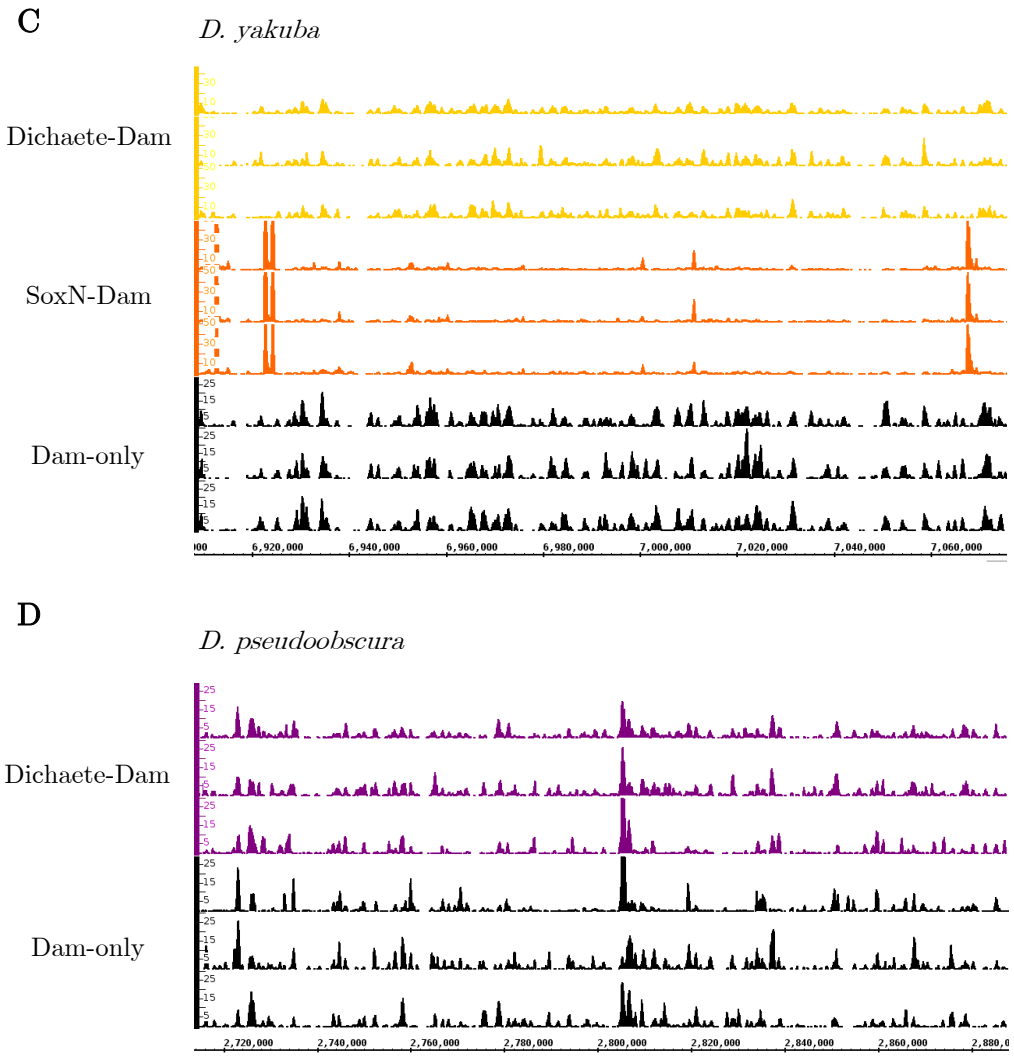


Figure 4.1: Reproducibility of biological replicate DamID samples. A.) *D. melanogaster* samples. Shown is a 180-kb region of chromosome 2L; all other tracks in other species genomes show orthologous regions. The bottom three tracks (black) are the three Dam-only control replicates, the middle three tracks (blue) are the three Dichaete-Dam replicates, and the top three tracks (light blue) are the three SoxN-Dam replicates. B.) *D. simulans* samples. The bottom three tracks (black) are the three Dam-only control replicates, the middle three tracks (green) are the three Dichaete-Dam replicates, and the top three tracks (light green) are the three SoxN-Dam replicates. C.) *D. yakuba* samples. The bottom three tracks (black) are the three Dam-only control replicates, the middle three tracks (orange) are the three Dichaete-Dam replicates, and the top three tracks (light orange) are the three SoxN-Dam replicates. D.) *D. pseudoobscura* samples. The bottom three tracks (black) are the three Dam-only control replicates, while the top three tracks (purple) are the three Dichaete-Dam replicates. All reads are scaled to a total library size of 1 million for visualization purposes. For *D. melanogaster*, the y-axes of the Dichaete-Dam and SoxN-Dam tracks range from 0-100 reads, while for all other species they range from 0-50 reads. The y-axes of all Dam-only tracks range from 0-30 reads in order to show the structure of these samples more closely.

Sample	Clean reads	Mapped reads	% Duplicate reads
D. mel D-Dam 1*	3,524,222	2,957,450	34.3
D. mel D-Dam 2	19,443,486	17,681,725	40.8
D. mel D-Dam 3	18,724,525	16,537,523	39.9
D. mel SoxN-Dam 1*	3,878,298	3,381,641	20.6
D. mel SoxN-Dam 2	18,114,056	15,864,372	43.4
D. mel SoxN-Dam 3	17,125,196	15,799,860	48.9
D. mel Dam 1*	5,165,334	2,198,072	25.2
D. mel Dam 2*	8,699,134	3,379210	28.6
D. mel Dam 3	18,225,579	14,970,090	33.5
D. sim D-Dam 1	11,506,247	9,238,360	38.3
D. sim D-Dam 2	13,729,540	10,492,499	32.2
D. sim D-Dam 3	12,839,381	9,842,133	27.6
D. sim SoxN-Dam 1	10,571,945	8,607,660	40.9
D. sim SoxN-Dam 2	11,933,942	9,300,216	50.9
D. sim SoxN-Dam 3	10,962,128	9,531,855	34.0
D. sim Dam 1	11,156,498	6,711,450	37.9
D. sim Dam 2	12,867,981	9,176644	32.5
D. sim Dam 3	12,351,232	9,719,920	30.0
D. yak D-Dam 1	14,791,084	12,244,800	37.8
D. yak D-Dam 2	13,712,518	11,662,356	44.6
D. yak D-Dam 3	13,483,629	11,018,173	44.4
D. yak SoxN-Dam 1	14,262,567	7,448,087	37.4
D. yak SoxN-Dam 2	13,678,011	10,119,667	26.7
D. yak SoxN-Dam 3	13,781,619	5,824,891	36.6
D. yak Dam 1	12,258,054	7,544,899	29.0
D. yak Dam 2	13,061,238	7,143,573	42.7
D. yak Dam 3	12,433,795	7,937,345	31.7
D. pse D-Dam 1	14,019,105	10,317,759	30.9
D. pse D-Dam 2	17,902,325	12,944,730	24.3
D. pse D-Dam 3	19,617,445	14,659,850	40.0
D. pse Dam 1	19,261,105	12,015,266	45.2
D. pse Dam 2	12,857,397	8,001,867	37.2
D. pse Dam 3	19,170,000	11,769,256	43.3

Table 4.1: Summary of reads for DamID-seq experiments on the Illumina HiSeq 2000 and Illumina MiSeq. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura*; D-Dam, Dichaete-Dam fusion protein; SoxN-Dam, SoxNeuro-Dam fusion protein; Dam, Dam-only control. * Samples sequenced on the MiSeq.

To facilitate the comparison of binding profiles for Dichaete-Dam and SoxN-Dam between *D. melanogaster* and each other species, I used the UCSC LiftOver utility to translate the mapped reads from each of the three non-*melanogaster* species to the *D. melanogaster* genome. It is possible to translate either the reads themselves or called binding intervals; I chose to do the translation at the read level to enable a more quantitative comparison of the binding profiles between species, as translating just the binding intervals results in loss of information about peak height and reproducibility between replicates (Bardet *et al.*, 2011). The translation process inevitably results in the loss of some reads, as not all genomic regions can be reliably translated to a single orthologous region in *D. melanogaster*. The tradeoff between the number of reads successfully translated and the quality of the resulting translated regions can be controlled with the -minMatch parameter, which determines the percentage of base pairs in the original read that must re-map in order for the translated read to be reported. Following the recommendations of Bardet *et al.* (2011), I used a minMatch value of 0.7 for translating *D. simulans* and *D. yakuba* reads to the *D. melanogaster* genome. For *D. pseudoobscura*, in order to account for the increasing phylogenetic distance and improve the percentage of translated reads, I used a minMatch value of 0.5.

The translated reads from each Sox fusion show broad similarities when plotted on the *D. melanogaster* genome, although differences in binding profiles are visible, which increase with evolutionary distance. For Dichaete-Dam, the translated *D. pseudoobscura* reads are considerably noisier and show fewer strong peaks than those of any other species (Figure 4.2A), while for SoxN-Dam, the translated *D. yakuba* reads show fewer strong peaks than those of *D. melanogaster* or *D. simulans* (Figure 4.2B). Table 4.2 shows the number of translated reads for each dataset. Bardet *et al.* (2011) calculated the percentages of all theoretical 36-bp reads that could be mapped in various species and then translated into the *D. melanogaster* genome. The percentages of reads that were translated for my datasets in *D. simulans* and *D. yakuba* were slightly lower than the calculated values (95.85% and 95.97%, respectively); however, they were still high. The percentages of reads that were translated for my *D. pseudoobscura* datasets were on average higher than the calculated value (61.33%); this is likely because Bardet *et al.* used a minMatch of 0.7 for *D. pseudoobscura* read translation, while I used a less stringent value of 0.5.

Sample	Mapped reads	Translated reads	% Translated
D. sim D-Dam 1	9,238,360	8,500,194	92.0
D. sim D-Dam 2	10,492,499	9,482,382	90.4
D. sim D-Dam 3	9,842,133	8,510,191	86.5
D. sim SoxN-Dam 1	8,607,660	7,922,971	92.0
D. sim SoxN-Dam 2	9,300,216	8,608,138	92.6
D. sim SoxN-Dam 3	9,531,855	8,575,810	90.0
D. sim Dam 1	6,711,450	5,833,352	87.0
D. sim Dam 2	9,176644	7,970,590	86.9
D. sim Dam 3	9,719,920	8,647,395	90.0
D. yak D-Dam 1	12,244,800	10,858,592	88.7
D. yak D-Dam 2	11,662,356	10,566,103	90.6
D. yak D-Dam 3	11,018,173	9,990,838	90.7
D. yak SoxN-Dam 1	7,448,087	6,655,088	89.4
D. yak SoxN-Dam 2	10,119,667	8,563,484	84.6
D. yak SoxN-Dam 3	5,824,891	5,039,416	86.5
D. yak Dam 1	7,544,899	6,699,792	88.8
D. yak Dam 2	7,143,573	6,404,045	89.6
D. yak Dam 3	7,937,345	7,067,736	89.0
D. pse D-Dam 1	10,317,759	7,751,401	75.1
D. pse D-Dam 2	12,944,730	9,770,702	75.5
D. pse D-Dam 3	14,659,850	8,428,062	57.5
D. pse Dam 1	12,015,266	9,068,101	75.5
D. pse Dam 2	8,001,867	6,119,021	76.5
D. pse Dam 3	11,769,256	8,974,292	76.3

Table 4.2: Reads translated from each sample in a non-*melanogaster* species into the *D. melanogaster* dm3 genome assembly. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura*; D-Dam, Dichaete-Dam fusion protein; SoxN-Dam, SoxNeuro-Dam fusion protein.

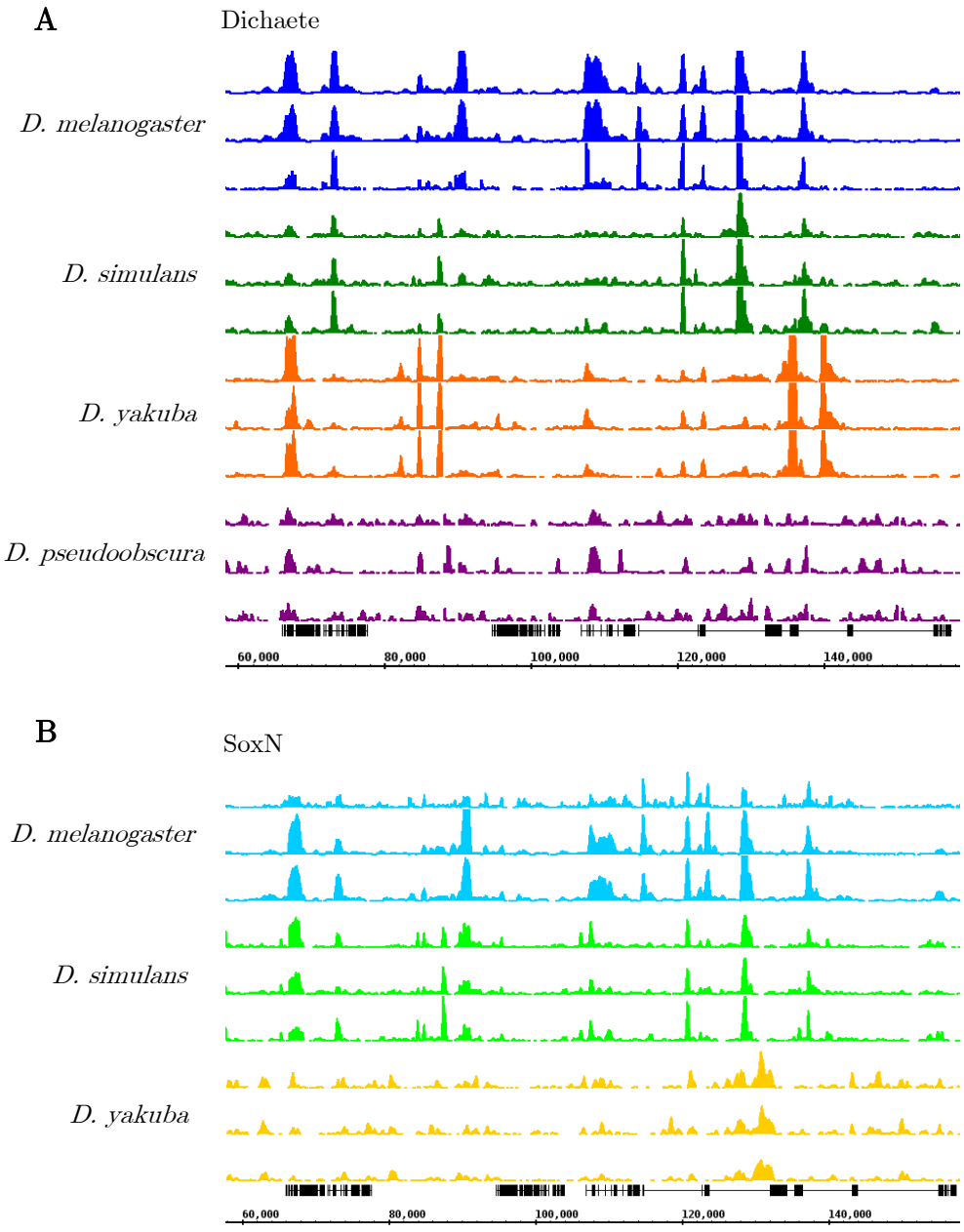


Figure 4.2: Translated reads for Sox fusion proteins in all species. A.) Dichaete-Dam reads from all biological replicates in *D. melanogaster* (blue), *D. simulans* (green), *D. yakuba* (orange) and *D. pseudoobscura* (purple) are plotted on the *D. melanogaster* genome. B.) SoxN-Dam reads from all biological replicates in *D. melanogaster* (light blue), *D. simulans* (light green) and *D. yakuba* (light orange) are plotted on the *D. melanogaster* genome. All reads are scaled to a total library size of 1 million after translation (if necessary) for visualization purposes. The y-axes of all tracks range from 0-50 reads. The same region of chromosome 2L is shown in both A and B.

In order to detect regions of enriched binding by the Sox fusions in comparison to the Dam-only controls, I identified all GATC sites in each genome and then counted the number of reads mapping to each GATC fragment for each replicate. I then used DESeq2 to test for differential enrichment of Sox fusion reads versus the controls in each GATC fragment (Love *et al.*, 2014). The log2 ratios between Sox fusion read counts and control read counts in each fragment represent normalized binding scores for each fusion protein, which were used in downstream analyses. Enriched fragments (adjusted p <0.05) that were within 100 bp of each other were merged to create peaks or binding intervals; it should be noted that these binding intervals are different from ChIP peaks, as they are based on the distribution of GATC fragments across the genome. Because DESeq2 uses the variance observed between biological replicates to evaluate confidence for each potentially enriched interval, noisier data will result in fewer enriched binding intervals being called. This effect can be seen in the different numbers of binding intervals called for Dichaete-Dam in *D. melanogaster, D. simulans* and *D. yakuba* in comparison to *D. pseudoobscura*, which had much noisier data. In *D. yakuba*, although a high number of binding intervals were called for Dichaete-Dam, the SoxN-Dam experiment failed to detect significant binding. This result was surprising, since indications from the preliminary work seemed to show that both experiments worked equally well, and the same SoxN-Dam fusion protein showed high levels of binding in *D. melanogaster* and *D. simulans*. Visual inspection of the binding profiles for each fusion protein in *D. yakub*a revealed that the SoxN-Dam replicates showed the same binding behavior as the Dam-only replicates (Figure 4.1C), suggesting that a mutation in the *SoxN* sequence may have rendered the protein nonfunctional. All further analysis in *D. yakuba* was performed with the Dichaete-Dam data only.

In *D. melanogaster*, I also identified enriched intervals at a more stringent threshold, with an adjusted p-value <0.01. For comparative analyses between species, I used the binding intervals identified at p <0.05 in all species; however, for all subsequent functional analyses within *D. melanogaster*, I used the more stringent p <0.01 intervals. The numbers of binding intervals called for each fusion protein in each species can be seen in Table 4.3.

For each non-*melanogaster* species, I also used the same procedure with the reads that had been translated to the *D. melanogaster* genome assembly to detect dif-

Species	**Dichaete-Dam**	**SoxNeuro-Dam**
D. melanogaster p <0.05	20848	22952
D. melanogaster p <0.01	17530	17833
D. simulans p <0.05	17833	17209
D. yakuba p <0.05	26563	681
D. yakuba p <0.01	21988	233
D. pseudoobscura p <0.05	2951	N/A

Table 4.3: Enriched binding intervals with indicated adjusted p-values called by DESeq2 for each fusion protein in each species. DamID for SoxNeuro-Dam was not performed in *D. pseudoobscura*.

ferential binding by each Sox fusion in comparison to the Dam-only controls and to identify binding regions. This strategy resulted in binding intervals that were directly comparable to those identified in the *D. melanogaster* DamID data; however, the total number of significantly enriched binding intervals called in the translated data decreased slightly in comparison to the number of binding intervals called in each species before translating reads (Table 4.4). I performed a crude pairwise comparison of the binding intervals detected in each species by intersecting the intervals called in *D. melanogaster* with the intervals called in the translated data from each other species. Because more binding intervals were called in *D. melanogaster* in most cases, the percentages of binding intervals present in *D. melanogaster* that overlap with binding intervals in other species are generally lower than the percentages of binding intervals in each other species that overlap with binding intervals in *D. melanogaster*. In *D. yakuba*, a similar number of binding intervals were called for Dichaete-Dam as in *D. melanogaster*. Accordingly, the percentages of overlapping intervals are very close for the two reciprocal comparisons. Considering the binding intervals in each non-*melanogaster* species that overlap with intervals in *D. melanogaster*, the percentages of shared intervals decrease with increasing phylogenetic distance, as expected (He *et al.*, 2011b; Paris *et al.*, 2013).

Sample	Binding intervals	Overlaps with *D. mel* binding intervals	% of *D. mel* intervals overlapping	% of non-*D. mel* intervals overlapping
D. sim D-Dam	16119	11647	55.9	72.3
D. sim SoxN-Dam	15142	11891	51.8	78.5
D. yak D-Dam	20964	14573	69.9	69.5
D. pse D-Dam	2020	1301	6.24	64.4

Table 4.4: Overlaps between binding intervals called on *D. melanogaster* data and binding intervals called on translated read data. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura*; D-Dam, Dichaete-Dam fusion protein; SoxN-Dam, SoxNeuro-Dam fusion protein.

4.3 Functional analysis of binding patterns in each species

4.3.1 Overlap between DamID-seq binding intervals and core Sox binding intervals

Previous work in the lab has generated a set of core binding intervals for both Dichaete and SoxN, based on a conservative integration of several ChIP-chip and DamID datasets for each transcription factor (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). In order to assess how well my binding data concur with these core intervals, I determined the overlaps between each of these datasets and the high-stringency DamID-seq binding intervals that I generated for each protein in *D. melanogaster* (Table 4.5). Since the core interval datasets were the result of high-confidence overlaps between other datasets, they include fewer intervals than the DamID-seq datasets and should be more conservative. Accordingly, there is a higher proportion of core intervals that are overlapped by a DamID-seq interval than DamID-seq intervals that are overlapped by a core interval. While the levels of overlap between my binding intervals and the Dichaete core intervals were reasonably good, for SoxN they were considerably lower. However, the SoxN core intervals are, on average, shorter than the Dichaete core intervals, which could artificially lower the agreement between the datasets, as nearby but slightly off-

set binding intervals are less likely to actually overlap with small core intervals than large ones. Additionally, both the SoxN and Dichaete core intervals are derived from experiments using embryos that were collected over a narrower range of developmental stages than the DamID-seq experiments (stages 8-11 for SoxN DamID-chip, stages 7-10 and 11-13 for SoxN ChIP-chip, stages 5-11 for Dichaete DamID-chip, and stage 4-5 and 0-11 for Dichaete ChIP-chip), meaning that some binding sites detected by DamID-seq may not have been bound or accessible during the stages represented in the core intervals.

	Core intervals	**Core intervals containing DamID-seq intervals**	**DamID-seq intervals containing core intervals**
Dichaete core	6720	4046 (60.2%)	3774 (21.5%)
SoxN core	5482	1893 (34.5%)	1683 (9.8%)

Table 4.5: Overlaps between DamID-seq binding intervals and core binding intervals for Dichaete and SoxN in *D. melanogaster* (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014).

4.3.2 Enriched motifs in binding intervals

I performed a motif analysis on the binding intervals to identify enriched sequence motifs relative to the genomic background using HOMER (Heinz *et al.*, 2010). A *de novo* motif analysis, which searches for any enriched 8-, 10- or 12-mers within binding intervals, uncovered a highly significantly enriched Dichaete motif (p = 1e-64) in the *D. melanogaster* Dichaete-Dam intervals (Figure 4.3A) (Aleksic *et al.*, 2013). Although the Dichaete motif was strongly enriched, it was the 14th ranked motif by p-value. The top 20 motifs identified by HOMER are listed in Table 4.6; these include a motif matching that of Ventral veins lacking (vvl, p = 1e-63), which is a known Dichaete cofactor in the midline (Aleksic *et al.*, 2013; Sánchez-Soriano and Russell, 1998).

A *de novo* motif analysis on the high-stringency SoxN-Dam intervals uncovered a significantly enriched match to a Dichaete motif (p = 1e-33), which was ranked 22nd by p-value. Additionally, the two top-scoring motifs were for Pangolin (Pan), a transcription factor that also contains an HMG box DNA binding domain. When I performed the same analysis on the high-stringency intervals before

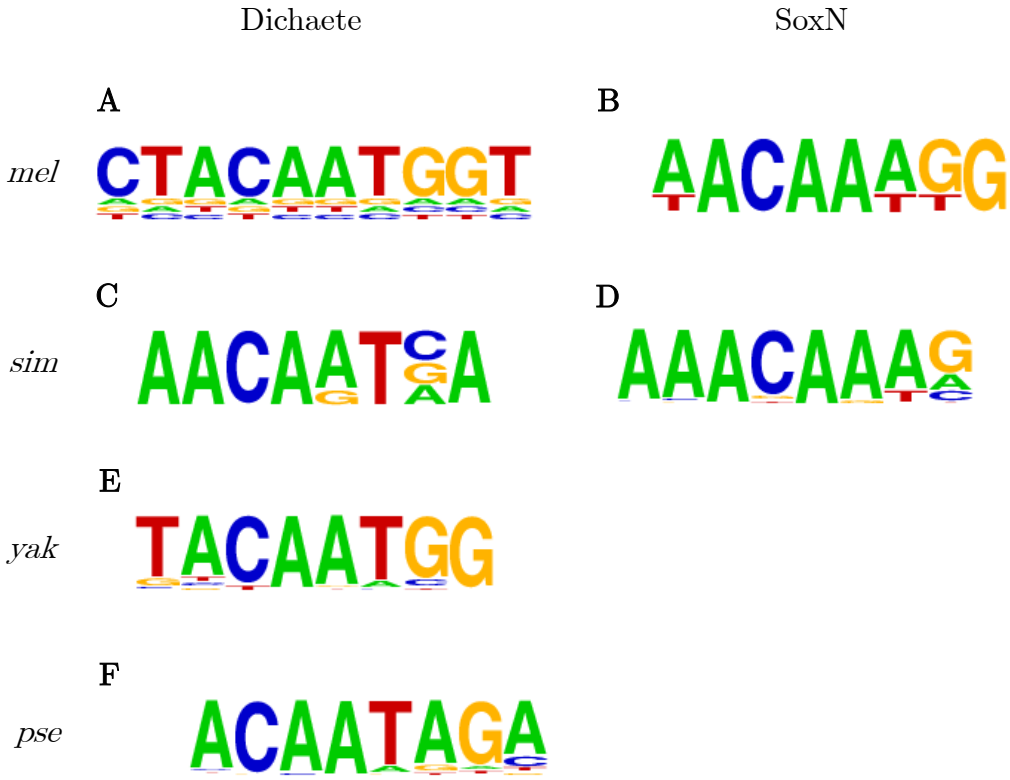


Figure 4.3: *De novo* Sox motifs discovered in DamID binding intervals. A.) Sox motif discovered in *D. melanogaster* Dichaete-Dam intervals. B.) Sox motif discovered in *D. melanogaster* SoxN-Dam intervals. C.) Sox motif discovered in *D. simulans* Dichaete-Dam intervals. D.) Sox motif discovered in *D. simulans* SoxN-Dam intervals. E.) Sox motif discovered in *D. yakuba* Dichaete-Dam intervals. F.) Sox motif discovered in *D. pseudoobscura* Dichaete-Dam intervals.

merging, so that the same sequences were considered but they were broken up into a greater number of fragments, a stronger match to a Dichaete motif was discovered, ranking 9th by p-value (Figure 4.3B, p = 1e-84); this motif was present in 16.8% of target sequences, a far greater percentage than for most other enriched motifs. It is also highly similar to the SoxN motif reported in an independent DamID experiment in *D. melanogaster* (Ferrero *et al.*, 2014). The top 20 motifs identified by HOMER in the unmerged intervals are listed in Table 4.7. For both Dichaete-Dam and SoxN-Dam, the top motif was predicted by HOMER to match Kni; however, since this motif contains the sequence GATC in both cases, it is likely the presence of this motif is due to the fact that the interval boundaries were determined by GATC sites, rather than reflecting true Kni binding.

Rank	Transcription Factor	Consensus Sequence	P-value
1	Kni	GATCHAWT	1E-124
2	Ftz	CCAAGGAGACCG	1E-116
3	Run	TTGYGGCTACAW	1E-104
4	Tin	TCCACCCGAAAT	1E-096
5	Prd	TAGACGGTCT	1E-094
6	Sd	ACTCCATTTTGC	1E-094
7	Nub	TCCTTTGSATDT	1E-093
8	Usp	CGGGGTCAACTA	1E-092
9	Su(H)	AGAATGTGAGTA	1E-091
10	CG11617	TTTACATCCAGA	1E-086
11	Br	TCTATTTCTATA	1E-078
12	Kni	CGACCCSGTTTW	1E-078
13	Onecut	ATTTAATCAATG	1E-072
14	D	CTACAATGGT	1E-064
15	Tag	TCTAACTYCA	1E-064
16	Vvl	ACTATCCACC	1E-063
17	Med	TCYCCGKCTGKC	1E-054
18	Abd-B	GGTGGCCATSMA	1E-051
19	Tin	TGAACTCTTGAT	1E-050
20	Btd	TGGAGGCBGAAT	1E-048

Table 4.6: Top 20 *de novo* motifs identified in p <0.01 *D. melanogaster* Dichaete-Dam intervals. The best match transcription factor, the consensus sequence, and the p-value are shown for each motif.

For all other species, I used HOMER to search for motifs in the binding intervals called in the original genomes, rather than in the intervals called after translation to the *D. melanogaster* genome. I used the sequences from each original genome because they contained the sites to which the fusion proteins actually bound *in vivo*, and also because any differences in the enriched motifs found between species might illustrate general differences in enhancer composition. In the *D. simulans* binding intervals, a *de novo* motif analysis of the Dichaete-Dam data uncovered a significantly enriched motif (p = 1e-20) matching the Sox consensus binding sequence (Figure 4.3C) ranked in the 23rd position by p-value. Other highly-ranked motifs included matches to Tll (p = 1e-30) and Kni (p = 1e-29). Additionally, a search for known motifs resulted in three hits to Sox motifs, corresponding to the vertebrate Sox2 (p = 1e-12), Sox3 (p = 1e-15) and Sox6 (p = 1e-13) motifs. A *de novo* motif analysis of the *D. simulans* SoxN-Dam binding

Rank	Transcription Factor	Consensus Sequence	P-value
1	Kni	ATCCGATC	1e-859
2	Tll	TTGCAACGTTAA	1E-162
3	Twi	CGCATATGCGAT	1E-143
4	Trl	AGAGTAGTKCCA	1E-135
5	Eip74EF	GGGAGAATTHTG	1E-127
6	Brk	MGTGCCSC	1E-112
7	B-H2	TGCCTATTAAST	1E-101
8	Slp1	GTCAATATTTAC	1E-090
9	D	CCTTTGTT	1E-084
10	Ap	GCCGCTAATCAG	1E-082
11	Hb	TTTTTTTTTTTT	1E-082
12	Med	CATAYTGCGS	1E-074
13	TATA-box	TTATAGGGAG	1E-073
14	Antp	YATAWTATRGGN	1E-072
15	Bap	TCTTGTTTAAGT	1E-071
16	Slbo	CTCWGTTGCTTG	1E-070
17	Antp	ATTCTGATTTGT	1E-060
18	Kni	AWATGGATCCAT	1E-057
19	Cad	CATAAAGA	1E-053
20	Trl	CACGACAGAG	1E-053

Table 4.7: Top 20 *de novo* motifs identified in unmerged p <0.01 *D. melanogaster* SoxN-Dam intervals. The best match transcription factor, the consensus sequence, and the p-value are shown for each motif.

intervals discovered a Sox motif as the 2nd-ranked enriched motif (Figure 4.3D, p = 1e-37), and a search for known motifs also resulted in hits for the vertebrate Sox2 (p = 1e-24), Sox3 (p = 1e-32) and Sox6 (p = 1e-25) motifs. Other high-ranking motifs included Pan (p = 1e-37) and Zeste (z, p = 1e-23), with matches to Vvl (p = 1e-17) and Tll (p = 1e-15) being found further down the list, at positions 24 and 27.

In *D. yakuba*, a *de novo* motif search of the Dichaete-Dam binding intervals uncovered a significantly enriched motif (p = 1e-39) matching the vertebrate Sox2 motif (Figure 4.3E), which was the 4th ranked motif by p-value and was present in 12.01% of target sequences. Other high-ranked motifs included Glial cells missing (Gcm, p = 1e-39), Tll (p = 1e-35), Slow border cells (Slbo, p = 1e-36), Ventral nervous system defective (Vnd, p = 1e-35) and Vvl (p = 1e-35). Additionally, a search for known motifs identified a Sox6 motif (p = 1e-21) and a

Sox3 motif (p = 1e-17) as the top two hits, with a Sox2 motif as the 6th-ranked hit (p = 1e-9).

Although a relatively small number of binding intervals were called for Dichaete-Dam in *D. pseudoobscura*, they still show enrichment for Sox motifs. A *de novo* motif search identified a significantly enriched motif (p = 1e-8) whose best match is to the vertebrate Sox9 motif (Figure 4.3F). Performing the same *de novo* search on the binding intervals before merging also uncovered a Sox9 motif; however, in this case the motif was a stronger match and had a lower p-value (p = 1e-21), as well as being present in a greater proportion of target sequences (29.53% versus 1.16% for the merged intervals). Other high-ranked motifs included Deformed epidermal autoregulatory factor-1 (Deaf1, p = 1e-32), Twi (p = 1e-23), Vnd (p = 1e-21) and Trl (p = 1e-18). For both the merged and unmerged intervals, a search for known motifs turned up the vertebrate Sox2 (p = 1e-7), Sox3 (p = 1e-6) and Sox6 (p = 1e-5) motifs as the top three hits.

The binding site analysis supports the view that the DamID experiments have identified genomic regions bound by Dichaete and SoxNeuro *in vivo*. Unlike ChIP-seq peaks, DamID-seq binding intervals are dependent on the distribution of GATC sites across the genome; as a result, they may contain flanking sequences that are not relevant to the factor-specific binding sites and thus lower target motif enrichments are expected. Although some sets of binding intervals, such as those for SoxN-Dam in *D. simulans*, show higher enrichment for Sox motifs than others, the fact that at least one Sox motif was identified in each dataset is encouraging. It is known that Dichaete and SoxN recognize very similar sequence motifs to the canonical Sox motif, A/T A/T CAAAG, a fact which is supported by my data (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). However, when comparing across species, certain patterns of preferences become visible; most clearly, in all four species assayed, the Sox motif found in Dichaete binding intervals contains a stronger match to a thymine residue at the fourth position of the core CAAAG motif, while each Sox motif found in SoxN binding intervals contains a stronger match to an adenine residue in that position (Figure 4.3). Additional motifs for transcription factors that have been shown to interact with both Dichaete and SoxN, such as Vvl, Vnd, Tll and Nub (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Sánchez-Soriano and Russell, 1998), were also discovered in binding intervals, as well as motifs for other developmentally-important TFs. Finally, the motif for

Trl, which was enriched in several binding datasets and has previously been shown to be associated with Dichaete binding (Aleksic *et al.*, 2013), is also a signature of highly occupied target (HOT) regions in the *Drosophila* genome, which are regions at which many TFs are reported to bind (Kvon *et al.*, 2012). Taken together, these results suggest that the Dichaete-Dam and SoxN-Dam binding intervals identify active and developmentally-relevant enhancers.

4.3.3 Gene and genomic annotation of binding intervals

I assigned each binding interval to the closest gene in the *D. melanogaster* genome within 10 kb upstream or downstream, using the intervals called on translated reads for all non-*melanogaster* species and the gene annotations from FlyBase release 5.55. A summary of the numbers of genes annotated to each dataset, as well as the number of genes that were commonly annotated to Dichaete-Dam binding intervals and SoxN-Dam binding intervals in *D. melanogaster* and *D. simulans*, can be seen in Table 4.8. In both species, a high percentage of all target genes are shared between the two TFs, although this percentage is slightly higher in *D. melanogaster* than in *D. simulans* (89% of Dichaete-Dam targets and 88% of SoxN-Dam targets versus 80% of Dichaete-Dam targets and 84% of SoxN-Dam targets) (Appendix A).

Combining previous ChIP-chip and DamID experiments with gene expression experiments has allowed us to identify direct targets of both Dichaete and SoxN in *D. melanogaster*; for Dichaete, this includes 1373 genes, while for SoxN it includes 544. I determined the number of these direct targets that are included in the genes annotated to each DamID-seq binding dataset. I also found the overlaps between genes annotated to the core binding intervals for each TF and the genes annotated to each DamID-seq dataset. A high percentage (70-90%) of both direct target genes and core interval genes are included in the gene annotations for Dichaete-Dam and SoxN-Dam binding intervals in all species but *D. pseudoobscura*. In both *D. melanogaster* and *D. simulans*, a slightly higher percentage of direct target and core interval genes are also annotated to the Dichaete-Dam binding intervals than to SoxN-Dam binding intervals. Among all the Dichaete-Dam datasets, the genes annotated to intervals translated from

D. yakuba contain the most core interval genes and direct target genes; this is likely because the *D. yakuba* Dichaete-Dam intervals had the largest number of gene annotations overall. These results shows good concordance between both the core and direct target genes for each TF and the DamID-seq target genes in all species, with the exception of *D. pseudoobscura*, which had a much lower number of binding intervals and target genes.

Dataset	**Genes annotated to binding intervals**	**Genes common to Dichaete and SoxN**	**Core interval genes annotated to binding intervals**	**Direct target genes annotated to binding intervals**
D. mel D-Dam	9400	8445	3433/4279 (80.2%)	1173/1373 (85.4%)
D. mel SoxN-Dam	9528	8445	2410/3246 (74.2%)	434/544 (80.0%)
D. sim D-Dam	9383	7524	3228/4279 (75.4%)	1111/1373 (80.9%)
D. sim SoxN-Dam	8948	7524	2326/3246 (71.7%)	412/544 (80.0%)
D. yak D-Dam	12192	N/A	3765/4279 (88.0%)	1249/1373 (91.0%)
D. pse D-Dam	1888	N/A	978/4279 (22.9%)	407/1373 (29.6%)

Table 4.8: Gene annotations for DamID-seq binding intervals. Numbers of genes common to Dichaete-Dam and SoxN-Dam are within each species. For core interval genes and direct target genes, shown are numbers annotated to intervals in each DamID-seq dataset over total number of core or direct target genes. Abbreviations: *D. mel, Drosophila melanogaster; D. sim, Drosophila simulans; D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura*; D-Dam, Dichaete-Dam fusion protein; SoxN-Dam, SoxNeuro-Dam fusion protein.

I performed enrichment analyses for Gene Ontology Biological Process (GO:BP) terms annotated to the genes hit by each TF in each species. In line with previous studies of Dichaete and SoxN binding (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014), targets of both TFs were highly enriched for general terms relating to organ and system development ($p < 1e\text{-}47$) and biological regulation ($p < 1e\text{-}44$). They were also both enriched for imaginal disc development ($p < 1e\text{-}19$), generation of neurons ($p < 1e\text{-}29$) and regulation of transcription ($p < 1e\text{-}9$). Enriched terms for

both proteins were highly similar across species (Appendix B). Notably, although there were many less genes hit by Dichaete-Dam in *D. pseudoobscura* than in the other species, the genes that were hit showed strong enrichment for similar GO:BP terms as in the other species, were strongly upregulated in the brain and larval CNS, and were highly associated with publications describing genes involved in the neural stem cell transcriptional network (p <1e-25), all of which are hallmarks of known Dichaete function (Aleksic *et al.*, 2013; Shen *et al.*, 2013; Sánchez-Soriano and Russell, 1998).

Using the same gene assignments, I calculated the distances between the end of each binding interval that was closest to each assigned gene and the transcription start site (TSS) of each gene to which it was assigned. It should be noted that, because DamID is dependent on the non-random distribution of GATC sites across the genome, it is difficult to know where a true binding site is located within a binding interval. Consequently, the distances between binding intervals and genomic features may not always accurately reflect the distances between the actual binding sites and those intervals. Nonetheless, they give an overview of where the DamID-fusion proteins bind relative to genes. I plotted these distances using the ChIPseeker R package (Yu, 2014) (Figure 4.4).

The overall distribution of distances is very similar for each transcription factor in each species. Although around 30% of binding intervals are located within 1 kb up- or downstream of TSSs, there is a clear skew towards downstream locations, suggesting a high amount of binding to regulatory regions in introns. Approximately 75% of binding intervals are located within 10 kb of the TSS in either direction, with the remaining 25% being located more distally downstream. I also used a custom pipeline in R (CHIPPAVI, Bettina Fischer) to plot the probability density of bound nucleotides around all TSSs; these plots used gene annotations following the same behavior as previously but with no upper limit on the distance between an interval and the closest nearby gene (Figure 4.5). These plots show a strong maximum at the TSS, with a skew towards downstream binding, although the skew is less pronounced than with the ChIPseeker plots. The distribution of binding around TSSs is very similar for both Dichaete and SoxN and across all species, although the plots for Dichaete-Dam in *D. pseudoobscura* are less smooth due to the lower number of binding intervals.

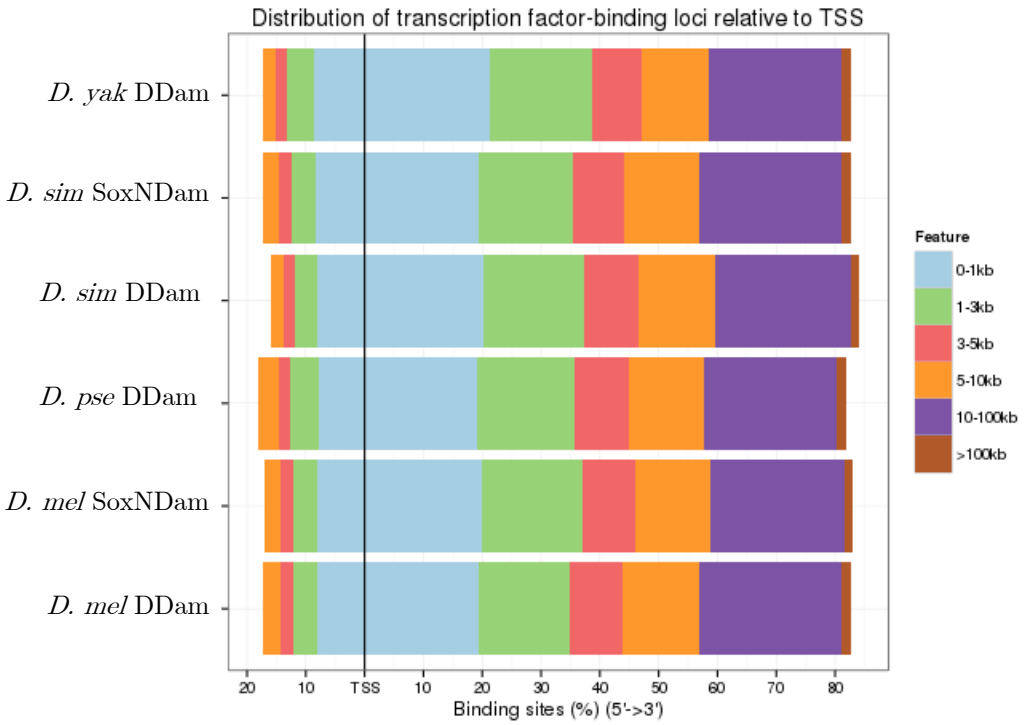


Figure 4.4: Distribution of Sox DamID binding intervals around TSS of annotated genes for each dataset. Distances were calculated from the end of each interval closest to its annotated gene (upstream or downstream). Sox DamID binding shows a clear skew towards positions downstream of TSSs.

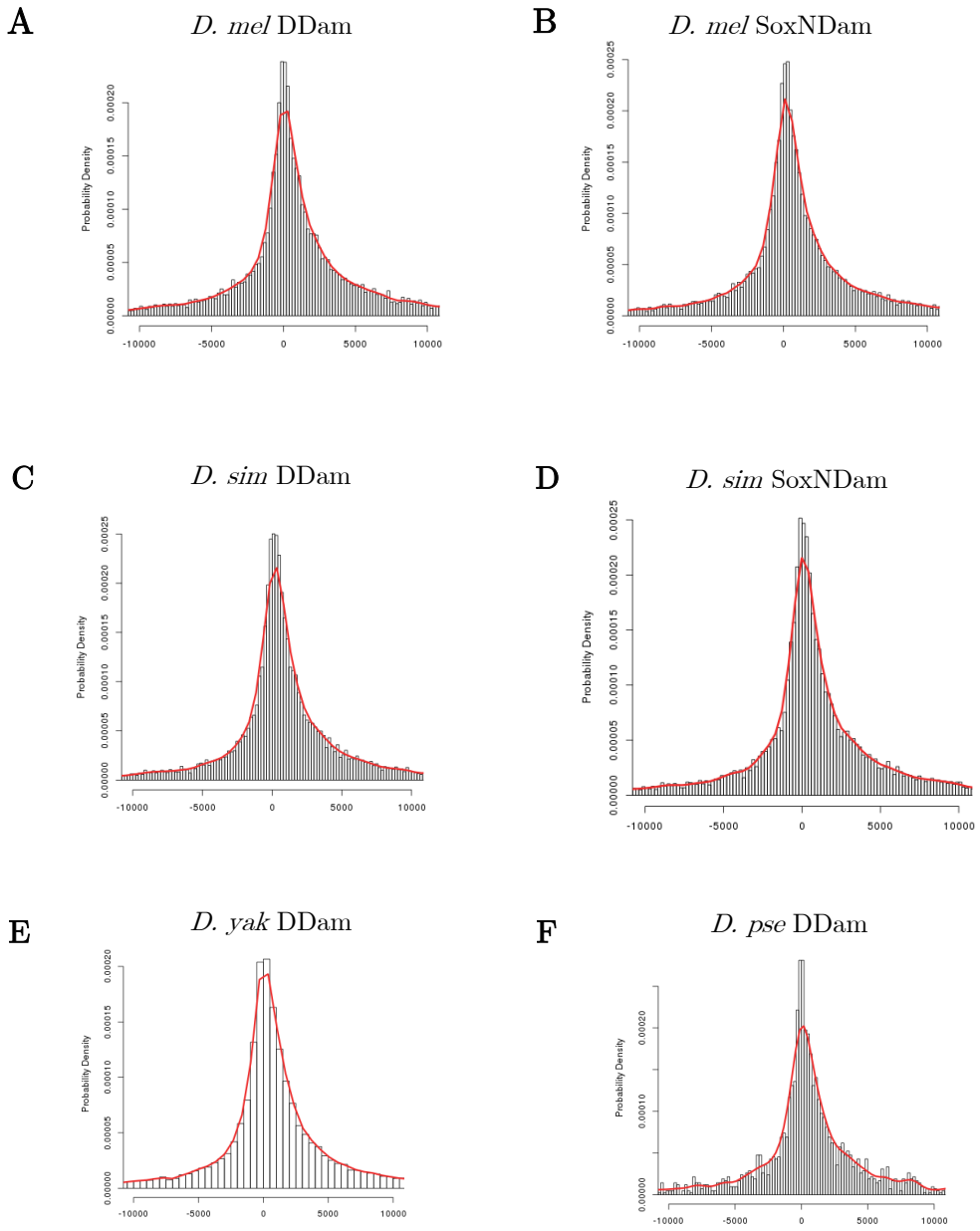


Figure 4.5: Distribution of Sox DamID binding intervals around TSSs for all genes in the *D. melanogaster* genome. A.) Probability density of bound nucleotides around TSSs for *D. melanogaster* Dichaete-Dam. B.) Probability density of bound nucleotides around TSSs for *D. melanogaster* SoxN-Dam. C.) Probability density of bound nucleotides around TSSs for *D. simulans* Dichaete-Dam. D.) Probability density of bound nucleotides around TSSs for *D. simulans* SoxN-Dam. E.) Probability density of bound nucleotides around TSSs for *D. yakuba* Dichaete-Dam. F.) Probability density of bound nucleotides around TSSs for *D. pseudoobscura* Dichaete-Dam.

I also annotated each binding interval in each dataset to a genomic feature category (exon, intron, 5' UTR, 3' UTR, promoter, immediate downstream, or intergenic) using the ChIPpeakAnno R package (Zhu *et al.*, 2010). Each interval could be annotated to multiple categories if it overlapped an annotated region corresponding to more than one feature (e.g. an interval partially overlapping an intron and an exon would be annotated to both). Again, the overall pattern of genomic feature annotation is quite similar for each transcription factor in each species. The most often hit feature is introns, which are hit by approximately 65% of binding intervals, followed by exons, which are hit by approximately 55% of binding intervals (Figure 4.6A-F). This is in agreement with the TSS distance distributions, which indicate that the majority of binding intervals are located downstream of TSSs. Approximately 30% of binding intervals are annotated to promoters. A higher percentage of the *D. yakuba* intervals (64%) are annotated to exons than in the other species. In *D. pseudoobscura*, a higher percentage of intervals are annotated to intergenic regions than in other species, while lower percentages of intervals are annotated to exons, introns, and UTRs. However, considering the differing quality of the *D. pseudoobscura* data and the lower number of intervals compared to the other species, it is difficult to interpret this as a biologically significant difference. In general, the patterns of binding to genomic features by Dichaete and SoxN appear very similar in all species studied.

4.3.4 High overlap with known enhancers

There are several resources containing data on known *Drosophila* enhancers or *cis*-regulatory elements (CRMs), based on a variety of different types of experimental evidence. I downloaded enhancers from REDFly, which contains 1864 manually-curated CRMs, and from the Janelia FlyLight database, which contains 7113 enhancers experimentally shown to drive expression of a GAL4 reporter construct in *D. melanogaster* embryos, including 4724 with specific activity in the embryonic CNS (Gallo *et al.*, 2010; Manning *et al.*, 2012). I used the BEDTools suite to find the overlaps between the high-confidence Dichaete-Dam and SoxN-Dam intervals in *D. melanogaster* and the annotated enhancers from each of these resources. I found that Dichaete-Dam and SoxN-Dam binding intervals overlap with a high number of REDFly and FlyLight enhancers (Table 4.9),

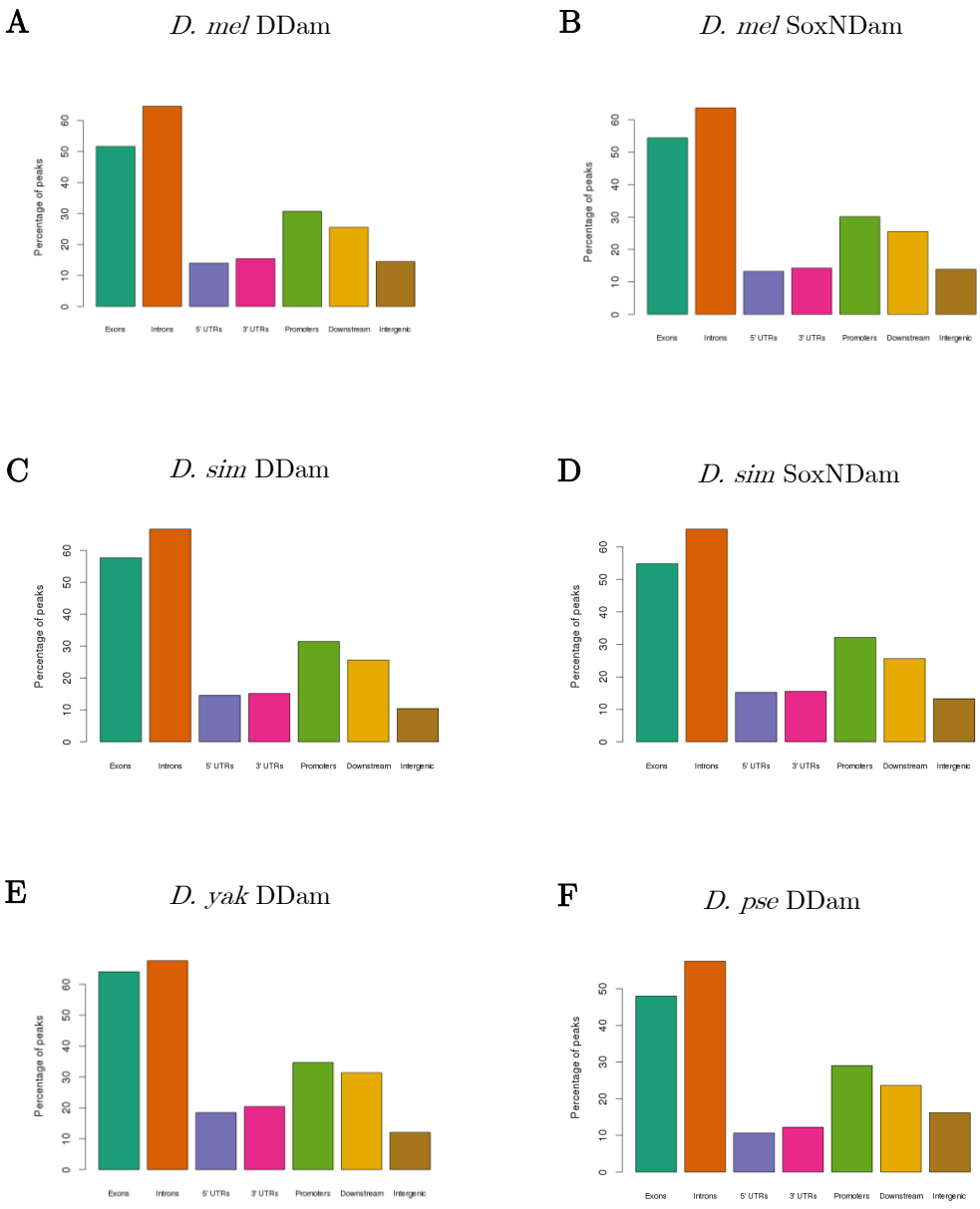


Figure 4.6: Genomic features annotated to Sox DamID binding intervals. Feature classes include exons, introns, 5 UTRs, 3 UTRs, promoters, immediate downstream and intergenic. A.) Percentages of *D. melanogaster* Dichaete-Dam intervals annotated to each genomic feature class. B.) Percentages of *D. melanogaster* SoxN-Dam intervals annotated to each genomic feature class. C.) Percentages of *D. simulans* Dichaete-Dam intervals annotated to each genomic feature class. D.) Percentages of *D. simulans* SoxN-Dam intervals annotated to each genomic feature class. E.) Percentages of *D. yakuba* Dichaete-Dam intervals annotated to each genomic feature class. F.) Percentages of *D. pseudoobscura* Dichaete-Dam intervals annotated to each genomic feature class.

with a slightly greater proportion of CNS-specific FlyLight enhancers containing Dichaete or SoxN binding than for all embryonic enhancers. Overall, more enhancers from each resource contain Dichaete binding than SoxN binding, although a large fraction of all bound enhancers are bound by both Dichaete and SoxN. Although the DamID binding intervals do not correspond directly to enhancer elements, since their borders are dependent on the genomic locations of GATC sites, in certain cases visual inspection reveals a remarkably high correlation between annotated enhancers and peaks of DamID binding, such as at the *ind*, *vnd*, *dpp*, *lola* and *psq* loci (Figure 4.7). The highest overlaps are with the REDFly enhancers, which is encouraging considering that these enhancers have been curated using multiple supporting forms of evidence and can therefore be considered to be the highest-confidence dataset.

	Enhancers	**Enhancers overlapping with Dichaete binding**	**Enhancers overlapping with SoxN binding**	**Enhancers overlapping with Dichaete and SoxN binding**
REDFly	1864	1152 (61.8%)	1130 (60.6%)	1108 (59.4%)
FlyLight	7113	2999 (42.2%)	2784 (39.1%)	2681 (37.8%)
FlyLight CNS	4724	2077 (44.0%)	1935 (41.0%)	1857 (39.3%)
STARR-seq S2	2325	1092 (47.0%)	951 (40.9%)	912 (39.2%)
STARR-seq OSC	3341	1144 (34.2%)	1061 (31.8%)	973 (29.1%)

Table 4.9: Overlaps between Dichaete and SoxN binding and known enhancers in *D. melanogaster*. Numbers in parentheses are percentages of all enhancers of each category containing specified binding.

In addition to the enhancers in these databases, I downloaded peak calls from a STARR-seq assay that was recently performed to identify enhancer activity in both S2 cells and ovarian somatic cells (OSCs) using DNA libraries from *D. melanogaster, D. yakuba, D. ananassae, D. pseudoobscura* and *D. willistoni* (Arnold *et al.*, 2014). These data include both low-stringency, unfiltered peaks, where any enhancer activity was detected, as well as high-stringency, thresholded peaks, where at least a 3-fold enrichment was observed in the STARR-seq samples compared to the input samples. The STARR-seq data contains peaks called for two independent replicates in each condition as well as peaks called for merged

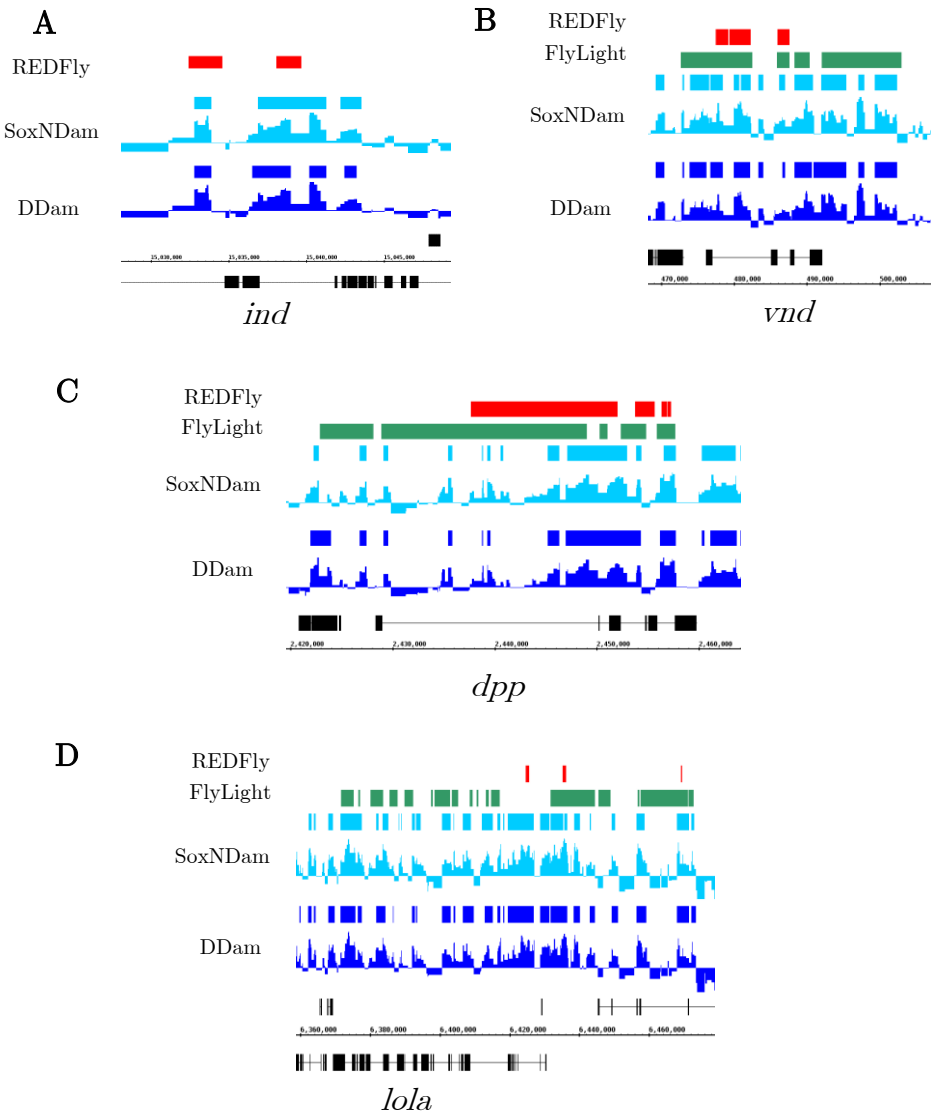


Figure 4.7: Concordance between *D. melanogaster* Sox DamID binding intervals and known enhancers from REDFly and FlyLight. Binding profiles represent the normalized log2 fold changes between Sox fusion binding and Dam-only control binding in each GATC fragment. Dichaete-Dam binding profiles and intervals are in blue, SoxN-Dam binding profiles and intervals are in light blue, REDFly enhancers are in red and FlyLight enhancers are in green. A.) Overlaps between Dichaete-Dam and SoxN-Dam binding intervals and two REDFly enhancers at the *ind* locus. B.) Overlaps between Dichaete-Dam and SoxN-Dam binding intervals and several REDFly and FlyLight enhancers at the *vnd* locus. C.) Overlaps between Dichaete-Dam and SoxN-Dam binding intervals and several REDFly and FlyLight enhancers at the *dpp* locus. D.) Overlaps between Dichaete-Dam and SoxN-Dam binding intervals and several REDFly and FlyLight enhancers at the *lola* and *psq* loci.

replicate data. To get a summary of high-confidence STARR enhancer activity, I focused on the thresholded, merged peaks from both S2 cells and OSCs. This is in line with the analysis of Arnold *et al.*, who first showed that the independent replicates had high levels of reproducibility and then performed further analysis on merged data (Arnold *et al.*, 2014). The peaks were reported as summits; that is, the single genomic coordinate corresponding to the highest point of enrichment for each peak. Following the analysis from Arnold *et al.*, I converted them to intervals by extending the coordinates 250 bp in either direction from the summits, resulting in 501-bp peaks. These enhancer peaks show a comparable overlap with Dichaete and SoxN binding as the FlyLight enhancers, with, again, a slightly greater percentage overlapping with Dichaete binding intervals than with SoxN binding intervals (Table 4.9).

It is interesting to note that, while a higher fraction of the S2 enhancer set is overlapped by Dichaete or SoxN binding intervals than with the OSC enhancers, in terms of absolute numbers this trend is reversed. The enhancers containing DamID-seq binding in the two cell types are mostly different; only 314 Dichaete-bound enhancers and 282 SoxN-bound enhancers are shared between S2 cells and OSCs. Although S2 cells are derived from an embryonic, haemocyte-like lineage, like OSCs they represent a specific, differentiated cell type. The DamID-seq binding intervals are derived from whole embryos containing diverse cell types over a wide range of developmental stages, meaning that they likely represent a regulatory landscape with both similarities to and differences from that of either S2 cells or OSCs. Therefore, it is not surprising that some DamID binding intervals would fall within enhancers characterized in the two different cell types, or that some would not overlap with any active enhancers in either S2 cells or OSCs.

Most of the enhancers characterized in the FlyLight and REDFly databases are considerably longer than 500 bp; it is therefore likely that the overlaps between the 501-bp STARR enhancers and the DamID-seq binding intervals are overly conservative. It is possible that some binding intervals that fall just outside of a 501-bp STARR enhancer are still located within a broader *cis*-regulatory region corresponding to that enhancer. In order to address this issue, I assigned all Dichaete-Dam and SoxN-Dam binding intervals to the nearest unfiltered STARR-seq peak (56220 in S2 cells and 44601 in OSCs). I then filtered the list of intervals

to contain only those that are up to 1 kb away from the nearest peak. I also filtered out any STARR-seq peaks with a p-value greater than 0.001, resulting in a list of binding intervals assigned to nearby peaks that are high-confidence but include strong as well as weak enhancer activity. For Dichaete, this resulted in 2799 assignments in S2 cells, representing 16.0% of high-confidence binding intervals, and 2662 assignments in OSCs, representing 15.2% of high-confidence binding intervals. 1314 binding intervals were assigned to enhancers in both S2 cells and OSCs, while 1485 were only assigned to enhancers in S2 cells and 1348 were only assigned to enhancers in OSCs, bringing the total number of unique binding intervals assigned to an enhancer to 4147, or 23.7% of high-confidence binding intervals. For SoxN, it resulted in 2663 assignments in S2 cells, representing 15.0% of high-confidence binding intervals, and 2693 assignments in OSCs, representing 15.1% of high-confidence binding intervals. 1167 binding intervals were assigned to enhancers in both S2 cells and OSCs, while 1496 were only assigned to enhancers in S2 cells and 1526 were only assigned to enhancers in OSCs, resulting in a total of 4189 unique intervals assigned to an enhancer, or 23.5% of high-confidence binding intervals (Table 4.10).

STARR-seq was also performed with the *D. yakuba* and *D. pseudoobscura* genomes in S2 cells and, in the case of *D. yakuba*, in OSCs. As with *D. melanogaster*, I found the overlaps between the Dichaete-Dam binding intervals and 501-bp peaks around each STARR-seq summit in each species (Table 4.11). In contrast to the *D. melanogaster* data, the *D. yakuba* Dichaete-Dam intervals overlap with a much higher percentage of STARR-seq enhancers in S2 cells than in OSCs, although in absolute numbers they still overlap more OSC enhancers. The overlap percentages are also higher overall in *D. yakuba*; however, it is difficult to draw conclusions from this, as the quality of the data may vary independently between the different species in the STARR-seq experiments and the DamID experiments. However, it is encouraging to see that a high number of the regions with enhancer activity in *D. yakuba* also contain Dichaete binding. Given the low number of enhancers overlapping with the Dichaete-Dam intervals in *D. pseudoobscura*, I decided to exclude this data from further analysis.

Following the same method as I used for the *D. melanogaster* binding intervals, I assigned all of the *D. yakuba* Dichaete-Dam intervals to the nearest unfiltered STARR-seq peak (55734 in S2 cells and 45762 in OSCs). I then filtered the list

Binding dataset	**Binding intervals assigned to S2 enhancers**	**Genes annotated to bound S2 enhancers**	**Binding intervals assigned to OSC enhancers**	**Genes annotated to bound OSC enhancers**	**Binding intervals shared between S2 and OSC enhancers**	**Genes shared between S2 and OSC enhancers**
D. mel Dichaete-Dam	2799	2112	2662	2099	1314	1104
D. mel SoxN-Dam	2663	2037	2693	2099	1167	1072
D. yak Dichaete-Dam	3641	2746	4847	3630	2135	1850

Table 4.10: Binding intervals assigned to a STARR-seq enhancer within 1kb and genes annotated to bound enhancers. Abbreviations: OSC, ovarian stem cells; *D. mel, Drosophila melanogaster; D. yak, Drosophila yakuba*; D-Dam, Dichaete-Dam fusion protein; SoxN-Dam, SoxNeuro-Dam fusion protein.

of intervals to contain only those that are up to 1 kb away from the nearest peak and are annotated to a STARR-seq peak with a p-value less than 0.001. This resulted in 3641 assignments in S2 cells, representing 17.4% of binding intervals, and 4847 assignments in OSCs, representing 23.1% of binding intervals. 2135 intervals were assigned to enhancers in both S2 cells and OSCs, while 1506 were assigned to enhancers only in S2 cells and 2712 were assigned to intervals only in OSCs, bringing the total of unique Dichaete-Dam binding intervals annotated to a STARR-seq enhancer to 6353, or 30.3% of all intervals (Table 4.10).

In order to examine the function of the STARR-seq enhancers that I was able to annotate with Sox binding intervals, I assigned each of them to the closest gene located within 10 kb upstream or downstream. Similar number of genes were annotated to S2 and OSC enhancers for both datasets in *D. melanogaster*, while in *D. yakuba*, nearly 1000 more genes were annotated to OSC enhancers. A summary of the gene annotations for each dataset can be seen in Table 4.10. For both Dichaete-Dam and SoxN-Dam in *D. melanogaster*, approximately half of the total genes annotated to enhancers in each cell type are shared between

	Enhancers	**Enhancers overlapping with Dichaete binding**
STARR-seq S2 *D. yak*	2293	1392 (60.7%)
STARR-seq OSC *D. yak*	3461	1647 (47.6%)
STARR-seq S2 *D. pse*	3233	148 (4.6%)

Table 4.11: Overlaps between Dichaete binding and STARR-seq enhancers in *D. yakuba* and *D. pseudoobscura*. Abbreviations: OSC, ovarian stem cells; *D. yak, Drosophila yakuba; D. pse, Drosophila pseudoobscura.*

the two cell types. In *D. yakuba* approximately half of the genes assigned to OSC enhancers and around two-thirds of the genes assigned to S2 enhancers are shared between the two cell types. Although the same number of genes are annotated to OSC enhancers bound by Dichaete-Dam and SoxN-Dam in *D. melanogaster*, they are not all the same genes. 1890 genes are shared between the two datasets, representing about 90% of the total genes for each factor. Similarly, 1902 genes are shared between S2 enhancers bound by Dichaete-Dam and SoxN-Dam. The proportion of shared genes between enhancers containing Dichaete-Dam and SoxN-Dam binding is similar to the proportion shared between all genes annotated with Dichaete-Dam binding and all genes annotated with SoxN-Dam binding. Comparing between species for enhancers bound by Dichaete-Dam, 1362 genes are annotated to S2 enhancers bound in both *D. melanogaster* and *D. yakuba*, and 1438 genes are annotated to OSC enhancers bound in both species.

Using the gene expression data from FlyAtlas, the genes annotated to S2 enhancers from all datasets are most highly upregulated in S2 cells, showing that these enhancers are indeed active in that particular cellular environment. Other tissues in which these genes are upregulated include the larval CNS, larval hindgut, hindgut, crop and head. The patterns of upregulation are similar for enhancers bound by both Dichaete-Dam and SoxN-Dam in *D. melanogaster* and for enhancers bound by Dichaete-Dam in both *D. melanogaster* and *D. yakuba*, although the genes annotated to S2 enhancers in the *D. yakuba* data are also somewhat upregulated in the ovary. The genes that are annotated to OSC enhancers from all datasets show the strongest upregulation in the ovary, as expected, and in the larval CNS, although they also show some upregulation in S2 cells.

All of the sets of genes annotated to Sox-bound STARR-seq enhancers show similar enrichments for GO:BP terms, including general terms related to biological regulation, development and morphogenesis. Pathways that are enriched for genes in each dataset include signalling pathways and axon guidance (pathway data is from KEGG and Reactome), and dorso-ventral axis formation is enriched for genes in all datasets except the genes annotated to S2 enhancers bound by Dichaete-Dam in *D. yakuba*. The regulation of beta-cell development and regulation of gene expression in beta cells pathways are also enriched for genes annotated to S2 enhancers bound by both Dichaete-Dam and SoxN-Dam in *D. melanogaster*; this may reflect specific functions of genes active in the S2 cell lineage.

4.4 Common and unique binding by Dichaete and SoxNeuro in *D. melanogaster* and *D. simulans*

In order to study the relationship between Dichaete binding and SoxNeuro binding in each species, I used DiffBind to identify intervals that are commonly bound and intervals that are differentially bound between the two transcription factors (Ross-Innes *et al.*, 2012). DiffBind takes the total set of binding intervals called in any sample included in the analysis and considers the read densities in those intervals for all samples in order to detect quantitative differences in binding. It also offers three different methods for normalizing all samples together, edgeR, DESeq and DESeq2, eliminating the problems that arise when using different enrichment thresholds for different samples. For all DiffBind analyses, I used the set of binding intervals with an adjusted p-value of <0.05 for all relevant DamID datasets. In *D. melanogaster*, a total of 24329 intervals are bound by Dichaete-Dam, SoxN-Dam, or both. Comparing each Dichaete-Dam replicate with each SoxN-Dam replicate, the correlations between the binding profiles of the two fusion proteins are quite high overall, reflecting the high level of similarity in their binding patterns. The biggest outlier in the data is replicate 1 of SoxN-Dam (Figure 4.8).

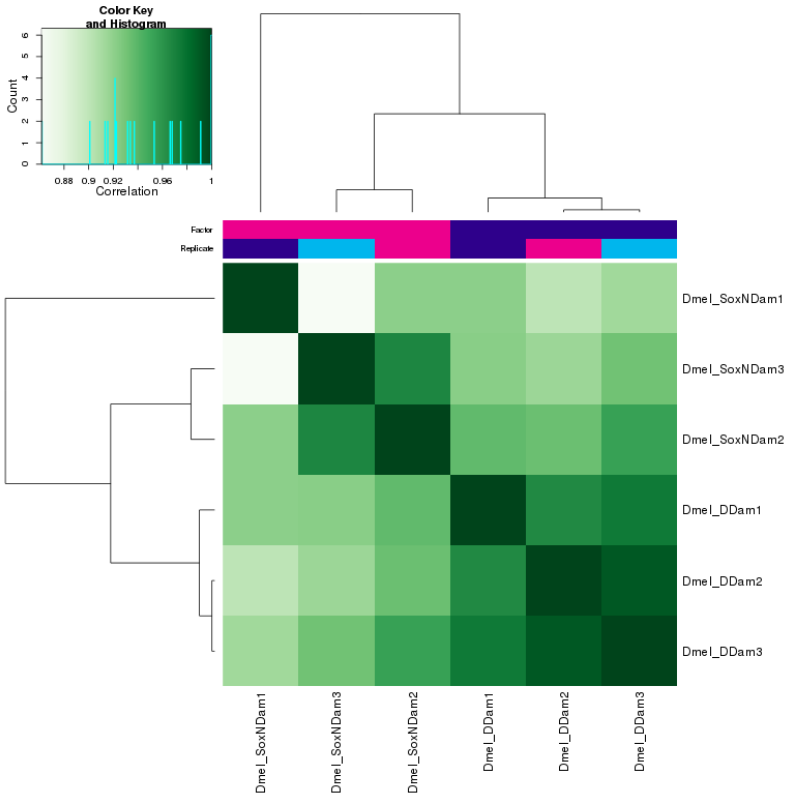


Figure 4.8: Clustering of *D. melanogaster* Dichaete-Dam and SoxN-Dam samples by binding affinity scores in all bound intervals. Both the Dichaete-Dam replicates and the SoxN-Dam replicates cluster together, while the biggest outlier is SoxN-Dam replicate 1. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

Correlations between replicates for the same fusion protein range from 0.86 - 0.99, while correlations between replicates for different fusion proteins range from 0.90 - 0.95. Using DESeq2 normalization, DiffBind identifies 3001 intervals are that differentially bound between Dichaete-Dam and SoxN-Dam at FDR10 and 1048 at FDR1 (Figure 4.9). The FDR1 differentially bound intervals represent 5.0% of all Dichaete-Dam bound intervals and 4.6% of all SoxN-Dam bound intervals. Of the 1048 FDR1 intervals, 675 are preferentially bound by Dichaete-Dam and 373 are preferentially bound by SoxN-Dam.

All of the intervals that are differentially bound by Dichaete-Dam were called as bound by Dichaete in the single-factor DESeq2 analysis. 459 of these were also

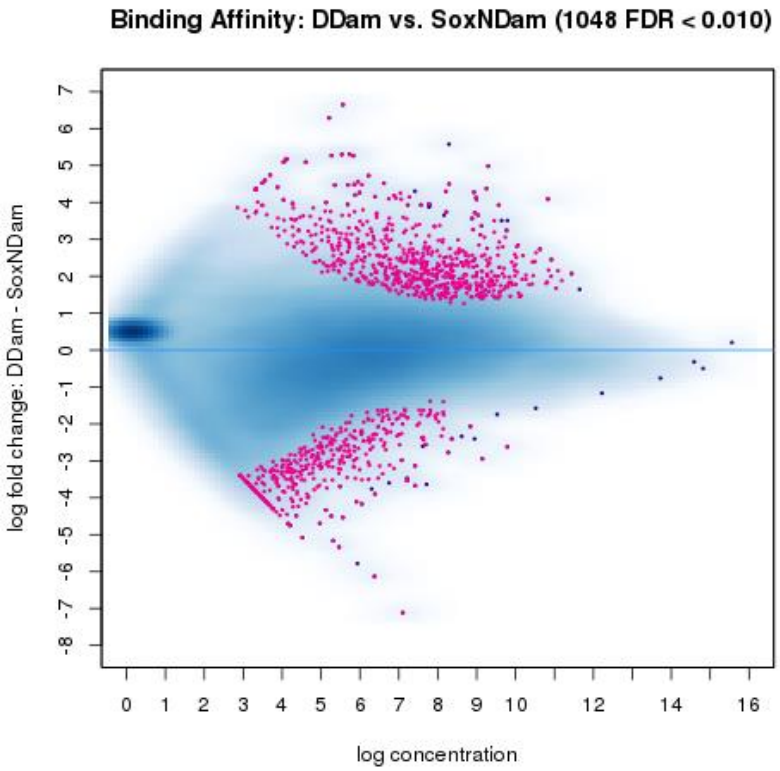


Figure 4.9: MA plot showing differentially bound intervals with FDR <0.01 between *D. melanogaster* Dichaete-Dam and SoxN-Dam. All intervals are plotted; differentially bound intervals are highlighted in pink.

called as bound by SoxN-Dam, while 216 were unique to Dichaete-Dam. All but one of the 373 intervals differentially bound by SoxN-Dam were called as bound by SoxN in the single-factor DESeq2 analysis, and of these, 114 were also bound by Dichaete-Dam, while 258 were unique to SoxN-Dam. The difference between the pattern of preferential binding by Dichaete-Dam and preferential binding by SoxN-Dam can be seen in a clustering of differential intervals by the log of normalized read counts within each interval (affinity score); many of the intervals preferentially bound by Dichaete-Dam are also strongly bound by SoxN-Dam, while a higher proportion of the intervals preferentially bound by SoxN-Dam are not bound or are weakly bound by Dichaete-Dam (Figure 4.10).

In *D. simulans*, a total of 19661 intervals are bound by Dichaete-Dam, SoxN-Dam, or both. In comparison to *D. melanogaster*, the *D. simulans* SoxN-Dam binding profiles within these intervals are much more similar to each other and more different from the Dichaete-Dam binding profiles. The correlations between replicates for the same fusion protein range from 0.91 - 0.98, and the correlations between replicates for different fusion proteins range from 0.47 - 0.62 (Figure 4.11). In agreement with the lower correlations between Dichaete-Dam and SoxN-Dam binding profiles, a DiffBind analysis with DESeq2 normalization identified many more differentially bound intervals between Dichaete and SoxN in *D. simulans* than in *D. melanogaster*. 8807 differential binding intervals were identified at FDR10, and 4881 were identified at FDR1 (Figure 4.12). The FDR1 differentially bound intervals represent 30.3% of all Dichaete binding intervals and 32.2% of all SoxN binding intervals. Of the 4881 FDR1 intervals, 2294 are preferentially bound by Dichaete-Dam, while 2587 are preferentially bound by SoxN-Dam. All but one of the intervals called as preferentially bound by Dichaete-Dam were identified as bound intervals in the single-factor DESeq2 analysis. Of these, 782 were also called as bound by SoxN-Dam, while roughly twice as many, 1511, were unique to Dichaete-Dam. All of the intervals called as preferentially bound by SoxN-Dam were identified as bound intervals in the single-factor DESeq2 analysis. 1502 of these were also called as bound by Dichaete-Dam, while 1085 were unique to SoxN-Dam. This is in contrast to the *D. melanogaster* data, where the majority of the SoxN-Dam differentially bound peaks were unique to SoxN-Dam. This pattern can clearly be seen in a heatmap clustering differentially bound sites by affinity score (Figure 4.13).

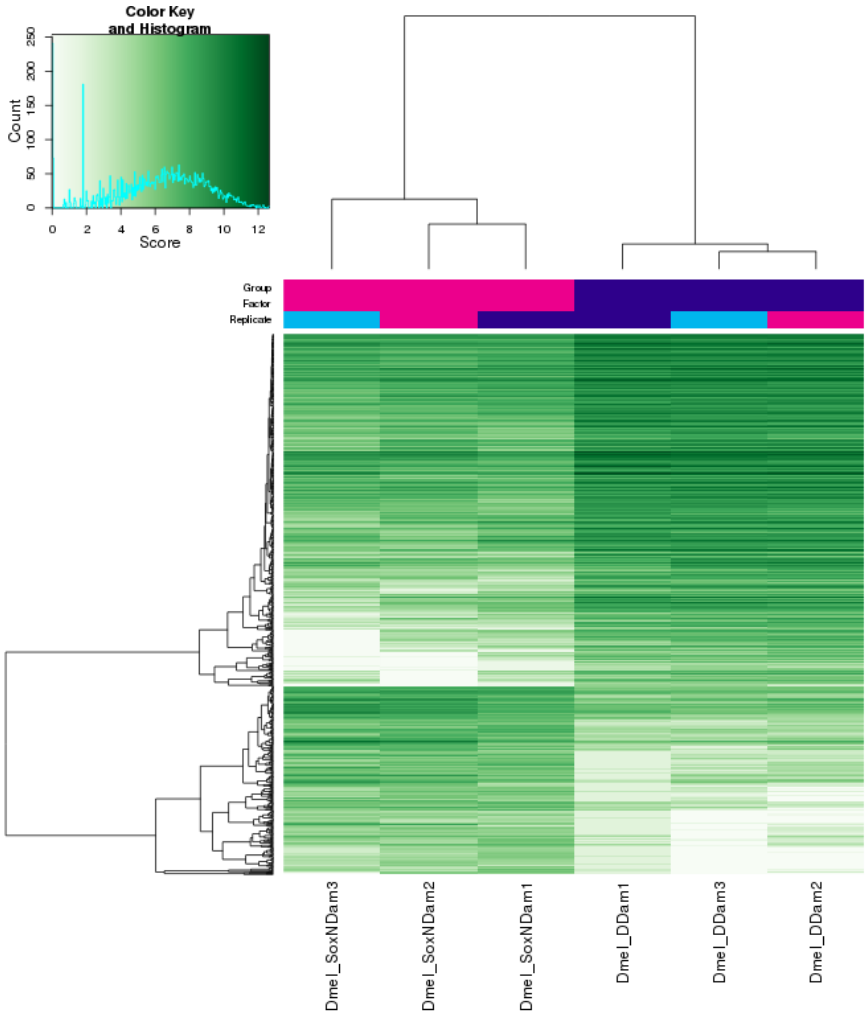


Figure 4.10: Clustering of *D. melanogaster* Dichaete-Dam and SoxN-Dam differentially bound intervals by binding affinity scores. A block of intervals that are bound more highly by Dichaete is present in the upper right quadrant, while a block of intervals that are bound more highly by SoxN is present in the lower left quadrant. The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding.

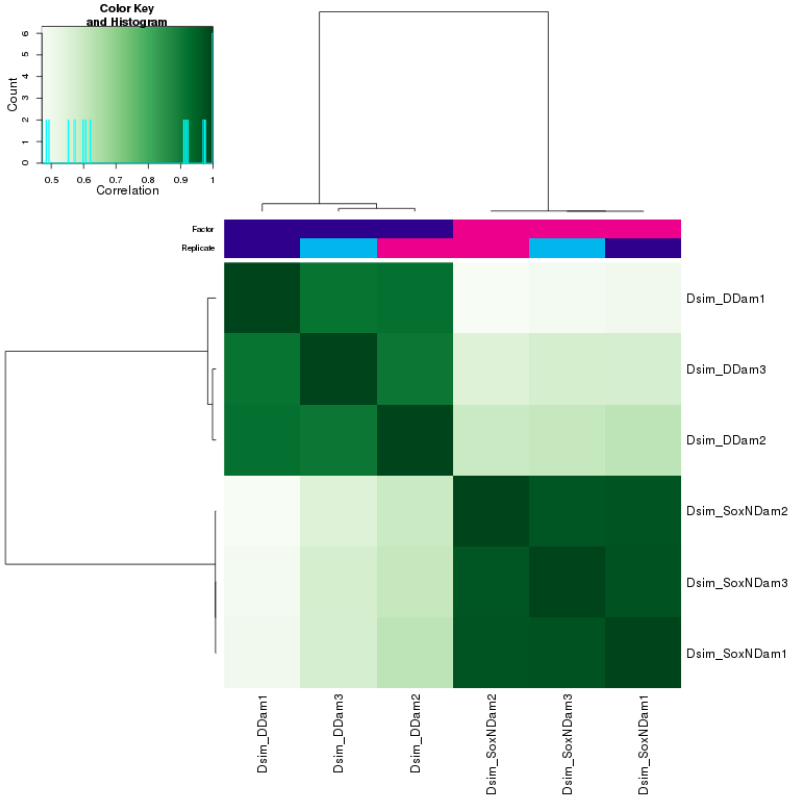


Figure 4.11: Clustering of *D. simulans* Dichaete-Dam and SoxN-Dam samples by binding affinity scores in all bound intervals. Both the Dichaete-Dam replicates and the SoxN-Dam replicates cluster strongly together. There is a greater differentiation visible between the two transcription factors than in *D. melanogaster*. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

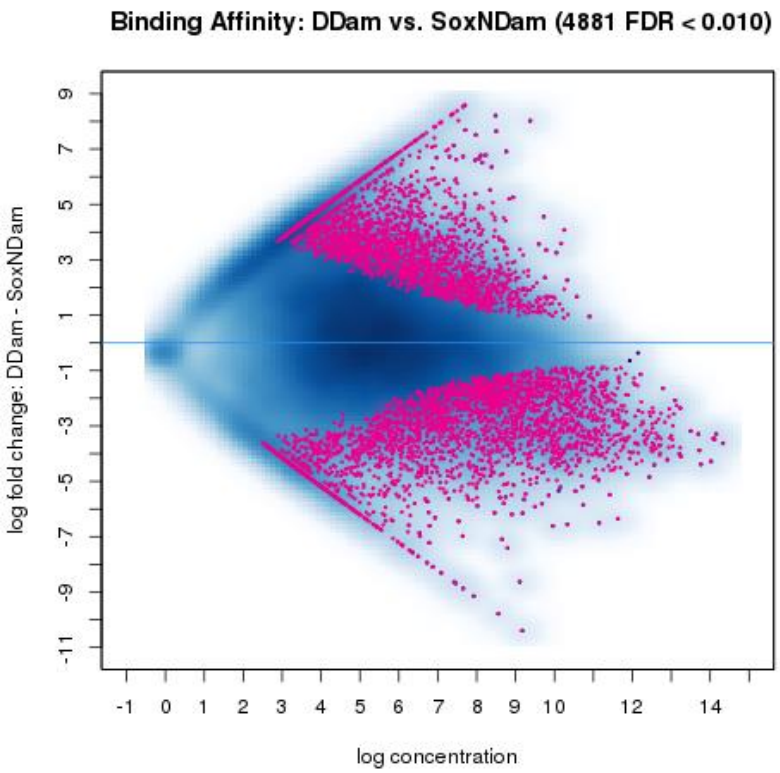


Figure 4.12: MA plot showing differentially bound intervals with FDR <0.01 between *D. simulans* Dichaete-Dam and SoxN-Dam. All intervals are plotted; differentially bound intervals are highlighted in pink.

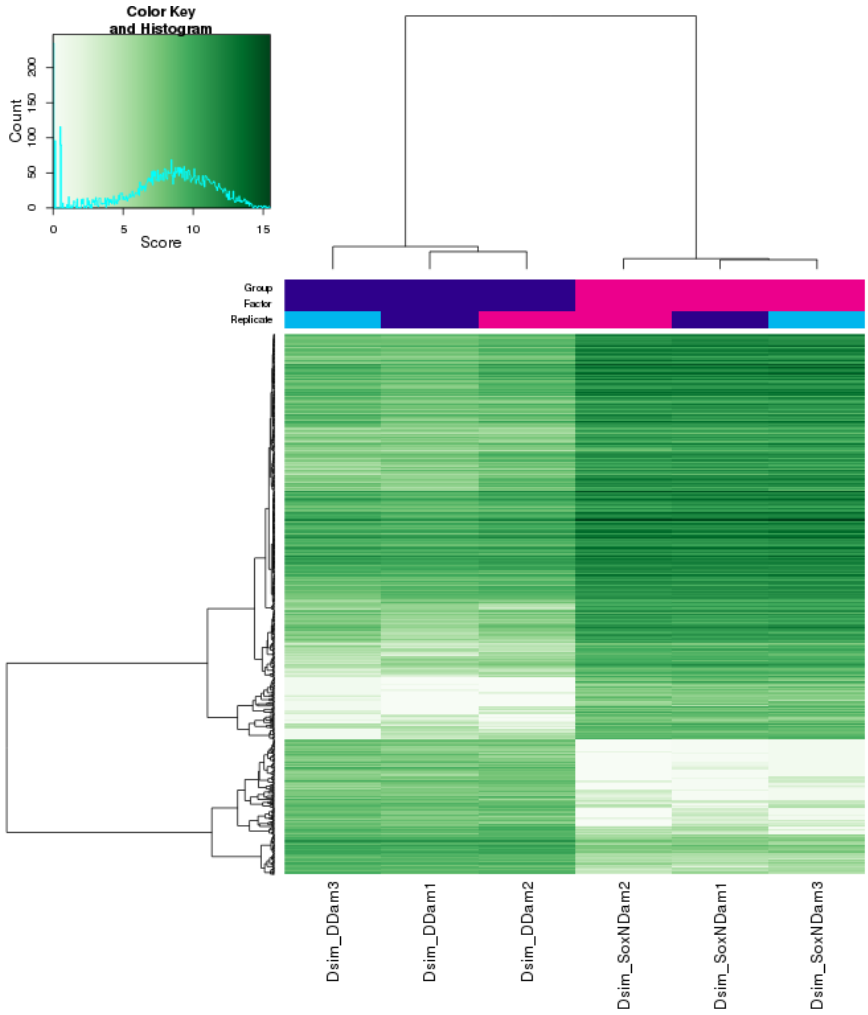


Figure 4.13: Clustering of *D. simulans* Dichaete-Dam and SoxN-Dam differentially bound intervals by binding affinity scores. A block of intervals that are bound more highly by Dichaete is present in the lower left quadrant, while a block of intervals that are bound more highly by SoxN is present in the upper right quadrant. A greater number of intervals are more highly bound by SoxN than by Dichaete, which is the opposite of the case in *D. melanogaster*. The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding.

There are some notable differences between the patterns of Dichaete and SoxN binding in the *D. melanogaster* data as opposed to the *D. simulans* data. The binding profiles of Dichaete and SoxN are more highly correlated in *D. melanogaster* than in *D. simulans*, which is a surprising finding. Accordingly, a much larger number of sites that are differentially bound between the two proteins was identified in *D. simulans* than in *D. melanogaster*. Figure 4.14 shows an example of a region where there are more differentially bound intervals between SoxN and Dichaete in *D. simulans* compared to *D. melanogaster*. It should be recognized that, while using DiffBind to examine differences in each species permits normalization of the Dichaete-Dam and SoxN-Dam data within each species, it does not allow for normalization of the data between species, meaning that the thresholded intervals called as differentially bound in each species are not directly comparable. Nonetheless, examining the binding profiles shows that there are differences between the ways in which the two TFs bind relative to one another in the two species, which can be compared on an overall level; this question will be revisited in more detail in Chapter 5.

For each species, I assigned the intervals that are preferentially bound by either Dichaete or SoxN, as well as those that are uniquely bound by either Dichaete or SoxN, to the closest gene within 10 kb upstream or downstream. The preferentially bound and uniquely bound intervals are not mutually exclusive; rather, the uniquely bound intervals are a subset of the preferentially bound intervals. In *D. melanogaster*, this resulted in 826 genes that are preferential targets of Dichaete and 251 that are unique targets of Dichaete, as well as 498 genes that are preferential targets of SoxN and 371 that are unique targets of SoxN. In *D. simulans*, 2180 genes are preferential and 1507 are unique targets of Dichaete, while 2295 genes are preferential and 995 are unique targets of SoxN. I then analyzed the resulting lists of target genes using FlyMine (Lyne *et al.*, 2007). The GO:BP term enrichments are quite similar for both the preferential and unique targets of each transcription factor in each species. In *D. melanogaster*, SoxN targets are broadly enriched for terms relating to development and biological regulation. Dichaete targets are also enriched for developmental terms, but, additionally, they show enrichment for signalling ($p < 0.02$), cell communication ($p < 0.02$) and response to stimulus ($p < 1e\text{-}20$). Interestingly, the Dichaete unique targets also show enrichment for synapse assembly ($p = 0.013$) and synaptic transmis-

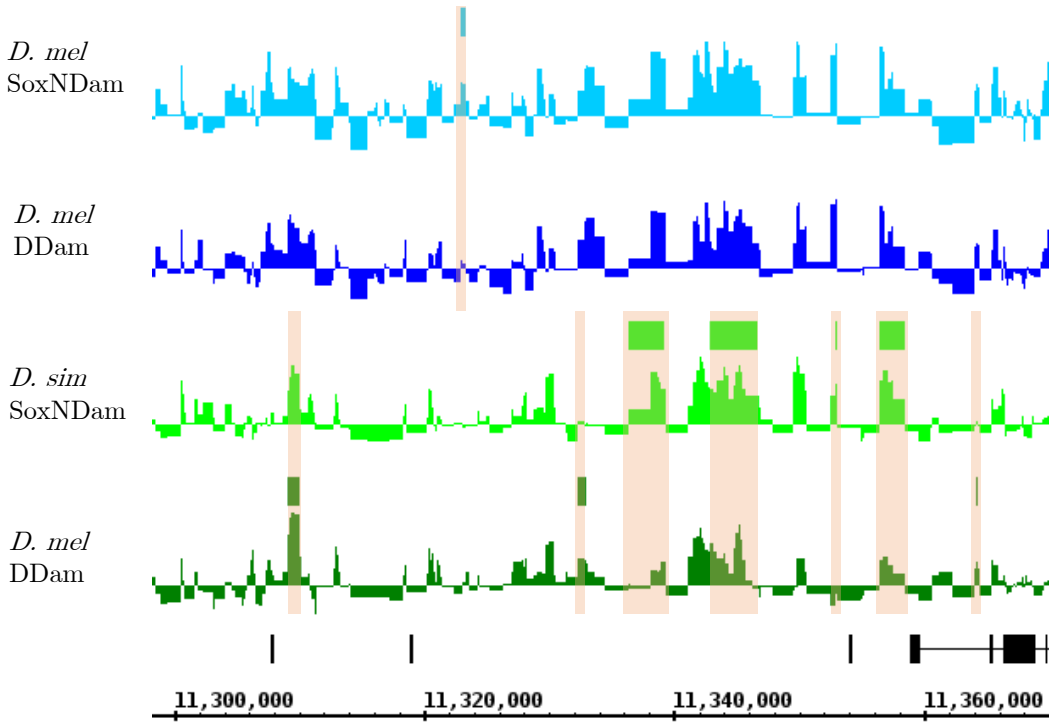


Figure 4.14: Quantitative differences in binding by Dichaete-Dam and SoxN-Dam in *D. simulans* versus in *D. melanogaster*. From the bottom, the *D. simulans* Dichaete-Dam binding profile is in green, the *D. simulans* SoxN-Dam binding profile is in light green, the *D. melanogaster* Dichaete-Dam binding profile is in blue and the *D. melanogaster* SoxN-Dam binding profile is in light blue. Intervals that are differentially bound by one TF in each species are represented by color-coded blocks above each respective binding profile. Three intervals that are preferentially bound by SoxN in *D. simulans* but not in *D. melanogaster*, as well as four intervals that are preferentially bound by Dichaete in *D. simulans* but not in *D. melanogaster*, are shown. In the same region, only one interval is preferentially bound by SoxN in *D. melanogaster* and none are preferentially bound by Dichaete. All intervals that are preferentially bound by one TF in comparison to the other are highlighted in tan.

sion (p = 0.023). In *D. simulans*, both Dichaete and SoxN targets are highly enriched for developmental terms (p <1E-11), but Dichaete targets have higher enrichments for imaginal disc morphogenesis (p <1E-8), while SoxN targets have higher enrichments for terms related to the regulation of transcription (p <1E-31). In both species, the Dichaete targets tend to be upregulated in the brain, head, hindgut, larval hindgut and larval CNS according to the FlyAtlas expression data, while the SoxN targets are only highly upregulated in the larval CNS and, in *D. melanogaster*, the ovary.

4.5 Discussion of results

In this chapter I have presented the major datasets that I generated during my Ph.D., using DamID-seq to measure *in vivo* genome-wide binding of the two group B Sox proteins Dichaete and SoxN in four species of *Drosophila*. This approach was successful in *D. melanogaster* and *D. simulans*, the two most closely-related species in my study. In *D. yakuba*, I successfully generated a high-quality *in vivo* binding profile for Dichaete; however, the DamID experiment for SoxN in this species failed, likely due to a mutation rendering the SoxN portion of the fusion protein nonfunctional. In *D. pseudoobscura*, I was unable to generate a transgenic line for SoxN-Dam, despite multiple injection attempts. The Dichaete-Dam experiment in this species was successful in that it uncovered binding intervals that show multiple indications of being biologically functional *in vivo*; however, the higher variance observed between replicates in comparison to other species resulted in significantly less binding intervals being called. Given the highly similar expression and sequence data for Dichaete between *D. pseudoobscura* and the other species studied, it is unlikely that this reflects any underlying biological difference in Dichaete function. The amino acid sequences of the HMG-box DNA-binding domains of Dichaete are essentially identical in *D. melanogaster* and *D. pseudoobscura*; however, there are some potentially significant differences in the N- and C-terminal regions (McKimmie *et al.*, 2005). It is possible that these differences are responsible for the increased variability in binding observed when the *D. melanogaster* protein is expressed in *D. pseudoobscura*, as the fusion protein may have a reduced ability to have its binding stabilized through inter-

actions with cofactors compared to the endogenous Dichaete. Nonetheless, these datasets represent the first assay of group B Sox binding in *Drosophila* species other than *D. melanogaster* and, to my knowledge, the first time that DamID-seq has been used in a comparative binding experiment.

I used multiple types of analyses to assess the functionality of the identified binding intervals, including discovery of *de novo* and known TF binding motifs, annotation to genes and genomic features, overlap with previously defined high-confidence *D. melanogaster* core intervals for each TF, overlap with known *D. melanogaster* enhancers, and gene ontology enrichment analysis of potential target genes. For each dataset in each species, these analyses show good agreement with previous group B Sox binding data and indicate that the binding intervals are likely to represent true instances of gene regulation by Dichaete and SoxN. In *D. pseudoobscura*, the Dichaete-Dam binding intervals show much lower overlap with known enhancers, and their putative target genes show lower overlap with direct target genes or targets of core binding intervals, than in other species. However, this is likely to be at least partially due to the much lower number of binding intervals identified overall; both the motif and Gene Ontology enrichments show good indications of known Dichaete function.

On a broad scale, the binding patterns that I identified for Dichaete-Dam and SoxN-Dam indicate that the functions of these proteins are largely conserved among the *Drosophila* species studied. There are no notable differences between species in the enriched GO:BP terms identified in sets of genes associated with Dichaete-Dam or SoxN-Dam binding intervals, suggesting that the two TFs have maintained their roles in early CNS specification, neural development and morphogenesis, and regulation of other developmentally important transcription factors. Several verified Dichaete target genes are identified in *D. simulans, D. yakuba* and *D. pseudoobscura*, including *nubbin (nub), grainy head (grh), miranda (mira)* and *POU domain protein 2 (pdm2)*, which are involved in maintenance of neuroblast self-renewal as well as differentiation; *decapentaplegic (dpp), wingless (wg), brother of odd with entrails limited (bowl), drumstick (drm), bagpipe (bap), Delta (Dl), outstretched (os), faint sausage (fas)* and *ribbon (rib)*, which are targets of Dichaete in the hindgut; and *slit (sli)*, a Dichaete target in the midline (Aleksic *et al.*, 2013). Similarly, direct targets of SoxN that are conserved in *D. simulans* include the proneural gene *asense (ase)* and the proneural repressor

hairy (h); *Drop (Dr)*, a gene involved in DV patterning in the CNS; *Kruppel (Kr), nub, pdm2, castor (cas), inscuteable (insc), numb, sanpodo (spdo), snail (sna), worniu (wor)*, and *escargot (esg)*, which are involved in specifying neuroblast identity and controlling neuroblast self-renewal and asymmetric divisions (Buescher *et al.*, 1998; Cai *et al.*, 2001; Isshiki *et al.*, 2001; Kraut *et al.*, 1996; Maurange and Gould, 2005; O'Connor-Giles and Skeath, 2003; Skeath and Doe, 1998; Van Doren *et al.*, 1994); and *cut (ct), dawdle (daw), knot (kn), longitudinals lacking (lola), midline (mid), nervous fingers 1 (nerfin-1)* and *Sema-1a*, all SoxN targets involved in morphogenesis of axons and dendrites (Ferrero *et al.*, 2014; Giniger *et al.*, 1994; Jinushi-Nakao *et al.*, 2007; Kuzin *et al.*, 2005; Liu *et al.*, 2009; Parker *et al.*, 2006; Yu *et al.*, 1998).

As with many other developmentally important regulators, Dichaete and SoxN bind extensively across the genome in all species studied. It has previously been suggested that Dichaete can bind at highly occupied target (HOT) regions in the *D. melanogaster* genome, where it may facilitate the formation of complexes of other regulatory factors by causing DNA bending (Aleksic *et al.*, 2013). I found a highly enriched motif for Trl, a marker of HOT regions, in both Dichaete-Dam and SoxN-Dam binding intervals, indicating that both proteins may be able to play this role. Several motifs for potential cofactors of Dichaete and SoxN were also enriched in binding intervals in multiple species. A motif for the known Dichaete cofactor Vvl is enriched in both the *D. melanogaster* and *D. yakuba* Dichaete-Dam binding intervals, as well as the *D. simulans* SoxN-Dam binding intervals. A motif for Vnd, a transcription factor that plays an important role in the specification of the CNS (Ferrero *et al.*, 2014; Zhao and Skeath, 2002), is enriched in Dichaete-Dam binding intervals in both *D. yakuba* and *D. pseudoobscura*. Dichaete binding has been shown to overlap significantly with Twi and Kni binding in *D. melanogaster* (Aleksic *et al.*, 2013); a Twi motif was also found to be enriched in the *D. pseudoobscura* Dichaete-Dam binding intervals, while a Kni motif was found to be enriched in the *D. simulans* as well as the *D. melanogaster* Dichaete-Dam binding intervals. Tll is a target of Dichaete in the hindgut and may work with it to regulate hindgut development (Aleksic *et al.*, 2013); a Tll motif is enriched in Dichaete-Dam binding intervals in *D. simulans* and *D. yakuba*, as well as in SoxN-Dam binding intervals in *D. melanogaster* and *D. simulans*. Taken together, these results suggest that the transcriptional

networks in which Dichaete and SoxN are embedded are also highly conserved between *Drosophila* species.

Although DamID-seq provides less spatial resolution and accuracy in measuring TF binding than ChIP-seq, because the distribution of bound fragments depends on the distribution of GATC sites in the genome, rather than on randomly sheared chromatin, certain features of Dichaete and SoxN binding patterns can be observed from the DamID-seq datasets. In all species studied, both proteins show a tendency to bind in introns; this is evidenced both by the high percentages of binding intervals that are annotated directly to introns and the downstream skew of binding intervals relative to TSSs. This pattern appears to hold across all species, although Dichaete-Dam binding intervals in *D. pseudoobscura* show a higher tendency to be annotated to intergenic regions than in other species. Both Dichaete-Dam and SoxN-Dam intervals also show high overlaps with known functional enhancers. These two enrichments are not mutually exclusive, as many developmentally important enhancers in *Drosophila* are known to be located within introns, including the enhancer to which Dichaete has been shown to bind in the midline gene *sli* (Aleksic *et al.*, 2013; Ma *et al.*, 2000). The STARR-seq enhancers which are bound by Dichaete and SoxN in *D. melanogaster* and *D. yakuba* are found near genes that are enriched for similar functions as the general sets of target genes for Dichaete and SoxN binding, providing evidence that binding at these loci is not incidental but is linked to regulatory function.

Using the data generated in *D. melanogaster* and *D. simulans*, I have also investigated the similarities and differences between Dichaete and SoxN binding patterns in two species. Prior to this work, it was known that Dichaete and SoxN showed highly similar patterns of binding *in vivo* in *D. melanogaster*, yet that their binding patterns were not identical (Ferrero *et al.*, 2014). My datasets confirm this view. Generally speaking, the substantial overlap between Dichaete and SoxN binding that was observed in *D. melanogaster* is a conserved feature of group B Sox binding in *D. simulans*. However, somewhat surprisingly, the preliminary analysis of Dichaete and SoxN binding suggests they may be more differentiated in *D. simulans*, both at the level of sequencing read correlations and binding interval locations. Since the fusion proteins expressed in both species were identical and derived from the gene sequences of *D. melanogaster*, any difference between species must be due to differences in the nuclear environment in *D. simulans*; that

is, either sequence changes in *cis*-regulatory elements where Dichaete and SoxN bind or changes in *trans* affecting the overall transcriptional regulatory network. While it is possible that Dichaete and SoxN function in a more independent manner in *D. simulans* than in *D. melanogaster*, one intriguing hypothesis is that the targets that are common to the two proteins in both species may be the sites of more functionally important binding events. Dichaete and SoxN recognize very similar sequence motifs, contributing to the similarity of their binding profiles; however, it is likely that not all of the observed binding events are functional. Perhaps expressing proteins from *D. melanogaster* in *D. simulans* allows for a de-coupling of functional binding and binding driven by incidental Sox motifs; this hypothesis should be tested in the future by performing transgenic assays of enhancer function. In the following chapter, I will examine the relationship between common and unique binding by Dichaete and SoxN in both species in greater depth.

Although these DamID binding datasets provide a rich resource for the comparative study of group B Sox binding, they also have some limitations. The material used for DamID, whole embryos from overnight collections, represented both a mix of various tissue types and a broad range of developmental stages. The binding intervals identified therefore reflect an average picture of group B Sox binding during development, rather than the exact binding profiles in any one cell type at a given time. As discussed previously the limit of spatial resolution possible with DamID is dependent on the location of GATC sites, making the exact identification of binding sites more difficult than with ChIP. Nonetheless, DamID can give a reasonably accurate view of binding patterns; the average binding interval lengths for these datasets range from 589 bp to 1474 bp, which are shorter than many relevant genomic features, such as genes or even introns. As with ChIP, DamID identifies all regions where the fusion protein binds *in vivo*. However, these binding events may not all be functional in the sense of contributing to transcriptional activation or repression. It is possible to identify direct targets of a TF by combining *in vivo* binding data with gene expression data in a mutant background, which has been done for Dichaete and SoxN in *D. melanogaster*. However, these functional binding events typically only make up a fraction of the genes that are bound by a TF, raising the question of what the effect of binding at other loci is. Some of this binding may simply be due to the thermodynamic

affinity of TFs for DNA (Biggin, 2011; Fisher *et al.*, 2012; Kaplan *et al.*, 2011), although, for TFs like Dichaete and SoxN which can induce DNA bending, it might help create enhancer loops to bring other regulatory factors together (Ghavi-Helm *et al.*, 2014). In this view, loss of Sox binding may result in variable effects on gene expression or increase expression noise due to perturbation of the regulatory network, both of which are difficult to detect with standard genomic expression analysis. Such a role, and observed variable effects on gene expression during early segmentation, have been reported for Dichaete (Russell *et al.*, 1996). One way to more specifically address this question is to examine which binding events have been conserved during evolution, as functional binding is more likely to be constrained by natural selection (Biggin, 2011).

In general, the genomic features associated with Dichaete and SoxN binding, including sequence motifs and putative gene targets, appear to be quite similar between species of *Drosophila*. This finding supports the expectation that Dichaete and SoxN have broadly similar roles during development across the drosophilids and, indeed, as far distant as vertebrates (Uwanogho *et al.*, 1995; Wood and Episkopou, 1999). However, a significant number of binding intervals differ between *D. melanogaster* and the other species examined, raising the question of the evolutionary significance of these differences in binding patterns. Thus far, I have only performed a crude comparison of the binding patterns between different species. In the following chapter, I will dissect the differences and similarities in binding on a quantitative basis, including the relationships between Dichaete and SoxN binding, and examine the possible mechanisms of binding site turnover and evolution within the phylogenetic distances that I have studied.

CHAPTER 5

Evolutionary Patterns of Group B Sox Binding in *Drosophila*

5.1 Overview and motivation

While many aspects of Dichaete and SoxN function can be understood by examining the DamID datasets that I generated in each species independently, a comparative approach that looks at binding through the lens of natural selection has the potential to reveal a more nuanced view. The turnover of transcription factor binding sites between species is a well-documented phenomenon in both flies and vertebrates, although, most likely due to the compact genome and large effective population sizes, a greater percentage of binding sites are generally conserved between different species of *Drosophila* than between mammals at a similar evolutionary distance (Bradley *et al.*, 2010; He *et al.*, 2011b; Odom *et al.*, 2007; Schmidt *et al.*, 2010; Stefflova *et al.*, 2013; Villar *et al.*, 2014). This is a useful feature, as it facilitates the identification of binding sites that have potentially been subject to selective pressure. In theory, binding events that are

more functionally important will be subject to greater constraint under purifying selection and will therefore tend to be conserved between species. However, not all non-conserved sites are non-functional; the evolution of new binding sites can be driven by positive selection either to compensate for the loss of a site elsewhere or in connection with a new function (Arnoult *et al.*, 2013; Frankel *et al.*, 2012; He *et al.*, 2011a; Kalay and Wittkopp, 2010). For Dichaete and SoxN, a number of different questions can be asked using comparative binding data, including whether certain genomic features or functional categories of genes are associated with conserved binding events for each TF, what is the relationship is between Dichaete and SoxN binding overlap and conservation of binding between species, and how changes in the sequence or number of Sox motifs within intervals are associated with binding conservation or divergence. I used comparisons between the binding patterns of the two TFs as well as the binding patterns of each TF between multiple species to try to answer these questions.

The evolutionary conservation of transcription factor binding sites can be studied on several levels: the qualitative presence or absence of a binding interval, quantitative measures of binding affinity, or the underlying DNA sequence and motifs. I have attempted to address all of these levels of conservation to build a detailed picture of the evolutionary dynamics of Dichaete and SoxN binding in *Drosophila*. In this chapter, I start by performing quantitative pairwise comparisons of Dichaete and SoxN binding patterns between *D. melanogaster* and each of the other species for which I have data. This provides a more detailed view of the similarities and differences between binding datasets than a simple intersection of binding intervals. I also perform a three-way quantitative comparison of Dichaete-Dam binding in *D. melanogaster, D. simulans* and *D. yakuba*. In order to address the relationship between Dichaete and SoxN binding, I use both qualitative and quantitative measures to examine whether there is any difference in conservation rates between intervals that are bound uniquely by Dichaete-Dam or SoxN-Dam in different species and those that are bound commonly by both TFs. Zooming out to the gene level, I examine possible instances of binding site turnover and compensatory evolution within gene loci. If Dichaete and SoxN binding at known enhancers and core intervals is truly functional, one would expect it to show increased levels of conservation; I also examine this hypothesis using qualitative measures of binding conservation.

Finally, I search for Sox motifs located within intervals and examine the relationship between motif number and quality and binding conservation. This analysis focuses primarily on the Dichaete-Dam binding intervals, as there is a Dichaete-Dam dataset available in all four species studied. Although the number of fully conserved binding intervals is small compared to the total number of binding intervals identified, in part due to the smaller number of intervals identified in *D. pseudoobscura*, the functional analyses performed in the previous chapter indicate that these are high-confidence binding intervals enriched for high-quality Sox motifs, and are thus useful for drawing conclusions about the effect of motif presence and quality on TF binding. In addition to searching for Sox motifs in conserved and non-conserved binding intervals, I also perform multiple alignments of binding intervals to examine positional and nucleotide-level motif conservation in the context of both qualitative and quantitative changes in binding. Details of the computational methods used to perform evolutionary analyses can be found in Chapter 2.

5.2 Pairwise comparison of binding between *D. melanogaster* and non-model species

Finding the overlap between binding intervals gives a rough estimate of the conservation of binding events between species; however, this does not take into account the strength of binding within intervals. It is also likely to be overly conservative due to the fact that the variance differs between the different sets of replicates, resulting in different effective thresholds for detection of enriched binding at a given p-value. In order to get a more nuanced view of conserved and differential binding between *D. melanogaster* and each other species, I used DiffBind to perform a comparative analysis, using the translated reads and peaks called from translated reads for all non-*melanogaster* species (Ross-Innes *et al.*, 2012).

5.2.1 Quantitative comparison of binding between *D. melanogaster* and *D. simulans*

Using DiffBind to cluster both the Dichaete-Dam and SoxN-Dam binding interval datasets for *D. melanogaster* and *D. simulans*, one can observe that, in general, the SoxN-Dam replicates from both species cluster together and the Dichaete-Dam replicates from both species cluster together (Figure 5.1). Replicate 1 of SoxN-Dam in *D. melanogaster* is an exception to this pattern, as it clusters more closely to replicate 1 of Dichaete-Dam in *D. melanogaster*; this may be due to an effect of the sequencing platform, since these two replicates were sequenced on a MiSeq while all of the others were sequenced on a HiSeq. In general, the binding profiles for Dichaete-Dam and SoxN-Dam in *D. melanogaster* appear more similar than the binding profiles for Dichaete-Dam and SoxN-Dam in *D. simulans*.

Examining the data for one transcription factor at a time highlights the differences between species (Figure 5.2). For Dichaete-Dam, a total of 23985 binding intervals were considered in both species. There is a clear division between the two species. The three *D. melanogaster* replicates are highly similar, while in *D. simulans*, replicate 3 is a slight outlier. Comparing each *D. melanogaster* replicate with each *D. simulans* replicate, the correlations of binding profiles at all bound peaks between the two species range from 0.62 - 0.72, while the correlations of replicates within a single species range from 0.92 - 0.93 for *D. simulans* and from 0.97 - 0.99 for *D. melanogaster*. 11596 binding intervals were identified as differentially enriched between the two species at FDR10 using DESeq2 normalization. For further analysis, I decided to use a stringent set of binding intervals that were called as differentially bound between *D. melanogaster* and *D. simulans* at FDR1. This set contains 7246 binding intervals, representing 45.0% of all *D. simulans* bound intervals and 34.8% of all *D. melanogaster* bound intervals (Figure 5.3).

Of these, 4039 are preferentially bound in *D. simulans* and 3207 are preferentially bound in *D. melanogaster*. Clustering the differentially bound intervals by affinity scores reveals two large blocks of intervals that are highly bound in only either *D. melanogaster* or *D. simulans* (Figure 5.4); these could be considered as gains or losses of binding events in each lineage. In addition, there are smaller clusters of intervals that have high affinity scores in both species, but are more highly

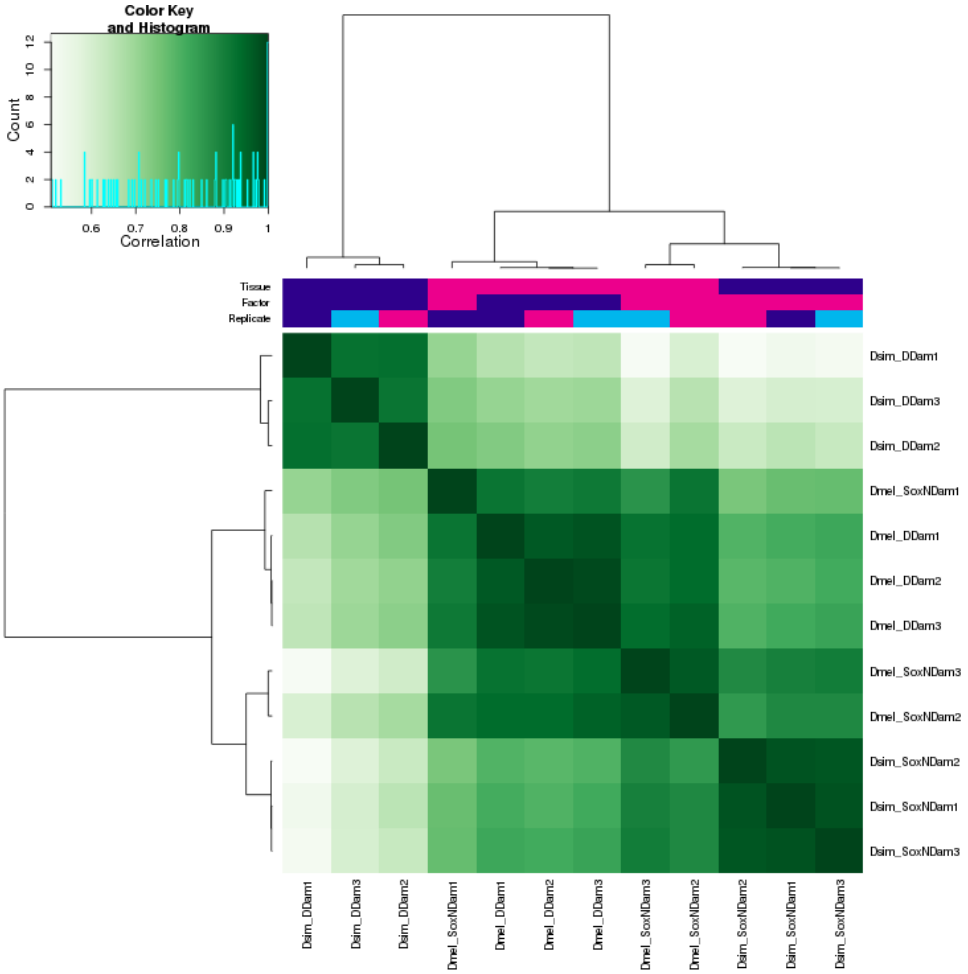
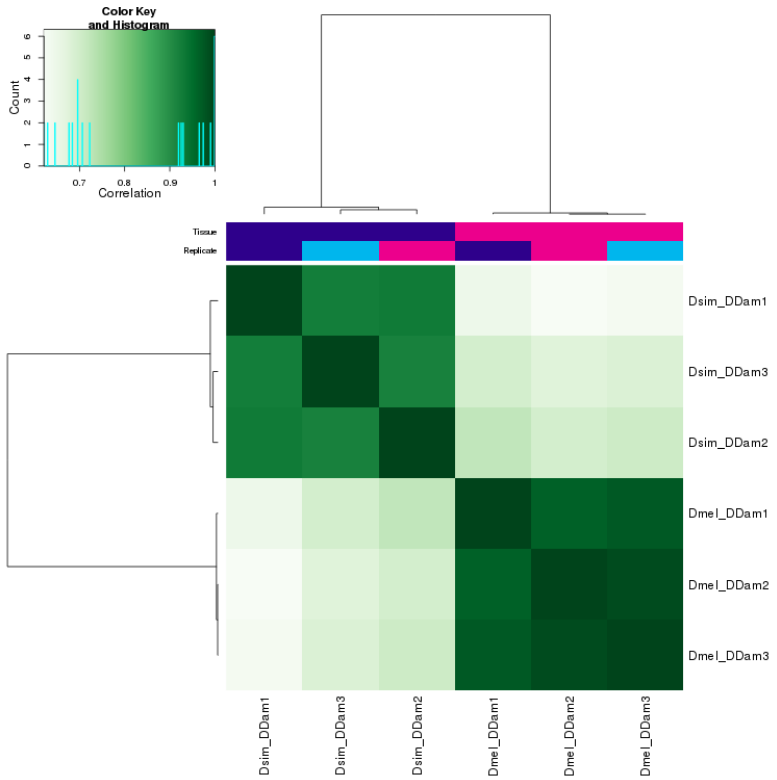


Figure 5.1: Clustering of *D. melanogaster* and *D. simulans* Dichaete-Dam and SoxN-Dam samples by binding affinity scores in all bound intervals. Both the Dichaete-Dam replicates from both species and the SoxN-Dam replicates from both species tend to cluster together, with the exception of *D. melanogaster* SoxN-Dam replicate 1. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.



[H]

Figure 5.2: Clustering of *D. melanogaster* and *D. simulans* Dichaete-Dam samples by binding affinity scores in all bound intervals. Biological replicates from each species cluster strongly together. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

bound in one than the other. These intervals represent binding events present in both species whose strength has changed quantitatively during evolution. All of the intervals called as preferentially bound in *D. simulans* were also identified as bound by *D. simulans* in the single-species DESeq2 analysis. Of these intervals, 1885 were also called as binding intervals at FDR5 in *D. melanogaster* in the single-species analysis, while 2154 were not. All but three of the intervals called as preferentially bound in *D. melanogaster* were called as binding intervals in the single-species analysis; this slight discrepancy results from normalizing the reads from both species together before performing the differential analysis. 1026 of these were also called as binding intervals at FDR5 in *D. simulans* in the single-species analysis, while 2178 were not.

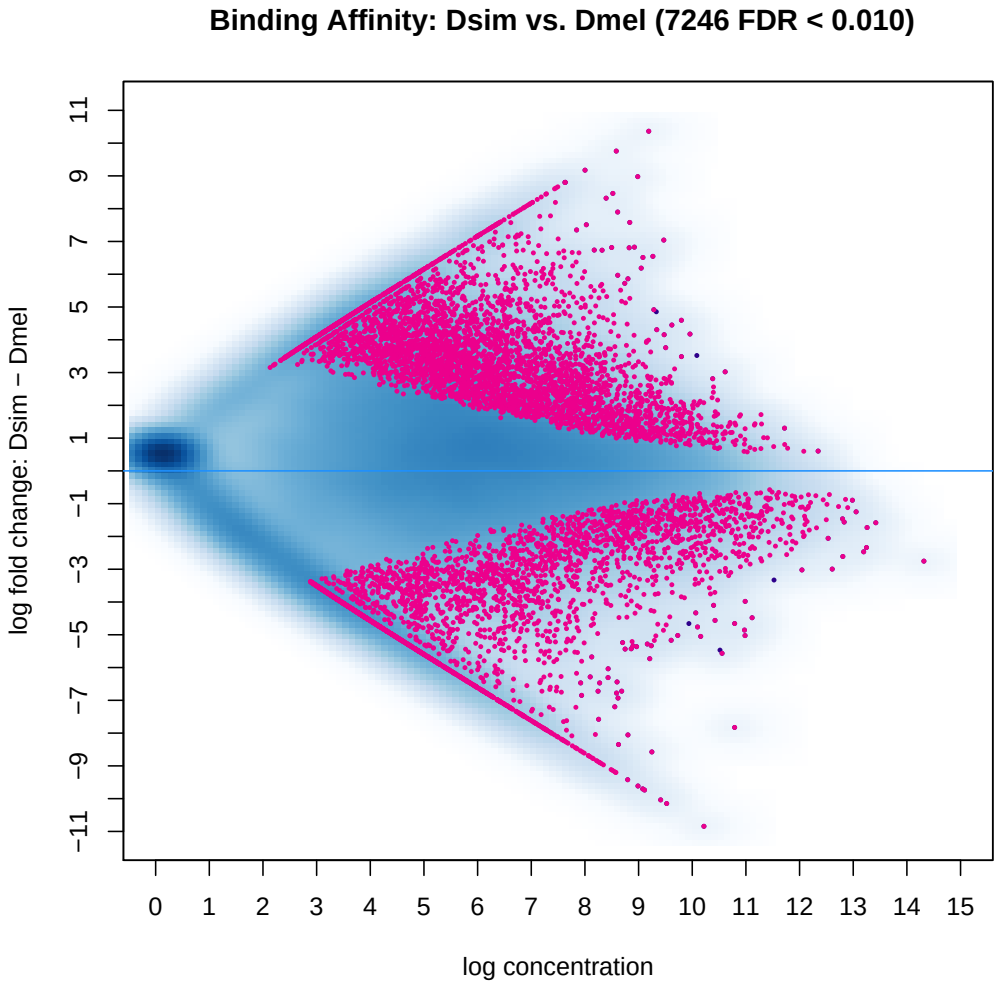


Figure 5.3: MA plot showing differentially bound intervals with FDR <0.01 between *D. melanogaster* Dichaete-Dam and *D. simulans* Dichaete-Dam. Intervals that are bound more strongly in *D. simulans* have a positive log fold change, while intervals that are bound more strongly in *D. melanogaster* have a negative log fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

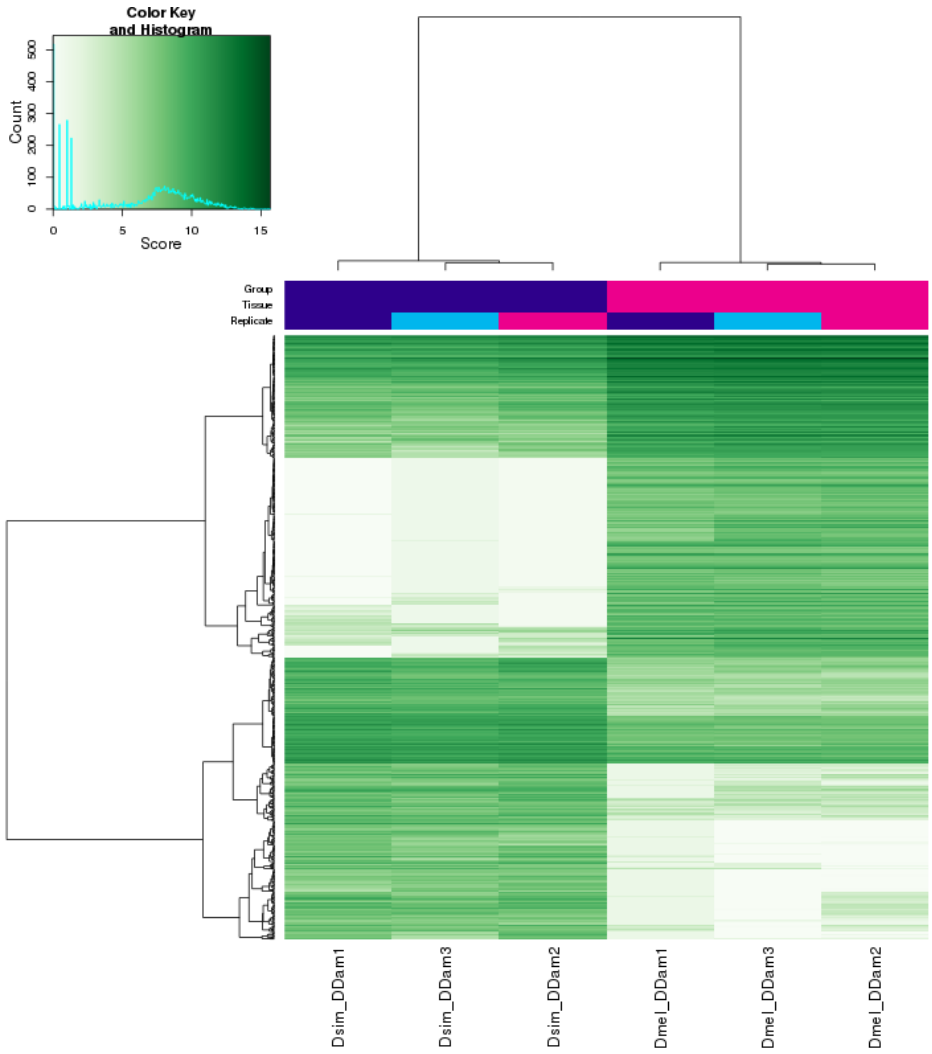


Figure 5.4: Clustering of *D. simulans* and *D. melanogaster* Dichaete-Dam differentially bound intervals by binding affinity scores. Roughly similar numbers of intervals are preferentially bound in each species. The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding.

For SoxN-Dam, a total of 24794 binding intervals were considered in both species. As with Dichaete-Dam, the replicates from each species cluster closely together. In this case, the three *D. simulans* replicates are the most similar, while the biggest outlier is replicate 1 for *D. melanogaster* (Figure 5.5). The correlations between binding profiles between *D. melanogaster* and *D. simulans* range from 0.75 - 0.90, while the correlations for replicates within a species range from 0.97 - 0.98 for *D. simulans* and from 0.87 - 0.97 for *D. melanogaster*. Using DESeq2 normalization, DiffBind identifies 9278 differentially bound intervals between the two species at FDR10. A high-confidence set of differentially bound intervals identified at FDR1 contains 4923 binding intervals (Figure 5.6), representing 32.5% of all *D. simulans* bound intervals and 21.4% of all *D. melanogaster* bound intervals. Of these, 2412 are preferentially bound in *D. simulans*, while 2511 are differentially bound in *D. melanogaster*. Clustering the differentially bound intervals by affinity score reveals that, as with Dichaete, the largest groups of intervals are highly bound only in either *D. melanogaster* or *D. simulans*; smaller blocks of intervals are bound in both species but more highly in one than in the other (Figure 5.7). Of the 2412 intervals that are preferentially bound in *D. simulans*, 2407 were called as binding intervals at FDR5 in the single-species analysis. 1101 of these were also called as bound at FDR5 in *D. melanogaster*, while 1306 were not. In this case, all 2511 intervals that are preferentially bound in *D. melanogaster* were called as binding intervals in the single-species analysis. Only 363 of these were also called as bound at FDR5 in *D. simulans*, while 2148 were not.

5.2.2 Quantitative comparison of Dichaete binding between *D. melanogaster* and *D. yakuba*

A total of 25072 unique binding intervals were considered by DiffBind in the comparison between Dichaete-Dam binding in *D. melanogaster* and *D. yakuba*. As with the Dichate-Dam data in *D. simulans*, there is a clear division between the binding intervals in *D. yakuba* and those in *D. melanogaster*, with the replicates from each species clustering closely together (Figure 5.8). The correlation coefficients between replicates from different species range from 0.68 - 0.70, while the correlation coefficients between replicates from the same species range from

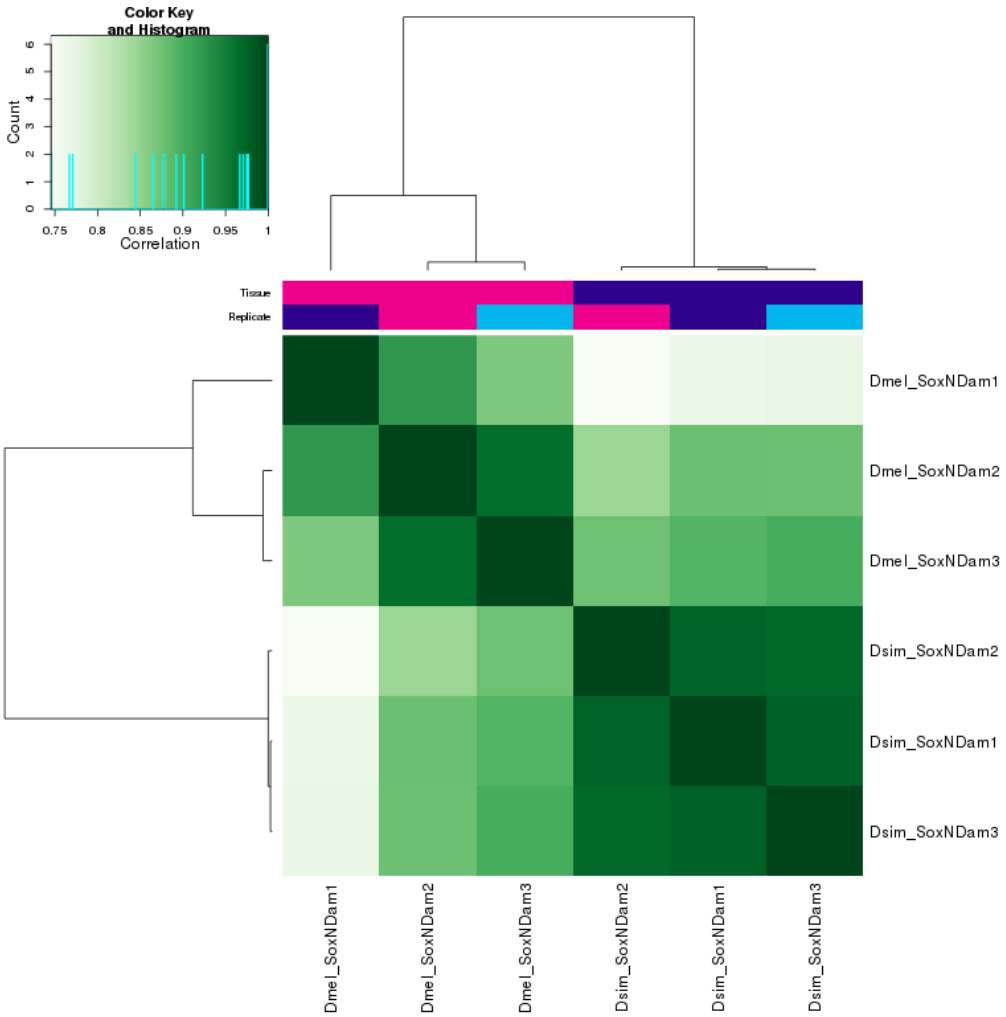


Figure 5.5: Clustering of *D. melanogaster* and *D. simulans* SoxN-Dam samples by binding affinity score in all bound intervals. Biological replicates from each species cluster together, although the *D. simulans* replicates show stronger correlations than the *D. melanogaster* replicates. The biggest outlier is *D. melanogaster* replicate 1. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

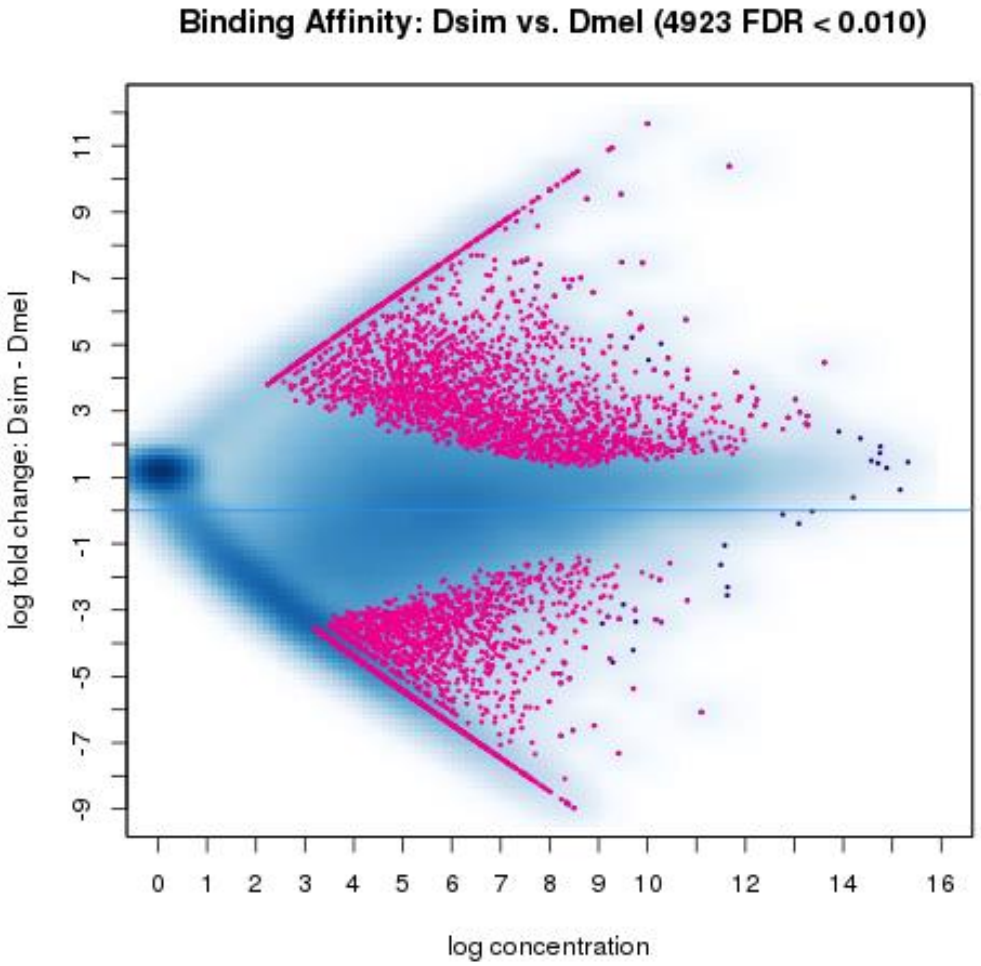


Figure 5.6: MA plot showing differentially bound intervals with FDR <0.01 between *D. simulans* SoxN-Dam and *D. melanogaster* SoxN-Dam. Intervals that are bound more strongly in *D. simulans* have a positive log fold change, while intervals that are bound more strongly in *D. melanogaster* have a negative log fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

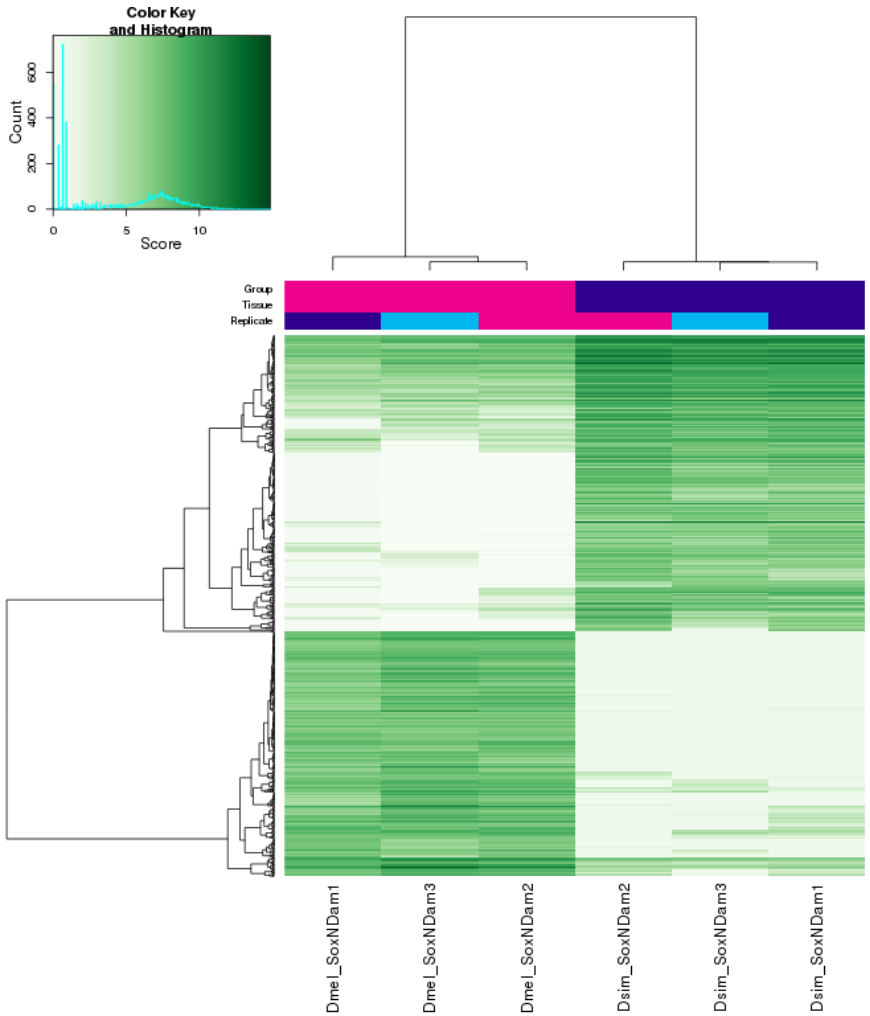


Figure 5.7: Clustering of *D. melanogaster* and *D. simulans* SoxN-Dam differentially bound intervals by binding affinity scores. Roughly similar numbers of intervals are preferentially bound in each species. The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding.

0.96 - 0.99, showing a very high degree of reproducibility. 13620 binding intervals were identified as differentially bound between the two species using DESeq2 normalization at FDR10. A more stringent, high-confidence set of differentially bound intervals at FDR1 contains 9205 binding intervals (Figure 5.9), representing 43.9% of all *D. yakuba* bound intervals and 44.2% of all *D. melanogaster* bound intervals. Of these, 4383 are preferentially bound in *D. yakuba* and 4822 are preferentially bound in *D. melanogaster*. Clustering the differentially bound intervals by affinity score reveals that, of those intervals preferentially bound in *D. yakuba*, the majority are also bound at a lower level in *D. melanogaster*, while a smaller number are unique to *D. yakuba* (Figure 5.10). Conversely, of those intervals preferentially bound in *D. melanogaster*, the majority are unique to that species, while a smaller number are also bound by *D. yakuba* at a lower level. Of the Dichaete-Dam intervals preferentially bound in *D. yakuba*, all but 2 were called as bound intervals at FDR5 in the single-species DESeq2 analysis. 2301 of these were also called as bound at FDR5 by *D. melanogaster*, while 2080 were not. Of the 4822 intervals preferentially bound in *D. melanogaster*, 4783 were called as bound intervals at FDR5 in the single-species analysis. 2193 of these were also called as bound at FDR5 in *D. yakuba*, while 2590 were unique to *D. melanogaster*.

5.2.3 Quantitative comparison of Dichaete binding between *D. melanogaster* and *D. pseudoobscura*

A total of 21294 unique binding intervals were considered by DiffBind in the comparison between Dichaete-Dam binding in *D. pseudoobscura* and *D. melanogaster*. The three replicates from each species cluster together; however, as expected given the noise present in the *D. pseudoobscura* data, the correlations between *D. pseudoobscura* replicates are much lower than those between *D. melanogaster* replicates (Figure 5.11). The biggest outlier is clearly replicate 1 from *D. pseudoobscura*. The correlation coefficients between replicates from *D. pseudoobscura* range from 0.46 - 0.79, while the correlation coefficients between replicates from *D. melanogaster* range from 0.96 - 0.99. The correlation coefficients between replicates from different species range from 0.46 - 0.72.

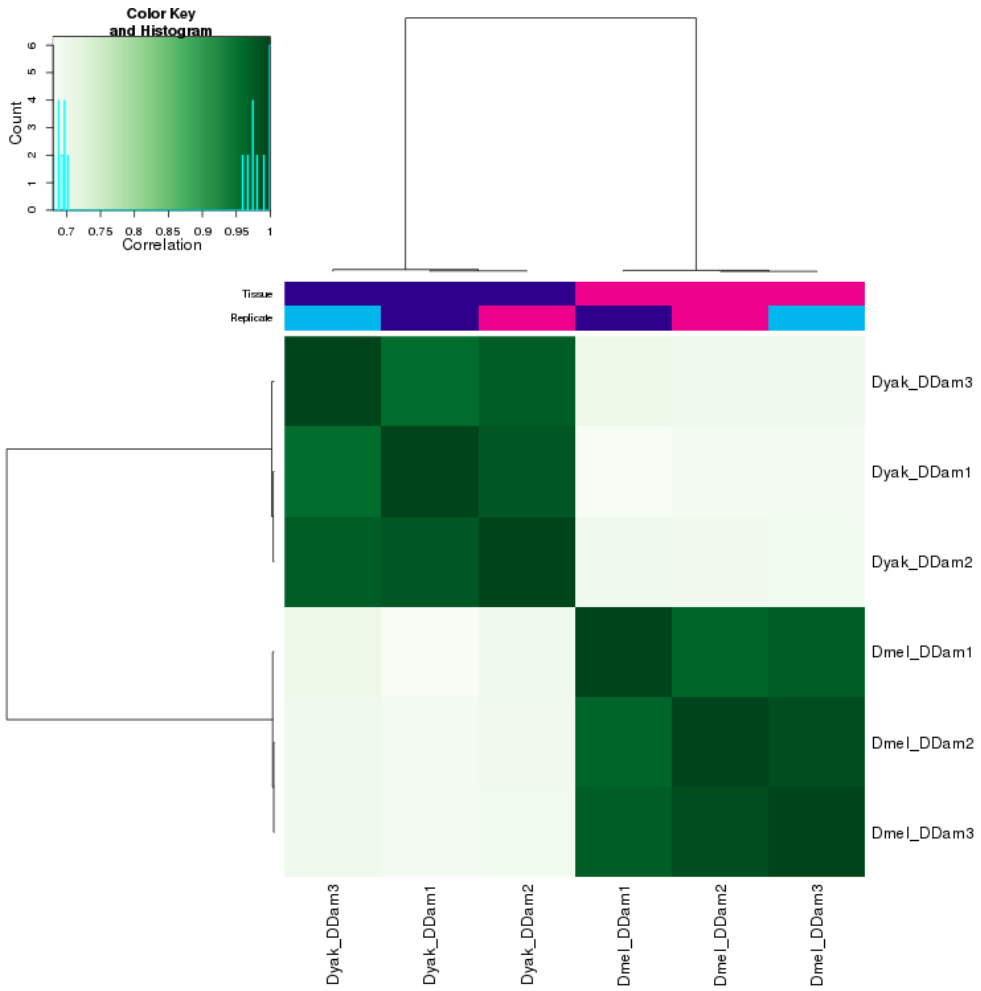


Figure 5.8: Clustering of *D. melanogaster* and *D. yakuba* Dichaete-Dam samples by binding affinity scores in all bound intervals. Biological replicates from each species cluster strongly together. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

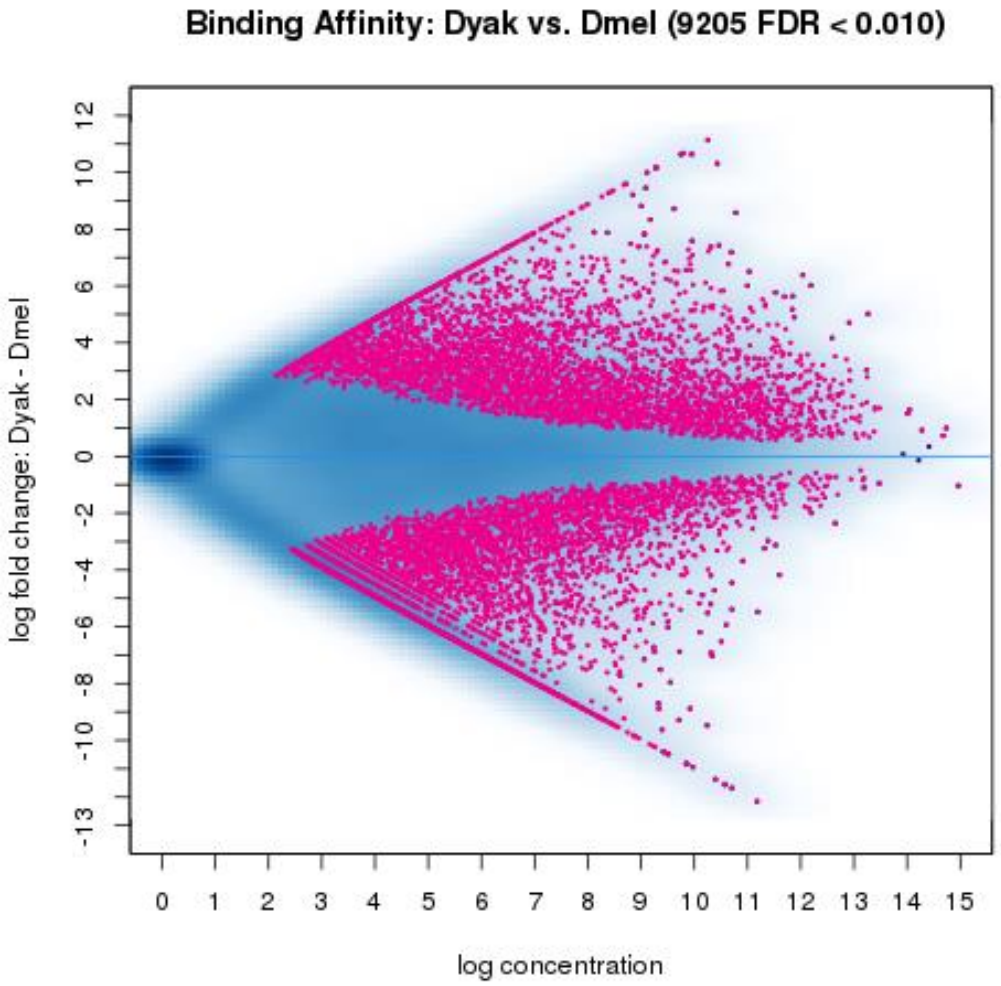


Figure 5.9: MA plot showing differentially bound intervals with FDR <0.01 between *D. melanogaster* Dichaete-Dam and *D. yakuba* Dichaete-Dam. Intervals that are bound more strongly in *D. yakuba* have a positive log fold change, while intervals that are bound more strongly in *D. melanogaster* have a negative log fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

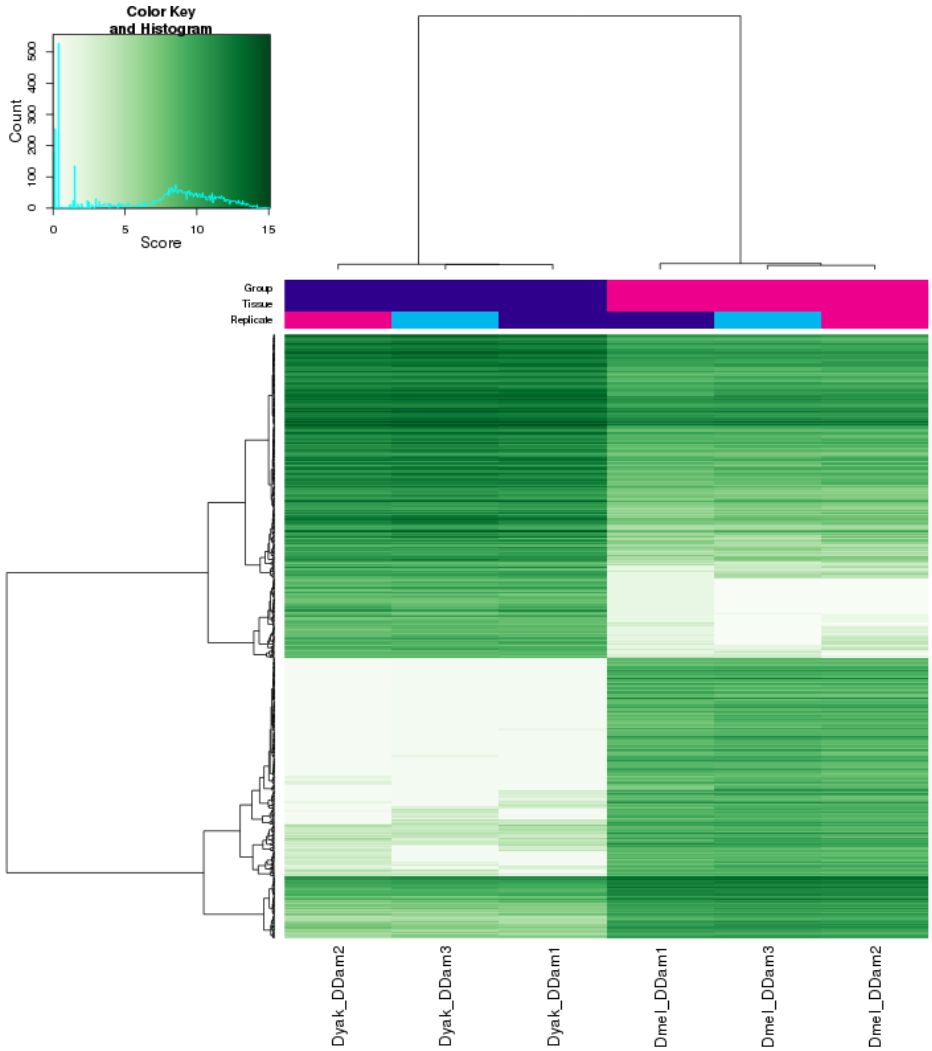


Figure 5.10: Clustering of *D. yakuba* and *D. melanogaster* Dichaete-Dam differentially bound intervals by binding affinity scores. Roughly similar numbers of intervals are preferentially bound in each species; however, the majority of intervals that are preferentially bound by *D. melanogaster* are not bound in *D. yakuba* (bottom half), while the majority of intervals that are preferentially bound in *D. yakuba* are also bound in *D. melanogaster* (top half). The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding.

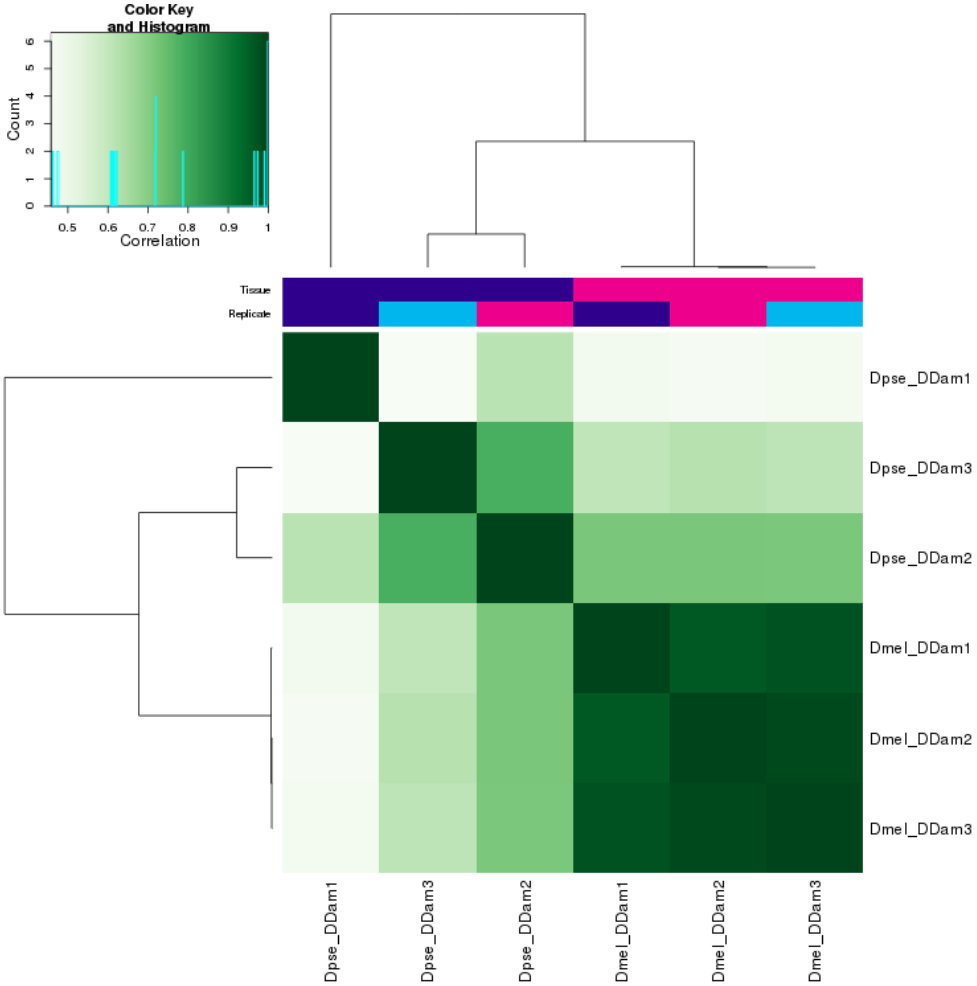


Figure 5.11: Clustering of *D. melanogaster* and *D. pseudoobscura* Dichaete-Dam samples by binding affinity scores in all bound intervals. Biological replicates from each species cluster together, although the *D. melanogaster* replicates are much more highly correlated than the *D. pseudoobscura* replicates. The strongest outlier is *D. pseudoobscura* replicate 1. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

This comparison presents a challenge for analysis, as, unlike the other pairwise comparisons between species, the two datasets are not of similar quality and do not contain similar numbers of binding intervals. In the previous DiffBind analyses using DESeq2, the sample affinity scores were normalized using only the reads present within binding intervals, as it could be assumed that the similar numbers of intervals present in the different samples reflected biological reality, and that most of the reads outside of those intervals represented background noise. However, that assumption was not valid in the case of the *D. pseudoobscura* data. Since there are many less binding intervals in the *D. pseudoobscura* dataset, normalizing by the number of reads within binding intervals would artificially inflate the affinity scores within those intervals relative to the ones present in the *D. melanogaster* dataset. Accordingly, the sample affinity scores were normalized using the total library sizes of each sample. This may result in an underestimation of the number of significant preferentially enriched binding intervals in *D. pseudoobscura*; however, it is the more conservative approach, and, as such, those intervals that are identified can be interpreted with high confidence.

Using this method, 12227 binding intervals were identified as differentially bound between the two species at FDR10, and 8105 high-confidence intervals were identified at FDR1 (Figure 5.12). Of the FDR1 intervals, only 321 were preferentially bound in *D. pseudoobscura*, while 7784 were preferentially bound in *D. melanogaster*. Of the intervals preferentially bound in *D. pseudoobscura*, 261 were called as binding intervals at FDR5 in the single-species DESeq2 analysis. 30 of these were also called as binding intervals at FDR5 in *D. melanogaster*, while 231 were unique to *D. pseudoobscura*. All of the intervals identified as preferentially bound in *D. melanogaster* were called as binding intervals in the single-species DESeq2 analysis. Of these, 338 were also called as binding intervals at FDR5 in *D. pseudoobscura*, while 7446 were unique to *D. melanogaster*. Almost all of the intervals that are preferentially bound in one species are not bound in the other, which can be seen when all of the differentially bound intervals are clustered by affinity score (Figure 5.13). This is in contrast to the pairwise comparisons for Dichaete-Dam in the other two species, where a sizeable proportion of differentially bound intervals in one species are also bound in the other species, albeit at a lower level.

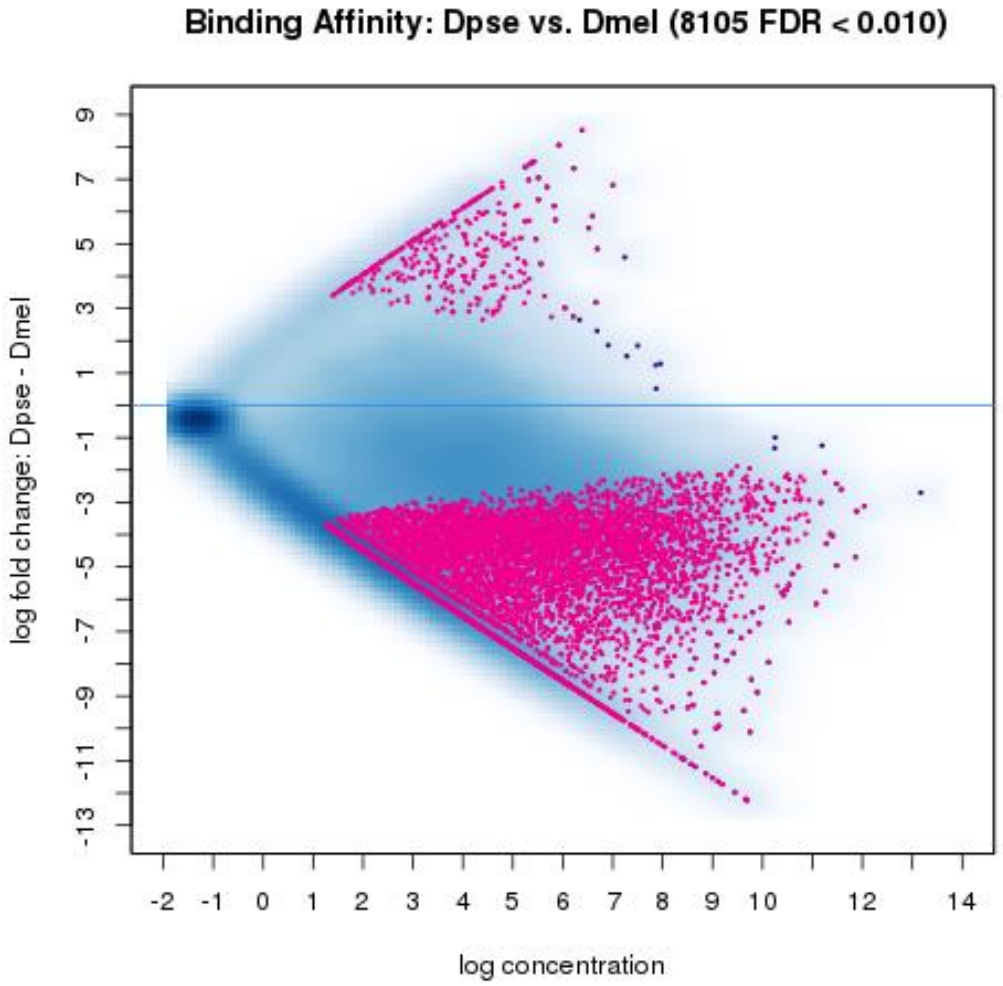


Figure 5.12: MA plot showing differentially bound intervals with FDR <0.01 between *D. melanogaster* Dichaete-Dam and *D. pseudoobscura* Dichaete-Dam. Intervals that are bound more strongly in *D. pseudoobscura* have a positive log fold change, while intervals that are bound more strongly in *D. melanogaster* have a negative log fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

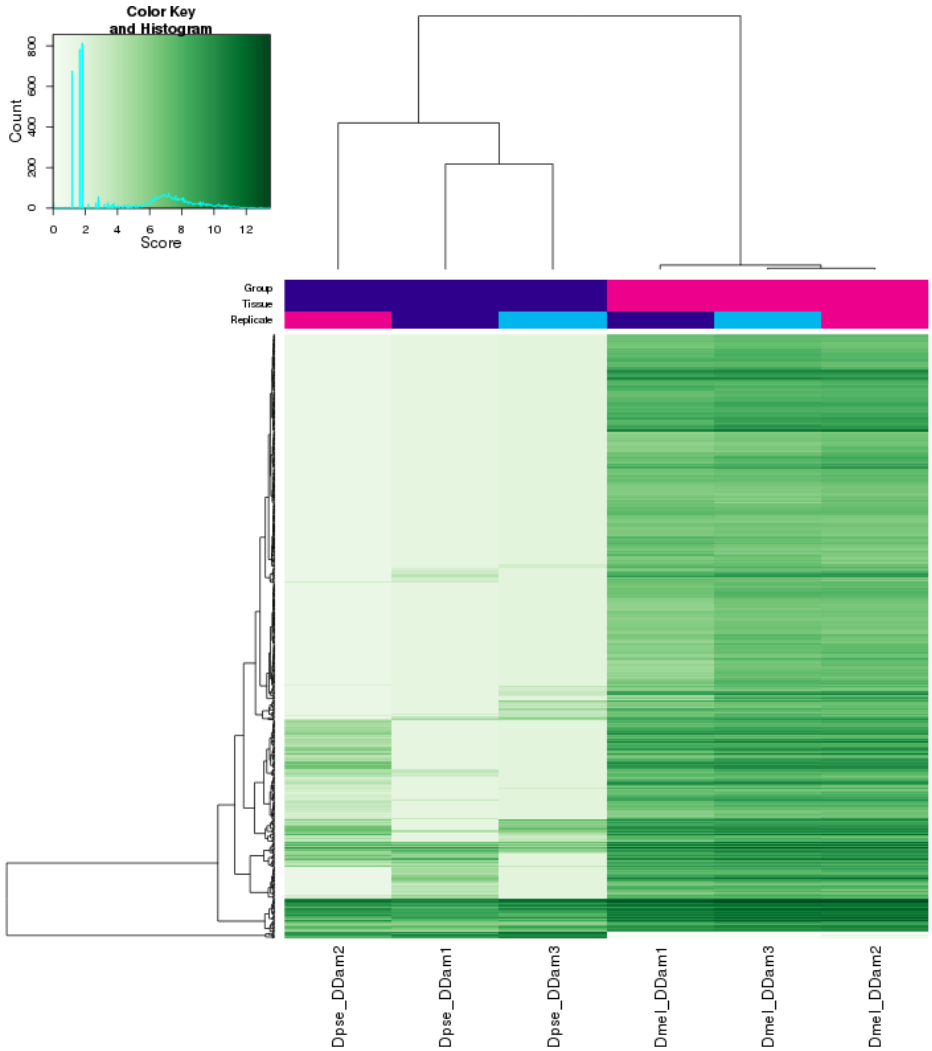


Figure 5.13: Clustering of *D. pseudoobscura* and *D. melanogaster* Dichaete-Dam differentially bound intervals by binding affinity scores. Many more intervals are preferentially bound in *D. melanogaster* (right side) than in *D. pseudoobscura* (left side) due to the higher noise and smaller number of intervals identified in *D. pseudoobscura*. The color key and histogram shows the distribution of binding affinity scores (log of normalized read counts), in all bound intervals in each sample. Darker green corresponds to higher affinity scores or stronger binding, while lighter green corresponds to lower affinity scores or weaker binding

5.2.4 Summary of pairwise binding divergence

The numbers of Dichaete-Dam binding intervals called as differentially enriched between *D. melanogaster* and each other species increase with phylogenetic distance from *D. simulans* to *D. yakuba*. Although less differential intervals are identified at FDR1 between *D. melanogaster* and *D. pseudoobscura* than between *D. melanogaster* and *D. yakuba*, the percentage of intervals that are unique to *D. melanogaster* in comparison to *D. pseudoobscura* is greater, and the number of intervals that are unique to *D. pseudoobscura* are likely underestimated due to the normalization method employed. The total numbers of intervals identified as differentially bound in each comparison are summarized in Table 5.1. Interestingly, the proportions of binding intervals that are qualitatively absent in non-*melanogaster* species versus intervals that are present but have a quantitative change in binding strength vary between different species as well as between Dichaete and SoxN. For Dichaete-Dam, roughly equal percentages of the total binding intervals called in *D. melanogaster* (20848) are qualitatively absent and present but quantitatively changed, either increasing or decreasing in binding strength, in *D. simulans* (10.4% and 10.6%, respectively). However, while a similar proportion are qualitatively absent in *D. yakuba* (12.4%), roughly double the percentage of intervals are present but have quantitative changes in binding affinity (21.6%). The proportion of *D. melanogaster* intervals that are qualitatively absent in *D. pseudoobscura*, 35.7%, is likely exaggerated by the lower quality of the *D. pseudoobscura* data; however, it is interesting that a much lower percentage of *D. melanogaster* intervals are present but show quantitative changes in *D. pseudoobscura* (1.8%). It should be noted that these percentages are lower than the percentage of non-overlapping intervals arrived at by simply intersecting binding intervals in *D. melanogaster* and *D. pseudoobscura*; this is due to the effect of normalizing all of the samples together. For SoxN-Dam, only one comparison was possible, between *D. melanogaster* and *D. simulans*. Of the total number of binding intervals called for SoxN-Dam in *D. melanogaster* (22952), 9.4% are qualitatively absent in *D. simulans* and only 6.4% are present but show quantitative changes in binding affinity between the two species. Overall, pairwise comparisons for each TF reveal a significant contribution of quantitative binding divergence at bound loci as well as gain or loss of binding intervals, with the

proportion of *D. melanogaster* intervals that are not bound in each other species increasing with evolutionary distance.

Comparison	**Total Differential Intervals (1% FDR)**	***D. mel* unique intervals**	**Other species unique intervals**	**Shared intervals**
Dichaete *D. mel* vs. *D. sim*	7246	2178	2154	2914
Dichaete *D. mel* vs. *D. yak*	9205	2590	2080	4535
Dichaete *D. mel* vs. *D. pse*	8105	7446	231	428
SoxN *D. mel* vs. *D. sim*	4923	2148	1306	1469

Table 5.1: Summary of quantitative differences in binding for Dichaete and SoxN between each pair of species. *D. mel* unique intervals are those that are only called as bound in *D. melanogaster*, while other species unique intervals are those that are only called as bound in each other species. Shared intervals are called as bound in both species in a comparison, but quantitatively bound more highly in one.

5.3 Three-way comparison of Dichaete binding patterns

Using the three best Dichaete-Dam binding datasets, from *D. melanogaster, D. simulans* and *D. yakuba*, I undertook a three-way comparison of Dichaete-Dam binding patterns using DiffBind. A total of 26117 unique binding intervals, which were bound in a least one of the three species, were considered. On a qualitative level, a core set of 7739 binding intervals are present and conserved between all three species. A total of 6119 binding intervals are conserved between any two species, and 12259 are unique to a single species (Figure 5.14).

D. yakuba has the highest percentage of unique binding intervals (32%) and the lowest percentage of 3-way conserved binding intervals (42%), while *D. simulans* has the lowest percentage of unique binding intervals (15%) and the highest percentage of 3-way conserved binding intervals (63%). The correlation coefficients

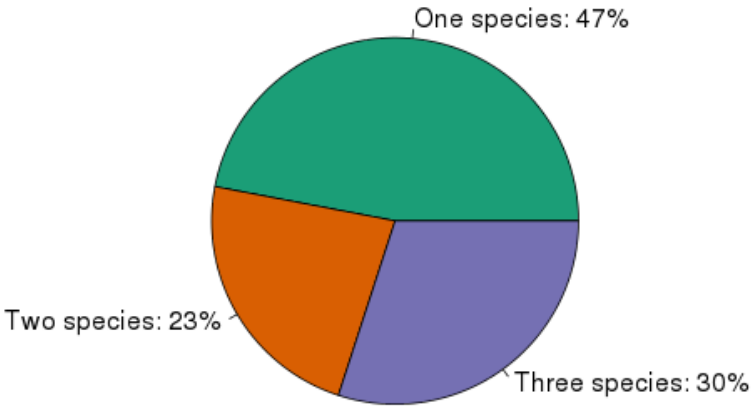


Figure 5.14: Proportions of all Dichaete-Dam binding intervals identified that are qualitatively conserved in one, two, and three species.

for read counts within all intervals between replicates within each species were quite high, ranging from 0.97 - 0.99 for *D. melanogaster*, from 0.92 - 0.93 for *D. simulans*, and from 0.96 - 0.98 for *D. yakuba*. When comparing replicates between species, the correlations decrease to 0.65 - 0.74 between *D. melanogaster* and *D. simulans*, 0.69 - 0.71 between *D. melanogaster* and *D. yakuba*, and 0.64 - 0.72 between *D. simulans* and *D. yakuba* (Figure 5.15). Because of the greater variance between the *D. simulans* replicates, it is difficult to determine whether Dichaete-Dam binding in *D. melanogaster* is more similar to binding in *D. simulans* or *D. yakuba* based on the coefficients of correlations alone. To get a better idea of the overall similarities and differences between the different datasets, I also performed a principal component analysis (PCA) on all of the samples (Figure 5.16).

The first principal component, which explains 41% of the variation at bound intervals, separates *D. melanogaster* and *D. simulans* from *D. yakuba*. The second principal component, which explains 35% of the variation at bound intervals, separates *D. melanogaster* from *D. simulans* and *D. yakuba*. This indicates that the primary driver of variation between the three species corresponds to changes in binding in *D. yakuba* relative to the other two species, which is in line with the

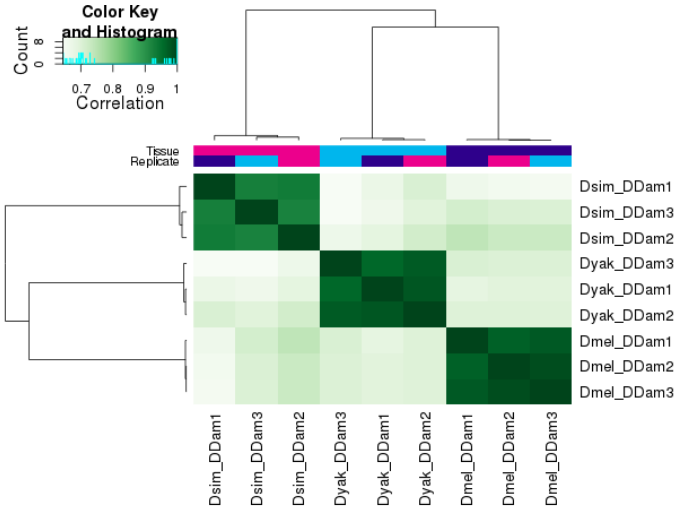


Figure 5.15: Clustering of *D. melanogaster*, *D. simulans* and *D. yakuba* Dichaete-Dam samples by binding affinity scores in all bound intervals. Biological replicates from each species cluster strongly together. The color key and histogram shows the distribution of correlation coefficients for affinity scores in each pair of samples. Darker green corresponds to a higher correlation between samples, while lighter green corresponds to a lower correlation.

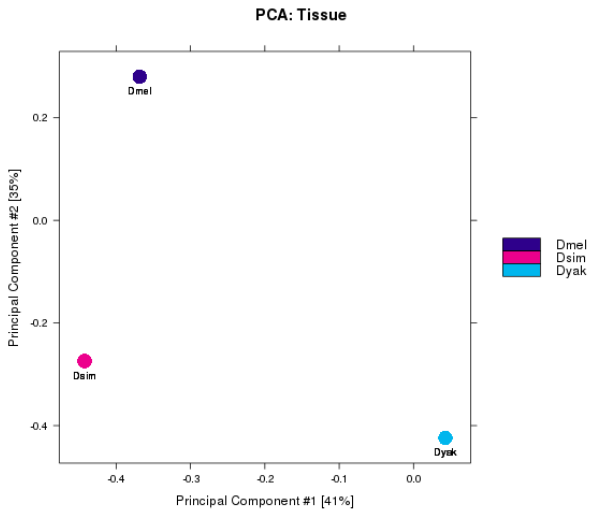
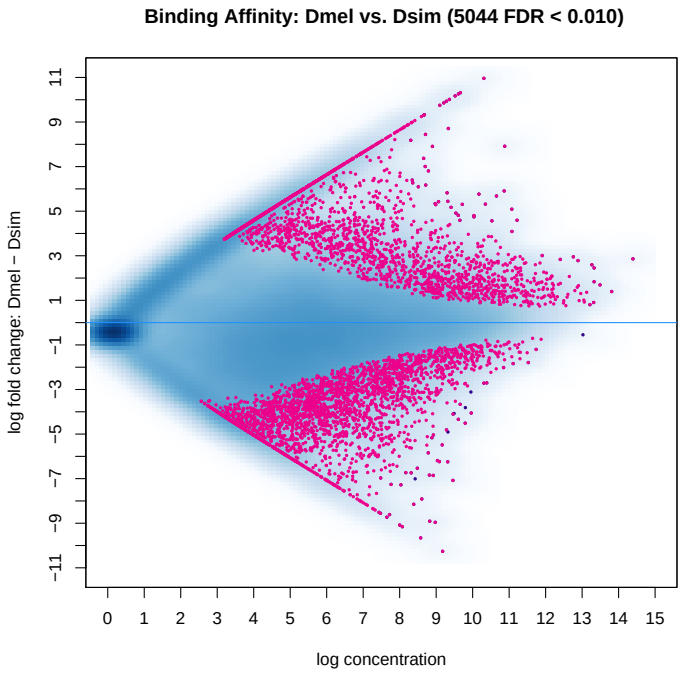
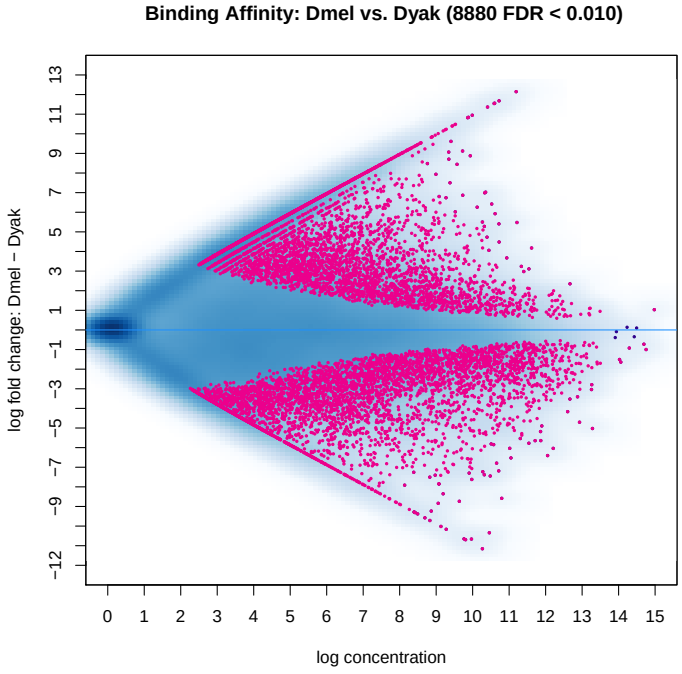


Figure 5.16: Prinicipal component analysis of binding affinity scores in bound intervals for *D. melanogaster*, *D. simulans* and *D. yakuba* Dichaete-Dam samples. The first principal component separates *D. yakuba* from the other two species, while the second principal component separates *D. melanogaster* from *D. simulans* and *D. yakuba*.

expectation of neutral evolution along the *Drosophila* phylogeny, as *D. yakuba* is the most distant from *D. melanogaster* (Russo *et al.*, 1995). In agreement with this observation, DiffBind identifies 5044 binding intervals that are differentially bound between *D. melanogaster* and *D. simulans* (Figure 5.17A), 8880 that are differentially bound between *D. melanogaster* and *D. yakuba* (Figure 5.17B), and 6324 that are differentially bound between *D. simulans* and *D. yakuba* (Figure 5.17C). Although these numbers are different from the numbers of differentially bound intervals detected in pairwise comparisons by DiffBind, since the pairwise comparisons started with different total sets of intervals and normalized the affinity scores between different sets of samples, the two analyses are broadly in agreement.

As with the pairwise comparisons, the percentages of all *D. melanogaster* binding intervals that are identified as divergent in another species using a three-way comparison increase with phylogenetic distance. According to this analysis, 6.9% of all *D. melanogaster* intervals are qualitatively absent in *D. simulans*, while 9.6% are present but show quantitative changes in binding affinity. 11.6% of *D. melanogaster* intervals are qualitatively absent in *D. yakuba*, while 20.4% are present but show quantitative changes in binding affinity. The proportion of divergent intervals that are due to quantitative changes also increases with phylogenetic distance, from 58.0% in *D. simulans* to 63.7% in *D. yakuba*. By normalizing the data from all three species together, this three-way comparison presents a more generalized picture of how Dichaete binding varies both qualitatively and quantitatively across the *melanogaster* clade and confirms the phylogenetic patterns observed in pairwise comparisons.


Binding Affinity: Dmel vs. Dsim (5044 FDR < 0.010)
log fold change: Dmel – Dsim
11
9
7
5
3
1
−1
−3
−5
−7
−9
−11
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
log concentration


Binding Affinity: Dmel vs. Dyak (8880 FDR < 0.010)
log fold change: Dmel – Dyak
13
11
9
7
5
3
1
−1
−3
−5
−7
−9
−12
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
log concentration

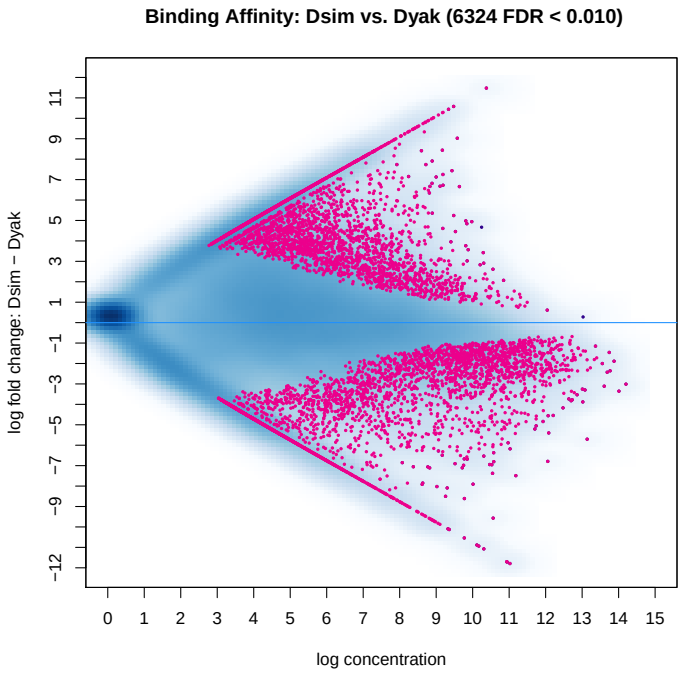


Figure 5.17: MA plots showing differentially bound Dichaete-Dam intervals with FDR <0.01 between pairs of species using normalization between three species. A.) Differentially bound intervals between *D. melanogaster* and *D. simulans*. Intervals that are more strongly bound in *D. melanogaster* have a positive log fold change, while intervals that are more strongly bound in *D. simulans* have a negative log fold change. B.) Differentially bound intervals between *D. melanogaster* and *D. yakuba*. Intervals that are more strongly bound in *D. melanogaster* have a positive fold change, while intervals that are differentially bound in *D. yakuba* have a negative fold change. C.) Differentially bound intervals between *D. simulans* and *D. yakuba*. Intervals that are more strongly bound in *D. simulans* have a positive fold change, while intervals that are more strongly bound in *D. yakuba* have a negative fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

5.4 Binding site turnover within gene loci

It has been hypothesized that, as the percentage of conserved binding events at orthologous positions decreases between more distantly related species, new binding events at the same gene loci should evolve in order to maintain the same level of gene expression; this is often referred to as binding site turnover or compensatory evolution (Arnold *et al.*, 2014; Bradley *et al.*, 2010; He *et al.*, 2011a;

Moses *et al.*, 2006). In order to detect instances where a binding interval that is lost in one species might be compensated for by the gain of a new binding interval at the same gene locus in another species, I first took the set of all binding intervals called for each factor in each species, then did pairwise comparisons to find those intervals in one species that did not overlap with any binding intervals in the other. I did this for Dichaete-Dam between *D. melanogaster* and *D. simulans* and between *D. melanogaster* and *D. yakuba*, as well as for SoxN-Dam between *D. melanogaster* and *D. simulans*; I excluded *D. pseudoobscura* because of the highly mismatched numbers of binding intervals called between it and *D. melanogaster*. I then took the resulting lists of intervals and assigned them all to the nearest genes within 10 kb upstream or downstream, as described previously. Finally, I found every instance where two non-overlapping binding intervals, one from each species, were annotated to the same gene. I considered these intervals to show compensatory, rather than positional, conservation between each pair of species. An example of compensatory conservation at the *reduced ocelli (rdo)* locus between *D. melanogaster* and *D. yakuba* is shown in Figure 5.18.

For Dichaete-Dam, I detected 5351 intervals in *D. melanogaster* that show compensatory conservation relative to *D. simulans*, and 3226 intervals in *D. simulans* that show compensatory conservation relative to *D. melanogaster*. In total, these pairs of intervals are located at 2457 unique genes. The greater number of compensatory intervals detected in *D. melanogaster* may be due to the fact that more binding intervals were called in *D. melanogaster* overall. I detected 4924 intervals for Dichaete-Dam in *D. melanogaster* that show compensatory conservation relative to *D. yakuba*, and 5083 that show compensatory conservation in *D. yakuba* relative to *D. melanogaster*, altogether located at 2806 unique genes. For SoxN-Dam, I detected 5497 binding intervals that show compensatory conservation in *D. melanogaster* relative to *D. simulans*, and 2939 that show compensatory conservation in *D. simulans* relative to *D. melanogaster*. These occur at 2393 unique genes.

Compensatory evolution has also been detected for active enhancers identified via STARR-seq; approximately 19% of *D. melanogaster* enhancers showed compensatory conservation in *D. yakuba*, and this percentage increased with evolutionary distance, as the percentage of positionally conserved enhancers decreased (Arnold *et al.*, 2014). In the case of Dichaete-Dam, 23.6% of *D. melanogaster* binding in-

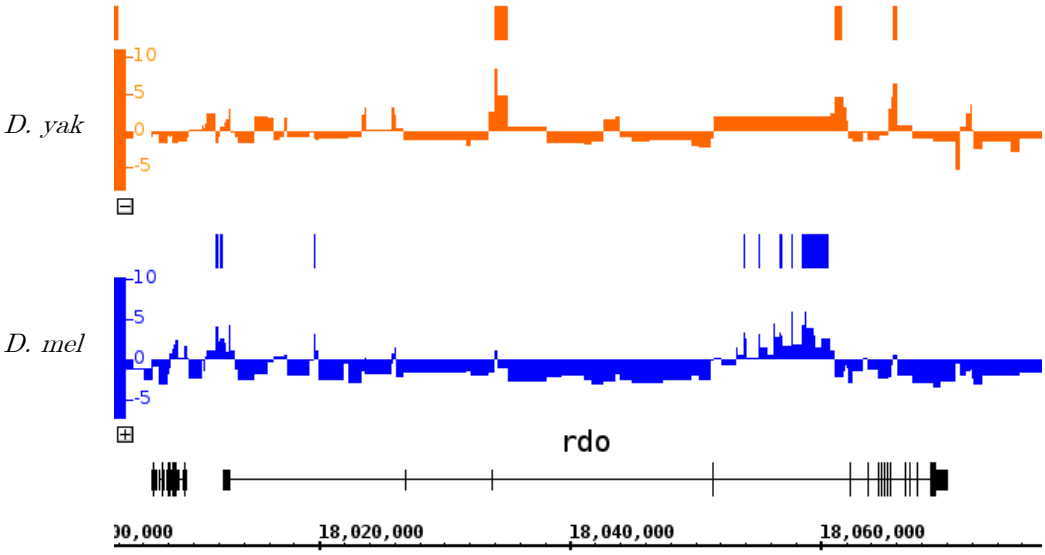


Figure 5.18: Dichaete-Dam binding site turnover between *D. melanogaster* and *D. yakuba* at the *reduced ocelli (rdo)* locus. Tracks are, from bottom, gene models (black), *D. melanogaster* Dichaete-Dam binding profile (blue), *D. melanogaster* Dichaete-Dam FDR5 bound intervals (blue bars), *D. yakuba* Dichaete-Dam binding profile (orange) and *D. yakuba* Dichaete-Dam FDR5 bound intervals (orange bars). For clarity, bound intervals that are positionally conserved between both species are not shown. Strong binding is observed in the third, fourth and eleventh introns in *D. yakuba*; these binding sites are lost in *D. melanogaster*, but several binding intervals are gained in the first and fourth introns. Binding profiles represent normalized log2 ratios of Dichaete-Dam binding to Dam-only binding in each GATC fragment.

tervals show compensatory conservation in *D. yakuba*, a slightly higher rate than for STARR-seq enhancers. In order to determine whether turnover of binding intervals is correlated with turnover of active enhancers, I followed the same protocol to identify pairs of compensatory enhancers between *D. melanogaster* and *D. yakuba*. For both S2 and OSC STARR-seq enhancers, I found all enhancers in one species that do not overlap with any enhancer in the other; I then assigned these to the closest genes within 10kb upstream and downstream and found all instances where two non-overlapping enhancers from different species were annotated to the same gene. Starting with the unfiltered lists of STARR-seq enhancers, this resulted in 21105 S2 enhancers in *D. melanogaster* that show compensatory conservation relative to *D. yakuba* and 22444 in *D. yakuba* that show compensatory conservation relative to *D. melanogaster*. These pairs of enhancers are annotated to 7514 unique genes. For OSCs, it resulted in 17843 enhancers that show compensatory conservation in *D. melanogaster* relative to *D. yakuba* and 20207 in *D. yakuba* that show compensatory conservation relative to *D. melanogaster*. These pairs of enhancers are annotated to 6941 unique genes.

Of the Dichaete-Dam intervals that are compensatory in *D. melanogaster* relative to *D. yakuba*, 233 were previously annotated to a STARR-seq enhancer in S2 cells and 326 were annotated to a STARR-seq enhancer in OSCs. 53 of these S2 enhancers and 105 of these OSC enhancers also show compensatory conservation in *D. melanogaster*. Conversely, of the Dichaete-Dam intervals that are compensatory in *D. yakuba* relative to *D. melanogaster*, 398 were previously annotated to a STARR-seq enhancer in S2 cells and 655 were annotated to a STARR-seq enhancer in OSCs. Only 90 of these S2 enhancers and 157 of these OSC enhancers also show compensatory conservation in *D. yakuba*. This result was somewhat surprising, as I expected that the same forces driving turnover of enhancer function between the two species would also drive turnover of group B Sox binding. However, it shows that, while some instances of the evolution of a new, compensatory Dichaete binding site in one species are located within compensatory enhancers in that species, the majority of Dichaete binding turnover events happen either in active enhancers that are positionally conserved in both species or outside of annotated STARR-seq enhancers. Because Dichaete and SoxN show such strong overlap in binding and an ability to compensate for each others loss (Ferrero *et al.*, 2014), it is possible that a SoxN binding site might evolve to

compensate for the loss of a Dichaete binding site within some enhancers; unfortunately, I was unable to test this without *in vivo* SoxN binding data in *D. yakuba*.

5.5 Binding conservation and regulatory function

5.5.1 Binding conservation at known enhancers

In my analysis of Dichaete-Dam and SoxN-Dam binding intervals in *D. melanogaster*, I found a high rate of overlap between group B Sox binding and known enhancers from the REDFly and FlyLight databases (Gallo *et al.*, 2010; Manning *et al.*, 2012). In order to test the hypothesis that conservation of binding between species should reflect functionality, I examined the proportion of binding intervals that are qualitatively conserved both within these known enhancers and outside of them. For Dichaete-Dam, I used the binding intervals identified in the three-way binding comparison in *D. melanogaster, D. simulans*, and *D. yakuba*. I considered three-way binding site conservation as well as pairwise conservation between *D. melanogaster* and each other species. For SoxN-Dam, I used the binding intervals identified in the pairwise comparison between *D. melanogaster* and *D. simulans*. I compared the conservation status of binding intervals that overlap with known REDFly and FlyLight enhancers and those that do not.

While a relatively low number of Dichaete-Dam binding intervals overlap REDFly enhancers (751 total intervals), 64.4% of these show three-way conservation between all species, compared to only 44.8% of all binding intervals that do not overlap a REDFly enhancer (Figure 5.19A). Only 9.6% of binding intervals overlapping a REDFly enhancer are unique to *D. melanogaster*, while 27.6% of those that do not overlap an enhancer are unique. Looking at pairwise conservation, being located within an enhancer does not have much of an effect; 3.6% of intervals within a REDFly enhancer are conserved between *D. melanogaster* and *D. simulans* compared to 7.2% of intervals that are not within an enhancer, while 22.2% of intervals within a REDFly enhancer are conserved between *D.*

melanogaster and *D. yakuba* compared to 20.3% of intervals that are not within a REDFly enhancer. However, performing a chi-squared test on this data reveals that, overall, the difference in conservation between binding intervals that do or do not overlap a REDFly enhancer is highly significant (χ^2 = 161.9, d.f. = 3, p-value = 7.06e-35).

A similar pattern can be seen with the FlyLight enhancers, although the effect is slightly smaller (Figure 5.19B). In this case, a total of 2531 Dichaete-Dam binding intervals overlap with an enhancer. 54.7% of these intervals show three-way conservation between *D. melanogaster, D. simulans* and *D. yakuba*, compared to 44.2% of binding intervals that do not overlap with an enhancer. 5.9% of binding intervals located within a FlyLight enhancer show pairwise conservation between *D. melanogaster* and *D. simulans* compared to 7.2% of intervals that are not located within an enhancer, while 22.6% of binding intervals located within a FlyLight enhancer show pairwise conservation between *D. melanogaster* and *D. yakuba* compared to 20.0% of intervals outside of an enhancer. Only 16.7% of intervals overlapping a FlyLight enhancer are unique to *D. melanogaster*, while 28.6% of intervals not overlapping an enhancer are unique. Again, a chi-squared test shows that the effect of being located within a FlyLight enhancer on conservation is highly significant (χ^2 = 177.3, d.f. = 3, p-value = 3.38e-38).

For SoxN, a set of binding intervals with three-way conservation was not available; however, even at the level of two-way conservation between *D. melanogaster* and *D. simulans*, binding intervals that are located within known enhancers are much more likely to be conserved. The effect is particularly strong for REDFly enhancers. A total of 799 SoxN-Dam binding intervals are located within a REDFly enhancer; 79.0% of these are qualitatively conserved between *D. melanogaster* and *D. simulans*, while only 46.5% of binding intervals located outside of a REDFly enhancer are conserved. Conversely, 21.0% of intervals within a REDFly enhancer are unique to *D. melanogaster*, compared to 53.5% of binding intervals not within a REDFly enhancer (Figure 5.19C). This effect is highly significant according to a chi-squared test (χ^2 = 323.6, d.f. = 1, p-value = 2.40e-72). The effect of being located within a FlyLight enhancer on conservation is slightly less strong but still substantial. 2844 SoxN-Dam binding intervals overlap a FlyLight enhancer; of these, 65.3% are conserved between *D. melanogaster* and *D. simulans*, compared to 45.1% of intervals outside of a FlyLight enhancer. 34.7%

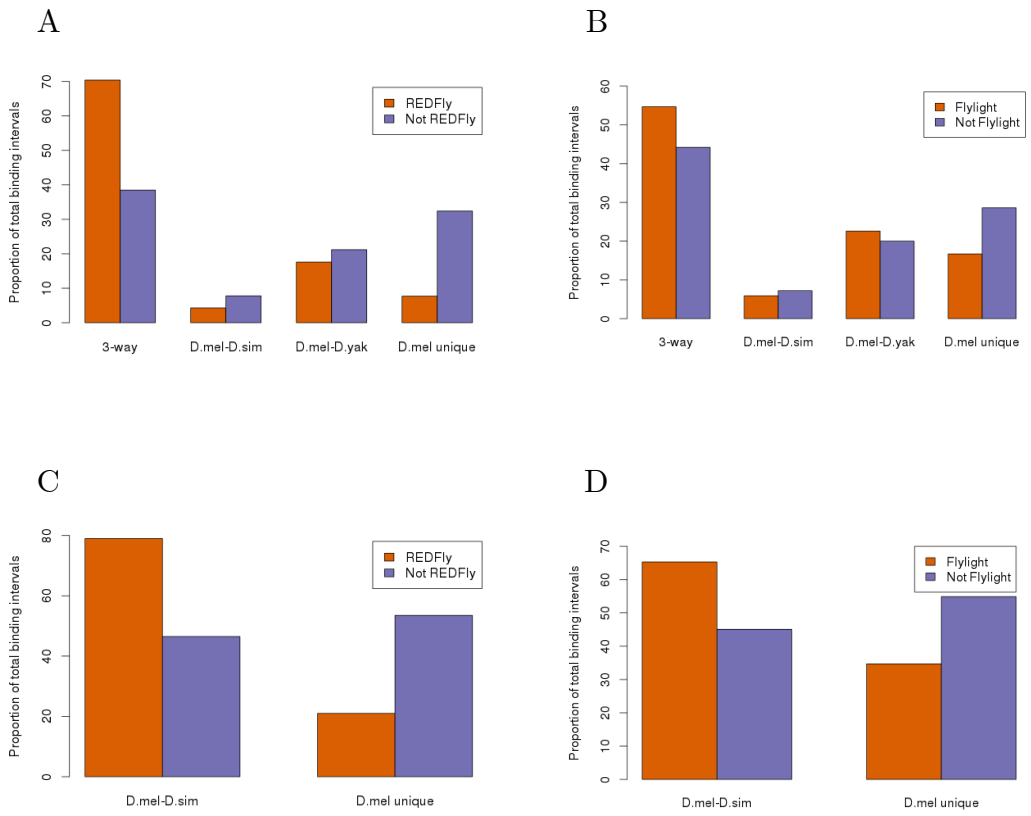


Figure 5.19: DamID binding intervals that overlap an annotated enhancer are more likely to be conserved than those that do not. A.) Dichaete-Dam binding intervals that overlap with a REDFly enhancer are more likely to show three-way conservation between *D. melanogaster, D. simulans* and *D. yakuba* (3-way) and are less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a REDFly enhancer (p = 7.06e-35). B.) Dichaete-Dam binding intervals that overlap with a FlyLight enhancer are more likely to show three-way conservation between *D. melanogaster, D. simulans* and *D. yakuba* (3-way) and are less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a FlyLight enhancer (p = 3.38e-38). C.) SoxN-Dam binding intervals that overlap with a REDFly enhancer are more likely to be conserved between *D. melanogaster* and *D. simulans* (D. mel - D. sim) and less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a REDFly enhancer (p = 2.40e-72). D.) SoxN-Dam binding intervals that overlap with a FlyLight enhancer are more likely to be conserved between *D. melanogaster* and *D. simulans* (D. mel - D. sim) and less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a FlyLight enhancer (p = 7.27e-12).

of binding intervals within a FlyLight enhancer are unique to *D. melanogaster*, while 54.9% of those not within a FlyLight enhancer are unique (Figure 5.19D). Again, a chi-squared test shows that this is a significant effect (χ^2 = 46.95, df = 1, p-value = 7.27e-12). The strong increase in conservation observed for binding intervals within enhancers indicates that these binding sites are likely under balancing selection to maintain their effect on gene regulation and confirms the hypothesis that functionally important binding events should show a high rate of evolutionary conservation.

5.5.2 Binding conservation at group B Sox core intervals

I was also interested in testing whether the core Dichaete and SoxN binding intervals that were previously identified in *D. melanogaster* were highly conserved between species. Since these intervals were identified in multiple experiments, we have high confidence that they are functional in the sense that they are truly bound *in vivo* in *D. melanogaster*; however, they are not necessarily the sites of direct gene regulatory activity by Dichaete and SoxN. Interestingly, I found that the FDR1 Dichaete-Dam binding intervals in *D. melanogaster* showed a much greater overlap with Dichaete core intervals than the FDR1 SoxN-Dam binding intervals did with SoxN core intervals.

Of the binding intervals identified in the three-way comparison for Dichaete-Dam, a total of 3855 overlap a Dichaete core interval. Of these, 70.4% show three-way conservation between *D. melanogaster, D. simulans* and *D. yakuba*, while only 38.5% of binding intervals that do not overlap a Dichaete core interval show three-way conservation (Figure 5.20A). Only 7.7% of binding intervals that overlap a Dichaete core interval are unique to *D. melanogaster*, compared to 32.4% of binding intervals that do not overlap a core interval. Somewhat surprisingly, binding intervals that overlap core intervals are not more likely to show two-way conservation; 4.3% show conservation between *D. melanogaster* and *D. simulans* compared to 7.8% of binding intervals that do not overlap a core interval, and 17.6% show conservation between *D. melanogaster* and *D. yakuba* compared to 21.2% of binding intervals that do not overlap a core interval. Nonetheless, the effect of overlapping a core Dichaete binding interval on the pattern of conserva-

tion is highly significant according to a chi-squared test (χ^2 = 1408.6, d.f. = 3, p-value = 4.10-305).

In the case of SoxN-Dam, a total of 2111 binding intervals overlap a SoxN core interval. Of these, 75.7% show two-way conservation between *D. melanogaster* and *D. simulans*, while 44.7% of binding intervals that do not overlap SoxN core interval show conservation (Figure 5.20B). Conversely, only 24.3% of binding intervals that overlap a core interval are unique to *D. melanogaster*, compared to 55.3% of binding intervals that do not overlap a core interval. Again, this effect is highly significant according to a chi-squared test (χ(2) = 733.1, d.f. = 1, p-value = 1.90e-161).

These results suggest that binding events at the Dichaete and SoxN core intervals, although they may not all represent direct gene regulation events, are nonetheless functionally important and subject to evolutionary constraint. It is somewhat surprising that, in the case of the core intervals as well as known enhancers, a strong effect on conservation between *D. melanogaster* and *D. simulans* is observed for SoxN, but this effect is missing for Dichaete. In contrast, for Dichaete, a strong effect is only observed on three-way binding conservation, with a smaller effect being observed for conservation between *D. melanogaster* and *D. yakuba* in the case of known enhancers. This may be because, over the relatively short evolutionary distance between *D. melanogaster* and *D. yakuba*, the majority of the binding events that are under selective pressure have been conserved between all three species, whereas few binding events are selectively constrained in only two out of the three lineages. For SoxN, where only data from two species are available, the conserved binding events likely also encompass many binding intervals that would be conserved in *D. yakuba* as well. Unfortunately, it is not possible to directly test this hypothesis with the current datasets.

5.5.3 Binding conservation at Dichaete and SoxN direct targets

A list of genes that are direct targets of both Dichaete and SoxN has previously been compiled by integrating gene expression data and *in vivo* binding data (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). I expected binding events at these genes to

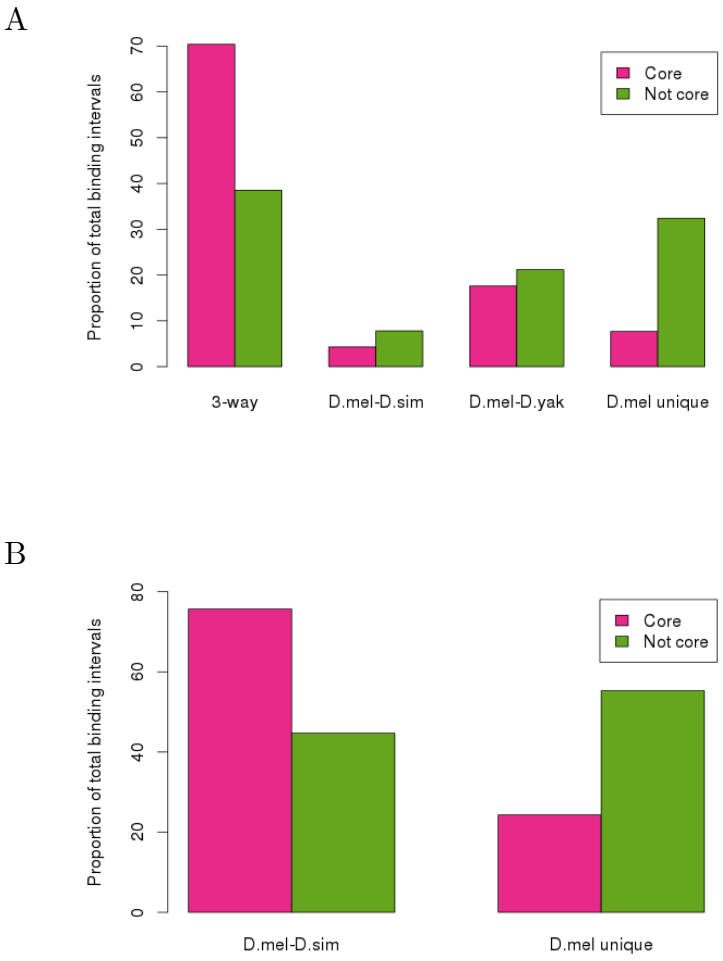


Figure 5.20: DamID binding intervals that overlap a Dichaete or SoxN core binding interval are more likely to be conserved than those that do not. A.) Dichaete-Dam binding intervals that overlap with a core Dichaete interval are more likely to show three-way conservation between *D. melanogaster, D. simulans* and *D. yakuba* (3-way) and are less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a core interval (p = 4.10-305). B.) SoxN-Dam binding intervals that overlap with a core SoxN interval are more likely to be conserved between *D. melanogaster* and *D. simulans* (D. mel - D. sim) and less likely to be unique to *D. melanogaster* (D. mel unique) than those that do not overlap with a core interval (p = 1.90e-161).

be highly conserved as well, since they are functional targets in *D. melanogaster*. For Dichaete-Dam, of the total binding intervals identified in a three-way comparison between *D. melanogaster, D. simulans* and *D. yakuba*, 4301 are annotated to a Dichaete direct target gene. This includes instances where multiple intervals are annotated to the same gene. 53.7% of these are conserved in all three species, compared to 43.1% of intervals that are not annotated to a direct target gene (Figure 5.21A). There is very little difference in the rates of two-way conservation between intervals annotated to direct targets and those that are not. Of the intervals at direct target genes, 20.4% are unique to *D. melanogaster*, compared to 29.0% of intervals not at direct target genes. Although these differences are significant according to a chi-squared test ($\chi^2 = 57.3$, d.f. $= 3$, p-value $=$ 2.3e-12), the effect on conservation rates is smaller than for binding intervals that overlap a core Dichaete interval.

In the case of SoxN-Dam, there is even less of an effect. Of the total SoxN-Dam binding intervals identified in a comparison between *D. melanogaster* and *D. simulans*, 1849 are annotated to a SoxN direct target gene. The fact that less binding intervals are located at direct targets is likely because less direct target genes have been identified for SoxN than for Dichaete. Of these, 53.3% are conserved between the two species, compared to 47.3% of binding intervals that are not annotated to a direct target (Figure 5.21B). Conversely, 47.6% of intervals annotated to direct targets are unique to *D. melanogaster*, while 52.7% of intervals not annotated to direct targets are unique to that species. These differences are significant according to a chi-squared test ($\chi^2 = 17.2$, d.f. $= 1$, p-value $=$ 3.4e-05). Initially, it was somewhat surprising to see that binding intervals at direct target genes are less likely to be conserved than those at core intervals, since binding at direct targets should be functional by definition. However, in many cases multiple intervals are annotated to a single target gene. Some of these binding events may be less important for gene regulation than others, perhaps representing shadow enhancers, and may therefore be less constrained by natural selection (Ludwig *et al.*, 2011; Perry *et al.*, 2010). The presence of these binding intervals in the dataset under consideration likely reduces the overall rate of conservation of intervals annotated to direct target genes.

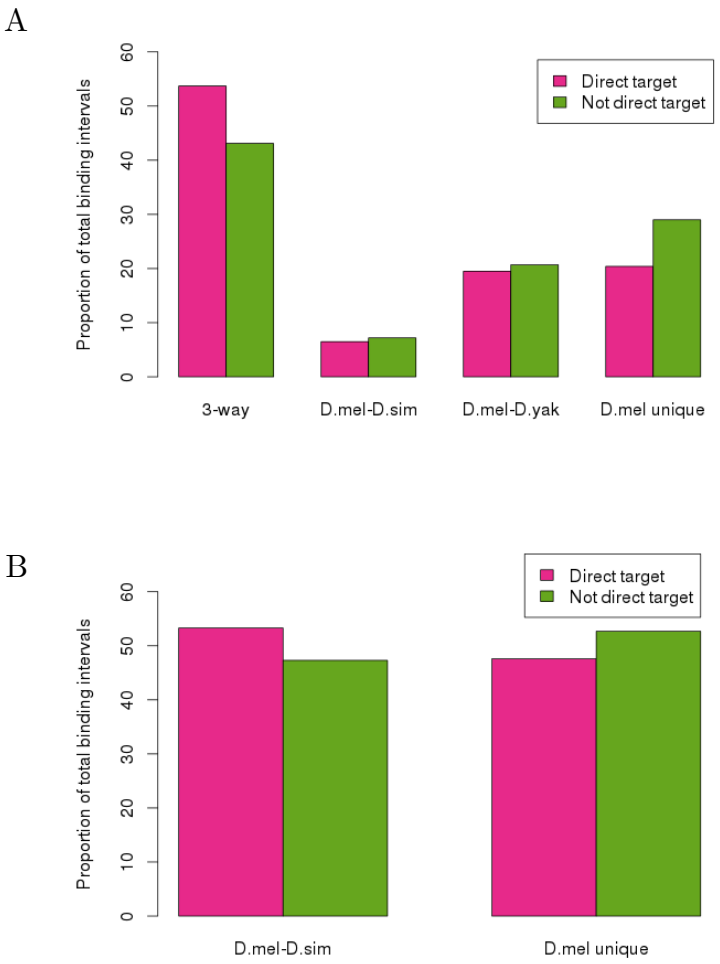


Figure 5.21: DamID binding intervals that are annotated to a Dichaete or SoxN direct target gene are more likely to be conserved that those that are not. A.) Dichaete-Dam binding intervals that are annotated to a Dichaete direct target gene are more likely to show three-way conservation between *D. melanogaster*, *D. simulans* and *D. yakuba* (3-way) and are less likely to be unique to *D. melanogaster* (D. mel unique) than those that are not annotated to a direct target (p = 2.3e-12). B.) SoxN-Dam binding intervals that are annotated to a SoxN direct target gene are more likely to be conserved between *D. melanogaster* and *D. simulans* (D. mel - D. sim) and less likely to be unique to *D. melanogaster* (D. mel unique) than those that are not annotated to a direct target (p = 3.4e-05).

5.6 Evolutionary perspective on common and unique binding by Dichaete and SoxNeuro

As was briefly discussed in the previous chapter, there are differences in the relationship between Dichaete-Dam and SoxN-Dam binding in *D. melanogaster* and *D. simulans*; most obviously, Dichaete-Dam and SoxN-Dam show considerably higher overlap in their binding patterns in *D. melanogaster*. I wanted to understand these differences better by examining the relationship between common and unique binding by Dichaete and SoxN and binding conservation. On a qualitative level, 15900 binding intervals are common to Dichaete and SoxN in *D. melanogaster*, while 9114 are common to Dichaete and SoxN in *D. simulans*; 7415 of these are commonly bound in both species, representing 46.7% of commonly bound intervals in *D. melanogaster* and 81.4% of commonly bound intervals in *D. simulans* (Figure 5.22). In *D. melanogaster*, 2634 binding intervals are unique to Dichaete-Dam, while in *D. simulans*, 4293 are unique to Dichaete-Dam. Only 338 of these are present and uniquely bound in both species, representing a much lower rate of conservation. The case is similar for SoxN-Dam; out of 4079 uniquely-bound intervals in *D. melanogaster* and 3741 uniquely-bound intervals in *D. simulans*, only 300 are present and uniquely bound in the two species. This suggests that binding intervals where both Dichaete and SoxN bind may be under greater evolutionary constraint than intervals where only one of the two proteins binds. Additionally, there are considerably more intervals that are uniquely bound by either protein in *D. simulans* and are commonly bound in *D. melanogaster* than the inverse (3372 versus 750), supporting the observation that Dichaete and SoxN binding patterns appear to be more differentiated in *D. simulans* than in *D. melanogaster*.

Taking just the binding intervals identified in *D. melanogaster*, there is a clear association between common binding by Dichaete and SoxN and binding conservation. 76.5% of binding intervals that are commonly bound are conserved in *D. simulans*, while 23.5% of them are not. In contrast, 40.1% of intervals that are uniquely bound by Dichaete are conserved in *D. simulans*, while 59.9% are not, and only 33.7% of intervals that are uniquely bound by SoxN are conserved in *D. simulans*, while 66.3% are not (Figure 5.23A). The difference in conservation

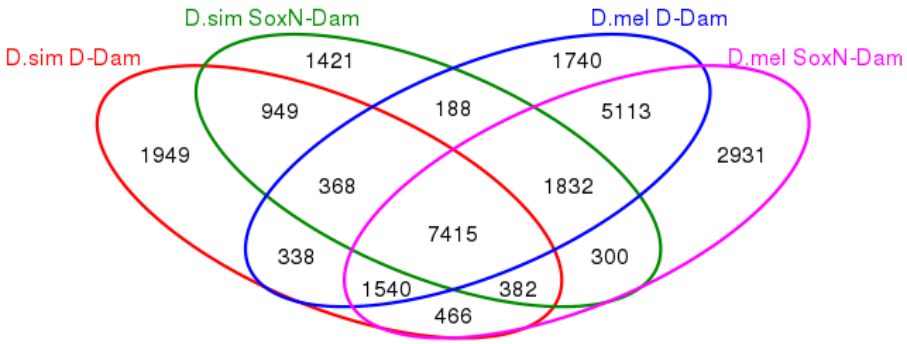


Figure 5.22: Venn diagram showing overlaps between Dichaete-Dam and SoxN-Dam binding intervals in *D. melanogaster* and *D. simulans*. A greater proportion of conserved intervals are uniquely bound by either Dichaete or SoxN in *D. simulans* but commonly bound by both TFs in *D. melanogaster* than the reverse.

between commonly and uniquely bound intervals is highly significant according to a chi-squared test ($\chi^2 = 3398.3$, d.f. $= 2$, p-value <2.2e-16 [approaches 0]). When the same analysis is performed from the perspective of binding intervals identified in *D. simulans*, a striking amount of intervals that are commonly bound by Dichaete and SoxN, 94.1%, are conserved in *D. melanogaster*, while just 5.9% are unique to *D. simulans* (Figure 5.23B). Unlike the *D. melanogaster* intervals, while the uniquely bound intervals in *D. simulans* show less conservation than the commonly bound intervals, they are still more likely to be conserved than not. 62.8% of intervals that are uniquely bound by Dichaete are conserved in *D. melanogaster*, while 37.2% are not, and 70.1% of intervals that are uniquely bound by SoxN are conserved in *D. melanogaster*, while 29.9% are not. However, the overall difference in conservation rates between commonly and uniquely bound intervals is still highly significant ($\chi^2 = 2488.9$, d.f. $= 2$, p-value <2.2e-16 [approaches 0]).

Assigning these conserved, commonly bound intervals to the closest genes within 10 kb upstream or downstream results in the identification of 5966 conserved core group B Sox targets in *D. melanogaster* and *D. simulans* (Appendix C). These gene targets have a profile that is consistent with the classical picture of group B Sox function. They are primarily upregulated in the larval CNS according to FlyAtlas, and they are enriched for GO:BP terms related to biological regulation (p

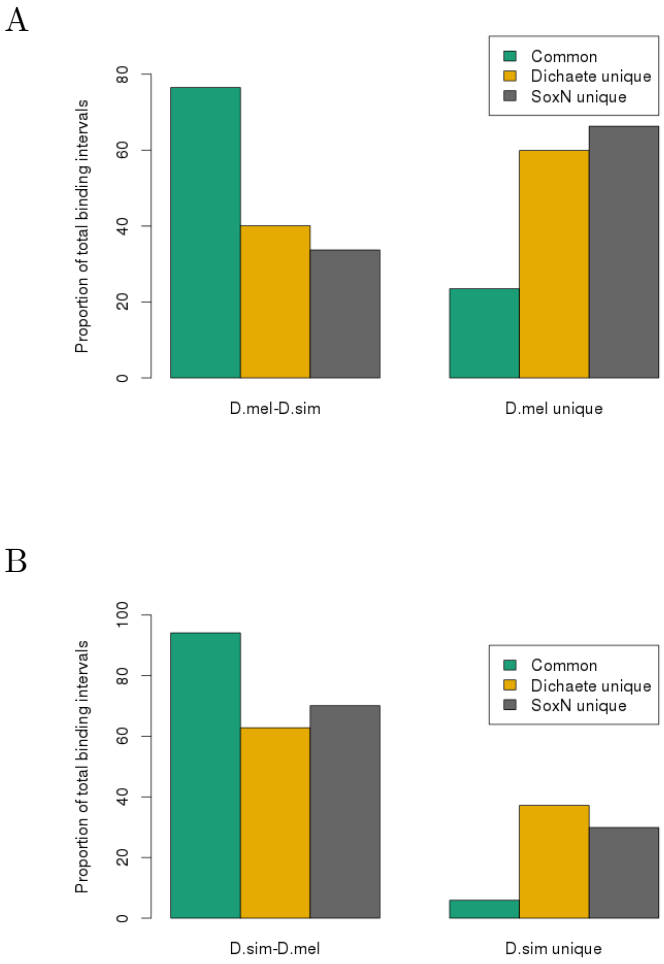


Figure 5.23: Intervals that are commonly bound between Dichaete-Dam and SoxN-Dam are more likely to be conserved between *D. melanogaster* and *D. simulans* than intervals that are uniquely bound by either Dichaete-Dam or SoxN-Dam. A.) Of all *D. melanogaster* binding intervals, those that are bound by both Dichaete-Dam and SoxN-Dam (Common) are more likely to also be bound in *D. simulans* (D. mel - D. sim) and less likely to be unique to *D. melanogaster* (D. mel unique) than those that are uniquely bound by either Dichaete-Dam or SoxN-Dam ($p < 2.2e\text{-}16$). B.) Of all *D. simulans* binding intervals, those that are bound by both Dichaete-Dam and SoxN-Dam (Common) are more likely to also be bound in *D. melanogaster* (D. sim - D. mel) and less likely to be unique to *D. simulans* (D. sim unique) than those that are uniquely bound by either Dichaete-Dam or SoxN-Dam ($p < 2.2e\text{-}16$).

= 1.48e-49), system development (p=2.86e-37), anatomical structure morphogenesis (p=6.41e-32), generation of neurons (p=4.55e-31) and neuron differentiation (p=4.92e-28) (Appendix D). To examine the evolutionary conservation of these target genes on an expanded scale, I compared them with targets of Sox2, Sox3 and Sox11, a group C Sox protein, identified in the mouse. I mapped the lists of bound genes discovered by Bergsland *et al.* to their *D. melanogaster* orthologues, resulting in 1301 orthologous targets of Sox2 in mouse neural precursor cells (NPCs), 4213 orthologous targets of Sox3 in NPCs and 1485 orthologous targets of Sox11 in NPCs (Bergsland *et al.*, 2011). 589 of the Sox2 target orthologues, 1730 of the Sox3 target orthologues and 595 Sox11 orthologues are conserved and commonly bound by Dichaete and SoxN. These deeply conserved Sox target genes represent around 40-45% of the targets of each mouse Sox protein but a smaller fraction of the commonly bound Dichaete/SoxN targets, indicating that, while Dichaete and SoxN can both perform aspects of mammalian Sox group B and C function, they have also both acquired a considerably broader range of targets in flies. A previous study of shared targets of Sox2 and Dichaete or SoxN core targets, as well as shared targets of Sox11 and SoxN, found similar numbers of orthologous target genes shared between Sox2 and each fly Sox protein individually. However, roughly twice as many targets were found to be shared between Sox11 and SoxN alone as between Sox11 and common Dichaete/SoxN targets, suggesting that while both Dichaete and SoxN may equally contribute to homologous functions of mammalian group B1 genes, Sox11s role may be largely played by SoxN in the fly, rather than by both TFs (Ferrero *et al.*, 2014). Overall, there is a high overlap between targets of mouse Sox genes from both groups B and C and common, conserved targets of Dichaete and SoxN in the fly, supporting the deep evolutionary conservation of Sox function in the CNS (Table 5.2).

To examine the differential functions of Dichaete and SoxN in *D. melanogaster* and *D. simulans*, I assigned the lists of binding intervals that are uniquely bound by either Dichaete or SoxN in both species to the nearest *D. melanogaster* genes within 10 kb upstream or downstream. This resulted in 381 gene assignments for Dichaete (Appendix E) and 361 gene assignments for SoxN (Appendix G). Only 14 genes are shared between the two lists, meaning that, at the loci where Dichaete and SoxN bind uniquely in both species, they are largely regulating separate sets of genes. I used FlyMine to analyze the functional enrichments of

Mouse Sox protein	**Orthologues of mouse targets in *D. melanogaster***	**Overlap with Dichaete/ SoxN common targets**	**Overlap with core SoxN targets**	**Overlap with core Dichaete targets**
Sox2	1301	589	443	522
Sox3	4213	1730	1134	1590
Sox11	1485	595	1092	610

Table 5.2: Numbers of shared target genes between *Drosophila* orthologues of mouse group B and C Sox proteins and either common Dichaete/SoxN targets or core targets of Dichaete or SoxN in *D. melanogaster*.

these gene sets. The two sets of genes have clearly different spatial expression profiles according to the FlyAtlas gene expression data; the genes uniquely bound by SoxN are predominantly upregulated only in the larval CNS, while the genes uniquely bound by Dichaete are also upregulated in the larval hindgut, head, crop, brain and thoracicoabdominal ganglion. These results agree with the observed expression patterns of the unique targets of Dichaete-Dam and SoxN-Dam uncovered in both *D. melanogaster* and *D. simulans* by the single-species DiffBind analysis in Chapter 4. The two sets of genes have similar GO:BP enrichments, including terms related to morphogenesis, development, neuron differentiation and biological regulation (Appendices F and H). These results suggest that a primary difference between Dichaete and SoxN function may be that, while Dichaete and SoxN are involved in many similar functions during development, Dichaete has targets that are spatially expressed in a broader range of tissues, while the targets that are unique to SoxN are more limited to the developing CNS. Interestingly, the *Drosophila* orthologues of mouse Sox11 targets, which overlap with more SoxN targets than common Dichaete/SoxN targets, are also primarily upregulated in the CNS, as opposed to orthologues of Sox2 targets, which also show upregulation in the brain.

I also searched for *de novo* motifs in the intervals that are uniquely bound by both Dichaete and SoxN in both species in order to identify any potential co-regulators that might shape the unique functions of each TF. Several motifs corresponding to transcriptional regulators that play broad roles during development were identified in both sets of intervals, including DNA replication-related element factor (Dref, p=1e-8) and Tramtrack (Ttk, p=1e-6). One of the top motifs identified

in the unique SoxN intervals corresponds to Ultraspiracle (Usp, p=1e-10), a TF involved in several aspects of neuron morphogenesis (Lee *et al.*, 2000; Parrish, 2006). Interestingly, one of the top hits in the unique Dichaete intervals is a Brachyenteron (Byn) motif (p=1e-9). Byn is a transcription factor that is critical for the development of the hindgut (Kispert *et al.*, 1994; Murakami *et al.*, 1999). The presence of this motif supports the idea that one of the primary unique functions of Dichaete, which is conserved in *D. melanogaster* and *D. simulans*, is its role regulating hindgut development.

I used DiffBind again to get a picture of the quantitative differences in Dichaete and SoxN binding in *D. melanogaster* and *D. simulans*. An analysis of differential binding between the two TFs using samples from both species with species as a blocking factor reveals 1257 differentially bound binding intervals at FDR1, with 778 of these preferentially bound by Dichaete-Dam and 479 preferentially bound by SoxN-Dam (Figure 5.24). Of the intervals preferentially bound by Dichaete-Dam, 681 were called as bound by Dichaete in both species in a single-species analysis. All of the intervals preferentially bound by SoxN-Dam were called as bound in both species in a single-species analysis. I assigned these differentially bound intervals to the nearest genes within 10 kb upstream and downstream in the *D. melanogaster* genome. This resulted in 925 gene assignments for Dichaete-Dam and 526 for SoxN-Dam. 54 genes are annotated as targets in both datasets. Similar differences in spatial expression are observed among these sets of target genes as among the genes identified in the qualitative analysis of differential binding; Dichaete-Dam preferential targets are upregulated in a wider range of tissues including the brain, head, larval CNS, crop, adult eye, hindgut, and thoracicoabdominal ganglion, while SoxN-Dam preferential targets are strongly upregulated only in the larval CNS. The set of SoxN-Dam preferential targets is enriched for homeodomain proteins, which is not observed in the Dichaete-Dam preferential targets. Again, the top enriched GO:BP terms for the two sets of genes are quite similar. However, in this case the list of publications that are enriched for preferential SoxN-Dam targets contains some interesting hints as to their functions; several publications related to neural stem cell differentiation, self-renewal and transcriptional networks are top hits (p <1e-21). There is also an enrichment of SoxN-Dam preferential target genes in the Reactome pathway Role of Abl in Robo-Slit signalling, an important pathway in axon guidance.

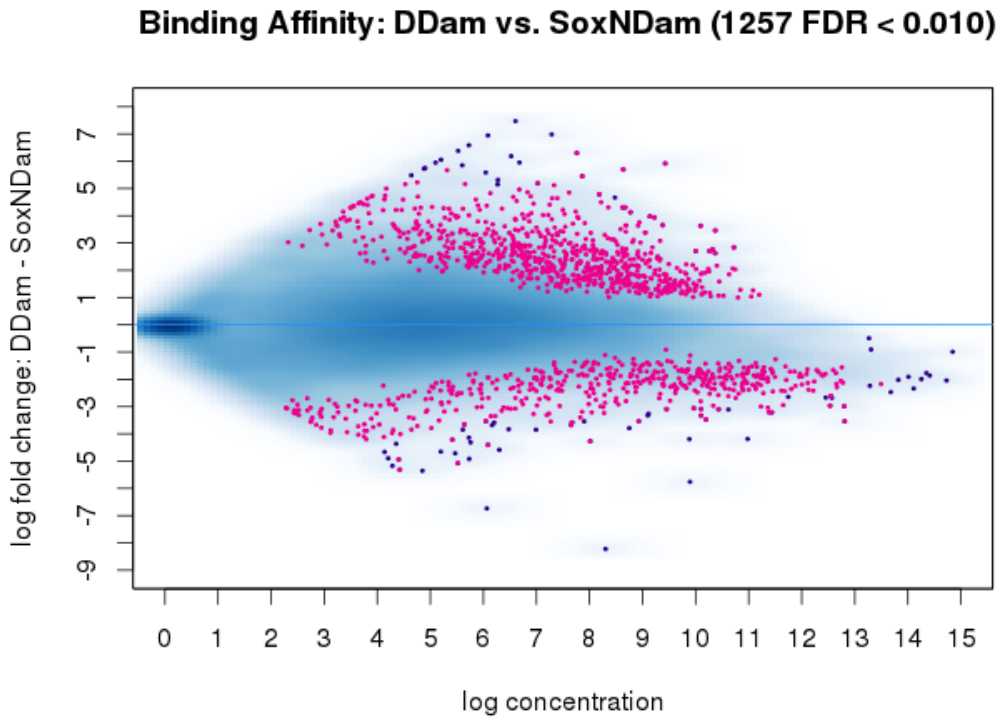


Figure 5.24: MA plot of differentially bound intervals with FDR <0.01 between Dichaete-Dam and SoxN-Dam in both *D. melanogaster* and *D. simulans*, with species as a blocking factor. Intervals that are more strongly bound by Dichaete-Dam have a positive fold change, while intervals that are more strongly bound by SoxN-Dam have a negative fold change. All intervals are plotted; differentially bound intervals are highlighted in pink.

I expected that, in cases where a TF bound preferentially in one species, it also bound preferentially in the other species; however, I wondered if there were certain binding intervals where the opposite was the case. In order to address this, I compared the binding intervals that were preferentially bound by either Dichaete-Dam or SoxN-Dam in each species separately. Of all of the intervals preferentially bound by either TF in *D. melanogaster* and *D. simulans*, only 347 overlap in the two species. Plotting the log2 fold changes for Dichaete-Dam binding versus SoxN-Dam binding at these intervals in both species reveals an interesting pattern (Figure 5.25A). The majority of the intervals shared between the two species are preferentially bound by Dichaete-Dam (fold change >0) in both species. A much smaller number are preferentially bound by SoxN-Dam (fold change <0)

in both species. Surprisingly, a similar number to those bound by SoxN-Dam are preferentially bound by Dichaete-Dam in one species and SoxN-Dam in the other, or vice versa. A linear regression of the fold changes in *D. simulans* versus *D. melanogaster* yields a positive but weak correlation of 0.19; this correlation is highly significant (p = 8.37e-18).

It is possible that the binding intervals with fold changes in the opposite direction in the two species represent extreme cases of Dichaete and SoxN's ability to compensate for each other, to the extent that they have effectively swapped binding functions during evolution. I decided to investigate these binding intervals in more detail. There are 15 intervals that are bound more strongly by Dichaete-Dam in *D. melanogaster* and more strongly by SoxN-Dam in *D. simulans*, and there are 27 that are bound more strongly by Dichaete-Dam in *D. simulans* and more strongly by SoxN-Dam in *D. melanogaster*. I annotated these intervals to the closest genes within 10 kb upstream and downstream in the *D. melanogaster* genome. Interestingly, some of them are annotated to known target genes with key roles in the developmental functions of Dichaete and SoxN. For example, a binding interval downstream of *tll*, a target of Dichaete in the hindgut, is more strongly bound by Dichaete in *D. melanogaster* but more strongly bound by SoxN in *D. simulans* (Figure 5.25B). In the opposite scenario, binding intervals located in an intron of *beat-IIa* (Figure 5.25C), also involved in axon guidance, are more strongly bound by Dichaete in *D. simulans* but more strongly bound by SoxN in *D. melanogaster*. However, the differences in binding strength between Dichaete and SoxN at these intervals is largely quantitative rather than qualitative, and they also tend to be located within or near genes that have other, additional binding intervals, making it unclear whether the differences observed are of functional significance in gene regulation.

Similarly, I wondered whether, in the intervals where binding has diverged between *D. melanogaster* and *D. simulans*, it has changed in the same direction for both Dichaete and SoxN. This type of correlated evolution has been found for the AP factors Bcd, Hb, Kr, Gt, Kni and Cad between *D. melanogaster* and *D. yakuba* and for Bcd, Hb, Kr and Gt between *D. melanogaster* and *D. pseudoobscura* as well as *D. virilis*, and has been linked to changes in chromatin accessibility as well as binding by the TF Zelda (Bradley *et al.*, 2010; Paris *et al.*, 2013). There are 2049 intervals that are differentially bound in either *D. melanogaster* or

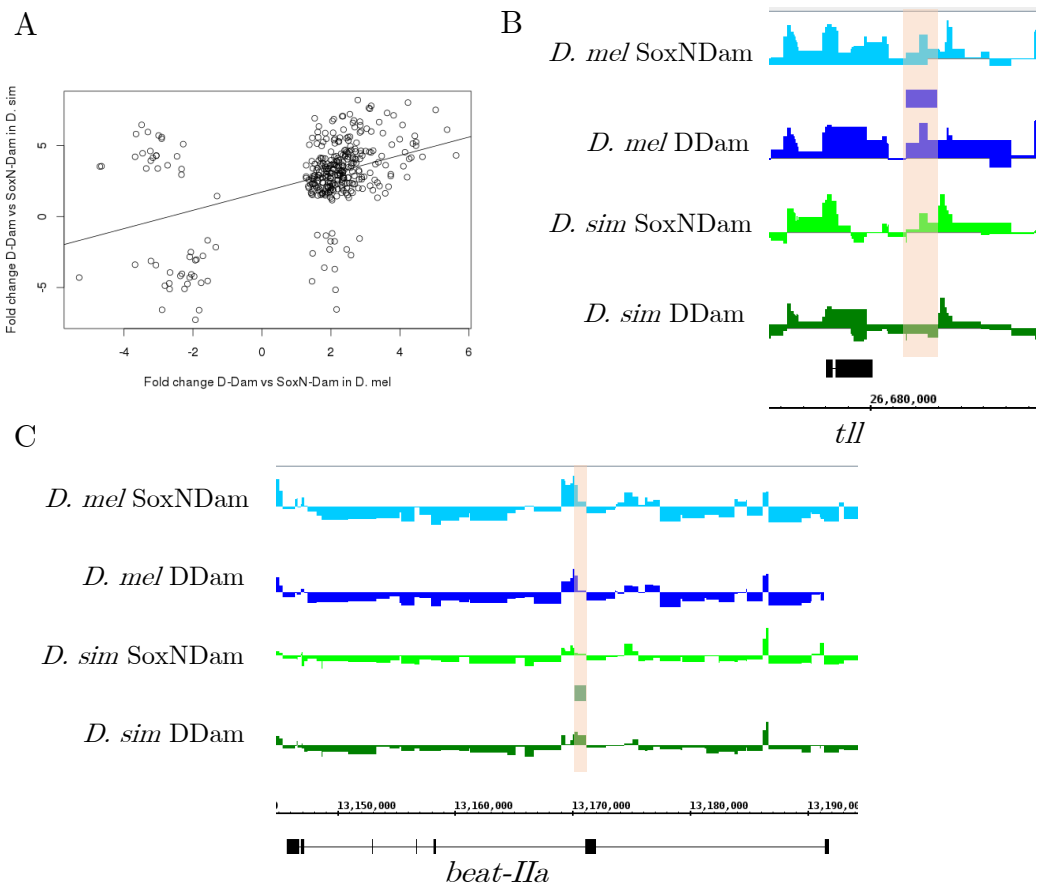


Figure 5.25: Differential binding between Dichaete-Dam and SoxN-Dam in *D. melanogaster* versus in *D. simulans*. A.) Scatter plot of fold changes between Dichaete-Dam and SoxN-Dam at differentially bound intervals in *D. melanogaster* versus fold changes at orthologous, differentially bound intervals in *D. simulans*. Positive fold changes indicate preferential binding by Dichaete-Dam in an interval, while negative fold changes indicate preferential binding by SoxN-Dam in an interval. The majority of intervals that are differentially bound in both species are preferentially bound by Dichaete-Dam in both species. Smaller numbers are preferentially bound by SoxN-Dam in both species or preferentially bound by Dichaete-Dam in one species and SoxN-Dam in the other ($R^2 = 0.19$, p = 8.37e-18). B.) A binding interval downstream of *tll* that is preferentially bound by Dichaete-Dam in *D. melanogaster* (dark blue) but preferentially bound by SoxN-Dam in *D. simulans* (light green). C.) A binding interval in an intron of *beat-IIa* that is preferentially bound by SoxN-Dam in *D. melanogaster* (dark green) but preferentially bound by Dichaete-Dam in *D. simulans* (light blue). Binding profiles represent normalized log2 ratios of Dam-fusion protein binding to Dam-only binding in each GATC fragment. Dichaete-Dam binding intervals are represented by blue or green bars above the profiles in which they are preferentially bound. Differentially bound intervals are highlighted in tan.

D. simulans and are bound by both TFs. Plotting the log2 fold changes for binding in *D. simulans* versus binding in *D. melanogaster* at these intervals for both TFs shows that, indeed, the majority of the changes in binding strength between species are in the same direction for both Dichaete and SoxN (Figure 5.26). Given the high degree of similarity between the binding profiles of the two TFs overall, this is not surprising. It indicates that most changes in Dichaete and SoxN binding between *D. melanogaster* and *D. simulans* are driven by factors common to both TFs, such as, potentially, chromatin accessibility or mutations in Sox motifs that are recognized by both proteins. A much smaller number of intervals show the opposite trend; these may be cases where a mutation in a motif has caused a specific gain in binding affinity for either Dichaete or SoxN in one species but not the other. Overall, the changes in binding strength between the two species for the two TFs are strongly correlated, and this correlation is highly significant ($R^2 = 0.73$, p <2.2e-16 [approaches 0]).

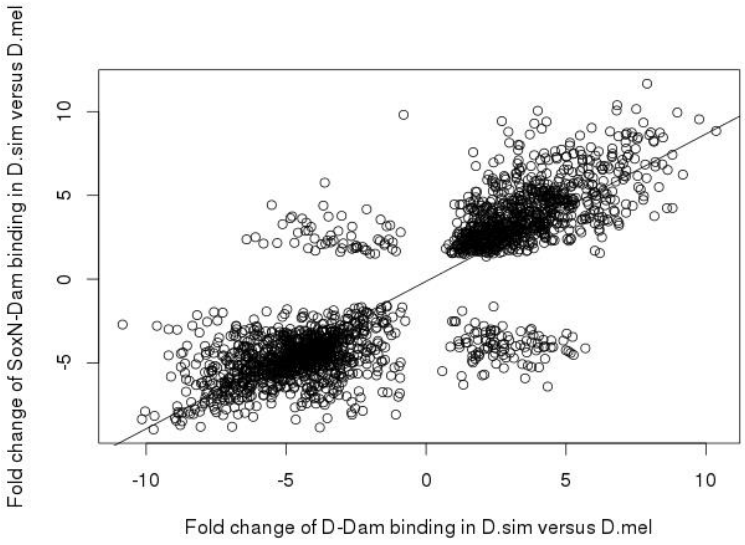


Figure 5.26: Scatter plot of fold changes between binding in *D. melanogaster* and *D. simulans* for Dichaete-Dam versus for SoxN-Dam in intervals bound by both TFs that are differentially bound between species. Positive fold changes indicate preferential binding in *D. simulans*, while negative fold changes indicate preferential binding in *D. melanogaster*. Most differentially bound intervals are more strongly bound in the same species by both Dichaete-Dam and SoxN-Dam, while a smaller number are more strongly bound in different species by each TF. Similar numbers of intervals are preferentially bound in each species overall.

5.7 Evolutionary analysis of Sox binding motifs

The availability of sequence data and *in vivo* binding data for each species facilitates an analysis of the contributions of sequence conservation within binding intervals and at TF-specific binding motifs to qualitative and quantitative binding conservation. In order to examine the patterns of motif conservation, I first identified all matches to the best *de novo* Sox motif discovered in each set of binding intervals using the tool FIMO (Grant *et al.*, 2011), with a p-value cutoff of 1e-4. I did the same with intervals that had been randomly shuffled to different locations in each genome using the BEDTools shuffle utility, as a control for each dataset (Quinlan and Hall, 2010). These shuffled intervals have the same lengths as the original binding intervals. The mean numbers of Sox motifs per binding interval range from 1.19 in the *D. pseudoobscura* Dichaete-Dam intervals to 2.75 in the *D. yakuba* Dichaete-Dam intervals. In all cases, there are significantly more Sox motifs in binding intervals than in randomly shuffled control intervals ($p < 4.03e\text{-}15$, Wilcoxon rank sum test with continuity correction). For the following analyses, I focused on Dichaete-Dam, comparing binding intervals showing four-way conservation between all species studied with intervals that are unique to *D. melanogaster*. I found 1896 *D. melanogaster* binding intervals that are conserved in all four species. These highly conserved intervals have significantly more Sox motifs on average (mean = 4.53) than do unique *D. melanogaster* binding intervals (mean = 1.29, p = 3.03e-193, Wilcoxon rank sum test with continuity correction) (Figure 5.27A).

Using PRANK, a phylogeny-aware aligner, I created multiple alignments of the orthologous sequences in each species for both the four-way conserved binding intervals and the unique *D. melanogaster* binding intervals (Löytynoja and Goldman, 2005, 2008). I only considered intervals for which a high-confidence orthologous sequence could be identified in all four species, which reduced the sets of intervals to 1064 showing four-way binding conservation and 1560 showing unique binding in *D. melanogaster*. These sequences should contain the enhancers or regulatory DNA to which Dichaete binds; however, they also contain flanking regions which may not be of functional relevance. Not surprisingly, the intervals showing four-way binding conservation do not display a higher rate of nucleotide conserva-

tion on average than the unique *D. melanogaster* intervals (He *et al.*, 2011b). In fact, the uniquely-bound intervals are slightly, but significantly, more conserved across their entire lengths (Wilcoxon p = 2.337e-20) (Figure 5.27B).

I scanned each set of multiple alignments for matches to the *de novo* Sox motifs, resulting in a count of the number of motifs in each binding interval that are positionally conserved in all four species as well as the nucleotide conservation within each motif. Within the set of intervals that show four-way binding conservation, 20.1% of all motifs identified in *D. melanogaster* are positionally conserved in all four species, with 19.5% of motifs in each interval being conserved on average. In the set of intervals that are only bound in *D. melanogaster*, 16.2% of all motifs identified in *D. melanogaster* are positionally conserved in all four species, with 16.1% of motifs in each interval being conserved on average (Figure 5.27C). A similar pattern holds when examining only those motifs that are both positionally conserved and that show 100% nucleotide conservation. In this case, for the intervals showing four-way binding conservation, 15.6% of all motifs identified in *D. melanogaster* show complete conservation (positional and sequence) in all four species, with 14.9% of motifs in each interval being conserved on average. For the intervals that are uniquely bound in *D. melanogaster*, 12.6% of all motifs identified in *D. melanogaster* show complete conservation, with 12.4% of motifs in each interval being conserved on average (Figure 5.27D). The differences in conservation rates between motifs in intervals showing four-way binding conservation and those uniquely bound in *D. melanogaster* are significant for both positional conservation (Wilcoxon p = 2.55e-24) and combined positional and nucleotide conservation (Wilcoxon p = 6.04e-28).

The Sox motifs identified in *D. melanogaster* binding intervals show high levels of nucleotide conservation in all four species overall, regardless of whether the orthologous sequences were identified as motif matches in other species or not. In the set of intervals showing four-way binding conservation, 14147 Sox motifs were identified in *D. melanogaster* that could be aligned without gaps to orthologous sequences in each other species. These motifs show an average of 74.5% nucleotide conservation across all four sequences in the multiple alignments. In the set of intervals showing unique binding in *D. melanogaster*, 5204 Sox motifs were identified in *D. melanogaster* that could be aligned without gaps to orthologous sequences in each other species. These motifs show a lower average

nucleotide conservation, 69.9%. The difference in motif nucleotide conservation rates between intervals showing four-way binding conservation and those uniquely bound in *D. melanogaster* is significant according to a Wilcoxon rank-sum test (p = 9.56e-36). As a further control, I randomly shuffled the columns in each PWM to produce a set of control motifs with the same GC content and length as the Sox motifs and searched for matches to each of them in the multiply aligned binding intervals. The average rates of nucleotide conservation are similar for shuffled control motifs in both sets of intervals, although they are slightly higher in intervals that display four-way binding conservation (70.7% versus 67.2% for unique *D. melanogaster* intervals, Wilcoxon p = 2.07e-19). The differences between average nucleotide conservation in Sox motifs and control motifs in both the four-way conserved binding intervals and the unique *D. melanogaster* binding intervals are also significant (Wilcoxon p = 1.62e-43 and p = 1.67e-9). Out of each set of motifs examined, the Sox motifs in intervals with four-way binding conservation clearly show the highest nucleotide conservation (Figure 5.27E).

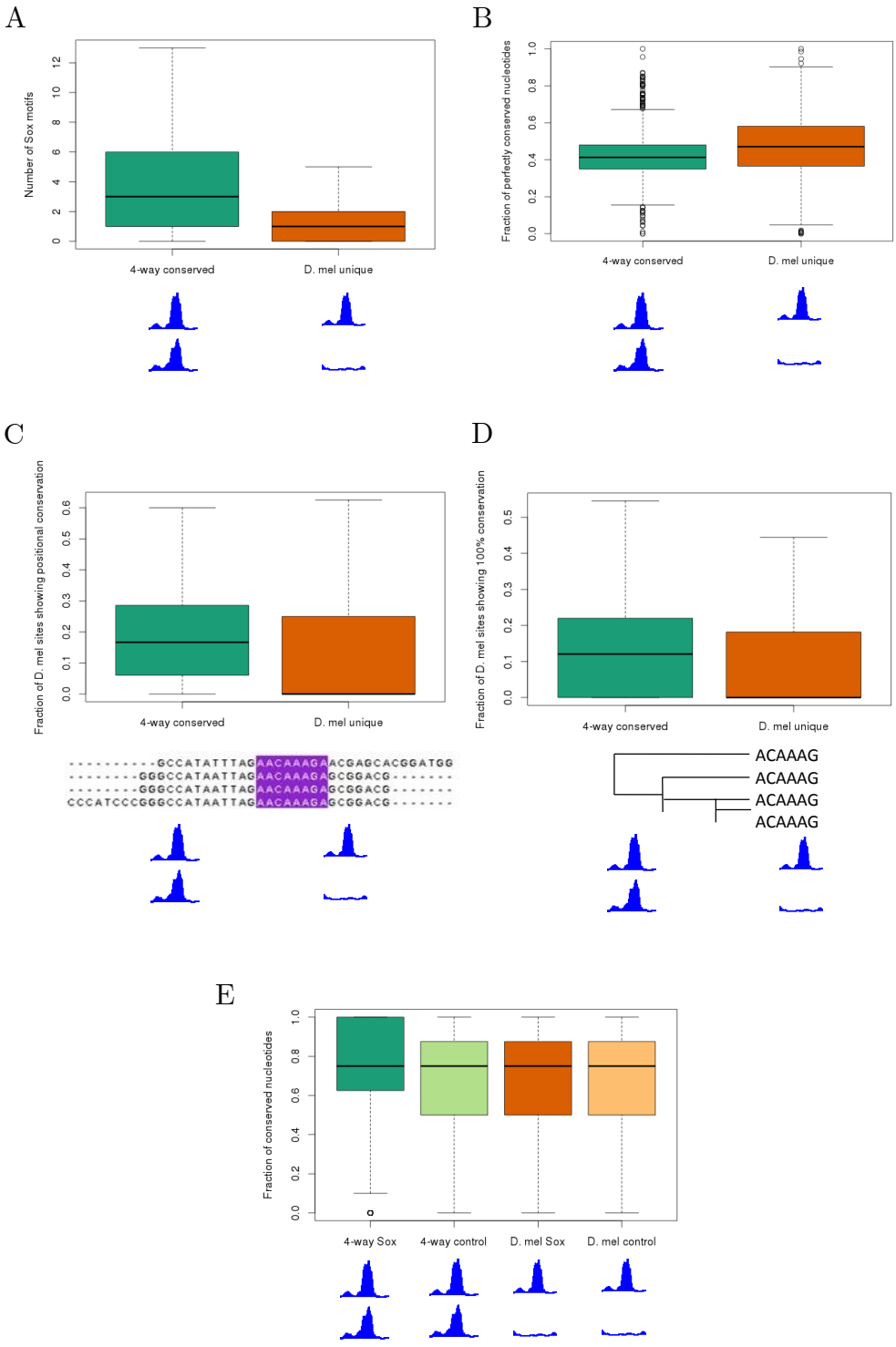

A
Number of Sox motifs
0
2
4
6
8
10
12
4-way conserved
D. mel unique
B
Fraction of perfectly conserved nucleotides
0.0
0.2
0.4
0.6
0.8
1.0
4-way conserved
D. mel unique
C
Fraction of D. mel sites showing positional conservation
0.0
0.1
0.2
0.3
0.4
0.5
0.6
4-way conserved
D. mel unique
----------GCCATATTTAGAACAAAGAACGAGCACGGATGG
--------GGGCCATAATTAGAACAAAGAGCGGACG-------
--------GGGCCATAATTAGAACAAAGAGCGGACG-------
CCCATCCCGGGCCATAATTAGAACAAAGAGCGGACG-------
D
Fraction of D. mel sites showing 100% conservation
0.0
0.1
0.2
0.3
0.4
0.5
4-way conserved
D. mel unique
ACAAAG
ACAAAG
ACAAAG
ACAAAG
E
Fraction of conserved nucleotides
0.0
0.2
0.4
0.6
0.8
1.0
4-way Sox
4-way control
D. mel Sox
D. mel control

Figure 5.27: Number and conservation of Sox motifs are associated with binding conservation. A.) On average, Dichaete-Dam binding intervals that are conserved between all four species have more Sox motifs than intervals that are unique to *D. melanogaster* (p = 3.03e-193). B.) Dichaete-Dam binding intervals that are conserved between all four species do not show a greater fraction of total nucleotide conservation on average than intervals that are unique to *D. melanogaster*. C.) On average, Dichaete-Dam binding intervals that show four-way binding conservation have more positionally conserved Sox motifs than intervals that are only bound in *D. melanogaster* (p = 2.55e-24). D.) On average, Dichaete-Dam binding intervals that show four-way binding conservation have more Sox motifs with 100% nucleotide conservation in addition to positional conservation than intervals that are only bound in *D. melanogaster* (p = 6.04e-28). E.) On average, Sox motifs in Dichaete-Dam intervals that are bound in all four species (4-way Sox) have a greater percentage of perfectly conserved nucleotides than either Sox motifs in intervals that are only bound in *D. melanogaster* (D. mel Sox, p = 9.56e-36), randomly shuffled control motifs in intervals that are bound in all four species (4-way control, p = 1.62e-43) or randomly shuffled control motifs in intervals that are only bound in *D. melanogaster* (D. mel control, p = 1.67e-9).

For binding intervals that are conserved but show quantitative changes in affinity in pairwise comparisons, I wanted to test whether motif quality was correlated with binding affinity. This has been shown to be the case for Bcd in a comparison between *D. melanogaster, D. yakuba, D. pseudoobscura* and *D. virilis*, as well as for Twi in a comparison between *D. melanogaster, D. simulans, D. yakuba, D. erecta, D. ananassae* and *D. pseudoobscura* (He *et al.*, 2011b; Paris *et al.*, 2013). I used two strategies to examine this question. First, I searched for Sox motifs in all Dichaete-Dam and SoxN-Dam intervals that were conserved but differentially bound between pairs of species using FIMO (Grant *et al.*, 2011). FIMO reports motif scores in the GFF output files which are calculated as -10*(log10(p-value)), thus reflecting the statistical confidence that a given sequence matches the consensus motif. In cases where more than one motif is predicted within an interval, it is difficult to determine *a priori* which motif(s) are primarily responsible for TF binding, since DamID binding intervals are not necessarily centered around the binding site. I therefore found both the average motif score and the total (cumulative) motif score within each interval in each species examined. Performing a linear regression of the log2 fold change in binding affinity between each pair of species at each interval versus the difference in either average or cumulative motif score at each interval reveals no significant correlations between motif quality and quantitative changes in binding, the one exception being for average motif

scores in differentially bound SoxN-Dam intervals between *D. melanogaster* and *D. simulans*, where a weak but significant correlation is present ($R^2 = 0.0027$, p = 0.0057) (Figure 5.28).

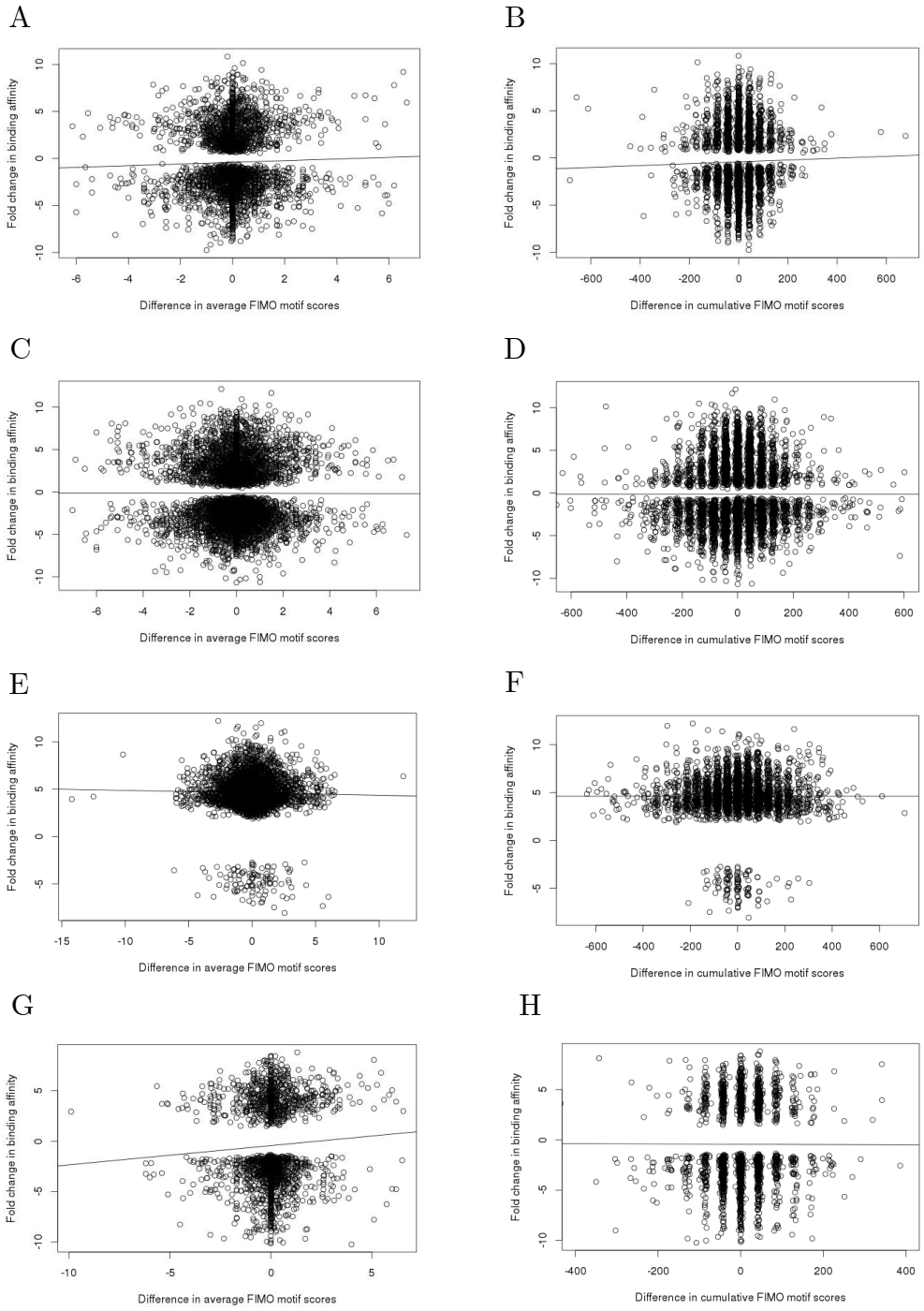


Figure 5.28: Changes in Sox motif quality within binding intervals between species do not correlate with changes in group B Sox binding affinity. Differences in cumulative or average motif scores between species are plotted on the x-axis and differences in binding affinity are plotted on the y-axis. Positive binding affinity and motif score differences represent increased binding or motif quality in *D. melanogaster*, while negative fold changes and negative motif score differences represent increased binding or motif quality in each other species. A.) Log2 fold change of Dichaete-Dam binding in *D. melanogaster* versus *D. simulans* plotted against the difference between average *D. melanogaster* motif scores and average *D. simulans* motif scores in each interval. R^2 = 0.00046, p = 0.099. B.) Log2 fold change of binding in *D. melanogaster* versus *D. simulans* plotted against the difference between cumulative *D. melanogaster* motif scores and cumulative *D. simulans* scores in each interval. R^2 = 0.00014, p = 0.21. C.) Log2 fold change of Dichaete-Dam binding in *D. melanogaster* versus *D. yakuba* plotted against the difference between average *D. melanogaster* motif scores and average *D. yakuba* motif scores in each interval. R^2 = -0.00016, p = 0.86. D.) Log2 fold change of Dichaete-Dam binding in *D. melanogaster* versus *D. yakuba* plotted against the difference between cumulative *D. melanogaster* motif scores and cumulative *D. yakuba* motif scores in each interval. R^2 = -7.95e-05, p = 0.47. E.) Log2 fold change of Dichaete-Dam binding in *D. melanogaster* versus *D. pseudoobscura* plotted against the difference between average *D. melanogaster* motif scores and average *D. pseudoobscura* motif scores in each interval. R^2 = 0.00014, p = 0.22. F.) Log2 fold change of Dichaete-Dam binding in *D. melanogaster* versus *D. pseudoobscura* plotted against the difference between cumulative *D. melanogaster* motif scores and cumulative *D. pseudoobscura* motif scores in each interval. R^2 = -0.00028, p = 0.85. G.) Log2 fold change of SoxN-Dam binding in *D. melanogaster* versus *D. simulans* plotted against the difference between average *D. melanogaster* motif scores and average *D. simulans* motif scores in each interval. R^2 = 0.0027, p = 0.0057). H.) Log2 fold change of SoxN-Dam binding in *D. melanogaster* versus *D. simulans* plotted against the difference between cumulative *D. melanogaster* motif scores and cumulative *D. simulans* motif scores in each interval. R^2 = -7.41e-05, p = 0.37).

As a secondary strategy, I found the scores assigned by RSAT to all positionally conserved motifs identified in four-way conserved Dichaete-Dam binding intervals that had been multiply aligned. The majority of these motifs have identical sequences and thus the same score in all species; however, there are some cases where a mutation in one species leads to a better or worse match to the consensus motif. For each pairwise comparison between species, I divided the intervals by whether they showed differential binding in one species or the other and counted the number of motifs in each group of intervals that had a higher score in each species. I compared the number of motifs in intervals that have a higher score in the species in which the interval is preferentially bound with the number of

motifs that have a lower score in the species in which the interval is preferentially bound. Comparing *D. melanogaster* and *D. simulans*, there are more total Sox motifs that score higher in *D. melanogaster* than Sox motifs that score lower in *D. melanogaster* in intervals that are preferentially bound in that species (26 versus 14), and the same holds true for *D. simulans* (5 versus 2). However, this pattern does not hold for the comparison of *D. melanogaster* and *D. yakuba*; in intervals preferentially bound in *D. melanogaster*, there are 35 motifs that score more highly in that species and 40 that score more highly in *D. yakuba*, while in intervals preferentially bound in *D. yakuba*, there are 26 motifs that score more highly in that species and 25 that score more highly in *D. melanogaster*. Comparing *D. melanogaster* and *D. pseudoobscura*, in intervals that are more highly bound in *D. melanogaster*, there are 189 motifs that score more highly in that species versus 173 that score more highly in *D. pseudoobscura*, while only one motif scoring more highly in each species was found in intervals that are preferentially bound in *D. pseudoobscura*. Counting the number of motifs within each interval, rather than the total number of motifs found for each group of intervals, there are no significant differences between the numbers of higher-scoring motifs and lower-scoring motifs in intervals that are preferentially bound in any pairwise species comparison (*D. melanogaster* vs. *D. simulans*, $p = 0.48$; *D. melanogaster* vs. *D. yakuba*, $p = 0.45$; *D. yakuba* vs. *D. melanogaster*, $p = 0.18$; *D. melanogaster* vs. *D. pseudoobscura*, $p = 0.64$).

Given the correlation between motif conservation and qualitative binding conservation, it is somewhat surprising to find no detectable correlation between motif quality and quantitative changes in binding affinity. However, in the case of four-way positionally conserved motifs, such an effect might be obscured by the high overall degree of quality and nucleotide conservation in the motifs examined. The FIMO motif analysis, in which all motifs in four-way conserved binding intervals were scored, should uncover a broader range of variability in motif quality; however, in this case, the fact that the scores of all motifs in each interval were taken into account, either through averaging or examining the cumulative motif scores, could obscure a signal from one or a few motifs that have a more direct effect on TF binding. The Twi study, which used ChIP-seq data, focused on the quality of motifs within a 151-bp window around the binding peak summit (He *et al.*, 2011b), allowing for a more focused assessment of the effect of motif quality on

binding. However, in DamID binding intervals, the highest scoring nucleotide does not necessarily correspond to the center of TF binding, due to the non-random distribution of GATC sites in the genome. This makes it difficult to narrow down binding regions in order to identify motifs that might be the most relevant for binding. Additionally, the technical differences between DamID and ChIP may make ChIP a more sensitive measure of quantitative binding affinity than DamID; this has not been tested experimentally. Nonetheless, the analyses of qualitative binding conservation show that both the number of Sox motifs in an interval and the positional and nucleotide conservation levels of those motifs are correlated with conserved binding.

5.8 Discussion of results

In this chapter, I set out to analyze the binding patterns of Dichaete and SoxN in the context of *Drosophila* evolution, using the DamID datasets that I generated. First, I performed quantitative binding comparisons between each pair of species for each TF. Normalizing the read counts from all samples in each pair of species together allowed me to reduce the effects of comparing separately thresholded samples, which can lead to an underestimate of similarity. Since the identification of differentially bound intervals by DiffBind requires a list of bound intervals in each sample as input, there is still some potential for thresholding effects. Nonetheless, pooling the bound intervals from all samples before the differential enrichment analysis should minimize this problem. The percentages of *D. melanogaster* binding intervals detected as qualitatively or quantitatively divergent between *D. melanogaster* and each other species range from 21.4% for SoxN-Dam in *D. simulans* to 44.2% for Dichaete-Dam in *D. yakuba*. On the level of read counts, the binding affinity score correlations in bound regions for the same TF between *D. melanogaster* and *D. simulans* range from 0.62 - 0.72 for Dichaete-Dam and from 0.75 - 0.90 for SoxN-Dam. For Dichaete-Dam between *D. melanogaster* and *D. yakuba*, they range from 0.68 - 0.70, and for Dichaete-Dam between *D. melanogaster* and *D. pseudoobscura*, they are more variable, ranging from 0.46 - 0.72. These numbers are in line with correlations of AP factor binding between *D. melanogaster* and *D. yakuba*, which range from 0.57 for

Cad to 0.75 for Kr (Bradley *et al.*, 2010). The quantitative changes in Dichaete binding between each pair of species examined follow the known *Drosophila* phylogeny (Russo *et al.*, 1995), with greater differences detected between more distant species, which follows an expectation of neutral evolution according to a molecular clock mechanism (He *et al.*, 2011b).

The reduced quality of the *D. pseudoobscura* Dichaete-Dam samples compared to those from the other species and the consequent lower number of binding intervals called in *D. pseudoobscura* posed a challenge for further analysis. In order to make a quantitative comparison between Dichaete-Dam binding in *D. pseudoobscura* and *D. melanogaster*, I decided to normalize the samples by total library size rather than by read counts in bound regions, since the numbers of bound regions were so different between the two species. This approach is a conservative one and may have underestimated the number of intervals that are preferentially bound in *D. pseudoobscura*; however, it prevented the over-inflation of binding signal in the *D. pseudoobscura* samples. For the subsequent analyses of binding conservation in relation to functional annotations, I decided to focus on a three-way comparison of Dichaete-Dam binding in *D. melanogaster, D. simulans* and *D. yakuba*, as these datasets are of comparable quality and offer the most unbiased view of evolutionary differences across the *melanogaster* clade. Normalizing samples from these three species together revealed a similar pattern of binding divergence as that seen in pairwise comparisons and confirmed the fact that changes in Dichaete-Dam binding correspond with the *Drosophila* phylogeny.

Besides quantitative and qualitative conservation of binding at orthologous loci between species, a comparative study of TF binding can be used to address the question of compensatory evolution or binding site turnover, when binding events evolve at different positions between species but regulate the same gene. Without gene expression data or functional enhancer assays, it is impossible to prove which binding events have a direct effect on gene regulation; nonetheless, by annotating binding events that are not positionally conserved to genes, I identified potential instances of Dichaete and SoxN binding site turnover between *D. melanogaster* and *D. simulans* as well as between *D. melanogaster* and *D. yakuba*. These binding events represent a large proportion of non-positionally conserved binding intervals between species. In *D. simulans*, out of 4472 Dichaete-Dam binding intervals that are not positionally conserved in *D. melanogaster*,

3226 or 72.1% are potential instances of binding site turnover, while conversely, 58.2% of *D. melanogaster* Dichaete-Dam intervals that are not positionally conserved in *D. simulans* could represent binding site turnover. For SoxN-Dam, the non-positionally conserved intervals that are identified as showing compensatory conservation represent 90.4% of *D. simulans* binding intervals and 61.4% of *D. melanogaster* binding intervals. Comparing Dichaete-Dam binding in *D. melanogaster* and *D. yakuba*, they represent 79.5% of *D. yakuba* non-positionally conserved intervals and 78.5% of *D. melanogaster* non-positionally conserved intervals. This supports the view that Dichaete and SoxN have very similar gene targets in each species studied, since in the majority of instances where binding has been lost in one species, a separate binding site has been gained at the same gene locus.

The availability of STARR-seq enhancer activity maps in several species of *Drosophila* allows for an interesting comparison of TF binding conservation with overall enhancer conservation (Arnold *et al.*, 2013). Applying the same criteria that I used for Dichaete and SoxN binding sites, I identified STARR-seq enhancers in *D. melanogaster* and *D. yakuba* that show compensatory conservation. Although a large number of enhancers show evidence of turnover, relatively few of them are bound by Dichaete-Dam in either *D. melanogaster* or *D. yakuba*. Even fewer contain a Dichaete-Dam binding site in either species that also shows compensatory conservation. Of all the Dichaete-Dam binding intervals that are potential instances of binding site turnover, only 1.1% and 2.1% of *D. melanogaster* intervals are located in STARR-seq enhancers that show turnover in S2 cells or OSCs, respectively, and only 1.8% and 3.1% of *D. yakuba* intervals are located in STARR-seq enhancers that show turnover in S2 cells or OSCs. Although binding site turnover at gene loci appears to be a common mode of evolution for Dichaete and SoxN in *Drosophila*, new binding events are not generally gained in newly-evolved enhancers, but rather in enhancer regions whose regulatory activity is conserved between species. This finding, although initially surprising, is in line with the hypothesis that turnover of TF binding events should operate to maintain the level of transcriptional output of their target enhancers under balancing selection.

The other sources of annotated enhancers that I used in this study, REDFly and FlyLight, do not have comparative data available (Gallo *et al.*, 2010; Man-

ning *et al.*, 2012). However, given the high overlap that I found between *D. melanogaster* binding intervals and these enhancers, I was curious about the relationship between binding in validated enhancers and binding conservation. I found a strong correlation between the two; for Dichaete-Dam, *D. melanogaster* binding events that overlap with an enhancer from either database are more likely to be conserved in both *D. simulans* and *D. yakuba*, while for SoxN-Dam, they are more likely to be conserved in *D. simulans*. For both TFs, binding intervals that overlap with a REDFly or FlyLight enhancer are less likely to be unique to *D. melanogaster* (Figure 5.19). This result confirms that binding to known functional regions is subject to selective pressure and is therefore more likely to be maintained during evolution. Interestingly, an even stronger effect was observed for binding intervals that overlap with a Dichaete or SoxN core interval (Figure 5.20) (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). While the core intervals were defined by overlapping multiple *D. melanogaster* genome-wide binding datasets, including ChIP-seq and DamID, their evolutionary conservation has not previously been assessed. The fact that DamID binding events that overlap with core intervals are much more likely to be conserved across all species studied than those that do not provides strong evidence for the importance of these binding intervals in group B Sox function. A smaller, but still significant, effect was found for binding intervals that are annotated to known Dichaete and SoxN target genes (Figure 5.21) (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). It should be noted that very conservative criteria were used to define direct target genes; for example, bound genes that show modest expression changes in mutant embryos were excluded. Such genes would include *bona fide* Sox targets whose expression is rescued by functional compensation. Hence it is likely that the true fraction of conserved target genes is higher than reported here.

The availability of a sequenced genome for all four species studied allowed me to examine the connection between binding intervals sequence motif content and *in vivo* binding conservation Clark *et al.* (2007). Previous comparative studies using ChIP-seq have found that overall nucleotide conservation is not significantly elevated in binding intervals that are conserved between species (Bradley *et al.*, 2010; He *et al.*, 2011b). Since DamID binding intervals tend to be wider than ChIP-seq intervals and are not centered on the binding site, my expectation was that this would hold true in my DamID data. Indeed, I found no correlation

between binding conservation and increased nucleotide conservation across the entire intervals. However, I did find significant correlations between binding conservation and several measures of Sox motif content. First, conserved intervals have more matches to Sox motifs on average than non-conserved intervals or control intervals whose genomic coordinates were randomly shuffled. This indicates that an increased density of recognizable motifs may contribute to group B Sox binding function and be important for its conservation. On the level of individual motifs, intervals with conserved binding contain more matches to Sox motifs that show positional conservation within the interval as well as showing 100% nucleotide conservation between species. Matches to Sox motifs within intervals that show conserved binding also have a higher percentage of conserved nucleotides than those in intervals that do not show conserved binding or than matches to control motifs whose columns were randomly shuffled. It is difficult to say whether higher quality Sox motifs lead to more functional binding or whether functional binding leads to selective pressure on Sox motifs; however, it seems likely that a feedback mechanism might act to maintain the observed correlation between highly conserved motif matches and *in vivo* binding conservation.

In the previous chapter, I began to address the question of common and unique binding by Dichaete and SoxN in *D. melanogaster* and *D. simulans*, noting that the binding patterns of the two Dam fusions appear more similar in *D. melanogaster*. Here I have explored the evolutionary relationship between these two TFs in more detail. The most striking observation from my analysis is the strong association between common binding by Dichaete and SoxN and binding conservation. Intervals that show binding by both TFs in one species are much more likely to be bound in the other species, and specifically to also be bound by both TFs in that species. Conversely, of the intervals that are uniquely bound by either Dichaete or SoxN in one species, relatively few are also uniquely bound by the same TF in the other species, suggesting that unique binding by one TF is less constrained by selection than common binding. A large number of intervals that are bound by one TF in *D. simulans* are bound by both in *D. melanogaster*, supporting the original observation that Dichaete-Dam and SoxN-Dam binding patterns are more similar in *D. melanogaster*. It is unclear why this is the case; however, given the fact that a higher percentage of *D. simulans* commonly-bound intervals are conserved in *D. melanogaster* than *vice versa*, the intervals that are

commonly bound in both species could represent a core set of binding intervals are likely to be of key functional importance.

Although I found a high level of conservation of common binding by Dichaete and SoxN, there are also examples of unique binding that are conserved in both species, which can be used to examine the conserved functions that are specific to each TF. Using both these data and a DiffBind analysis that searched for differentially bound intervals between the two TFs using data from both species normalized together, I identified target genes that are uniquely bound by either Dichaete or SoxN in both species. Functional enrichment analyses using FlyMine indicated that the major differences between these sets of target genes are in the tissues in which they are expressed, rather than biological processes in which they play a role. Specifically, conserved, unique Dichaete targets are upregulated in a broader range of tissues than conserved, unique SoxN targets, including the head, brain, crop and hindgut. The presence of unique Dichaete targets in the brain and hindgut is particularly interesting, as Dichaete is known to play a role in the development of these tissues (Sánchez-Soriano and Russell, 2000). Additionally, a strong motif for Byn, a TF that is critical for hindgut development (Kispert *et al.*, 1994; Murakami *et al.*, 1999), was found in conserved, uniquely-bound Dichaete intervals, suggesting that Dichaete and Byn might work together to regulate target gene expression in the hindgut. This motif was not identified in single-species analyses or in an analysis of the entire set of Dichaete-bound intervals, highlighting the power of a comparative analysis to detect specific features of regulatory function. In the case of SoxN, conserved, uniquely-bound targets are largely expressed in the CNS. The presence of a Usp motif in these intervals and the enrichment of target genes in the Robo-Slit signalling pathway highlight a conserved role for SoxN in axon morphogenesis and pathfinding (Ferrero *et al.*, 2014; Girard *et al.*, 2006; Lee *et al.*, 2000; Parrish, 2006).

Taken together, these results suggest a model whereby Dichaete and SoxN binding, while subject to turnover during evolution, is highly conserved at loci where both TFs can bind and at potentially functional sites, including annotated enhancers and core Dichaete and SoxN intervals. Given the similarity of the motifs recognized by these two TFs and the number of target genes that they share, it may be easier for natural selection to maintain sites where both Dichaete and SoxN can bind, rather than maintaining independent binding sites for each. From

another perspective, sequences that can be bound and contribute to functional regulation by both Dichaete and SoxN may experience a double dose of selective constraint. If this were true, then why would sites that are uniquely bound by one TF be conserved at all? The comparative data suggest that conserved, uniquely bound targets are largely expressed in different tissues, corresponding to the differences in expression patterns shown by Dichaete and SoxN themselves. At loci where Dichaete and SoxN have evolved new, independent regulatory functions, unique binding could be driven by external factors such as differences in chromatin accessibility between embryonic tissues or the availability of cofactors that might interact specifically only one Sox protein, which would prevent the other protein from binding in tissues in which they are commonly expressed. In the following chapters, I will explore the question of chromatin accessibility in *D. pseudoobscura* and its relationship to group B Sox binding and then conclude by attempting to synthesize the information gained from my comparative studies of chromatin accessibility and transcription factor binding.

CHAPTER 6

Chromatin Accessibility During Development in *Drosophila pseudoobscura*

6.1 Experimental Motivation and Design

Despite having distinct DNA binding domains and preferences for specific sequence motifs, many developmental transcription factors show surprisingly similar genome-wide binding patterns in *D. melanogaster* embryos, differing primarily in quantitative levels of occupancy at a highly-overlapping set of genomic regions (MacArthur *et al.*, 2009). Both experimental evidence and computational modelling have revealed an important role for chromatin accessibility in determining these overlapping bound regions (Kaplan *et al.*, 2011; Li *et al.*, 2011). Patterns of chromatin accessibility in embryonic nuclei change throughout development as cells take on more committed fates, allowing transcription factors access to different regions of regulatory DNA and ultimately contributing to overall body patterning (Thomas *et al.*, 2011). The importance of chromatin accessibility in directing patterns of transcription factor binding has also been observed in

Drosophila imaginal discs as well as in mammalian cells (John *et al.*, 2011; McKay and Lieb, 2013; Neph *et al.*, 2012). Since a major goal of this thesis was to examine differences in transcription factor binding between *Drosophila* species, I was interested in measuring chromatin accessibility during development of non-model drosophilids in order to determine whether observed differences in TF binding could be correlated with differences in accessibility.

Two major techniques exist to detect genome-wide patterns of chromatin accessibility *in vivo*: DNase-seq and FAIRE-seq. DNase-seq relies on the non-specific digestion of chromatin by the enzyme DNaseI. Nuclei are isolated and immediately treated with DNaseI, which cleaves DNA wherever it is accessible. Short DNA fragments resulting from these cleavages are then recovered and sequenced, leading to the identification of DNase-hypersenstive sites (DHS) (Thomas *et al.*, 2011). Although this technique has been used extensively, there is some evidence that DHS datasets may suffer from bias due to sequence preferences of DNaseI, which may vary depending on the experimental conditions (Koohy *et al.*, 2013). An alternative technique is FAIRE-seq (Formaldehyde-Assisted Identification of Regulatory Elements). In FAIRE-seq, nuclei are isolated and fixed with formaldehyde. The chromatin is then sonicated, breaking the more accessible regions into small fragments, and purified using phenol-chloroform extractions. This results in only DNA from accessible regions being recovered, as inaccessible, compacted chromatin is left in the organic phase during the extractions (Giresi and Lieb, 2009; Simon *et al.*, 2012). Although DNase-seq and FAIRE-seq do not perfectly recapitulate each other, as DNAse-seq tends to detect a higher signal at promoter regions while FAIRE-seq tends to detect a higher signal at distal regulatory regions, overall the two techniques show good correspondence (Koohy *et al.*, 2013; McKay and Lieb, 2013).

I decided to use FAIRE-seq to study chromatin accessibility and to focus on one species, *D. pseudoobscura*, which is the most distant species to *D. melanogaster* of those that I studied and which shows the greatest difference in chromosomal structure and arrangement (Clark *et al.*, 2007; Richards, 2005). I performed FAIRE-seq on *D. pseudoobscura* embryonic chromatin from five developmental stages, stage 5, stage 9, stage 10, stage 11 and stage 14, chosen to provide a comparison with *D. melanogaster* DNase-seq data from Thomas et al. (2011). The timing of each developmental stage in *D. pseudoobscura* was calculated according

to Kuntz and Eisen (2013); more details are available in Chapter 2. I sequenced three biological replicates from each stage. Although input chromatin can be used as a control for FAIRE-seq, as with ChIP-seq, it is not strictly necessary (Simon *et al.*, 2012). Indeed, as one of the sources of the non-random patterns of reads observed in input controls is chromatin accessibility, it is possible that using such a control with FAIRE-seq would reduce the detection of true FAIRE signal. For my FAIRE-seq experiments, I did not sequence matched input controls for each developmental stage, but rather used GC-content and mappability data calculated from the *D. pseudoobscura* genome to correct for potential biases in the data during analysis. A detailed description of the methods used in the FAIRE protocol and for processing the sequencing data can be found in Chapter 2.

6.2 Overview of FAIRE-seq results

A summary of the clean and uniquely mapped reads as well as the rate of duplicate reads for each sample can be found in Table 6.1. The rates of duplication for these datasets range from 10.5 - 16.6%; however, the majority of the duplicates are only 2-fold, indicating high-quality, high-complexity libraries overall. Visualization of the read density scores across the genome shows a mix of regions with very strong, high peaks and regions with a much lower signal-to-noise ratio. High reproducibility is observed between biological replicates from the same stage, as well as a high degree of similarity between stages (Figure 6.1).

After mapping and extending reads, peaks were called for each replicate separately using MOSAiCS (Chung *et al.*, 2012). Peaks from each replicate were combined to make a high-confidence set for each stage by keeping all peaks that are present in 2 out of 3 replicates at FDR10. After merging peaks that were overlapping or immediately adjacent, this resulted in a set of 6348 total unique peaks across all stages. The number of peaks called at FDR5 and FDR10 for each replicate, as well as the combined peaks and unique peaks for each developmental stage are shown in Table 6.2. A large number of peaks for each stage are mapped to unassembled regions (chrU); while these peaks are likely to represent legitimate regions of accessible chromatin, they could map to repetitive regions

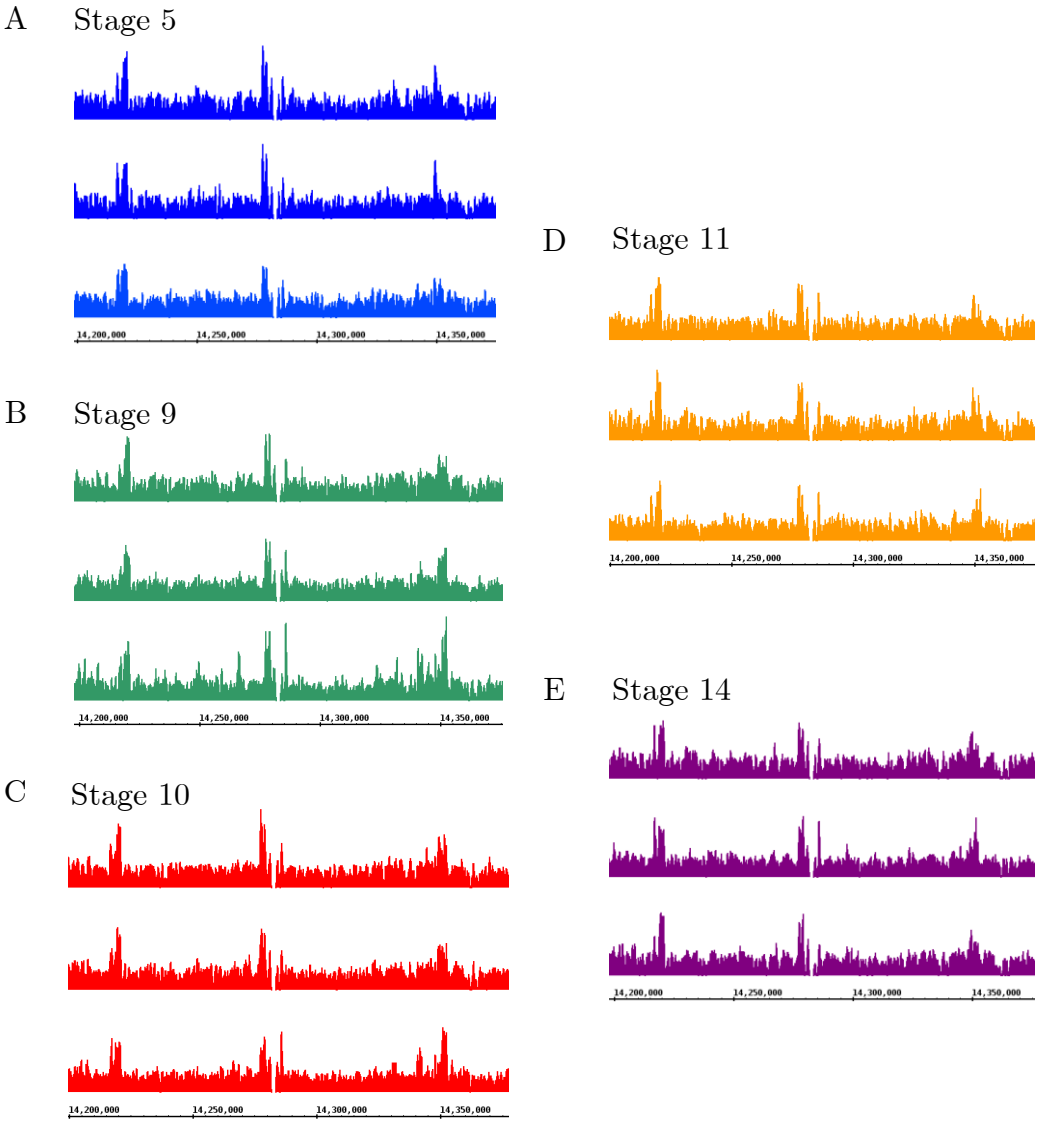


Figure 6.1: Comparison of read profiles between FAIRE biological replicates and developmental stages. The same 175-kb region of chromosome 2 in *D. pseudoobscura* is shown for each stage. In all samples, a relatively small number of strong, highly reproducible peaks are present, as well as smaller, less reproducible peaks which may constitute background noise. In all cases, reads have been normalized to a total library size of 1,000,000 for visualization purposes. The y-axes range from 0-15 reads. A.) Stage 5, three biological replicates. B.) Stage 9, three biological replicates. C.) Stage 10, three biological replicates. D.) Stage 11, three biological replicates. E.) Stage 14, three biological replicates.

Sample	Clean reads	Mapped reads	% Duplicate reads
Stage 5_1	14,316,011	7,693,543	12.72
Stage 5_2	7,160,440	5,073,937	10.52
Stage 5_3	11,116,549	9,115,838	12.67
Stage 9_1	10,665,720	8,708,489	12.73
Stage 9_2	10,451,702	8,782,095	12.00
Stage 9_3	10,476,067	8,891,655	12.79
Stage 10_1	12,937,965	7,749,216	13.03
Stage 10_2	11,122,744	9,629,801	12.47
Stage 10_3	10,356,274	8,987,571	12.36
Stage 11_1	11,459,640	9,950,997	13.81
Stage 11_2	11,148,754	9,635,232	13.35
Stage 11_3	8,897,559	7,727,392	12.64
Stage 14_1	10,336,055	8,988,933	13.42
Stage 14_2	9,947,079	8,639,987	12.94
Stage 14_3	14,637,840	12,291,220	15.65

Table 6.1: Summary of reads produced for FAIRE-seq libraries

and are, unfortunately, more difficult to annotate than peaks in assembled chromosomes.

In general the read density profiles for replicate stages are highly correlated; however, there is some between-replicate variability. In particular, replicate 2 for stage 5 is an outlier compared to the other stage 5 replicates. Clustering of each replicate by read counts within peaks shows that the two later stages are highly similar and can be differentiated from the earlier stages, with the stage 11 and stage 14 replicates clustering together and the stage 5, stage 9 and stage 10 replicates (except for Stage 5_2) clustering together (Figure 6.2). A similar effect can be seen in a principal component analysis (PCA) plot (Figure 6.3); the stage 11 and stage 14 replicates form a cluster together which is separable from the stage 9 samples, while the stage 5 and stage 10 replicates, which show greater within-stage variability, are spread across both principal component axes. Using DiffBind to identify differentially enriched sites between each stage and then clustering the replicates based only on those differential sites reveals a tighter clustering within the stage 11 samples and the stage 9 samples but greater variability within the stage 5, stage 10 and stage 14 samples (Figure 6.4) (Ross-Innes *et al.*, 2012).

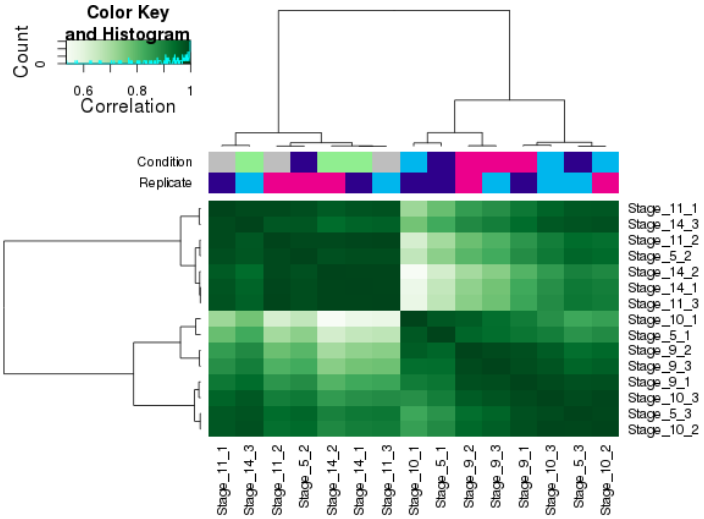


Figure 6.2: Heatmap showing clustering of all FAIRE-seq samples by affinity scores in every FAIRE interval. None of the stages has all three biological replicates clustered together. However, a division is visible between earlier stages (5, 9 and 10, in lower right) and later stages (11 and 14, in upper left), with the exception of stage 5 replicate 2, which clusters with the later stages. The color key and histogram show the distribution of pairwise correlations between sample affinity scores. Darker green corresponds to a higher correlation, while lighter green corresponds to a lower correlation.

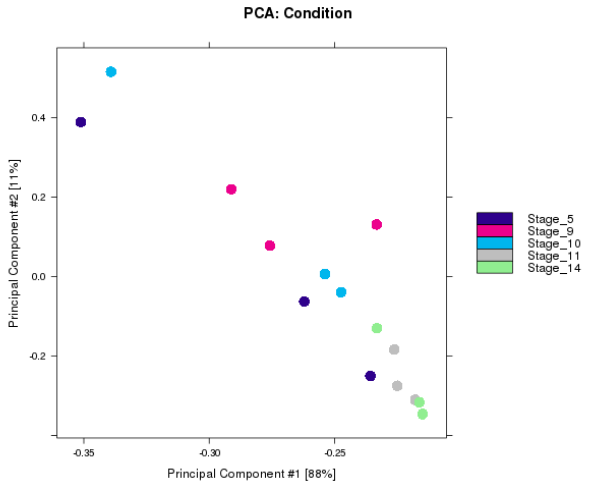


Figure 6.3: Principal component analysis of all FAIRE-seq samples. The first principal component, which explains 88% of the variation among samples, is plotted on the x-axis, and the second principal component, which explains 11% of the variation among samples, is plotted on the y-axis. As with the heatmap, a division is visible between earlier stages (5, 9 and 10) and later stages (11 and 14), although the replicates from later stages cluster more tightly than the replicates from earlier stages.

Sample	FDR5 peaks	FDR10 peaks	Combined peaks	Unique peaks
Stage 5_1	4600	5208		
Stage 5_2	4209	4685		
Stage 5_3	4527	5100		
Stage 5 combined			**4607**	**212**
Stage 9_1	4714	5326		
Stage 9_2	4087	4490		
Stage 9_3	7010	7979		
Stage 9 combined			**5165**	**475**
Stage 10_1	6102	6966		
Stage 10_2	4175	4599		
Stage 10_3	4697	5238		
Stage 10 combined			**5102**	**316**
Stage 11_1	4520	5059		
Stage 11_2	4008	4423		
Stage 11_3	4378	4837		
Stage 11 combined			**4444**	**185**
Stage 14_1	4514	4988		
Stage 14_2	4355	4794		
Stage 14_3	4287	4739		
Stage 14 combined			**4674**	**288**

Table 6.2: Peaks called in each FAIRE-seq replicate as well as combined and unique peaks for each developmental stage.

Individual sites show a high level of persistence between developmental stages, although some sites are gained and lost at each stage (Figure 6.5). Stage 9 has the highest percentage of sites that are not present in the previous stage and therefore originate in that stage (21.5%). Of the high-confidence accessible sites present at stage 14, 84% are present in stage 5, with 5.1% originating in stage 9, 1.8% originating in stage 10, 2.5% originating in stage 11, and 6.2% originating uniquely in stage 14. In accordance with this high persistence of accessible sites throughout development, DiffBind found relatively low numbers of sites with differential enrichment of read counts between stages, in particular when comparing between early stages (5, 9, and 10) or between late stages (11 and 14) (Table 6.3). It should be noted, however, that because these differentially enriched sites are based on read counts that are normalized between samples, while peaks were

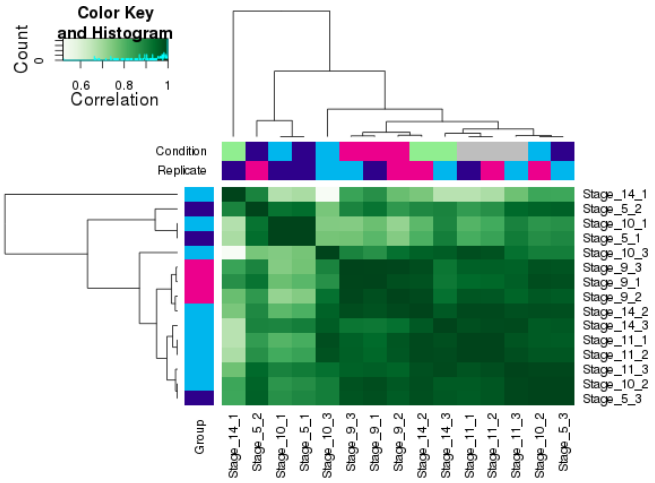


Figure 6.4: Heatmap showing clustering of all FAIRE-seq samples by affinity scores in FAIRE intervals that are differentially enriched in each stage in relation to the others. The three biological replicates from stage 9 cluster together along with one stage 14 replicate, while the other stages show greater variability between replicates. The color key and histogram show the distribution of pairwise correlations between sample affinity scores. Darker green corresponds to a higher correlation, while lighter green corresponds to a lower correlation.

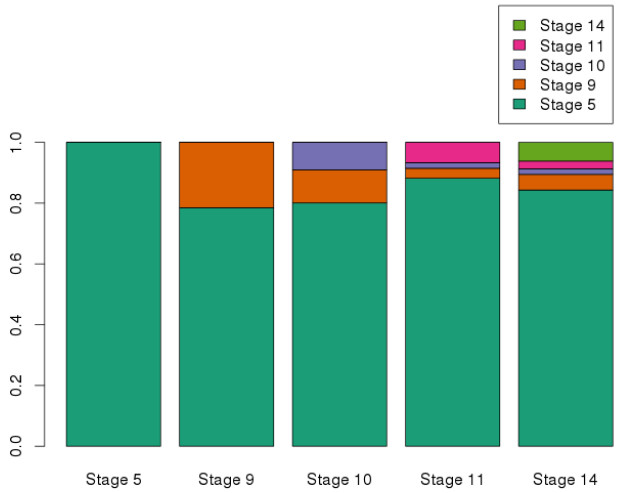


Figure 6.5: FAIRE intervals in each stage by stage of origin. For all intervals in each developmental stage, the earliest stage in which the interval is present was determined. The majority of FAIRE sites are present starting in stage 5 through stage 14, and a smaller proportion originate in each stage. The proportion of sites originating in each stage after stage 5 decreases over the course of development.

initially called based on the read density profiles from each sample independently, the numbers of unique peaks and the numbers of differentially enriched peaks for each stage are not directly comparable.

1st stage	2nd stage	Differential peaks
Stage 5	Stage 9	4
Stage 5	Stage 10	0
Stage 5	Stage 11	23
Stage 5	Stage 14	111
Stage 9	Stage 10	2
Stage 9	Stage 11	1581
Stage 9	Stage 14	1028
Stage 10	Stage 11	608
Stage 10	Stage 14	488
Stage 11	Stage 14	5

Table 6.3: Peaks with differential enrichment in pairwise comparisons between developmental stages.

6.3 Functional analysis of FAIRE peaks

6.3.1 Genomic annotation of FAIRE peaks

One of the challenges of working with non-model species is the relative lack of genome annotations available. In order to examine the genomic distribution of FAIRE peaks, I downloaded gene predictions made by both Genscan and GeneID from the UCSC Table Browser (Burge and Karlin, 1997; Karolchik, 2004; Karolchik *et al.*, 2014; Parra, 2000). I used these to annotate each FAIRE peak to either an exon, exon border, intron, gene border (5' or 3'), or intergenic region. The percentages of peaks annotated to each type of genomic region are very similar between all stages using each set of gene predictions. Of the total unique peaks, using the Genscan predictions, 3.7% fall entirely within exons, 30% fall on exon borders, 14.6% fall entirely within introns, 15.5% fall on gene borders and 36.1% fall entirely within intergenic regions (Figure 6.6A). Using the GeneID predictions, 2.8% fall entirely within exons, 20% fall on exon borders, 28.8% fall entirely within introns, 11.8% fall on gene borders and 36.5% fall entirely within

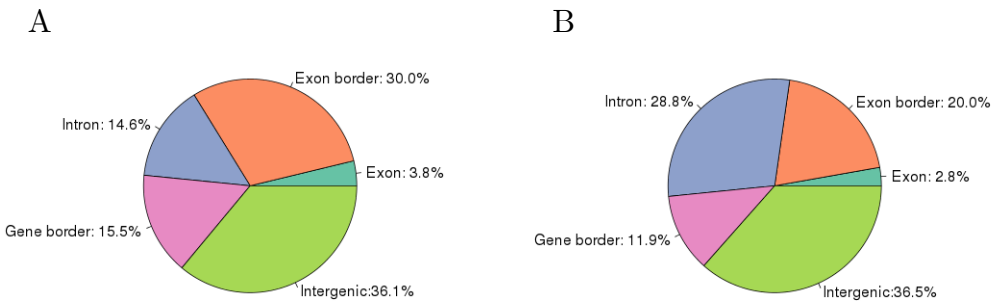


Figure 6.6: Genomic annotation of FAIRE sites. For each set of annotations, five genomic feature categories were considered: intron, exon, exon border, gene border and intergenic. All unique FAIRE sites, from all stages, were annotated. A.) Annotations based on Genscan gene models. The majority of FAIRE sites fall into intergenic regions, followed by exon borders and gene borders. Both the intergenic and gene border categories may include promoters. B.) Annotations based on GeneID gene models. The proportion of FAIRE sites in intergenic regions and exons is similar to that for Genscan annotations, but more sites are annotated to introns and less to gene borders and exon borders.

intergenic regions (Figure 6.6B). The main difference between annotations made with the Genscan and GeneID predictions is between the exon border, intron, and gene border categories, suggesting that while the exact locations of gene and exon predictions vary between the two predicted gene sets, the overall proportion of FAIRE peaks hitting genes and exons is similar.

6.3.2 Enriched motifs in FAIRE peaks

To identify enriched sequence motifs within FAIRE peaks in an unbiased way, I used HOMER to perform scans for both *de novo* motifs and known motifs (Heinz *et al.*, 2010). All stages showed similar motif enrichments. For each stage, the top hits of known motifs included a helix-loop-helix (HLH) motif (p <1e-26), a basic leucine zipper (bZIP) motif (p <1e-21) and a zinc-finger domain (zf) motif (p <1e-11), as well as several motifs flagged as promoters, including the TATA-box motif (p <1e-6). 15-16% of peaks in all stages contained a TATA-box motif, indicating a strong presence of promoter regions in the recovered FAIRE intervals.

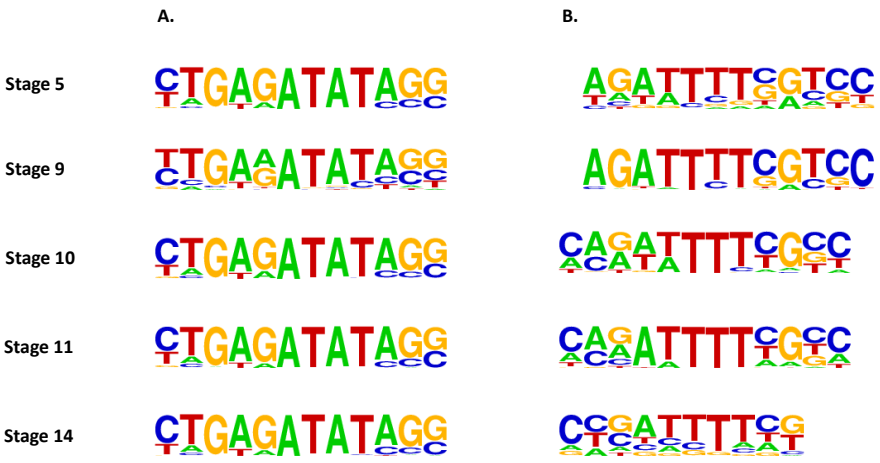


Figure 6.7: Two of the top *de novo* motifs identified in FAIRE intervals at every stage. A.) A highly significant GAGATATA motif, which may correspond to Cf1 binding in promoter regions, is found in all stages (p <= 1e-135). B.) A motif potentially matching Kni, which may represent nuclear hormone receptor activity, is found in all stages (p <= 1e-136).

Although none of the promoter motifs identified in the Thomas *et al.* DNase-seq data were enriched in my datasets, the FAIRE peaks at each stage were enriched for multiple known promoter sequences according to HOMER (Thomas *et al.*, 2011). The TATA motif was also flagged as enriched in *de novo* motif analyses, with one of the top-hit and most highly significant motifs in each stage being a GAGATATA motif (p <= 1e-135) (Figure 6.7A). The best match for this motif among curated *Drosophila* motifs found using STAMP is Chorion factor 1 (Cf1), a zinc-finger transcription factor with a functional annotation of RNA polymerase II core promoter proximal region sequence-specific DNA binding activity (Mahony and Benos, 2007). Another highly significant motif found in *de novo* analysis was difficult to assign to a known *Drosophila* transcription factor; however, its closest match was for Knirps (Kni) (p <= 1e-136). As Kni is a member of the nuclear hormone receptor (NHR) family, this motif may reflect general NHR activity in FAIRE intervals (Figure 6.7B).

Many of the top *de novo* motifs are predicted by HOMER to match non-*Drosophila* TFs, including vertebrate, yeast, and plant TFs. However, these motifs are highly

significant, and, given the high percentages of reads that mapped to the *D. pseudoobscura* genome, they are unlikely to be the result of contamination by DNA from other species. The predicted TF reported by HOMER for each motif is only a best guess and is dependent on the motifs available in the databases queried, meaning that true matches could be missed. It is possible that these motifs represent promoter elements or other deeply conserved features of open chromatin. A summary of the top ten *de novo* motifs identified in each stage can be found in Table 6.4.

Stage	**Rank**	**Predicted TF**	**Consensus sequence**	**P-value**
Stage 5	1	Pan	AGATTTTSGTCC	1E-163
	2	E2F3	YGYGAKCGGAAR	1E-162
	3	SIG1	CCTATATCTCAG	1E-146
	4	STP3	GCTAGAGCAACG	1E-136
	5	Hand1::Tcfe2a	AGTCTGGATC	1E-136
	6	Hbp1_2	CGAAAATGGG	1E-130
	7	MATA1	GCATCCACAATT	1E-127
	8	XBP1	CTCAAAGACTAT	1E-117
	9	Ovo	CTTCTGTKAKAT	1E-111
	10	ZmHOX2a	AGGGCCCGATCG	1E-109
Stage 9	1	ARR10	AGATTTTCGTCC	1E-158
	2	SIG1	CCTATATYTCAR	1E-142
	3	Smad3_1	RGAKCCAGACTS	1E-142
	4	MET31	TTCGCCSCACTY	1E-132
	5	Btd	CTTCCGCCCCCA	1E-132
	6	ZmHOX2a	CGATCGGGCCCT	1E-115
	7	MATA1	GCATCCACAATT	1E-110
	8	CST6	GTRACATC	1E-101
	9	Pros	KAGTCMTGCC	1E-099
	10	STB1	GATWCGAGAAAA	1E-086
Stage 10	1	SIG1	CTGAGATATAGG	1E-166
	2	Kni	CARWTTTTCGCC	1E-161
	3	AtLEC2	SCATNCACAAWW	1E-149
	4	TATA-box	TATTAAAGCTAG	1E-144

	5	Spdef_1	VAGTCTGGATCY	1E-141
	6	EGR1	CTTCCGMCCCCR	1E-138
	7	Zpf691_2	AATGAGNCTCAT	1E-125
	8	ZmHOX2a	CGATCGGGCCCT	1E-107
	9	Hth	GTRACATC	1E-101
	10	Pros	TAGCCATGCC	1E-092
Stage 11	1	Kni	CARATTTTYGYC	1E-149
	2	Btd	CTTCCGCCCCCA	1E-149
	3	Pan	KCGAGATTTTGA	1E-139
	4	SIG1	SCTATATCTCAG	1E-135
	5	MET4	GCATCCACAATT	1E-134
	6	Arid5a_1	ATSYCACTRTWA	1E-121
	7	MAC1	GCTAGAGCAACG	1E-113
	8	XBP1	CTCARAGACTAT	1E-097
	9	STB1	GTTTTCTCGAAT	1E-094
	10	ZmHOX2a	TTTCGTGATCGG	1E-088
Stage 14	1	Btd	CTTCCGCCCCCA	1E-166
	2	Gata5_2	CCTATATCTCAG	1E-157
	3	CG34031	CTATWARAGCTA	1E-139
	4	Kni	CYSATTTTCK	1E-136
	5	DAL82	CGAAATTTTG	1E-135
	6	FOXH1	GCATCCACAA	1E-131
	7	Smad3_1	GAGTCTGGAT	1E-128
	8	ZmHOX2a	CGATCGGGCCCT	1E-107
	9	Hb	GTTTTCTYGAAT	1E-095
	10	SOX10	KASTCATTGT	1E-087

Table 6.4: Top ten *de novo* motifs by p-value in FAIRE peaks from each stage. For each motif, the TF predicted by HOMER, the consensus sequence and the p-value are shown.

6.3.3 Relationship between FAIRE peaks and TF binding

I downloaded peaks from six ChIP-seq datasets for the transcription factors Pipsqueak (Psq), Trithorax-like (Trl), Kruppel (Kr), Giant (Gt), Bicoid (Bcd) and Hunchback (Hb) in 0-4 hour or blastoderm-stage *D. pseudoobscura* embryos (GEO accession numbers GSE25666, GSE25667 and GSE50771) and examined the patterns of Stage 5 FAIRE-seq tag enrichment within regions 2.5 kb upstream and downstream of peak centers (Figure 6.8). Kr, Gt, Bcd and Hb are anterior-posterior (AP) TFs whose binding has been shown to correlate well with chromatin accessibility measured by both DNaseI digestion and FAIRE in *D. melanogaster* embryos (McKay and Lieb, 2013; Li *et al.*, 2011). Trl, which is also known as the GAGA-binding factor or GAF, is a maternally-contributed factor that plays a role in chromatin remodelling as well as regulating RNA polymerase II activity and is thought to be important for establishing open chromatin during zygotic genome activation (ZGA) in the very early *Drosophila* embryo (Darbo *et al.*, 2013). Psq, on the other hand, is involved in chromatin silencing through binding to polycomb response elements (PREs) (Huang *et al.*, 2002). Interestingly, there is an increase in average FAIRE tag density at the center of peaks for all of these factors (Figure 6.8). This is true for all three biological FAIRE replicates; however, in replicates 1 and 2, the presence of a few very high peaks of FAIRE signal dominate the distributions for some TFs, creating a jagged appearance. Therefore, I focused on the FAIRE scores from replicate 3 for visualization and assessing the relative strength of FAIRE signal in peaks from each TF. The peaks with the highest FAIRE scores are those for Psq. Although it seems surprising that there is an enrichment of open chromatin at binding sites for a factor involved in chromatin silencing, an enrichment for PREs in FAIRE peaks has also been observed in *D. melanogaster* (McKay and Lieb, 2013). This could reflect the fact that Psq binds in open chromatin in order to repress neighboring regions, possibly setting up boundaries for chromatin domains. Of the AP factors that have been studied in *D. pseudoobscura*, the relative enrichments of FAIRE tags in peaks follow the same order as those observed in *D. melanogaster* embryos, with Bcd showing the greatest enrichment, followed by Gt, Kr and Hb (McKay and Lieb, 2013).

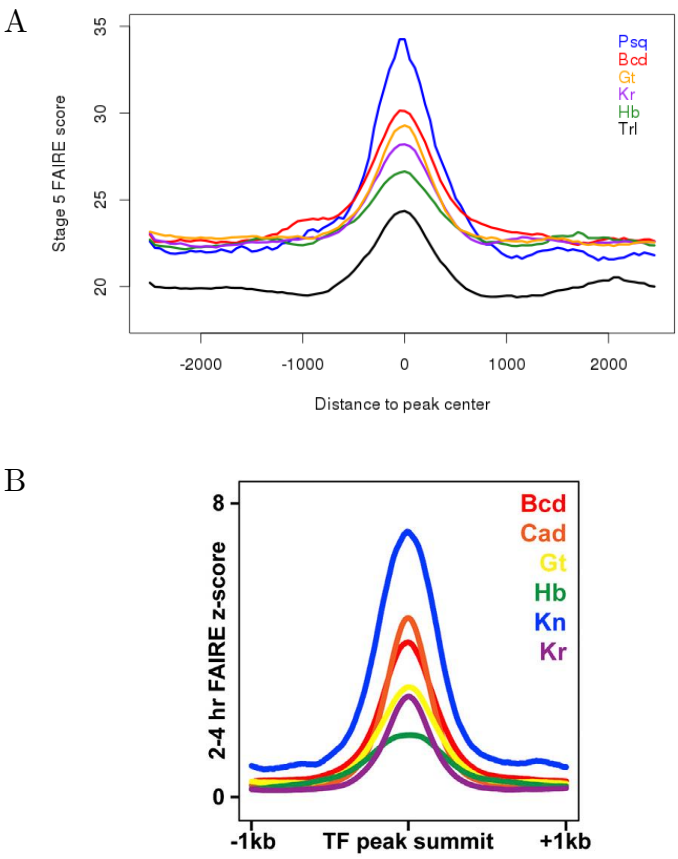


Figure 6.8: Enrichment of FAIRE read counts in TF binding peaks in *D. pseudoobscura* and *D. melanogaster*. A.) In *D. pseudoobscura*, a 5-kb region around the center of each ChIP-seq peak for Psq, Bcd, Gt, Kr, Hb and Trl was considered, and the number of FAIRE-seq reads overlapping 50-bp bins in each interval was counted. FAIRE scores represent the average counts from Stage 5 replicate 3 in all peaks. A local maximum of FAIRE accessibility is seen at the center of the intervals for all TFs, with the highest scores varying between TFs. The highest FAIRE scores are found in Psq peaks, while the lowest are found in Trl peaks. B.) The order of the AP factors Bcd, Gt, Kr and Hb by FAIRE scores is the same as that found in *D. melanogaster*. Figure reproduced from McKay and Lieb (2013).

6.3.4 Relationship between FAIRE accessibility and Dichaete binding in *D. pseudoobscura*

I was also curious to investigate the relationship between chromatin accessibility and Dichaete binding as measured by the DamID experiment that I performed in *D. pseudoobscura*. There is evidence to suggest that, in *D. melanogaster*, Dichaete binds to HOT regions, which are associated with Trl binding and open chromatin (Aleksic *et al.*, 2013; Kvon *et al.*, 2012). If Dichaete binding is partially driven by patterns of chromatin accessibility, then changes in accessibility between *D. melanogaster* and *D. pseudoobscura* may underscore some changes in binding between the two species, possibly leading to new functional binding events. In order to examine this relationship, I followed the same procedure as above, finding the center of each Dichaete-Dam binding interval and calculating the coverage of FAIRE-seq reads in 50-bp bins extending 2.5 kb on either side. This approach is not ideal for DamID, as the center of each binding interval does not necessarily correspond to the actual location of TF binding. However, by examining 5-kb intervals, the majority of true binding sites should be captured, as most GATC fragments are shorter than 5 kb. I calculated the average FAIRE scores in all intervals using all three biological replicates from each developmental stage separately, since the Dichaete-Dam experiment used embryos spanning all of the stages that were assayed using FAIRE-seq.

The profiles of average FAIRE scores within Dichaete-Dam binding intervals are quite jagged and contain multiple peaks, which may be reflective of the fact that the strongest binding does not necessarily take place at the center of intervals. However, two main peaks of FAIRE accessibility are visible at each developmental stage, one located at around 1500 bp upstream of the interval centers and one coinciding with the center of the intervals (Figure 6.9). The relative heights of these peaks vary with developmental stage; the peak at the center of the intervals is the highest at stages 9 and 10 and decreases in stages 11 and 14. The absolute FAIRE scores around the center of binding intervals are also highest in stages 9 and 10, suggesting that Dichaete binding correlates best with chromatin accessibility during these stages. The overall shapes of the FAIRE profiles in Dichaete-Dam binding intervals are similar for all stages, which is not surprising given the high correlations between FAIRE accessibility profiles at all stages.

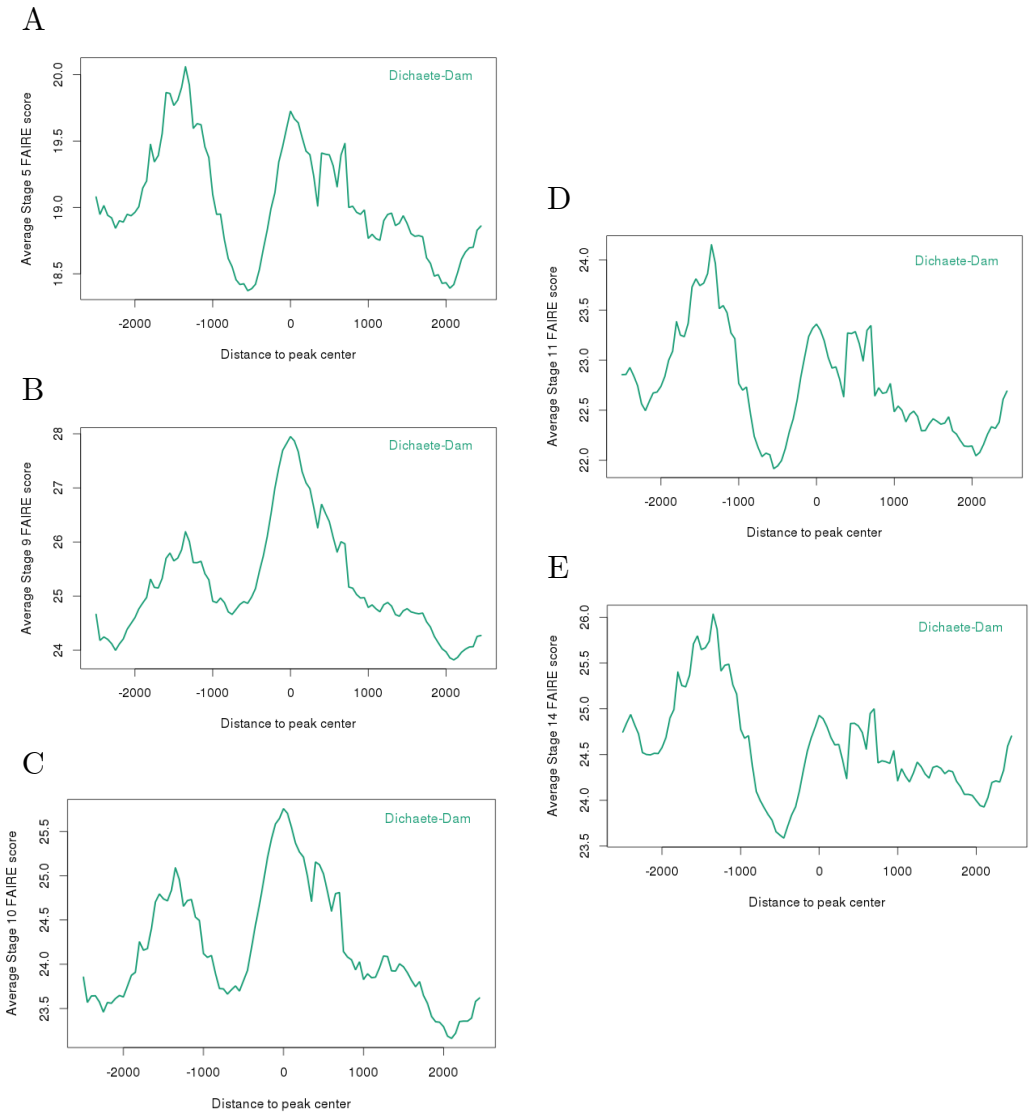


Figure 6.9: Enrichment of FAIRE read counts in Dichaete-Dam binding intervals in *D. pseudoobscura*. A 5-kb region around the center of each binding interval was considered, and the number of FAIRE-seq reads overlapping 50-bp bins in each interval was counted. FAIRE scores represent the average counts from three replicates at each stage in all Dichaete-Dam intervals. In all stages, a peak of FAIRE accessibility is visible at about 1500 bp upstream of the center of the binding intervals, and another is visible at the center of the binding intervals. The peak at the center of the binding intervals is the strongest in stages 9 and 10. A.) Stage 5 FAIRE scores in Dichaete-Dam intervals. B.) Stage 9 FAIRE scores in Dichaete-Dam intervals. C.) Stage 10 FAIRE scores in Dichaete-Dam intervals. D.) Stage 11 FAIRE scores in Dichaete-Dam intervals. E.) Stage 14 FAIRE scores in Dichaete-Dam intervals.

The peaks of chromatin accessibility found in Dichaete-Dam intervals do not have as high FAIRE scores as for some other TFs, including Psq, Bcd and Gt; however, they are consistent with the FAIRE scores found in Trl and Hb peaks (Figure 6.8). In contrast to the pattern of FAIRE signal in Dichaete-Dam binding intervals, the intervals that are bound more highly by the Dam-only control in *D. pseudoobscura* show lower FAIRE scores overall in each stage and do not show a peak of enrichment around the interval centers (Figure 6.10). These results suggest that, although FAIRE accessibility shows a complex pattern of enrichment within Dichaete-Dam binding intervals, functional Dichaete binding does correlate with chromatin accessibility to some extent in *D. pseudoobscura*.

Because DamID peaks do not necessarily contain true binding sites at their centers, I repeated the preceding analysis using peaks defined as 2.5kb up- and downstream from the coordinates of the highest-scoring match to a Sox motif in each interval. This definition of a peak is also somewhat problematic, as there were often more than one Sox motif with equally high or very close scores, and it is impossible to know from this dataset whether only one or more than one motif was bound in each binding interval. Again, the profiles of FAIRE accessibility scores in these peaks are quite jagged, although they do show a peak of enrichment near the center. This centered accessibility is most noticeable in stages 9 and 10; in both stage 5 and later stages, it appears to shift slightly downstream of the center, while another, more upstream peak of FAIRE accessibility is dominant (Figure 6.11). The FAIRE scores are lower overall in these motif-defined Dichaete-Dam peaks compared to the peaks defined around the centers of binding intervals, suggesting that, although there is some enrichment of accessibility at high-scoring Sox motifs, this approach does not capture the most accessible chromatin. Since the motif scores are based on a PWM constructed from the average of all motifs found, it is possible that the best-scoring motifs do not reflect actual motif usage by Dichaete, which may differ subtly in different populations of cells or different developmental stages. The variation in accessibility profiles observed in different stages might be a function of such differential motif usage.

I also used BedTools to find all intersections between Dichaete-Dam binding intervals and FDR10 FAIRE intervals in *D. pseudoobscura* at each developmental stage. Although this approach is likely to underestimate the correlation between Dichaete-Dam binding and FAIRE accessibility, it allowed me to determine a set

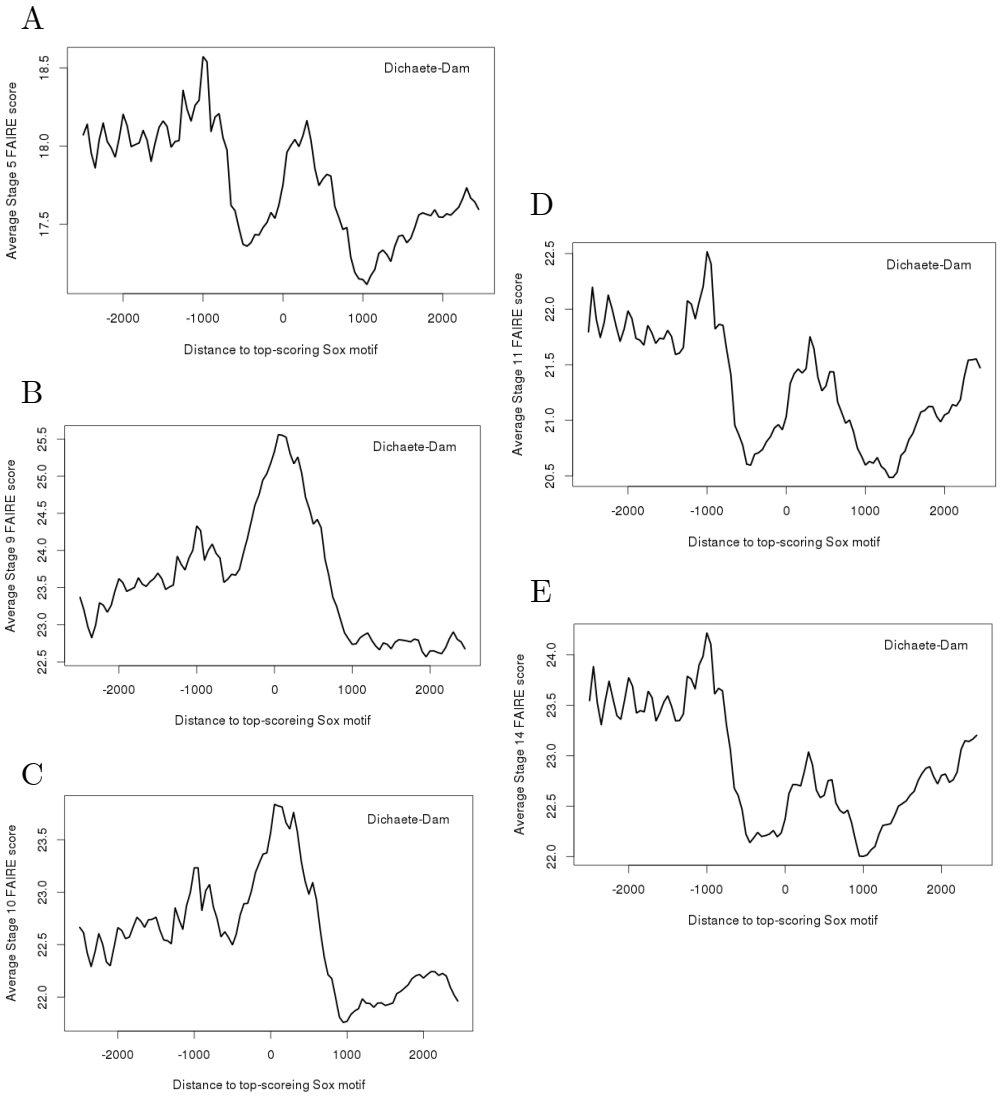


Figure 6.10: Average FAIRE scores in Dam-only binding intervals in *D. pseudoobscura*. A 5-kb region around the center of each binding interval was considered, and the number of FAIRE-seq reads overlapping 50-bp bins in each interval was counted. FAIRE scores represent the average counts from three replicates at each stage in all Dichaete-Dam intervals. For each stage, the FAIRE scores in Dam-only control intervals are lower than the corresponding FAIRE scores in Dichaete-Dam intervals. Several local peaks of enrichment are present, but they are not located at the center of intervals. A.) Stage 5 FAIRE scores in Dam-only intervals. B.) Stage 9 FAIRE scores in Dam-only intervals. C.) Stage 10 FAIRE scores in Dam-only intervals. D.) Stage 11 FAIRE scores in Dam-only intervals. E.) Stage 14 FAIRE scores in Dam-only intervals.

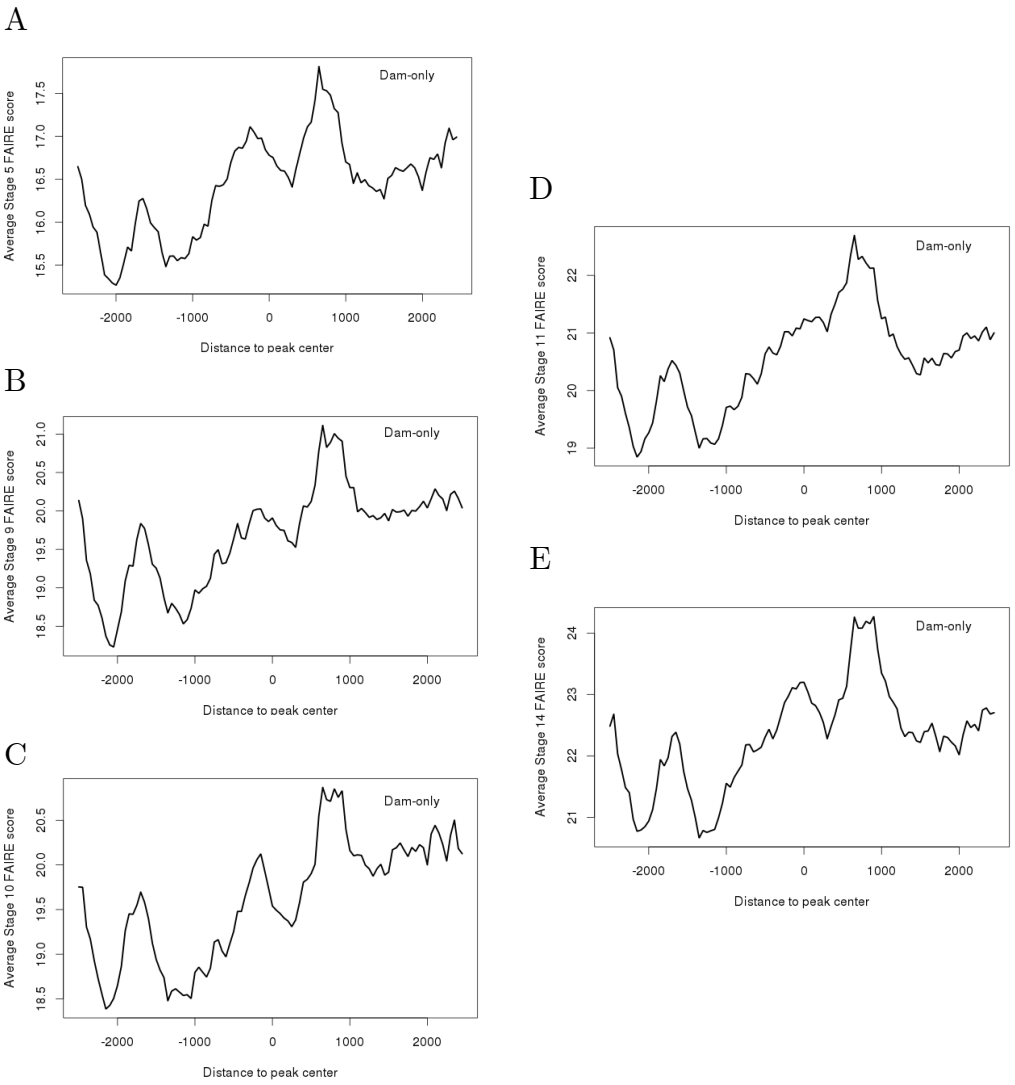


Figure 6.11: Enrichment of FAIRE read counts around Sox motifs in Dichaete-Dam binding intervals in *D. pseudoobscura*. A 5-kb region centered around the best-scoring Sox motif in each binding interval was considered, and the number of FAIRE-seq reads overlapping 50-bp bins in each interval was counted. FAIRE scores represent the average counts from three replicates at each stage in all Dichaete-Dam intervals. In all stages, a peak of FAIRE accessibility is visible at about 1000 bp upstream of the Sox motifs, and another is visible centered around the Sox motifs. The peak centered around Sox motifs is the strongest in stages 9 and 10; however, in all stages it is weaker than the peak observed at the center of binding intervals. A.) Stage 5 FAIRE scores in Dichaete-Dam intervals. B.) Stage 9 FAIRE scores in Dichaete-Dam intervals. C.) Stage 10 FAIRE scores in Dichaete-Dam intervals. D.) Stage 11 FAIRE scores in Dichaete-Dam intervals. E.) Stage 14 FAIRE scores in Dichaete-Dam intervals.

of Dichaete-Dam intervals that are definitively located in open chromatin. The numbers of Dichaete-Dam intervals that overlap with a FAIRE interval in each stage correspond to the average FAIRE scores in Dichaete-Dam intervals in each stage, with the highest numbers of overlaps present in stages 9 and 10 and the lowest numbers in stages 11 and 14 (Table 6.5). In total there are 257 unique Dichaete-Dam intervals detected that are located within a FAIRE interval, representing 8.7% of all *D. pseudoobscura* Dichaete-Dam binding intervals.

Developmental Stage	**Overlaps between Dichaete-Dam and FAIRE intervals**	**Percent of FAIRE intervals overlapping**	**Percent of Dichaete-Dam intervals overlapping**
Stage 5	96	2.1%	3.3%
Stage 9	196	3.8%	6.6%
Stage 10	188	3.7%	6.4%
Stage 11	48	1.1%	1.6%
Stage 14	50	1.1%	1.7%

Table 6.5: Overlaps between Dichaete-Dam binding intervals and FAIRE intervals in *D. pseudoobscura* embryos at five developmental stages.

In order to evaluate whether Dichaete-Dam binding intervals in FAIRE intervals tend to be unique to *D. pseudoobscura* or conserved across species, I first found the set of binding intervals that are qualitatively conserved between *D. melanogaster* and *D. pseudoobscura*, as well as those that are unique to *D. pseudoobscura*, and then translated their genomic coordinates to the *D. pseudoobscura* genome assembly. This resulted in the loss of some intervals, as not all coordinates could be uniquely re-mapped; however, it was necessary in order to examine the effect of FAIRE accessibility in the *D. pseudoobscura* genome. 1111 conserved intervals and 447 unique intervals were translated; 151 of the conserved intervals are located in a FAIRE interval, while only 12 of the unique *D. pseudoobscura* intervals are located in a FAIRE interval (Figure 6.12). While the total proportion of Dichaete-Dam binding intervals located in FAIRE intervals is low, these binding intervals are significantly more likely to be conserved in *D. melanogaster* than to be unique to *D. pseudoobscura* (Chi-squared test with Yates continuity correction, $\chi^2 = 39.3$, d.f. $= 1$, p-value = 3.6e-10). This suggests that, rather

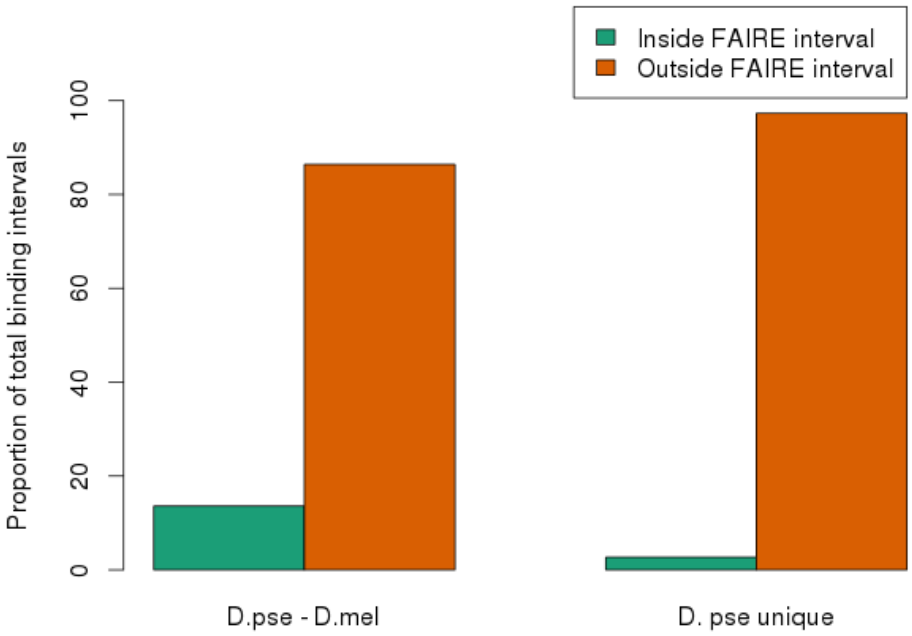


Figure 6.12: Dichaete-Dam binding intervals located within FAIRE intervals are more likely to be conserved than those located outside FAIRE intervals. Although the majority of Dichaete-Dam intervals do not overlap with a FAIRE interval in *D. pseudoobscura*, those that do are significantly more likely to also be bound by Dichaete-Dam in *D. melanogaster* compared to those that do not (p-value = 3.6e-10). Abbreviations: D. pse - D. mel, conserved in both *D. pseudoobscura* and *D. melanogaster*; D. pse unique, bound uniquely in *D. pseudoobscura*.

than the evolution of new binding events being driven by changes in chromatin accessibility, Dichaete-Dam binding sites in accessible chromatin tend to be conserved across species. In *D. melanogaster* embryos, regions of open chromatin as measured by both DNase-seq and FAIRE-seq are bound by multiple regulatory factors and are associated with developmental regulatory genes (McKay and Lieb, 2013; Thomas *et al.*, 2011). Considering the functional importance of these regions, there is likely to be selective pressure to maintain open chromatin domains at key regulatory loci during evolution, despite the chromosomal rearrangements that have occurred between *D. melanogaster* and *D. pseudoobscura*.

6.4 Comparison with chromatin accessiblity data in *D. melanogaster*

The two chromatin accessibility datasets in *D. melanogaster* that offer the most direct comparison to my *D. pseudoobscura* FAIRE-seq datasets are the DNase-seq data generated in five matching developmental stages by Thomas *et al.* (2011) and the FAIRE-seq data generated by McKay *et al.* (McKay and Lieb, 2013) at 2-4 hours after egg laying, 6-8 hours after egg laying and 16-18 hours after egg laying. The McKay *et al.* FAIRE-seq data are a better match in terms of technique, while the Thomas *et al.* data are a more precise match in terms of temporal specificity; however, interesting comparisons can be made with both datasets. A simple comparison of the number of FAIRE and DHS peaks called reveals that I found significantly fewer peaks for every stage in *D. pseudoobscura* than were found in *D. melanogaster* for either technique. This is particularly the case for the DNase-seq dataset, in which 20,000-30,000 DHS peaks were called for each stage, or 5-6 times more than I found. McKay *et al.* found 11,000-13,000 FAIRE peaks for each stage in *D. melanogaster*, which are considerably fewer than the DHS peaks but still more than twice the number of FAIRE peaks called for *D. pseudoobscura*. It is unclear why this is the case; it seems unlikely that the the *D. pseudoobscura* genome genuinely has 2-6 times less open chromatin than the *D. melanogaster* genome, as they have similar sizes and gene densities (Richards, 2005).

The FAIRE peaks in *D. melanogaster* also overlap with a higher proportion of TF binding peaks than those identified in *D. pseudoobscura* (McKay and Lieb, 2013); this may be due to an under-identification of FAIRE accessible regions in *D. pseudoobscura* embryos, rather than a difference in the relationship between TFs and accessible chromatin between species. These differences could result in part from differences in the analytical methods used to process the data and call peaks; however, inspecting the read density profiles by eye suggests that the salient peaks have been successfully called for each dataset. Finally, the difference in numbers of peaks could be due to technical variation in the FAIRE protocol. It is possible that the *D. pseudoobscura* embryos were underfixed, leading to a relative homogenization of signal and loss of peaks. On the other hand, if

the embryos were overfixed, genuinely accessible regions might not have been recovered. Nonetheless, and encouragingly, despite the decreased numbers of peaks called in *D. pseudoobscura*, the peaks that are called share similar properties to those identified in *D. melanogaster* in terms of genomic annotations and TF binding patterns.

To measure the similarities between FAIRE-seq and DNase-seq datasets at each stage, I downloaded the Thomas *et al.* DNase-seq data from the NCBI Sequence Read Archive [SRA:SRX020691, SRA:SRX020692, SRA:SRX020693, SRA:SRX020694, SRA:SRX020695, SRA:SRX020696, SRA:SRX020697, SRA:
SRX020698, SRA:SRX020699, SRA:SRX020700] and mapped the data against the *D. melanogaster* genome. I then calculated the correlations between reads from each set of DNase-seq biological replicates and FAIRE-seq biological replicates that had been translated to the *D. melanogaster* genome in the full set of DNase accessible regions, a total of 65536 intervals (Thomas *et al.*, 2011). For each stage, the DNase-seq replicates are highly correlated (R^2 >0.98), as are the FAIRE-seq replicates (R^2 >0.94). The correlations between FAIRE-seq replicates and DNase-seq replicates range from 0.55 to 0.71, with the highest correlations being present at stage 9. The differences between these samples encompass both technical differences between FAIRE-seq and DNase-seq as well as differences in patterns of accessibility between species; however, the coefficients of correlation are similar to those calculated for Dichaete-Dam binding between *D. melanogaster* and *D. pseudoobscura*, showing that a similar amount of inter-species variation is captured by examining chromatin accessibility as by examining the binding patterns of a single TF. Clustering all replicates from both techniques at all stages shows that FAIRE-seq samples are more similar across stages than DNase-seq samples (Figure 6.13). For some stages, such as stage 14, there are higher correlations between samples from the two techniques than there are between samples from different stages; however, overall, the high degree of similarity between FAIRE-seq stages means that the FAIRE-seq samples show similar correlations to DNase-seq samples at all stages.

I performed the same type of analysis with the McKay *et al.* FAIRE-seq samples from *D. melanogaster* embryos, which I also downloaded from the NCBI Sequence Read Archive [SRA:SRX155022, SRA:SRX155023, SRA:SRX155024] and mapped against the *D. melanogaster* genome. I calculated the correlations be-

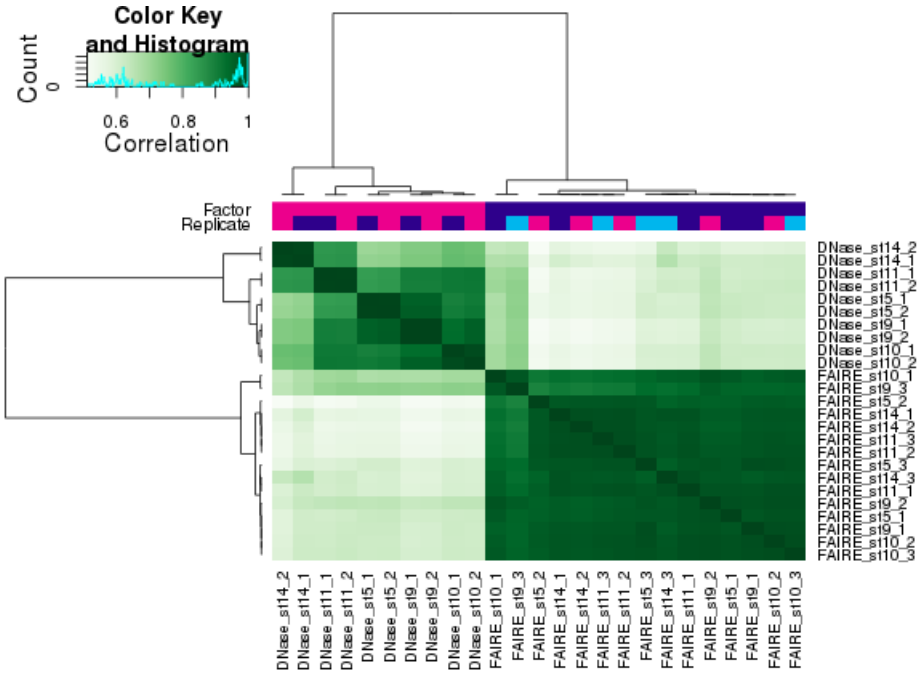
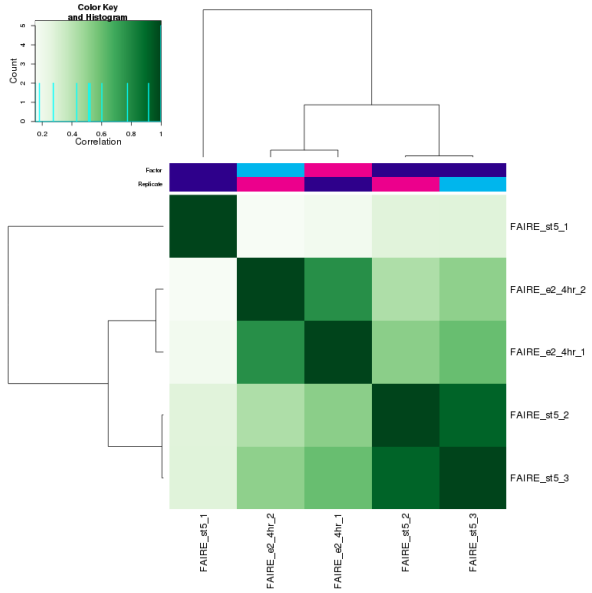


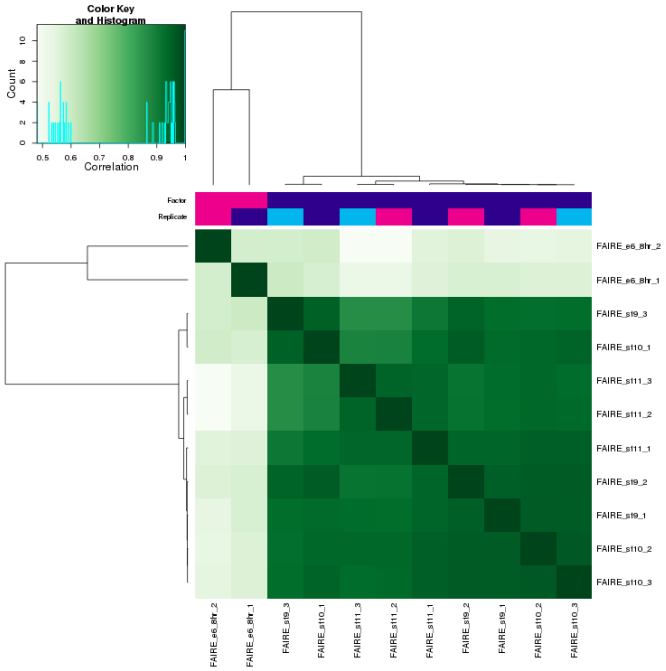
Figure 6.13: Heatmap showing correlations between translated FAIRE-seq sample read counts and DNase-seq sample read counts within all DNase accessible sites in five developmental stages in *D. melanogaster*. FAIRE-seq samples show higher correlations between stages than DNase-seq samples. The highest correlations between techniques are for FAIRE-seq samples from stages 9 and 10 with all DNase-seq samples. The color key and histogram show the distribution of pairwise correlations between sample affinity scores in all DNase accessible regions. Darker green corresponds to a higher correlation, while lighter green corresponds to a lower correlation.

tween reads from each McKay *et al.* FAIRE-seq sample and each of my translated FAIRE-seq samples from overlapping stages (stage 5 versus 2-4 hour embryos, stage 9, 10 and 11 versus 6-8 hour embryos and stage 14 versus 16-18 hour embryos) within the McKay FAIRE accessible regions. In each comparison, the *D. pseudoobscura* FAIRE-seq replicates show considerably higher correlations, ranging from 0.92 - 0.97, than the McKay FAIRE-seq replicates, whose correlations ranged from 0.58 - 0.77. The one outlier for the *D. pseudoobscura* FAIRE-seq samples was stage 5 replicate 1, which showed poor correlations with the other stage 5 replicates in the regions examined. With the exception of that sample, the *D. pseudoobscura* FAIRE-seq samples show similar but slightly lower levels of correlation with the McKay FAIRE-seq samples compared to the DNase-seq samples, ranging from 0.43 - 0.60 for 2-4 hour embryos (Figure 6.14A), 0.48 - 0.60 for 6-8 hour embryos (Figure 6.14B) and 0.36 - 0.56 (Figure 6.14C). For the latest stage embryos, replicate 1 correlates more closely with the *D. pseudoobscura* stage 14 samples than does replicate 2. While it is somewhat surprising that the samples from two FAIRE-seq experiments are less correlated than samples from a FAIRE-seq experiment and a DNase-seq experiment, even between different stages, this may be due to the fact that fewer peaks were identified in the McKay FAIRE-seq data and so less of the data was included in calculating the coefficients of correlation, which may have resulted in the exclusion of some relevant genomic regions.

A


Color Key and Histogram
Count
Correlation
Factor
Replicate
FAIRE_st5_1
FAIRE_e2_4hr_2
FAIRE_e2_4hr_1
FAIRE_st5_2
FAIRE_st5_3

B


Color Key and Histogram
Count
Correlation
Factor
Replicate
FAIRE_e6_8hr_2
FAIRE_e6_8hr_1
FAIRE_st9_3
FAIRE_st10_1
FAIRE_st11_3
FAIRE_st11_2
FAIRE_st11_1
FAIRE_st9_2
FAIRE_st9_1
FAIRE_st10_2
FAIRE_st10_3

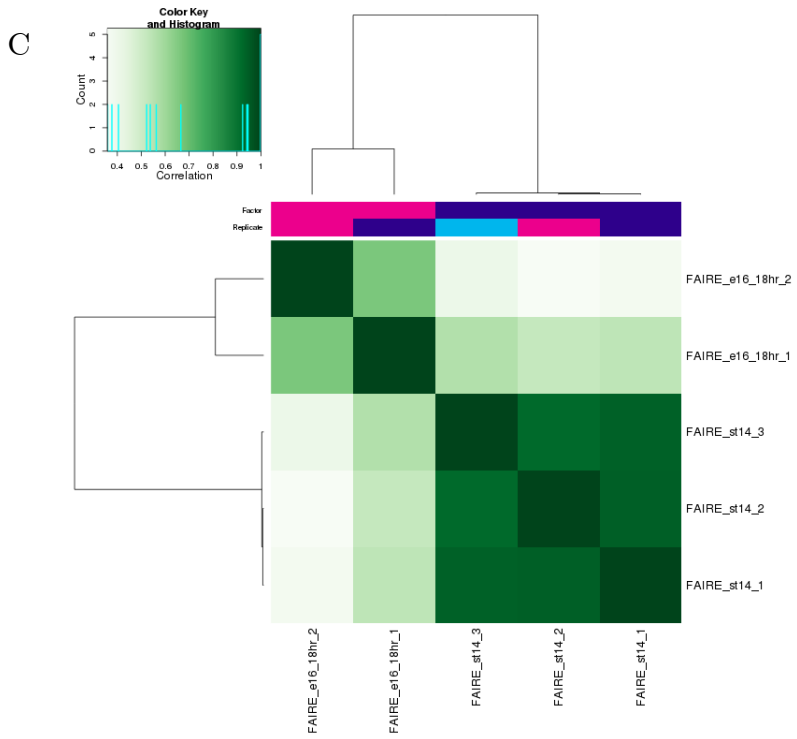


Figure 6.14: Heatmaps showing correlations between translated *D. pseudoobscura* FAIRE-seq sample read counts and *D. melanogaster* FAIRE-seq sample read counts from McKay *et al.* (2013) within all *D. melanogaster* embryonic FAIRE accessible sites. The color key and histogram show the distribution of pairwise correlations between sample affinity scores in all FAIRE accessible regions. Darker green corresponds to a higher correlation, while lighter green corresponds to a lower correlation. A.) Comparison of stage 5 translated FAIRE-seq samples from *D. pseudoobscura* and 2-4 hour FAIRE-seq samples from *D. melanogaster*. *D. pseudoobscura* stage 5 replicate 1 is a clear outlier. B.) Comparison of stage 9, stage 10 and stage 11 translated FAIRE-seq samples from *D. pseudoobscura* and 6-8 hour FAIRE-seq samples from *D. melanogaster*. The *D. pseudoobscura* samples show higher correlations, even across stages, than do the *D. melanogaster* replicates. C.) Comparison of stage 14 translated FAIRE-seq samples from *D. pseudoobscura* and 16-18 hour FAIRE-seq samples from *D. melanogaster*. *D. melanogaster* replicate 1 is more similar to all of the *D. pseudoobscura* samples than is replicate 2.

6.5 Discussion of results

The FAIRE-seq datasets which I generated for five developmental stages in *D. pseudoobscura* embryos show very high levels of reproducibility between repli-

cates, as well as high correlations between developmental stages. The majority of accessible sites identified originate in stage 5 and are maintained throughout development, although some developmentally dynamic sites originate and are lost in later stages. These samples show significant differences from publicly available chromatin accessibility datasets in *D. melanogaster*, some of which seem more likely to be due to technical differences in sample preparation than to biological differences. My *D. pseudoobscura* FAIRE-seq data contains 2-6 times fewer highly accessible regions in each developmental stage, and these regions contain fewer overlaps with TF binding intervals in *D. pseudoobscura*, measured either through ChIP-seq or DamID. Additionally, there are fewer peaks of accessibility identified as unique to each stage in *D. pseudoobscura* than in *D. melanogaster*; however, again, this may be due to technical differences resulting in a loss of more developmentally dynamic accessible regions. On the level of read counts, however, the *D. pseudoobscura* samples show similar levels of correlation with *D. melanogaster* DNase-seq samples as do *D. pseudoobscura* Dichaete-Dam samples with *D. melanogaster* Dichaete-Dam samples. They show slightly lower levels of correlation with *D. melanogaster* FAIRE-seq samples, which, although they were detected using the same technique, span different periods of developmental time. The read-level correlations suggest that, while a comparison of thresholded peaks may highlight technical differences, the overall accessibility profiles of *D. pseudoobscura* embryonic chromatin and *D. melanogaster* embryonic chromatin have evolved differences at a similar rate as the binding profiles of several transcription factors, as well as the insulator protein CTCF, between these two species (He *et al.*, 2011b; Ni *et al.*, 2012; Paris *et al.*, 2013).

In terms of annotation to genomic features and TF binding, the *D. pseudoobscura* FAIRE intervals show similar overall properties to the *D. melanogaster* FAIRE and DNase intervals. Although the quality and level of detail of gene model predictions available for *D. pseudoobscura* is lower than that for *D. melanogaster*, a similar proportion of the *D. pseudoobscura* FAIRE intervals and the *D. melanogaster* DNase intervals are annotated to intergenic DNA and, for the GeneID gene predictions, intronic DNA. The gene border category in *D. pseudoobscura* may include TSSs as well as 5' UTRs and 3' UTRs; the DNase intervals are annotated to these categories at a combined proportion that is close to that of *D. pseudoobscura* FAIRE intervals in gene borders. The biggest difference

between the genomic annotations in the two species is that a higher proportion of *D. melanogaster* DNase intervals are annotated to coding sequences; however, these may include intervals that partially overlap exons as well as introns, which were annotated to the exon border category in *D. pseudoobscura* (Thomas *et al.*, 2011).

Previous studies have indicated that DNase-seq tends to identify more open chromatin regions in promoters than FAIRE-seq (Koohy *et al.*, 2013). However, both the *D. pseudoobscura* FAIRE intervals and the *D. melanogaster* DNase intervals show a strong presence of promoter motifs, although different promoter motifs are enriched in each dataset (Thomas *et al.*, 2011). Many of the top known and *de novo* motifs found in the *D. pseudoobscura* FAIRE intervals were difficult to assign to a *Drosophila* TF. However, there is a strong enrichment for motifs corresponding to several major families of DNA binding domains, including the NHR, HLH, bZIP and zf families. This indicates that, in addition to promoters, FAIRE accessible regions include enhancers that are bound by a broad variety of regulatory factors. In support of this view, a peak of FAIRE signal was found in the center of ChIP-seq binding intervals for several AP factors in *D. pseudoobscura*, including Bcd, Gt, Kr and Hb, as well as the TFs Psq and Trl, which are both implicated in chromatin remodelling. Interestingly, the relative intensities of FAIRE signal in AP factor binding intervals follow the same order in *D. pseudoobscura* embryos as in *D. melanogaster* embryos (McKay and Lieb, 2013).

One of the main motivations behind generating a FAIRE-seq dataset in *D. pseudoobscura* was to investigate the relationship between accessible chromatin and conservation of group B Sox binding, using the DamID data that I acquired in *D. pseudoobscura* and *D. melanogaster*. Since I was not able to perform DamID for SoxNeuro in *D. pseudoobscura*, I focused this analysis on Dichaete-Dam binding. First, I examined the overall pattern of FAIRE accessibility in Dichaete-Dam binding intervals in *D. pseudoobscura*. I found that, although the FAIRE scores in Dichaete-Dam intervals are not as high as for some other TFs, there is an enrichment of FAIRE accessibility both in the center of Dichaete-Dam intervals and approximately 1.5 kb upstream of the center. The average profiles of FAIRE scores in Dichaete-Dam intervals are complex, reflecting the fact that DamID binding intervals are not necessarily centered around the true binding site, and vary with developmental stage; the highest peak of FAIRE signal in the center of

Dichaete-Dam intervals is present at stage 9. The intervals that are more highly bound by the Dam-only control also have a complex FAIRE signal profile; however, they do not show a peak of accessibility at their center, and their FAIRE scores are lower on average than those in Dichaete-Dam intervals, suggesting that accessibility is more strongly related to functional TF binding. I also found the overlaps between Dichaete-Dam binding intervals and FDR10 FAIRE intervals in each stage. Although only a small percentage of intervals are directly overlapping, the numbers of overlaps in each stage correspond to the FAIRE signal profiles, with the highest number overlaps also present at stage 9. These results suggest that, although Dichaete-Dam binding may not take place in the accessible regions that are most strongly identified by FAIRE-seq, there is a correlation between chromatin accessibility and Dichaete binding.

In the case of the Dichaete-Dam intervals that do overlap with FAIRE accessible regions in *D. pseudoobscura*, I wondered if these intervals were more likely to be uniquely bound in *D. pseudoobscura* or if binding was conserved at orthologous sites in *D. melanogaster*. Since the *D. pseudoobscura* genome has undergone substantial rearrangements since its split from a common ancestor with *D. melanogaster*, it seemed feasible that newly-evolved accessible regions in the *D. pseudoobscura* genome might underpin the evolution of lineage-specific TF binding events. However, I found that very few *D. pseudoobscura* Dichaete-Dam binding intervals located in FAIRE intervals are unique to *D. pseudoobscura*. On the contrary, they are significantly more likely to be conserved in *D. melanogaster*, while Dichaete-Dam binding intervals that are not located in FAIRE intervals are more likely to be uniquely bound (Figure 6.12). While it is still possible that the uniquely bound Dichaete-Dam intervals within FAIRE accessible regions evolved in tandem with rearrangements of chromatin domains, it appears that selective pressure on functional enhancers may act to maintain both open chromatin and Dichaete binding sites between species of *Drosophila*.

In this chapter and the preceding ones, I have presented the major datasets that I generated during my Ph.D., which consist of DamID binding datasets for Dichaete in four species of *Drosophila* and SoxNeuro in two species of *Drosophila* as well as FAIRE-seq datasets for five developmental stages in *D. pseudoobscura*. Transcription factor binding and chromatin accessibility have been shown to be highly correlated in *D. melanogaster*, with chromatin accessibility highlighted as

a potential driver of TF binding patterns (Kaplan *et al.*, 2011; Li *et al.*, 2011). While several comparative studies of TF binding have been performed in various *Drosophila* species, chromatin accessibility in non-model species has not previously been examined. Although my FAIRE-seq samples may have suffered from some technical problems resulting in the identification of significantly fewer peaks than for similar experiments in *D. melanogaster*, the biological replicates show extremely high reproducibility, suggesting that the peaks that were identified represent true open chromatin. Combining a comparative study of TF binding and chromatin accessibility allowed me to discover the fact that *D. pseudobscura* Dichaete-Dam binding intervals located in open chromatin are significantly more likely to be conserved in *D. melanogaster* compared to those that are not located in open chromatin, which supports the functional relationship between chromatin accessibility and TF binding. In the following chapter, I will discuss the conclusions that can be drawn from all of these datasets in the context of the ongoing debate over what constitutes functional regulatory DNA, as well as presenting my vision for future directions.

CHAPTER 7

Discussion and Future Directions

7.1 Regulatory function and evolution

In their rebuttal to the conclusions of the ENCODE consortium, Graur and colleagues write that "[f]rom an evolutionary viewpoint, a function can be assigned to a DNA sequence if and only if it is possible to destroy it ... Unless a genomic functionality is actively protected by selection, it will accumulate deleterious mutations and will cease to be functional (Graur *et al.*, 2013)." According to this definition of function, a transcription factor binding site and, by extension, a TF binding event, is functional not simply because it occurs but if it has a result that can be altered or broken by its loss. The classical, and most stringent, way to detect such functional binding events is to combine genome-wide studies of *in vivo* binding patterns with gene expression data in a mutant background to detect genes that are both bound by a TF and change expression levels upon its loss or overexpression. This approach has yielded fruitful results in the past with both *Dichaete* and *SoxN* in *D. melanogaster* (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Shen *et al.*, 2013). However, it can also be an overly conservative strategy,

since not all TF functions result in a direct change in expression of the nearest gene, particularly for TFs like Sox proteins that can bend DNA and potentially alter the local chromatin environment (Bowles *et al.*, 2000; Ferrari *et al.*, 1992; Giese *et al.*, 1992; Russell *et al.*, 1996). Additionally, the effects of the loss of one particular binding site can be masked by robustness from secondary shadow enhancers or other members of the regulatory network, particularly in the relatively stress-free lab environment (Aldana *et al.*, 2007; Ciliberti *et al.*, 2007; Ludwig *et al.*, 2011; Perry *et al.*, 2010). In this thesis, I have attempted to focus on the second part of Graur *et al.*'s definition, using the conservation of TF binding during evolution as a filter through which to refine our understanding of group B Sox function in *Drosophila*. In this final chapter, I will review the major findings from my analysis, present a model for SoxN and Dichaete binding that arises from the evolutionary patterns I have observed, and speculate on the origin of both redundancy and neofunctionalization between Sox genes in the vertebrate and invertebrate phylogenies.

7.2 Major conclusions of experimental results

As outlined in the introduction, I set out to study the conservation of group B Sox function on several functional levels, ranging from the DNA sequence of target regions to expression patterns and overall phenotypic effects. Starting from the highest level, I found that the roles of Dichaete and SoxN within the fly developmental regulatory network do not appear to have diverged significantly during the evolution of the *Drosophila* species examined. The gene targets and genomic annotations associated with Dichaete and SoxN function are largely conserved between *D. melanogaster, D. simulans, D. yakuba* and *D. pseudoobscura*, which is not surprising given the high degree of sequence similarity between the orthologous proteins in each species and their equivalent expression patterns during embryonic development. However, for both transcription factors, a comparison of *in vivo* binding patterns revealed turnover of binding sites at gene loci as well as quantitative divergence in binding affinity between species. In the case of Dichaete, for which binding was compared between four species, the proportion of *D. melanogaster* binding intervals that are not conserved in each other species

increases with phylogenetic distance. This observation is in line with previous comparative studies of other transcription factors in *Drosophila* and suggests that, as with other DNA binding proteins, the evolution of group B Sox binding may follow a molecular clock mechanism (Bradley *et al.*, 2010; He *et al.*, 2011b; Paris *et al.*, 2013). The range of binding divergence at the evolutionary scale studied, which is less than that between vertebrate species compared in similar studies with other TFs (Odom *et al.*, 2007; Schmidt *et al.*, 2010; Stefflova *et al.*, 2013; Villar *et al.*, 2014), enabled me to identify patterns of increased conservation compared to the background rate at certain functional categories of binding interval.

As expected, group B Sox binding is highly conserved at sites that are most likely to be involved in functional gene regulation, including known enhancers from the REDFly and FlyLight databases (Gallo *et al.*, 2010; Manning *et al.*, 2012), Dichaete and SoxN direct target genes, and the Dichaete and SoxN core binding intervals believed to represent very high confidence *in vivo* binding locations (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). These findings validate the hypothesis that a signature of selective constraint in the form of increased conservation can be found at functional sites. They also confirm that the Dichaete and SoxN binding events identified in multiple *in vivo* genome-wide studies are functionally important, whether through direct transcriptional regulation or an indirect architectural role (Russell *et al.*, 1996). Interestingly, binding at core intervals was shown to be more highly conserved than binding at direct target genes, suggesting that binding site turnover can occur even at direct targets. Integrating the FAIRE-seq chromatin accessibility data with the DamID-seq data reveals that not only is group B Sox binding associated with open chromatin in multiple species of *Drosophila*, binding in accessible chromatin is more likely to be conserved between species. This relationship likely reflects a feedback loop whereby chromatin accessibility patterns direct transcription factor binding and selection on functionally bound enhancer elements works to maintain open chromatin.

One of the primary questions of my work was whether common binding by Dichaete and SoxN is conserved to the same extent as specific binding by each protein at unique targets. A comparative analysis of DamID for both TFs in *D. melanogaster* and *D. simulans* revealed that, in fact, common binding is much

more likely to be conserved than unique binding by either Sox protein. This is true from the perspective of *D. melanogaster* binding intervals that are conserved in *D. simulans* as well as *vice versa*. Such a high rate of conservation of common binding strongly suggests that the ability of Dichaete and SoxN to bind to and regulate a set of common targets and to compensate for each other at those targets is an important aspect of their biological function. The targets of commonly bound, conserved binding intervals reflect the known functions of group B Sox proteins in the developing central nervous system. Target genes are primarily upregulated in the CNS, and they are enriched for Gene Ontology terms related to biological regulation, morphogenesis, and the specification and differentiation of neurons. They also include targets where Dichaete and SoxN have previously been shown to demonstrate compensation, such as the homeodomain DV-patterning genes *ind* and *vnd*, as well as targets where Dichaete and SoxN appear to have opposite regulatory effects, such as *ac* and *l'sc*, both proneural genes, or *pros*, a TF involved in neuroblast differentiation (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Overton *et al.*, 2002).

D. melanogaster and *D. simulans* also share smaller numbers of conserved binding intervals that are uniquely bound by either Dichaete or SoxN. The primary difference in the target genes annotated to these intervals is in their expression profiles. Uniquely bound Dichaete targets show expression in a broader range of tissues, including the brain and hindgut, where Dichaete is known to play a role (Sánchez-Soriano and Russell, 2000), while uniquely bound SoxN targets show strong upregulation only in the developing CNS. Conserved binding regions unique to Dichaete also contain a highly enriched motif for Byn, a transcription factor that is necessary for hindgut development, which may represent a new physical or genetic interaction specific to Dichaete in the hindgut (Kispert *et al.*, 1994; Murakami *et al.*, 1999). Although unique SoxN targets have a similar expression profile as common targets, other features of these targets, including their enrichment in the Robo-Slit signalling pathway and the presence of an enriched Usp motif in binding intervals, indicate that they may play important and unique roles in axon guidance. This confirms the functional importance of a number of previously discovered SoxN targets involved in later stages of neuronal differentiation as well as the large overlap observed between SoxN targets and fly orthologues of targets of mouse Sox11, a group C Sox protein primarily expressed

in differentiated neurons (Bergsland *et al.*, 2011; Ferrero *et al.*, 2014). These features of unique Dichaete and SoxN binding are more clearly apparent when analyzing data from two species, rather than from *D. melanogaster* alone. In the initial comparison of Dichaete and SoxN binding in *D. melanogaster* and *D. simulans*, it appeared that the two TFs had more differentiated binding patterns in *D. simulans*. Using evolutionary conservation as a filter may have reduced the effect of noise in these datasets, allowing me to home in on the truly unique functions of each protein.

One interesting effect of the use of a quantitative analysis of binding differences between TFs is that it allowed me to identify both a subset of genes that are uniquely bound by Dichaete or SoxN in multiple species and a subset of genes that are preferentially bound by each TF. These preferential targets show binding at the same regions of regulatory DNA by both Dichaete and SoxN across species, but they are consistently bound by one protein at a higher affinity than by the other. Many of these preferentially bound targets are identified as common targets of Dichaete and SoxN in a qualitative analysis of binding. Although for both Dichaete and SoxN, the preferential targets have similar expression profiles as the unique targets, the lists of preferentially and uniquely bound genes show relatively low overlap (76 genes for SoxN and 169 for Dichaete). The preferentially bound genes in each case may highlight binding sites where the regulatory function of Dichaete and SoxN has diverged, but their HMG domains remain similar enough that they can both recognize and bind to the same DNA sequences, particularly under the conditions of DamID, when both proteins are expressed uniformly at comparable levels. Preferential targets include genes whose regulation has been shown to be important for Dichaete and SoxN function, including *pros* in the case of Dichaete and *ase, ind* and *vnd* in the case of SoxN (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Overton *et al.*, 2002). These may also represent cases where Dichaete and SoxN can compensate for one another to increase the robustness of key regulatory networks.

A sequence-based analysis of the Sox motifs found in each set of DamID binding intervals revealed some subtle differences in the binding motifs preferred by each TF, primarily at position six of the consensus A/T A/T CAAAG motif. Previous studies have indicated that this nucleotide is more likely to be a thymine residue in Dichaete core intervals and an adenine residue in SoxN core intervals; however,

it was not known whether this difference reflected any underlying differences in the structures of the two proteins (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). The DamID approach employed here used the same fusion proteins, derived from the Dichaete and SoxN sequences in *D. melanogaster*, to assess binding in all species, meaning that binding differences due to evolutionary changes in orthologous proteins could not be detected. However, multiple alignments of the amino acid sequences show that there is a higher degree of sequence conservation, including a perfectly conserved HMG box domain, between each set of orthologous proteins than between Dichaete and SoxN in any one species. Consequently, it should be easier to detect potential differences in motif preference between paralogues than between orthologues. The fact that very similar differences in the Sox motifs for Dichaete and SoxN were found independently in each genome studied indicates that these sequence preferences are likely to be real and may reflect differences in the preferred binding modes of each protein.

Considering all variants of the Sox motif detected, intervals which show conserved binding in all four species contain more motifs on average than those that are only bound in one species; these motifs are also more highly conserved at both the nucleotide level and at the level of positional organization within regulatory regions. It should be noted that I did not perform any classical tests for selection on either the binding interval or Sox motif sequences. This is partly because, although methods such as the McDonald-Kreitman test have been adapted for use with non-coding DNA, it is difficult to establish an appropriate neutral reference against which to test for selection in putative functional sites (Zhen and Andolfatto, 2012). Testing for selection in entire enhancers is difficult because, while high-confidence TF binding sites may be identified, it is often unknown whether the rest of the sequence is functional or not. Detecting selection at specific motifs or binding sites is more feasible, and alternative methods have been proposed to do so (Moses, 2009); however, such tests still rely on the presence of substitutions and polymorphism, which were not found in many of the Sox motifs that I uncovered. Although I was unable to detect an effect of motif quality on quantitative binding affinity, the finding that Sox motifs in intervals that show binding conservation also show increased rates of conservation provides a link between group B Sox function and DNA sequence evolution, as well as a mechanism through which natural selection can act to maintain functional binding.

7.3 Toward a selection-based model of group B Sox binding

One of the primary findings of this work is the high rate of evolutionary conservation of Dichaete and SoxN binding at sites where both proteins can bind compared to sites where only one protein is bound *in vivo*. The implication of this is that such common binding is an important feature of group B Sox function. Many of the potential target genes annotated to these intervals are known Sox targets in the developing CNS; one hypothesis as to why these binding sites might be preferentially conserved is to confer robustness on cell fate decisions from the specification of the neuroectoderm through to neuroblast differentiation, axonogenesis and gliogenesis (Ferrero *et al.*, 2014; Wagner, 2005, 2008). This is supported by the partial functional redundancy between *Dichaete* and *SoxN* seen on a phenotypic level in single mutant embryos, as well as at certain loci where one protein can substitute for the binding of another in its absence (Ferrero *et al.*, 2014; Overton *et al.*, 2002). However, increased conservation is also seen at binding sites where Dichaete and SoxN have opposite regulatory functions, several of which are also critical for determining neuroblast fate (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Overton *et al.*, 2002). Why would natural selection preferentially maintain binding of factors with antagonistic functions at the same sites?

It is possible that in some situations, Dichaete and SoxN might directly compete with one another for binding. Although this has not been demonstrated, it is a feasible mechanism for establishing a balance between the up- or down-regulation of genes promoting neuroblast differentiation, for example, with the ultimate outcome dependent on the relative concentration of each TF within a cell. All of the group B Sox proteins have very similar HMG domains and can recognize similar consensus DNA sequences (McKimmie *et al.*, 2005), despite the discovery in this study of some possible differences in motif preference between Dichaete and SoxN. Another view for the explanation of common binding is simply that it is easier for natural selection to maintain Sox motifs in enhancers that can be bound by both TFs than to maintain a suite of slightly different motifs for each. Analogous to a gene duplication event, one might expect that a newly originated TF binding site

would often experience low selective pressure and quickly accumulate mutations, resulting in the maintenance of a minimal complement of sites. Such a mechanism could be self-reinforcing, as sites that are functionally bound by multiple TFs would experience a higher dose of selective constraint since mutations that disrupted binding would perturb the regulatory networks associated with both TFs. This could also explain the fact that group B Sox proteins have primarily diversified in regions other than their DNA-binding domains during evolution; strong selection on common binding sites would encourage the acquisition of new functions through interactions with specific binding partners or changes in the ability to modify the local chromatin environment.

Given the observed selective constraint on commonly bound sites, what is the explanation for the presence of highly conserved sites that are uniquely bound by either SoxN or Dichaete? The fact that target genes at these sites have different spatial expression profiles suggests a model whereby the different expression patterns of Dichaete and SoxN themselves, along with extrinsic factors in the nuclear environment, may shape the unique functions of these two TFs. Although Dichaete and SoxN expression patterns overlap to a great extent in the CNS, they are not identical; Dichaete is expressed uniquely in the midline, brain and hindgut, for example, while SoxN is expressed uniquely in the lateral column of delaminating neuroblasts and shows specific patterns of expression in the epidermis at later stages of development (Crémazy *et al.*, 2000; Overton *et al.*, 2007; Sánchez-Soriano and Russell, 2000, 1998). The chromatin landscape has been shown to differ between different tissues in the *Drosophila* embryo as well as over the course of development, both in terms of general accessibility and specific activating or repressing histone marks (Bonn *et al.*, 2012; McKay and Lieb, 2013). It is therefore likely that certain enhancers are only available to be bound in the tissues where Dichaete and SoxN are expressed uniquely, preventing common binding from ever being observed. A comparative analysis of chromatin accessibility and histone marks between the hindgut and the CNS would be a fascinating way to test this hypothesis with regard to unique and common binding by Dichaete in these tissues.

Another possible factor that could explain the unique, conserved binding patterns of Dichaete and SoxN in different tissues is the tissue-specific presence of certain cofactors. Sox proteins often bind to DNA as heterodimers with other TFs

(Ambrosetti *et al.*, 1997; Archer *et al.*, 2011; Bery *et al.*, 2013; Bonneaud *et al.*, 2003); Dichaete has previously been demonstrated to bind together with Vvl in the midline (Ma *et al.*, 2000; Sánchez-Soriano and Russell, 1998). As discussed in the introduction, although paralogous Hox proteins generally show greater specificity in their gene targets than Sox, it has been suggested that much of this specificity may arise from interactions with binding partners (Chan *et al.*, 1997; Mann *et al.*, 2009; Slattery *et al.*, 2011). The sequences of *Drosophila* group B Sox genes have diverged much more outside of their HMG domains than within them, although sections of the C-terminal regions of both Dichaete and SoxN still show good levels of conservation between fly species, suggesting that a major driver of their evolutionary diversification may have been the acquisition of new cofactors (McKimmie *et al.*, 2005). Although changes in target specificity due to binding with cofactors has not been demonstrated for Sox proteins in *Drosophila*, this is a feasible mechanism behind the specific binding of Dichaete and SoxN in different embryonic tissues. The identification of enriched motifs in Dichaete- and SoxN-specific binding intervals that are not present in common binding intervals, corresponding to TFs such as Byn in the case of Dichaete and Usp in the case of SoxN, may represent tissue-specific cofactors of these proteins, although physical interactions remain to be demonstrated.

The patterns of conservation of Dichaete and SoxN binding in *Drosophila* suggest a model whereby, despite slight differences in the consensus motifs bound by each protein, both group B Sox proteins can and do bind a majority of their sites in common in tissues where they are both expressed. These common binding sites are preferentially maintained during evolution in comparison to uniquely-bound sites, whether to increase the robustness of the regulatory networks shared by Dichaete and SoxN or due to the effect of selection favoring the re-use of binding sites in a dense, compact genome. At the same time, unique binding by Dichaete and SoxN is conserved at specific target genes that have largely different expression profiles, possibly reflecting the effect of tissue-specific chromatin landscapes or cofactor availability. These observations are supported by the fact that the majority of the sequence differences between Dichaete and SoxN can be found outside of their HMG domains, in protein domains that may be involved in interactions with other TFs as well as in their own regulatory regions, which determine the overlapping and unique expression patterns of each TF.

7.4 Implications for the evolution of Sox function and redundancy

Sox genes encode an ancient family of transcription factors that, despite numerous gene duplication events, continue to show functional redundancy between members of the same subgroups throughout their phylogeny (Bhattaram *et al.*, 2010; Ferri, 2004; Guth and Wegner, 2008; Matsui, 2006; Nishiguchi *et al.*, 1998; Okuda *et al.*, 2010; Overton *et al.*, 2002; Rizzoti *et al.*, 2004; Uchikawa *et al.*, 2011; Uwanogho *et al.*, 1995; Wegner and Stolt, 2005; Wood and Episkopou, 1999). While common binding patterns between two TFs do not necessarily imply redundancy or compensation, they are, if not required for it, then likely to facilitate it. Indeed, SoxN and Dichaete have a complex relationship in the fly embryo that includes functional compensation as well as interdependence and binding at some loci with opposite regulatory effects (Ferrero *et al.*, 2014; Overton *et al.*, 2002). Previous *in vivo* binding studies as well as studies of gene expression changes in mutants have identified large numbers of genes bound in common by Dichaete and SoxN (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). This study shows that such common binding has been conserved during the evolution of the drosophilids at a rate higher than that of unique binding by either protein, suggesting that it is a key feature of group B Sox function. It has been speculated that robustness may arise as a general property of complex gene regulatory networks, without any direct selective pressure (Aldana *et al.*, 2007). My results show that, in the case of *Drosophila* group B Sox genes, partially redundant binding patterns that can lead to increased robustness may also be specifically maintained by natural selection.

Given the apparently independent evolutionary trajectories of group B *Sox* genes in insects and mammals and the fact that it is difficult to assign direct orthology between individual members of each class, an obvious question is whether the partial redundancy seen among group B *Sox* genes in mammals is a shared ancestral feature or the result of convergent evolution (Bergsland *et al.*, 2011; Ferri, 2004; Nishiguchi *et al.*, 1998; Okuda *et al.*, 2010; Rizzoti *et al.*, 2004). Although the evolutionary models proposed by McKimmie *et al.* and Zhong *et al.* differ, they both suggest that at least one tandem duplication event occurred

at the ancestral group B *Sox* locus before the protostome/deuterostome split (Figure 7.1) (McKimmie *et al.*, 2005; Zhong *et al.*, 2011). According to Zhong and colleagues, this duplication gave rise to the proto-B1 and -B2 genes, which then expanded in the vertebrates through whole-genome duplications and in the arthropods through further tandem duplications (Zhong *et al.*, 2011). This model can account for the divergent functions seen in vertebrate group B1 and B2 genes (Uchikawa *et al.*, 1999); however, it does not explain the ability of both *Dichaete* and *SoxN* to play B1-like and B2-like roles. If the pattern seen in the vertebrate *Hox* gene expansions, in which *trans*-paralogues arising from genome duplications show greater functional similarity than co-linear *cis*-paralogues, holds any general applicability, then the McKimmie *et al.* model may be more consistent with functional data. In this model, *Dichaete* and *SoxN* ancestors arose via whole-genome duplication followed by a tandem duplication to create the *Sox21a*/B2-like ancestor, both prior to the protostome/deuterostome split. In the arthropod lineage, a further tandem duplication led to the origin of *Sox21b*, while in the vertebrate lineage, another genome duplication event filled out the complement of group B1 and group B2 genes (McKimmie *et al.*, 2005). If this is the case, then redundancy between the first group B paralogues resulting from a genome duplication may have been partially retained throughout evolution, while later paralogues split into group B1 and B2 functions in vertebrates or acquired partial neofunctionalizations in insects.

Such a model suggests that functional redundancy between group B *Sox* genes, particularly in the developing CNS, is a truly ancestral feature that has been refined and elaborated upon separately in different lineages. In vertebrates, multiple genome duplicates have given rise to a larger complement of *Sox* genes, which have apparently undergone a greater degree of subfunctionalization and neofunctionalization, while still retaining overlapping expression patterns and some degree of functional compensation. This particularly appears to be true of group B *Sox* genes, whose function in the CNS can be split both by temporal succession (Bergsland *et al.*, 2011) and by activator/repressor roles (Uchikawa *et al.*, 1999). While it is somewhat surprising that substantial functions in the CNS have not been discovered for *Sox21a* and *Sox21b* in *Drosophila*, this underscores the observation that, in insects, *Dichaete* and *SoxN* appear to direct virtually all aspects of neurogenesis and are the only *Sox* genes to do so (Ferrero *et al.*, 2014).

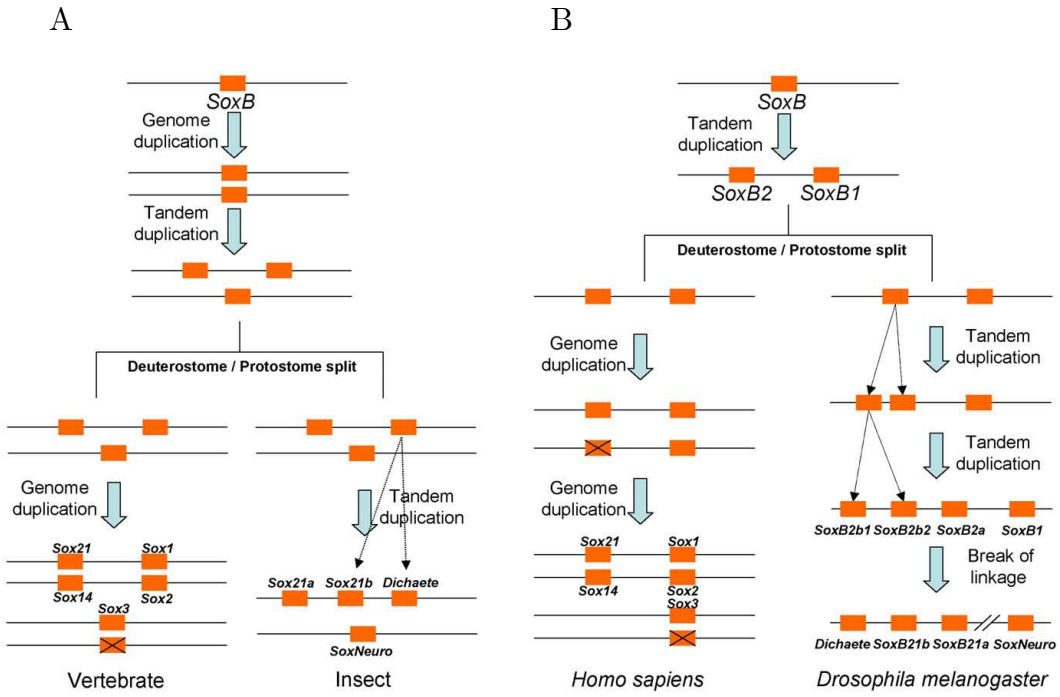


Figure 7.1: Two models of the evolution of group B *Sox* genes in vertebrates and insects. A.) The model proposed by McKimmie *et al.* In this model, an ancestral group B *Sox* gene gave rise to the *Dichaete* and *SoxN* ancestors via a whole-genome duplication. The *Sox21a* ancestor then arose through a tandem duplication before the protostome/deuterostome split. Finally, a further tandem duplication generated *Sox21b* in the insect lineage, while another whole genome duplication led to the origin of the remaining group B *Sox* genes in vertebrates. B.) The model proposed by Zhong *et al.* In this model, a single tandem duplication before the protostome/deuterstome split gave rise to the ancestral *SoxB1* and *SoxB2* genes. These then underwent two rounds of whole genome duplications in vertebrates to generate the full complement of group B *Sox* genes, while in insects two further tandem duplications led to the origin of *Sox21a* and *Sox21b*. Figure reproduced from Zhong *et al.* (2011).

Although *Dichaete* and *SoxN* have undoubtedly undergone partial neofunctionalization, which is reflected both in their expression patterns and in their unique binding targets, they have maintained a close and complex relationship comprising aspects of interdependence, antagonistic regulatory effects and compensation. If the integrated action of multiple Sox proteins, whether as opposing factors or to provide additional robustness, is a feature of CNS development that has been consistently selected for, it is possible that the presence of additional group B Sox proteins in vertebrates has led to a relaxation of this selective pressure and allowed them to specialize to a greater degree. Although unique binding by Dichaete and SoxN in *Drosophila* is less conserved than common binding, it still occurs at numerous loci throughout the genome. These binding events, while not necessarily functional, may represent opportunities for further neofunctionalization through an unconstrained exploration of the regulatory landscape.

7.5 Future work

Although this thesis has shed some light on the conserved functional relationship between *Dichaete* and *SoxN* in several species of *Drosophila*, the complex functions of group B *Sox* genes in invertebrates are far from completely understood. A number of experimental approaches could help to validate the model proposed in this thesis. The major drawbacks of DamID include its lack of tissue specificity and the fact that it does not measure TF binding in its native context, but rather by using a transgenic fusion protein expressed in addition to the endogenous protein. Using ChIP as a complementary technique can help reduce these problems; since each technique is subject to different sources of bias, a binding dataset derived from intersecting the two will be much more stringent than using either technique alone (Aleksic *et al.*, 2013). However, given the lack of success in performing ChIP using current antibodies for Dichaete and SoxN, this does not appear to be the most promising avenue for further research, unless new and more reliable antibodies for group B Sox proteins in insects can be derived. Fortunately, a targeted DamID technique (TaDa) has recently become available, which enables the measurement of binding in specific tissue or cell types (Southall *et al.*, 2013). Using TaDa to dissect Dichaete and SoxN binding patterns in tis-

sues where they are commonly expressed, such as the medial and intermediate columns of neuroblasts, versus tissues where only one is present, such as the hindgut or midline, would be a useful follow-up both to identify tissue-specific enhancers and target genes and to test whether unique binding is indeed primarily driven by tissue-specific factors. This could also provide *in vivo* binding data with a greater temporal resolution, as the expression patterns of Dichaete and SoxN change throughout developmental time, which could then be correlated with the detailed time course of gene expression changes in *Dichaete* and *SoxN* mutant backgrounds that is already available (Ferrero *et al.*, 2014).

The proposed sources of tissue-specific binding, namely chromatin accessibility and the presence of specific cofactors, could also be tested. The two primary techniques for assessing chromatin accessibility, DNase-seq and FAIRE-seq, can both feasibly be applied in dissected tissues, although FAIRE-seq has been more effective in this regard in *Drosophila* because embryos or larvae can be fixed before dissection, greatly facilitating the process of collecting material (McKay and Lieb, 2013). For tissues that cannot easily be dissected, BiTS-ChIP, a technique involving fluorescently sorting fixed nuclei that are tagged with a cell-type specific marker (Bonn *et al.*, 2012), could be used to study chromatin accessibility either in combination with FAIRE or by performing ChIP for a general marker of transcriptional activity such as RNA polymerase II (Pol II). Similarly, performing TaDa with a Pol II-Dam fusion protein can also yield data on cell-type specific chromatin landscapes (Southall *et al.*, 2013). The use of these techniques in tissues where Dichaete and SoxN are commonly or uniquely expressed would enable the discovery of regulatory regions that are only accessible in certain tissues, which could then be correlated with binding patterns. In order to determine whether tissue-specific cofactors can direct Dichaete and SoxN binding, *in vitro* methods such as co-immunoprecipitation could be used to test for physical interactions between group B Sox proteins and potential cofactors. *In vivo*, the dependence of group B Sox binding on candidate cofactors could be tested by measuring Dichaete or SoxN binding in a mutant background for the cofactor of interest. If performed in a tissue-specific manner, this experiment could yield convincing data either in support of or against a model whereby Dichaete and SoxN have acquired unique binding sites through interactions with other TFs that are only present in a subset of their spatial expression domains.

On the level of individual targets, a large number of instances of putative binding site turnover events have been identified for Dichaete and SoxN, where non-orthlogous regulatory regions for the same gene are bound *in vivo* in different species of *Drosophila*. It is hypothesized that such turnover is subject to constraint such that, in the absence of a gain of function driven by positive selection, the overall level of gene expression should be buffered (He *et al.*, 2011a; Spivakov *et al.*, 2012). This hypothesis could be tested by performing reporter assays in transgenic lines of *D. melanogaster* carrying putative enhancer sequences from other species (Hare *et al.*, 2008). Although a staining-based assay would not provide a quantitative measure of gene expression, it would allow for the detection of any differences in spatial expression patterns driven by species-specific enhancer elements.

In order to further refine our understanding of group B *Sox* function and evolution, it would be useful to expand the work done here into species more distant from *D. melanogaster*. Such a project is currently underway in the red flour beetle, *Tribolium castaneum*. It is not currently known whether the fifth group B *Sox* gene present in *Tribolium*, *SoxB3*, is functional or represents a pseudogene; its expression pattern has not yet been determined. If it is functional and expressed in the developing CNS, it could add yet another layer of complexity and potential compensation to the functional roles of group B *Sox* genes. However, sequence analysis suggests that, if *Tribolium SoxB3* is functional, it may have diverged sufficiently from its paralogues to have acquired a new, independent function. If so, then it would represent an exceptional case of neofunctionalization in the insect *Sox* clade. Although it becomes progressively more difficult to align genomes as the phylogenetic distance between two species being compared increases, complicating the assignment of orthology to putative enhancer regions and binding events, such a comparison has the potential to reveal stronger selective effects and more deeply conserved features of TF binding. The use of a more distant *Drosophila* species or another non-*Drosophila* dipteran whose genome is available, such as the scuttle fly *Megaselia*, would also be very useful in this regard, as it would enable a comparison of binding patterns in the context of greater sequence divergence but highly conserved mechanisms of embryonic patterning and development (Hare *et al.*, 2008).

Other experiments that would help to progress this work include a more detailed dissection of Dichaete and SoxN binding sites as well as an exploration of other factors that interact with Dichaete and SoxN in the *Drosophila* transcriptional regulatory network. Both DamID and conventional ChIP-seq have sufficient resolution to identify binding events on the scale of a few hundred base pairs, but as described in this thesis, those binding intervals often contain multiple matches to a TF's consensus binding site. Particularly for DamID, it is difficult to identify the actual DNA sequence to which the TF is bound *in vivo*. Even with ChIP-seq data, it can be difficult to distinguish between a single bound motif and multiple, closely-spaced bound motifs. ChIP-exo, in which ChIP DNA is treated with an exonuclease to digest away non-bound nucleotides before sequencing, is a technique that can help overcome these limitations and identify bound sites with very high resolution (Bardet *et al.*, 2013; Rhee and Pugh, 2011). Although it is also antibody-dependent, it would be very interesting to perform ChIP-exo for Dichaete and SoxN in multiple species of *Drosophila*, as it would enable a much more detailed comparison of the binding sites preferred by each protein and the evolutionary forces to which they are subject.

In addition to identifying new, unique cofactors for Dichaete and SoxN, it would also be informative to study factors with which they have already been suggested to interact in an evolutionary context. A comparative analysis of Dichaete binding intervals and binding profiles of 33 other TFs in *D. melanogaster* identified seven TFs whose profiles significantly overlapped with that of Dichaete. Four of these, Senseless (Sens), Prospero (Pros), Hunchback (Hb), and Kruppel (Kr), are known to be involved in CNS development (Aleksic *et al.*, 2013). Hb and Kr are the first two transcription factors expressed in a temporal series in embryonic neuroblasts as they differentiate into ganglion mother cells (GMCs), during the time that both Dichaete and SoxN are expressed in the developing neuroectoderm (Buescher *et al.*, 2002; Maurange and Gould, 2005; Overton *et al.*, 2002). Since *hb* and *Kr* were found to be targets of both Dichaete and SoxN in this and previous studies (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014), they may form part of a feed-forward loop with the group B *Sox* genes in the genetic regulatory network specifying neuroblast fate. Assessing the conservation of overlaps between Dichaete, SoxN, Hb and Kr binding patterns in multiple species of *Drosophila* could help clarify the targets that are commonly regulated by these factors.

Finally, in order to paint a complete picture of the evolution of insect *Sox* genes and their roles in development, it will be necessary to address the functions of the remaining two group B *Sox* genes, *Sox21a* and *Sox21b*. While *Sox21a* is expressed in the midline as well as the anlage of the foregut and hindgut, *Sox21b* is excluded from the CNS but is expressed in the ventral epidermis and the hindgut, where it overlaps with *Dichaete* expression (Crémazy *et al.*, 2001; McKimmie *et al.*, 2005; Phochanukul and Russell, 2010). Surprisingly, deletions of either of these genes individually or both together produce no observable phenotype in *D. melanogaster*. However, they show conservation at both a sequence level and in terms of genomic location across the insects, suggesting that they do provide some functionality (McKimmie *et al.*, 2005). Perhaps their role is largely limited to increasing the robustness of the *SoxN*- and *Dichaete*-driven regulatory networks in specific cell types, although it would be surprising if they had no independent functions and yet remained conserved. Additionally, the only other *Sox* gene known to be expressed in the developing CNS in *Drosophila* is *Sox102F*, the only insect group D *Sox* gene (Crémazy *et al.*, 2001; Phochanukul and Russell, 2010). Although RNAi-mediated knockdown of *Sox102F* results in severe CNS disruptions, its gene targets are not known (Phochanukul and Russell, 2010). Genome-wide *in vivo* binding studies of these three Sox proteins would help to fill in the gaps in our current knowledge of insect *Sox* biology and possibly provide new data on functional compensation by Sox proteins in the developing CNS.

7.6 Conclusions

Genetic redundancy is a curious phenomenon because it appears to violate the rule that a biological function exists if and only if it can be broken (Graur *et al.*, 2013). Indeed, redundancy among *Sox* genes was first described by researchers who were no doubt frustrated by observing that single mutants generated in model animals appeared phenotypically normal. Nonetheless, it has been observed throughout the *Sox* family tree, in multiple subgroups and species (Bhattaram *et al.*, 2010; Ferri, 2004; Matsui, 2006; Nishiguchi *et al.*, 1998; Okuda *et al.*, 2010; Rizzoti *et al.*, 2004). The hypothesis that redundancy can confer robustness on a genetic regulatory network suggests that redundancy itself may be a func-

tion that can be broken (Nowak *et al.*, 1997; Tautz, 1992; Wagner, 2005, 2008). While it does appear that compensation between Dichaete and SoxN may lend robustness to the developing CNS, their relationship is clearly much more complex than one of simple redundancy (Ferrero *et al.*, 2014). Evolutionary comparisons reveal that this relationship is conserved across species and that the functions of Dichaete and SoxN binding are intimately tied to one another. I hope that through this work I have demonstrated the value of studying transcription factor binding patterns through the lens of natural selection while refining the current model of the common and unique functions of group B *Sox* genes in *Drosophila* development.

List of appendices

The following datasets are included as appendices in the attached CD:

A. Genes annotated to DamID binding datasets

1. Genes bound by Dichaete-Dam in *D. melanogaster*
2. Genes bound by SoxN-Dam in *D. melanogaster*
3. Genes bound by Dichaete-Dam in *D. simulans*
4. Genes bound by SoxN-Dam in *D. simulans*
5. Genes bound by Dichaete-Dam in *D. yakuba*
6. Genes bound by Dichaete-Dam in *D. pseudoobscura*

B. GO:BP terms enriched in bound genes

1. GO:BP terms enriched in Dichaete target genes in *D. melanogaster*
2. GO:BP terms enriched in SoxN target genes in *D. melanogaster*
3. GO:BP terms enriched in Dichaete target genes in *D. simulans*
4. GO:BP terms enriched in SoxN target genes in *D. simulans*
5. GO:BP terms enriched in Dichaete target genes in *D. yakuba*
6. GO:BP terms enriched in Dichaete target genes in *D. pseudoobscura*

C. Genes annotated to commonly-bound, conserved binding intervals in *D. melanogaster* and *D. simulans*

D. GO:BP terms enriched in commonly-bound, conserved genes

E. Genes annotated to unique Dichaete-Dam binding intervals conserved in *D. melanogaster* and *D. simulans*

F. GO:BP terms enriched in Dichaete-unique, conserved target genes

G. Genes annotated to unique SoxN-Dam binding intervals conserved in *D. melanogaster* and *D. simulans*

H. GO:BP terms enriched in SoxN-unique, conserved target genes

Bibliography

Adams, M.D. The genome sequence of *Drosophila melanogaster*. *Science*, 287(5461):2185–2195, March 2000. ISSN 00368075, 10959203. doi: 10.1126/science.287.5461.2185. URL http://www.sciencemag.org/cgi/doi/10.1126/science.287.5461.2185.

Akiyama, H., Chaboissier, M.C., Martin, J.F., Schedl, A. and de Crombrugghe, B. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes & development*, 16(21):2813–2828, 2002. URL http://genesdev.cshlp.org/content/16/21/2813.short.

Aldana, M., Balleza, E., Kauffman, S. and Resendiz, O. Robustness and evolvability in genetic regulatory networks. *Journal of Theoretical Biology*, 245(3):433–448, April 2007. ISSN 00225193. doi: 10.1016/j.jtbi.2006.10.027. URL http://linkinghub.elsevier.com/retrieve/pii/S0022519306005170.

Aleksic, J. and Russell, S. ChIPing away at the genome: the new frontier travel guide. *Molecular BioSystems*, 5(12):1421, 2009. ISSN 1742-206X, 1742-2051. doi: 10.1039/b906179g. URL http://xlink.rsc.org/?DOI=b906179g.

Aleksic, J., Ferrero, E., Fischer, B., Shen, S.P. and Russell, S. The role of Dichaete in transcriptional regulation during *Drosophila* embryonic development. *BMC genomics*, 14(1):861, 2013. URL http://www.biomedcentral.com/1471-2164/14/861.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–

410, 1990. URL http://www.sciencedirect.com/science/article/pii/S0022283605803602.

Ambrosetti, D.C., Basilico, C. and Dailey, L. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein-protein interactions facilitated by a specific spatial arrangement of factor binding sites. *Molecular and cellular biology*, 17(11):6321–6329, 1997. URL http://mcb.asm.org/content/17/11/6321.short.

Archer, T.C., Jin, J. and Casey, E.S. Interaction of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis. *Developmental Biology*, 350(2):429–440, February 2011. ISSN 00121606. doi: 10.1016/j.ydbio.2010.12.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0012160610012546.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryn, L.M., Rath, M. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, January 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1232542. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1232542.

Arnold, C.D., Gerlach, D., Spies, D., Matts, J.A., Sytnikova, Y.A. *et al.* Quantitative genome-wide enhancer activity maps for five *Drosophila* species show functional enhancer conservation and turnover during *cis*-regulatory evolution. *Nature Genetics*, June 2014. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.3009. URL http://www.nature.com/doifinder/10.1038/ng.3009.

Arnoult, L., Su, K.F.Y., Manoel, D., Minervino, C., Magrina, J. *et al.* Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. *Science*, 339(6126):1423–1426, March 2013. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1233749. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1233749.

Bardet, A.F., He, Q., Zeitlinger, J. and Stark, A. A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, 7(1):45–61, December 2011. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2011.420. URL http://www.nature.com/doifinder/10.1038/nprot.2011.420.

Bardet, A.F., Steinmann, J., Bafna, S., Knoblich, J.A., Zeitlinger,

J. *et al.* Identification of transcription factor binding sites from ChIP-seq data at high-resolution. *Bioinformatics*, page btt470, 2013. URL http://bioinformatics.oxfordjournals.org/content/early/2013/08/24/bioinformatics.btt470.short.

Bauer, S., Grossmann, S., Vingron, M. and Robinson, P.N. Ontologizer 2.0–a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, 24(14):1650–1651, May 2008. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btn250. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btn250.

Bergsland, M., Ramskold, D., Zaouter, C., Klum, S., Sandberg, R. *et al.* Sequentially acting Sox transcription factors in neural lineage development. *Genes & Development*, 25(23):2453–2464, December 2011. ISSN 0890-9369. doi: 10.1101/gad.176008.111. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.176008.111.

Bery, A., Martynoga, B., Guillemot, F., Joly, J.S. and Retaux, S. Characterization of enhancers active in the mouse embryonic cerebral cortex suggests Sox/Pou *cis*-regulatory logics and heterogeneity of cortical progenitors. *Cerebral Cortex*, May 2013. ISSN 1047-3211, 1460-2199. doi: 10.1093/cercor/bht126. URL http://www.cercor.oxfordjournals.org/cgi/doi/10.1093/cercor/bht126.

Bhattaram, P., Penzo-Méndez, A., Sock, E., Colmenares, C., Kaneko, K.J. *et al.* Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nature Communications*, 1(1):1–12, April 2010. ISSN 2041-1723. doi: 10.1038/ncomms1008. URL http://www.nature.com/doifinder/10.1038/ncomms1008.

Biggin, M. Animal transcription networks as highly connected, quantitative continua. *Developmental Cell*, 21(4):611–626, October 2011. ISSN 15345807. doi: 10.1016/j.devcel.2011.09.008. URL http://linkinghub.elsevier.com/retrieve/pii/S1534580711004060.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A. *et al.* Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. *Nature Genetics*, 44(2):148–

156, January 2012. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.1064. URL http://www.nature.com/doifinder/10.1038/ng.1064.

Bonneaud, N., Savare, J., Berta, P. and Girard, F. SNCF, a SoxNeuro interacting protein, defines a novel protein family in drosophila melanogaster. *Gene*, 319:33–41, November 2003. ISSN 03781119. doi: 10.1016/S0378-1119(03)00795-9. URL http://linkinghub.elsevier.com/retrieve/pii/S0378111903007959.

Bowles, J., Schepers, G. and Koopman, P. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental Biology*, 227(2):239–255, November 2000. ISSN 00121606. doi: 10.1006/dbio.2000.9883. URL http://linkinghub.elsevier.com/retrieve/pii/S001216060099883X.

Bradley, R.K., Li, X.Y., Trapnell, C., Davidson, S., Pachter, L. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. *PLoS Biology*, 8(3):e1000343, March 2010. ISSN 1545-7885. doi: 10.1371/journal.pbio.1000343. URL http://dx.plos.org/10.1371/journal.pbio.1000343.

Buck, M.J. and Lieb, J.D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, 83(3):349–360, March 2004. ISSN 08887543. doi: 10.1016/j.ygeno.2003.11.004. URL http://linkinghub.elsevier.com/retrieve/pii/S0888754303003628.

Buescher, M., Yeo, S.L., Udolph, G., Zavortink, M., Yang, X. *et al.* Binary sibling neuronal cell fate decisions in the *Drosophila* embryonic central nervous system are nonstochastic and require inscuteable-mediated asymmetry of ganglion mother cells. *Genes & development*, 12(12):1858–1870, 1998. URL http://genesdev.cshlp.org/content/12/12/1858.short.

Buescher, M., Hing, F.S. and Chia, W. Formation of neuroblasts in the embryonic central nervous system of *Drosophila melanogaster* is controlled by SoxNeuro. *Development*, 129(18):4193–4203, 2002. URL http://dev.biologists.org/content/129/18/4193.short.

Burge, C. and Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology*, 268(1):78–94, 1997. URL http://www.sciencedirect.com/science/article/pii/S0022283697909517.

Cai, Y., Chia, W. and Yang, X. A family of snail-related zinc finger proteins regulates two distinct and parallel mechanisms that mediate *Drosophila* neuroblast asymmetric divisions. *The EMBO journal*, 20(7):1704–1714, April 2001. doi: 10.1093/emboj/20.7.1704.

Celniker, S.E., Wheeler, D.A., Kronmiller, B., Carlson, J.W., Halpern, A. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome biology*, 3(12):research0079, 2002. URL http://genomebiology.com/content/3/12/RESEARCH0079.

Chan, S.K., Ryoo, H.D., Gould, A., Krumlauf, R. and Mann, R.S. Switching the in vivo specificity of a minimal hox-responsive element. *Development*, 124 (10):2007–2014, 1997. URL http://dev.biologists.org/content/124/10/2007.short.

Chenna, R. Multiple sequence alignment with the clustal series of programs. *Nucleic Acids Research*, 31(13):3497–3500, July 2003. ISSN 1362-4962. doi: 10.1093/nar/gkg500. URL http://www.nar.oupjournals.org/cgi/doi/10.1093/nar/gkg500.

Chung, D., Kuan, P.F. and Keles, S. *mosaics: MOSAiCS (MOdel-based one and two Sample Analysis and Inference for ChIP-Seq)*, 2012. URL http://groups.google.com/group/mosaics_user_group. R package version 1.12.0.

Ciliberti, S., Martin, O.C. and Wagner, A. Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences*, 104(34):13591–13596, 2007. URL http://buonmathuot.vn/ws/r/www.pnas.org/content/104/34/13591.full.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B. *et al.* Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*, 450(7167):203–218, November 2007. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature06341. URL http://www.nature.com/doifinder/10.1038/nature06341.

Collignon, J., Sockanathan, S., Hacker, A., Cohen-Tannoudji, M., Norris, D.

et al. A comparison of the properties of Sox-3 with Sry and two related genes, Sox-1 and Sox-2. *Development*, 122(2):509–520, 1996. URL http://dev.biologists.org/content/122/2/509.short.

Contrino, S., Smith, R.N., Butano, D., Carr, A., Hu, F. *et al.* modMine: flexible access to modENCODE data. *Nucleic Acids Research*, 40(D1):D1082–D1088, November 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr921. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr921.

Coyne, J.A., Elwyn, S., Kim, S.Y. and Llopart, A. Genetic studies of two sister species in the *Drosophila melanogaster* subgroup, *D. yakuba* and *D. santomea*. *Genetical Research*, 84(1):11–26, August 2004. ISSN 0016-6723, 1469-5073. doi: 10.1017/S0016672304007013. URL http://www.journals.cambridge.org/abstract_S0016672304007013.

Crémazy, F., Berta, P. and Girard, F. Sox Neuro, a new *Drosophila* Sox gene expressed in the developing central nervous system. *Mechanisms of development*, 93(1):215–219, 2000. URL http://www.sciencedirect.com/science/article/pii/S0925477300002689.

Crémazy, F., Berta, P. and Girard, F. Genome-wide analysis of Sox genes in *Drosophila melanogaster*. *Mechanisms of development*, 109(2):371–375, 2001. URL http://www.sciencedirect.com/science/article/pii/S0925477301005299.

Darbo, E., Herrmann, C., Lecuit, T., Thieffry, D. and van Helden, J. Transcriptional and epigenetic signatures of zygotic genome activation during early *Drosophila* embryogenesis. *BMC genomics*, 14(1):226, 2013. URL http://www.biomedcentral.com/1471-2164/14/226?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Bmc%2FGenomics%2FLatestArticles+(BMC+Genomics+-+Latest+articles).

Dean, E.J., Davis, J.C., Davis, R.W. and Petrov, D.A. Pervasive and persistent redundancy among duplicated genes in yeast. *PLoS Genetics*, 4(7):e1000113, July 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000113. URL http://dx.plos.org/10.1371/journal.pgen.1000113.

Dohm, J.C., Lottaz, C., Borodina, T. and Himmelbauer, H. Substantial biases

in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105–e105, August 2008. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkn425. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkn425.

Downes, M. and Koopman, P. SOX18 and the transcriptional regulation of blood vessel development. *Trends in Cardiovascular Medicine*, 11(8):318–324, 2001. doi: 10.1016/S1050-1738(01)00131-1. URL http://www.tcmonline.org/article/S1050-1738(01)00131-1/abstract.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, September 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11247. URL http://www.nature.com/doifinder/10.1038/nature11247.

Espinosa-Soto, C. and Wagner, A. Specialization can drive the evolution of modularity. *PLoS computational biology*, 6(3):e1000719, 2010. URL http://dx.plos.org/10.1371/journal.pcbi.1000719.

Ferrari, S., Harley, V.R., Pontiggia, A., Goodfellow, P.N., Lovell-Badge, R. *et al.* SRY, like HMG1, recognizes sharp angles in DNA. *The EMBO journal*, 11(12):4497, 1992. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC557025/.

Ferrero, E., Fischer, B. and Russell, S. SoxNeuro orchestrates central nervous system specification and differentiation in *Drosophila* and is only partially redundant with Dichaete. *Genome Biology*, 15(5):R74, 2014. ISSN 1465-6906. doi: 10.1186/gb-2014-15-5-r74. URL http://genomebiology.com/2014/15/5/R74.

Ferri, A.L.M. Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development*, 131(15):3805–3819, June 2004. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.01204. URL http://dev.biologists.org/cgi/doi/10.1242/dev.01204.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D. *et al.* DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. *Proceedings of the National*

Academy of Sciences, 109(52):21330–21335, December 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1209589110. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1209589110.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.l. *et al.* Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151 (4):1531–1545, 1999. URL http://www.genetics.org/content/151/4/1531.short.

Foronda, D., de Navas, L.F., Garaulet, D.L. and Sanchez-Herrero, E. Function and specificity of Hox genes. *The International Journal of Developmental Biology*, 53(8-9-10):1404–1419, 2009. ISSN 0214-6282. doi: 10.1387/ijdb.072462df. URL http://www.intjdevbiol.com/paper.php?doi=072462df.

Frankel, N., Wang, S. and Stern, D.L. Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. *Proceedings of the National Academy of Sciences*, 109(51):20975–20979, November 2012. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1207715109. URL http://www.pnas.org/cgi/doi/10.1073/pnas.1207715109.

Gallo, S.M., Gerrard, D.T., Miner, D., Simich, M., Des Soye, B. *et al.* REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. *Nucleic Acids Research*, 39(Database):D118–D123, October 2010. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkq999. URL http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkq999.

Gerstein, M.B., Bruce, C., Rozowsky, J.S., Zheng, D., Du, J. *et al.* What is a gene, post-ENCODE? history and updated definition. *Genome Research*, 17(6):669–681, June 2007. ISSN 1088-9051. doi: 10.1101/gr.6339607. URL http://www.genome.org/cgi/doi/10.1101/gr.6339607.

Ghavi-Helm, Y. and Furlong, E.E.M. Analyzing transcription factor occupancy during embryo development using ChIP-seq. In Deplancke, B. and Gheldof, N., editors, *Gene Regulatory Networks*, volume 786, pages 229–245. Humana Press, Totowa, NJ, 2012. ISBN 978-1-61779-291-5, 978-1-61779-292-2. URL http://www.springerlink.com/index/10.1007/978-1-61779-292-2_14.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D. *et al.*

Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, July 2014. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature13417. URL http://www.nature.com/doifinder/10.1038/nature13417.

Giese, K., Cox, J. and Grosschedl, R. The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. *Cell*, 69(1):185 – 195, 1992. ISSN 0092-8674. doi: http://dx.doi.org/10.1016/0092-8674(92)90129-Z. URL http://www.sciencedirect.com/science/article/pii/009286749290129Z.

Giniger, E., Tietje, K., Jan, L.Y. and Jan, Y.N. *lola* encodes a putative transcription factor required for axon growth and guidance in drosophila. *Development*, 120(6):1385–1398, 1994. URL http://dev.biologists.org/content/120/6/1385.short.

Girard, F., Joly, W., Savare, J., Bonneaud, N., Ferraz, C. *et al.* Chromatin immunoprecipitation reveals a novel role for the *Drosophila* SoxNeuro transcription factor in axonal patterning. *Developmental Biology*, 299(2):530–542, November 2006. ISSN 00121606. doi: 10.1016/j.ydbio.2006.08.014. URL http://linkinghub.elsevier.com/retrieve/pii/S0012160606010840.

Giresi, P.G. and Lieb, J.D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (formaldehyde assisted isolation of regulatory elements). *Methods*, 48(3):233–239, July 2009. ISSN 10462023. doi: 10.1016/j.ymeth.2009.03.003. URL http://linkinghub.elsevier.com/retrieve/pii/S1046202309000504.

Gordon, K.L. and Ruvinsky, I. Tempo and mode in evolution of transcriptional regulation. *PLoS genetics*, 8(1):e1002432, 2012. URL http://dx.plos.org/10.1371/journal.pgen.1002432.

Grant, C.E., Bailey, T.L. and Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics*, 27(7):1017–1018, April 2011. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btr064. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btr064.

Graur, D., Zheng, Y., Price, N., Azevedo, R.B.R., Zufall, R.A. *et al.* On the immortality of television sets: "function" in the human genome according to the evolution-free gospel of ENCODE. *Genome Biology and Evolution*, 5(3): 578–590, March 2013. ISSN 1759-6653. doi: 10.1093/gbe/evt028. URL http://gbe.oxfordjournals.org/cgi/doi/10.1093/gbe/evt028.

Greer, J.M., Puetz, J., Thomas, K.R. and Capecchi, M.R. Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature*, 403(6770): 661–665, February 2000. ISSN 0028-0836. doi: 10.1038/35001077. URL http://dx.doi.org/10.1038/35001077.

Greil, F., Moorman, C. and van Steensel, B. [16] DamID: Mapping of in vivo proteingenome interactions using tethered DNA adenine methyltransferase. In *Methods in Enzymology*, volume 410, pages 342–359. Elsevier, 2006. ISBN 9780121828158. URL http://linkinghub.elsevier.com/retrieve/pii/S0076687906100166.

Guth, S.I.E. and Wegner, M. Having it both ways: Sox protein function between conservation and innovation. *Cellular and Molecular Life Sciences*, 65(19):3000–3018, May 2008. ISSN 1420-682X, 1420-9071. doi: 10.1007/s00018-008-8138-7. URL http://link.springer.com/10.1007/s00018-008-8138-7.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R. and Eisen, M.B. Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. *PLoS Genetics*, 4(6):e1000106, June 2008. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000106. URL http://dx.plos.org/10.1371/journal.pgen.1000106.

Harrison, M.M., Li, X.Y., Kaplan, T., Botchan, M.R. and Eisen, M.B. Zelda binding in the early *Drosophila melanogaster* embryo marks regions subsequently activated at the maternal-to-zygotic transition. *PLoS Genetics*, 7(10): e1002266, October 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002266. URL http://dx.plos.org/10.1371/journal.pgen.1002266.

He, B.Z., Holloway, A.K., Maerkl, S.J. and Kreitman, M. Does positive selection drive transcription factor binding site turnover? A test with *Drosophila cis*-regulatory modules. *PLoS Genetics*, 7(4):e1002053, April 2011a. ISSN

1553-7404. doi: 10.1371/journal.pgen.1002053. URL http://dx.plos.org/10.1371/journal.pgen.1002053.

He, Q., Bardet, A.F., Patton, B., Purvis, J., Johnston, J. *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nature Genetics*, 43(5):414–420, April 2011b. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.808. URL http://www.nature.com/doifinder/10.1038/ng.808.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C. *et al.* Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Molecular Cell*, 38(4):576–589, May 2010. ISSN 10972765. doi: 10.1016/j.molcel.2010.05.004. URL http://linkinghub.elsevier.com/retrieve/pii/S1097276510003667.

Horn, C. and Wimmer, E.A. A versatile vector set for animal transgenesis. *Development Genes and Evolution*, 210(12):630–637, 2000. URL http://link.springer.com/article/10.1007/s004270000110.

Huang, D.H., Chang, Y.L., Yang, C.C., Pan, I.C. and King, B. *pipsqueak* encodes a factor essential for sequence-specific targeting of a polycomb group protein complex. *Molecular and Cellular Biology*, 22(17):6261–6271, September 2002. ISSN 0270-7306. doi: 10.1128/MCB.22.17.6261-6271.2002. URL http://mcb.asm.org/cgi/doi/10.1128/MCB.22.17.6261-6271.2002.

Huang, J., Arsenault, M., Kann, M., Lopez-Mendez, C., Saleh, M. *et al.* The transcription factor sry-related HMG box-4 (SOX4) is required for normal renal development *in vivo*: *SOXC* genes during renal development. *Developmental Dynamics*, 242(6):790–799, June 2013. ISSN 10588388. doi: 10.1002/dvdy.23971. URL http://doi.wiley.com/10.1002/dvdy.23971.

Hueber, S.D., Bezdan, D., Henz, S.R., Blank, M., Wu, H. *et al.* Comparative analysis of Hox downstream genes in *Drosophila*. *Development*, 134(2):381–392, January 2007. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.02746. URL http://dev.biologists.org/cgi/doi/10.1242/dev.02746.

Hueber, S.D. and Lohmann, I. Shaping segments: *Hox* gene function in the genomic age. *BioEssays*, 30(10):965–979, October 2008. ISSN 02659247,

15211878. doi: 10.1002/bies.20823. URL http://doi.wiley.com/10.1002/bies.20823.

Isshiki, T., Pearson, B., Holbrook, S. and Doe, C.Q. *Drosophila* neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell*, 106(4):511, 2001. URL http://pearsonlab.ca/papers/2001_Cell106.511.pdf.

Jager, M., Quinnec, E., Houliston, E. and Manuel, M. Expansion of the SOX gene family predated the emergence of the bilateria. *Molecular Phylogenetics and Evolution*, 39(2):468–477, May 2006. ISSN 10557903. doi: 10.1016/j.ympev.2005.12.005. URL http://linkinghub.elsevier.com/retrieve/pii/S1055790305004148.

Jager, M., Quinnec, E., Chiori, R., Le Guyader, H. and Manuel, M. Insights into the early evolution of *SOX* genes from expression analyses in a ctenophore. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 310B(8):650–667, December 2008. ISSN 15525007, 15525015. doi: 10.1002/jez.b.21244. URL http://doi.wiley.com/10.1002/jez.b.21244.

Jinushi-Nakao, S., Arvind, R., Amikura, R., Kinameri, E., Liu, A.W. *et al.* Knot/Collier and Cut control different aspects of dendrite cytoskeleton and synergize to define final arbor shape. *Neuron*, 56(6):963–978, 2007. doi: 10.1016/j.neuron.2007.10.031. URL http://www.cell.com/neuron/abstract/S0896-6273(07)00827-6.

John, S., Sabo, P.J., Thurman, R.E., Sung, M.H., Biddie, S.C. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature Genetics*, 43(3):264–268, January 2011. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.759. URL http://www.nature.com/doifinder/10.1038/ng.759.

Kalay, G. and Wittkopp, P.J. Nomadic enhancers: Tissue-specific *cis*-regulatory elements of *yellow* have divergent genomic positions among *Drosophila* species. *PLoS Genetics*, 6(11):e1001222, November 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001222. URL http://dx.plos.org/10.1371/journal.pgen.1001222.

Kamachi, Y., Uchikawa, M., Collignon, J., Lovell-Badge, R. and Kondoh, H.

Involvement of Sox1, 2 and 3 in the early and subsequent molecular events of lens induction. *Development*, 125(13):2521–2532, 1998. URL http://dev.biologists.org/content/125/13/2521.short.

Kaplan, T., Li, X.Y., Sabo, P.J., Thomas, S., Stamatoyannopoulos, J.A. *et al.* Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. *PLoS Genetics*, 7(2):e1001290, February 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1001290. URL http://dx.plos.org/10.1371/journal.pgen.1001290.

Kappen, C. and Ruddle, F.H. Evolution of a regulatory gene family: HOM/HOX genes. *Current opinion in genetics & development*, 3(6):931–938, December 1993.

Karolchik, D. The UCSC table browser data retrieval tool. *Nucleic Acids Research*, 32(90001):493D–496, January 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh103. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkh103.

Karolchik, D., Barber, G.P., Casper, J., Clawson, H., Cline, M.S. *et al.* The UCSC genome browser database: 2014 update. *Nucleic Acids Research*, 42 (D1):D764–D770, January 2014. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1168. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1168.

Kispert, A., Herrmann, B.G., Leptin, M. and Reuter, R. Homologs of the mouse Brachyury gene are involved in the specification of posterior terminal structures in *Drosophila, Tribolium*, and *Locusta*. *Genes & Development*, 8(18):2137–2150, September 1994. ISSN 0890-9369. doi: 10.1101/gad.8.18.2137. URL http://www.genesdev.org/cgi/doi/10.1101/gad.8.18.2137.

Koohy, H., Down, T.A. and Hubbard, T.J. Chromatin accessibility data sets show bias due to sequence specificity of the DNase I enzyme. *PLoS ONE*, 8(7):e69853, July 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0069853. URL http://dx.plos.org/10.1371/journal.pone.0069853.

Kraut, R., Chia, W., Jan, L., Jan, Y. and Knoblich, J. Role of inscuteable in orienting asymmetric cell divisions in *Drosophila*. *Nature*, 383:50–55, 1996.

Kuntz, S.G. and Eisen, M.B. Native climate uniformly influences temperature-dependent growth rate in *Drosophila* embryos. *arXiv preprint arXiv:1306.5297*, 2013. URL http://arxiv.org/abs/1306.5297.

Kuzin, A., Brody, T., Moore, A.W. and Odenwald, W.F. Nerfin-1 is required for early axon guidance decisions in the developing *Drosophila* CNS. *Developmental Biology*, 277(2):347–365, January 2005. ISSN 00121606. doi: 10.1016/j.ydbio.2004.09.027. URL http://linkinghub.elsevier.com/retrieve/pii/S0012160604006633.

Kvon, E.Z., Stampfel, G., Yanez-Cuna, J.O., Dickson, B.J. and Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes & Development*, 26(9):908–913, April 2012. ISSN 0890-9369. doi: 10.1101/gad.188052.112. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.188052.112.

Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, 22(9):1813–1831, September 2012. ISSN 1088-9051. doi: 10.1101/gr.136184.111. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.136184.111.

Langmead, B. and Salzberg, S.L. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, March 2012. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth.1923. URL http://www.nature.com/doifinder/10.1038/nmeth.1923.

Larroux, C., Luke, G.N., Koopman, P., Rokhsar, D.S., Shimeld, S.M. *et al.* Genesis and expansion of metazoan transcription factor gene classes. *Molecular Biology and Evolution*, 25(5):980–996, February 2008. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msn047. URL http://mbe.oxfordjournals.org/cgi/doi/10.1093/molbev/msn047.

Larroux, C., Fahey, B., Liubicich, D., Hinman, V.F., Gauthier, M. *et al.* Developmental expression of transcription factor genes in a demosponge: insights into the origin of metazoan multicellularity. *Evolution & development*, 8(2):150–173, 2006. URL http://onlinelibrary.wiley.com/doi/10.1111/j.1525-142X.2006.00086.x/full.

Laudet, V., Stehelin, D. and Clevers, H. Ancestry and diversity of the HMG box superfamily. *Nucleic Acids Research*, 21(10):2493–2501, 1993. URL http://nar.oxfordjournals.org/content/21/10/2493.short.

Lee, T., Marticke, S., Sung, C., Robinow, S. and Luo, L. Cell-autonomous requirement of the USP/EcR-b ecdysone receptor for mushroom body neuronal remodeling in *Drosophila*. *Neuron*, 28(3):807–818, 2000. doi: 10.1016/S0896-6273(00)00155-0. URL http://www.cell.com/neuron/abstract/S0896-6273(00)00155-0.

Lefebvre, V., Dumitriu, B., Penzo-Méndez, A., Han, Y. and Pallavi, B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *The International Journal of Biochemistry & Cell Biology*, 39 (12):2195–2214, 2007. ISSN 13572725. doi: 10.1016/j.biocel.2007.05.019. URL http://linkinghub.elsevier.com/retrieve/pii/S1357272507001756.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, August 2009. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btp352. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp352.

Li, X.Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J.A. *et al.* The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. *Genome Biol*, 12(4):R34, 2011. URL http://www.biomedcentral.com/content/pdf/gb-2011-12-4-r34.pdf.

Liu, Q.X., Hiramoto, M., Ueda, H., Gojobori, T., Hiromi, Y. *et al.* Midline governs axon pathfinding by coordinating expression of two major guidance systems. *Genes & Development*, 23(10):1165–1170, May 2009. ISSN 0890-9369. doi: 10.1101/gad.1774209. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.1774209.

Love, M.I., Huber, W. and Anders, S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *bioRxiv*, 2014. doi: 10.1101/002832. URL http://dx.doi.org/10.1101/002832.

Löytynoja, A. and Goldman, N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, 320(5883):1632–1635, June 2008. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1158395. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1158395.

Löytynoja, A. and Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(30):10557–10562, 2005. URL http://www.pnas.org/content/102/30/10557.short.

Ludwig, M.Z., Manu, Kittler, R., White, K.P. and Kreitman, M. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. *PLoS Genetics*, 7(11):e1002364, November 2011. ISSN 1553-7404. doi: 10.1371/journal.pgen.1002364. URL http://dx.plos.org/10.1371/journal.pgen.1002364.

Lynch, M. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, November 2000. ISSN 00368075, 10959203. doi: 10.1126/science.290.5494.1151. URL http://www.sciencemag.org/cgi/doi/10.1126/science.290.5494.1151.

Lynch, M., O'Hely, M., Walsh, B. and Force, A. The probability of preservation of a newly arisen gene duplicate. *Genetics*, 159(4):1789–1804, 2001. URL http://www.genetics.org/content/159/4/1789.short.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A. *et al.* FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. *Genome biology*, 8(7):R129, 2007. URL http://www.biomedcentral.com/1465-6906/8/R129.

Ma, Y., Certel, K., Gao, Y., Niemitz, E., Mosher, J. *et al.* Functional interactions between *Drosophila* bHLH/PAS, Sox, and POU transcription factors regulate CNS midline expression of the *slit* gene. *The Journal of Neuroscience*, 20(12):4596–4605, 2000. URL http://www.jneurosci.org/content/20/12/4596.short.

MacArthur, S., Li, X.Y., Li, J., Brown, J.B., Chu, H.C. *et al.* 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(7):R80, 2009. URL http://www.biomedcentral.com/content/pdf/gb-2009-10-7-r80.pdf.

Maconochie, M., Nonchev, S., Morrison, A. and Krumlauf, R. Paralogous Hox genes: function and regulation. *Annual review of genetics*, 30:529–556, 1996. doi: 10.1146/annurev.genet.30.1.529.

Mahony, S. and Benos, P.V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Research*, 35(Web Server):W253–W258, May 2007. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm272. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkm272.

Malas, S., Duthie, S., Deloukas, P. and Episkopou, V. The isolation and high-resolution chromosomal mapping of human SOX14 and SOX21; two members of the SOX gene family related to SOX1, SOX2, and SOX3. *Mammalian genome : official journal of the International Mammalian Genome Society*, 10 (9):934–937, September 1999.

Mann, R.S., Lelli, K.M. and Joshi, R. Chapter 3 Hox specificity. In *Current Topics in Developmental Biology*, volume 88, pages 63–101. Elsevier, 2009. ISBN 9780123745293. URL http://linkinghub.elsevier.com/retrieve/pii/S0070215309880034.

Manning, L., Heckscher, E., Purice, M., Roberts, J., Bennett, A. *et al.* A resource for manipulating gene expression and analyzing *cis*-regulatory modules in the *Drosophila* CNS. *Cell Reports*, 2(4):1002–1013, October 2012. ISSN 22111247. doi: 10.1016/j.celrep.2012.09.009. URL http://linkinghub.elsevier.com/retrieve/pii/S2211124712002902.

Masui, S., Nakatake, Y., Toyooka, Y., Shimosato, D., Yagi, R. *et al.* Pluripotency governed by Sox2 via regulation of Oct3/4 expression in mouse embryonic stem cells. *Nature Cell Biology*, 9(6):625–635, June 2007. ISSN 1465-7392, 1476-4679. doi: 10.1038/ncb1589. URL http://www.nature.com/doifinder/10.1038/ncb1589.

Matsui, T. Redundant roles of Sox17 and Sox18 in postnatal angiogenesis in mice. *Journal of Cell Science*, 119(17):3513–3526, September 2006. ISSN 0021-9533,

1477-9137. doi: 10.1242/jcs.03081. URL http://jcs.biologists.org/cgi/doi/10.1242/jcs.03081.

Maurange, C. and Gould, A.P. Brainy but not too brainy: starting and stopping neuroblast divisions in *Drosophila*. *Trends in Neurosciences*, 28(1):30–36, January 2005. ISSN 01662236. doi: 10.1016/j.tins.2004.10.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0166223604003376.

McKay, D. and Lieb, J. A common set of DNA regulatory elements shapes *Drosophila* appendages. *Developmental Cell*, 27(3):306–318, November 2013. ISSN 15345807. doi: 10.1016/j.devcel.2013.10.009. URL http://linkinghub.elsevier.com/retrieve/pii/S1534580713006060.

McKimmie, C., Woerfel, G. and Russell, S. Conserved genomic organisation of group b Sox genes in insects. *BMC genetics*, 6(1):26, 2005. URL http://www.biomedcentral.com/1471-2156/6/26/.

Molin, L., Mounsey, A., Aslam, S., Bauer, P., Young, J. *et al.* Evolutionary conservation of redundancy between a diverged pair of forkhead transcription factor homologues. *Development*, 127(22):4825–4835, 2000. URL http://dev.biologists.org/content/127/22/4825.short.

Moses, A.M. Statistical tests for natural selection on regulatory regions based on the strength of transcription factor binding sites. *BMC Evolutionary Biology*, 9(1):286, 2009. ISSN 1471-2148. doi: 10.1186/1471-2148-9-286. URL http://www.biomedcentral.com/1471-2148/9/286.

Moses, A.M., Pollard, D.A., Nix, D.A., Iyer, V.N., Li, X.Y. *et al.* Large-scale turnover of functional transcription factor binding sites in *Drosophila*. *PLoS Computational Biology*, 2(10):e130, 2006. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.0020130. URL http://dx.plos.org/10.1371/journal.pcbi.0020130.

Murakami, R., Takashima, S. and Hamaguchi, T. Developmental genetics of the *Drosophila* gut: specification of primordia, subdivision and overt-differentiation. *Cellular and molecular biology (Noisy-le-Grand, France)*, 45 (5):661–676, July 1999.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E. *et al.* An

expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, September 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature11212. URL http://www.nature.com/doifinder/10.1038/nature11212.

Ni, X., Zhang, Y.E., Ngre, N., Chen, S., Long, M. *et al.* Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biology*, 10 (11):e1001420, November 2012. ISSN 1545-7885. doi: 10.1371/journal.pbio.1001420. URL http://dx.plos.org/10.1371/journal.pbio.1001420.

Nicol, J.W., Helt, G.A., Blanchard, S.G., Raja, A. and Loraine, A.E. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, October 2009.

Nishiguchi, S., Wood, H., Kondoh, H., Lovell-Badge, R. and Episkopou, V. Sox1 directly regulates the -crystallin genes and is essential for lens development in mice. *Genes & development*, 12(6):776–781, 1998. URL http://genesdev.cshlp.org/content/12/6/776.short.

Nowak, M.A., Boerlijst, M.C., Cooke, J. and Smith, J.M. Evolution of genetic redundancy. *Nature*, 388(6638):167–171, 1997. URL http://www.nature.com/nature/journal/v388/n6638/abs/388167a0.html.

O'Connor-Giles, K.M. and Skeath, J.B. Numb inhibits membrane localization of Sanpodo, a four-pass transmembrane protein, to promote asymmetric divisions in *Drosophila*. *Developmental Cell*, 5(2):231–243, 2003. doi: 10.1016/S1534-5807(03)00226-0. URL http://www.cell.com/developmental-cell/abstract/S1534-5807(03)00226-0.

Odom, D.T., Dowell, R.D., Jacobsen, E.S., Gordon, W., Danford, T.W. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, 39(6):730–732, May 2007. ISSN 1061-4036. doi: 10.1038/ng2047. URL http://www.nature.com/doifinder/10.1038/ng2047.

Okuda, Y., Ogura, E., Kondoh, H. and Kamachi, Y. B1 SOX coordinate cell specification with patterning and morphogenesis in the early zebrafish embryo. *PLoS*

Genetics, 6(5):e1000936, May 2010. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000936. URL http://dx.plos.org/10.1371/journal.pgen.1000936.

Overton, P.M., Chia, W. and Buescher, M. The *Drosophila* HMG-domain proteins SoxNeuro and Dichaete direct trichome formation via the activation of shavenbaby and the restriction of Wingless pathway activity. *Development*, 134(15):2807–2813, June 2007. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.02878. URL http://dev.biologists.org/cgi/doi/10.1242/dev.02878.

Overton, P. *The role of Sox genes in the development of* Drosophila melanogaster. PhD thesis, University of Cambridge, 2003.

Overton, P.M., Meadows, L.A., Urban, J. and Russell, S. Evidence for differential and redundant function of the Sox genes Dichaete and SoxN during CNS development in *Drosophila*. *Development*, 129(18):4219–4228, 2002. URL http://dev.biologists.org/content/129/18/4219.short.

Paris, M., Kaplan, T., Li, X.Y., Villalta, J.E., Lott, S.E. *et al.* Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. *PLoS Genetics*, 9(9):e1003748, September 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003748. URL http://dx.plos.org/10.1371/journal.pgen.1003748.

Park, D., Lee, Y., Bhupindersingh, G. and Iyer, V.R. Widespread misinterpretable ChIP-seq bias in yeast. *PloS one*, 8(12):e83506, 2013. URL http://dx.plos.org/10.1371/journal.pone.0083506.g009.

Parker, L., Ellis, J.E., Nguyen, M.Q. and Arora, K. The divergent TGF- ligand Dawdle utilizes an activin pathway to influence axon guidance in *Drosophila*. *Development*, 133(24):4981–4991, November 2006. ISSN 0950-1991, 1477-9129. doi: 10.1242/dev.02673. URL http://dev.biologists.org/cgi/doi/10.1242/dev.02673.

Parra, G. GeneID in *Drosophila*. *Genome Research*, 10(4):511–515, April 2000. ISSN 10889051. doi: 10.1101/gr.10.4.511. URL http://www.genome.org/cgi/doi/10.1101/gr.10.4.511.

Parrish, J.Z. Genome-wide analyses identify transcription factors required for proper morphogenesis of *Drosophila* sensory neuron dendrites. *Genes & De-*

velopment, 20(7):820–835, March 2006. ISSN 0890-9369. doi: 10.1101/gad.1391006. URL http://www.genesdev.org/cgi/doi/10.1101/gad.1391006.

Patel, N.H. Imaging neuronal subsets and other cell types in whole-mount *Drosophila* embryos and larvae using antibody probes. *Methods in cell biology*, 44:445–487, 1994.

Perry, M.W., Boettiger, A.N., Bothma, J.P. and Levine, M. Shadow enhancers foster robustness of *Drosophila* gastrulation. *Current Biology*, 20(17):1562–1567, September 2010. ISSN 09609822. doi: 10.1016/j.cub.2010.07.043. URL http://linkinghub.elsevier.com/retrieve/pii/S0960982210009450.

Phochanukul, N. and Russell, S. No backbone but lots of Sox: Invertebrate Sox genes. *The International Journal of Biochemistry & Cell Biology*, 42(3):453–464, March 2010. ISSN 13572725. doi: 10.1016/j.biocel.2009.06.013. URL http://linkinghub.elsevier.com/retrieve/pii/S1357272509001915.

Pioro, H.L. and Stollewerk, A. The expression pattern of genes involved in early neurogenesis suggests distinct and conserved functions in the diplopod *Glomeris marginata*. *Development Genes and Evolution*, 216(7-8):417–430, May 2006. ISSN 0949-944X, 1432-041X. doi: 10.1007/s00427-006-0078-3. URL http://link.springer.com/10.1007/s00427-006-0078-3.

Quinlan, A.R. and Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, March 2010. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btq033. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btq033.

Rhee, H. and Pugh, B. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, December 2011. ISSN 00928674. doi: 10.1016/j.cell.2011.11.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867411013511.

Riaz, F. *The application of mutant DNA adenine methyltransferase enzymes in Dam Identification*. PhD thesis, University of Cambridge, May 2009.

Richards, S. Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. *Genome Research*, 15(1):1–

18, January 2005. ISSN 1088-9051. doi: 10.1101/gr.3059305. URL http://www.genome.org/cgi/doi/10.1101/gr.3059305.

Rizzoti, K., Brunelli, S., Carmignac, D., Thomas, P.Q., Robinson, I.C. *et al.* SOX3 is required during the formation of the hypothalamo-pituitary axis. *Nature Genetics*, 36(3):247–255, March 2004. ISSN 1061-4036. doi: 10.1038/ng1309. URL http://www.nature.com/doifinder/10.1038/ng1309.

Ross-Innes, C.S., Stark, R., Teschendorff, A.E., Holmes, K.A., Ali, H.R. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, January 2012. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature10730. URL http://www.nature.com/doifinder/10.1038/nature10730.

Rozowsky, J., Euskirchen, G., Auerbach, R.K., Zhang, Z.D., Gibson, T. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature Biotechnology*, 27(1):66–75, January 2009. ISSN 1087-0156. doi: 10.1038/nbt.1518. URL http://www.nature.com/doifinder/10.1038/nbt.1518.

Russell, S.R., Sánchez-Soriano, N., Wright, C.R. and Ashburner, M. The *Dichaete* gene of *Drosophila melanogaster* encodes a SOX-domain protein required for embryonic segmentation. *Development*, 122(11):3669–3676, 1996. URL http://dev.biologists.org/content/122/11/3669.short.

Russo, C.A., Takezaki, N. and Nei, M. Molecular phylogeny and divergence times of drosophilid species. *Molecular biology and evolution*, 12(3):391–404, 1995. URL http://mbe.oxfordjournals.org/content/12/3/391.short.

Sánchez-Soriano, N. and Russell, S. The *Drosophila* SOX-domain protein Dichaete is required for the development of the central nervous system midline. *Development*, 125(20):3989–3996, 1998. URL http://dev.biologists.org/content/125/20/3989.short.

Sánchez-Soriano, N. and Russell, S. Regulatory mutations of the *Drosophila* Sox gene Dichaete reveal new functions in embryonic brain and hindgut development. *Developmental Biology*, 220(2):307–321, April 2000. ISSN

00121606. doi: 10.1006/dbio.2000.9648. URL http://linkinghub.elsevier.com/retrieve/pii/S0012160600996489.

Sand, O., Thomas-Chollier, M., Vervisch, E. and van Helden, J. Analyzing multiple data sets by interconnecting RSAT programs via SOAP web services—an example with ChIP-chip data. *Nature Protocols*, 3(10):1604–1615, September 2008. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2008.99. URL http://www.nature.com/doifinder/10.1038/nprot.2008.99.

Sandmann, T., Jakobsen, J.S. and Furlong, E.E.M. ChIP-on-chip protocol for genome-wide analysis of transcription factor binding in *Drosophila melanogaster* embryos. *Nature Protocols*, 1(6):2839–2855, January 2007. ISSN 1754-2189. doi: 10.1038/nprot.2006.383. URL http://www.nature.com/doifinder/10.1038/nprot.2006.383.

Satija, R. and Bradley, R.K. The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the *Drosophila* embryo. *Genome Research*, 22(4):656–665, January 2012. ISSN 1088-9051. doi: 10.1101/gr.130682.111. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.130682.111.

Schepers, G.E., Teasdale, R.D. and Koopman, P. Twenty pairs of Sox: extent, homology, and nomenclature of the mouse and human Sox transcription factor gene families. *Developmental cell*, 3(2):167–170, August 2002.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, 328(5981):1036–1040, April 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1186176. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1186176.

Searls, D.B. Linguistic approaches to biological sequences. *Computer applications in the biosciences: CABIOS*, 13(4):333–344, 1997. URL http://bioinformatics.oxfordjournals.org/content/13/4/333.short.

Searls, D.B. Reading the book of life. *Bioinformatics*, 17(7):579–580, 2001. doi: 10.1093/bioinformatics/17.7.579. URL http://bioinformatics.oxfordjournals.org/content/17/7/579.short.

Searls, D.B. The language of genes. *Nature*, 420(6912):211–217, November 2002.

ISSN 0028-0836. doi: 10.1038/nature01255. URL http://dx.doi.org/10.1038/nature01255.

Shen, S.P., Aleksic, J. and Russell, S. Identifying targets of the Sox domain protein Dichaete in the *Drosophila* CNS via targeted expression of dominant negative proteins. *BMC developmental biology*, 13(1):1, 2013. URL http://www.biomedcentral.com/1471-213X/13/1/.

Simon, J.M., Giresi, P.G., Davis, I.J. and Lieb, J.D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nature Protocols*, 7(2):256–267, January 2012. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2011.444. URL http://www.nature.com/doifinder/10.1038/nprot.2011.444.

Sinclair, A.H., Berta, P., Palmer, M.S., Hawkins, J.R., Griffiths, B.L. *et al.* A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*, 346(6281):240–244, July 1990. doi: 10.1038/346240a0. URL http://dx.doi.org/10.1038/346240a0.

Skeath, J.B. and Doe, C.Q. Sanpodo and Notch act in opposition to Numb to distinguish sibling neuron fates in the *Drosophila* CNS. *Development*, 125 (10):1857–1865, 1998. URL http://dev.biologists.org/content/125/10/1857.short.

Slattery, M., Ma, L., Ngre, N., White, K.P. and Mann, R.S. Genome-wide tissue-specific occupancy of the hox protein ultrabithorax and hox cofactor homothorax in drosophila. *PLoS ONE*, 6(4):e14686, April 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0014686. URL http://dx.plos.org/10.1371/journal.pone.0014686.

Smits, P., Li, P., Mandel, J., Zhang, Z., Deng, J.M. *et al.* The transcription factors L-Sox5 and Sox6 are essential for cartilage formation. *Developmental Cell*, 1(2):277–290, 2001. doi: 10.1016/S1534-5807(01)00003-X. URL http://www.cell.com/developmental-cell/abstract/S1534-5807(01)00003-X.

Sock, E., Rettig, S.D., Enderich, J., Bosl, M.R., Tamm, E.R. *et al.* Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Molecular and Cellular Biology*, 24(15):6635–6644, August

2004. ISSN 0270-7306. doi: 10.1128/MCB.24.15.6635-6644.2004. URL http://mcb.asm.org/cgi/doi/10.1128/MCB.24.15.6635-6644.2004.

Southall, T., Gold, K., Egger, B., Davidson, C., Caygill, E. *et al.* Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: Assaying RNA pol II occupancy in neural stem cells. *Developmental Cell*, 26(1): 101–112, July 2013. ISSN 15345807. doi: 10.1016/j.devcel.2013.05.020. URL http://linkinghub.elsevier.com/retrieve/pii/S1534580713003146.

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C. *et al.* Analysis of variation at transcription factor binding sites in drosophila and humans. *Genome Biol*, 13:R49, 2012. URL http://www.biomedcentral.com/content/pdf/gb-2012-13-9-r49.pdf.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U. *et al.* The *Trichoplax* genome and the nature of placozoans. *Nature*, 454(7207):955–960, August 2008. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature07191. URL http://www.nature.com/doifinder/10.1038/nature07191.

Stefflova, K., Thybert, D., Wilson, M., Streeter, I., Aleksic, J. *et al.* Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell*, 154(3):530–540, August 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.07.007. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867413008416.

Tanaka, S., Kamachi, Y., Tanouchi, A., Hamada, H., Jing, N. *et al.* Interplay of SOX and POU factors in regulation of the nestin gene in neural primordial cells. *Molecular and Cellular Biology*, 24(20):8834–8846, October 2004. ISSN 0270-7306. doi: 10.1128/MCB.24.20.8834-8846.2004. URL http://mcb.asm.org/cgi/doi/10.1128/MCB.24.20.8834-8846.2004.

Tautz, D. Redundancies, development and the flow of information. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 14(4): 263–266, April 1992. doi: 10.1002/bies.950140410.

The modENCODE Consortium, Roy, S., Ernst, J., Kharchenko, P.V., Kheradpour, P. *et al.* Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science*, 330(6012):1787–1797, Decem-

ber 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1198374. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1198374.

Thomas, S., Li, X.Y., Sabo, P.J., Sandstrom, R., Thurman, R.E. *et al.* Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. *Genome Biol*, 12(5):R43, 2011. URL http://www.biomedcentral.com/content/pdf/gb-2011-12-5-r43.pdf.

Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C. *et al.* RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Research*, 39(suppl):W86–W91, July 2011. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkr377. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkr377.

Toedling, J., Sklyar, O. and Huber, W. Ringo –an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, 8(1):221, 2007. ISSN 14712105. doi: 10.1186/1471-2105-8-221. URL http://www.biomedcentral.com/1471-2105/8/221.

Trinh, Q.M., Jen, F.Y.A., Zhou, Z., Chu, K.M., Perry, M.D. *et al.* Cloud-based uniform ChIP-seq processing tools for modENCODE and ENCODE. *BMC genomics*, 14(1):494, 2013. URL http://www.biomedcentral.com/1471-2164/14/494?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+Bmc%2FGenomics%2FLatestArticles+(BMC+Genomics+-+Latest+articles).

Turatsinze, J.V., Thomas-Chollier, M., Defrance, M. and van Helden, J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, 3(10):1578–1588, September 2008. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2008.97. URL http://www.nature.com/doifinder/10.1038/nprot.2008.97.

Uchikawa, M., Kamachi, Y. and Kondoh, H. Two distinct subgroups of group b sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. *Mechanisms of Development*, 84(12):103 – 120, 1999. ISSN 0925-4773. doi: http://dx.doi.org/10.1016/S0925-4773(99)00083-0. URL http://www.sciencedirect.com/science/article/pii/S0925477399000830.

Uchikawa, M., Yoshida, M., Iwafuchi-Doi, M., Matsuda, K., Ishida, Y. *et al.* B1 and B2 Sox gene expression during neural plate development in chicken and mouse embryos: Universal versus species-dependent features. *Development, Growth & Differentiation*, 53(6):761–771, 2011. ISSN 1440-169X. doi: 10.1111/j.1440-169X.2011.01286.x. URL http://dx.doi.org/10.1111/j.1440-169X.2011.01286.x.

Untergasser, A., Nijveen, H., Rao, X., Bisseling, T., Geurts, R. *et al.* Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Research*, 35(Web Server):W71–W74, May 2007. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkm306. URL http://www.nar.oxfordjournals.org/cgi/doi/10.1093/nar/gkm306.

Uwanogho, D., Rex, M., Cartwright, E.J., Pearl, G., Healy, C. *et al.* Embryonic expression of the chicken *Sox2, Sox3* and *Sox11* genes suggests an interactive role in neuronal development. *Mechanisms of development*, 49(1):23–36, 1995. URL http://www.sciencedirect.com/science/article/pii/0925477394002993.

Van Doren, M., Bailey, A.M., Esnayra, J., Ede, K. and Posakony, J.W. Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. *Genes & Development*, 8(22):2729–2742, November 1994. ISSN 0890-9369. doi: 10.1101/gad.8.22.2729. URL http://www.genesdev.org/cgi/doi/10.1101/gad.8.22.2729.

van Steensel, B., Delrow, J. and Henikoff, S. Chromatin profiling using targeted DNA adenine methyltransferase. *Nature Genetics*, 27(3):304–308, March 2001.

Vavouri, T., Semple, J.I. and Lehner, B. Widespread conservation of genetic redundancy during a billion years of eukaryotic evolution. *Trends in Genetics*, 24(10):485–488, 2008. doi: 10.1016/j.tig.2008.08.005. URL http://www.cell.com/trends/genetics/abstract/S0168-9525(08)00224-2.

Villar, D., Flicek, P. and Odom, D.T. Evolution of transcription factor binding in metazoans –mechanisms and functional implications. *Nature Reviews Genetics*, March 2014. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg3481. URL http://www.nature.com/doifinder/10.1038/nrg3481.

Vogel, M.J., Peric-Hupkes, D. and van Steensel, B. Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nature Protocols*, 2 (6):1467–1478, June 2007. ISSN 1754-2189, 1750-2799. doi: 10.1038/nprot.2007.148. URL http://www.nature.com/doifinder/10.1038/nprot.2007.148.

Wagner, A. Distributed robustness versus redundancy as causes of mutational robustness. *BioEssays*, 27(2):176–188, February 2005. ISSN 0265-9247, 1521-1878. doi: 10.1002/bies.20170. URL http://doi.wiley.com/10.1002/bies.20170.

Wagner, A. Gene duplications, robustness and evolutionary innovations. *BioEssays*, 30(4):367–373, April 2008. ISSN 02659247, 15211878. doi: 10.1002/bies.20728. URL http://doi.wiley.com/10.1002/bies.20728.

Wegner, M. SOX after SOX: SOXession regulates neurogenesis. *Genes & Development*, 25(23):2423–2428, December 2011. ISSN 0890-9369. doi: 10.1101/gad.181487.111. URL http://genesdev.cshlp.org/cgi/doi/10.1101/gad.181487.111.

Wegner, M. and Stolt, C.C. From stem cells to neurons and glia: a soxist's view of neural development. *Trends in Neurosciences*, 28(11):583–588, November 2005. ISSN 01662236. doi: 10.1016/j.tins.2005.08.008. URL http://linkinghub.elsevier.com/retrieve/pii/S0166223605002201.

Wei, L., Cheng, D., Li, D., Meng, M., Peng, L. *et al.* Identification and characterization of *Sox* genes in the silkworm, *Bombyx mori*. *Molecular Biology Reports*, 38(5):3573–3584, December 2010. ISSN 0301-4851, 1573-4978. doi: 10.1007/s11033-010-0468-5. URL http://link.springer.com/10.1007/s11033-010-0468-5.

Whyte, W., Orlando, D., Hnisz, D., Abraham, B., Lin, C. *et al.* Master transcription factors and Mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, April 2013. ISSN 00928674. doi: 10.1016/j.cell.2013.03.035. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867413003929.

Wilson, M.E., Yang, K.Y., Kalousova, A., Lau, J., Kosaka, Y. *et al.* The

HMG box transcription factor Sox4 contributes to the development of the endocrine pancreas. *Diabetes*, 54(12):3402–3409, December 2005. doi: 10.2337/diabetes.54.12.3402. URL http://diabetes.diabetesjournals.org/content/54/12/3402.abstract.

Wilson, M.J. and Dearden, P.K. Evolution of the insect Sox genes. *BMC Evolutionary Biology*, 8(1):120, 2008. ISSN 1471-2148. doi: 10.1186/1471-2148-8-120. URL http://www.biomedcentral.com/1471-2148/8/120.

Wilson, M.D., Barbosa-Morais, N.L., Schmidt, D., Conboy, C.M., Vanes, L. *et al.* Species-specific transcription in mice carrying human chromosome 21. *Science*, 322(5900):434–438, October 2008. doi: 10.1126/science.1160930. URL http://www.sciencemag.org/content/322/5900/434.abstract.

Wood, H.B. and Episkopou, V. Comparative expression of the mouse *Sox1*, *Sox2* and *Sox3* genes from pre-gastrulation to early somite stages. *Mechanisms of development*, 86(1):197–201, 1999. URL http://www.sciencedirect.com/science/article/pii/S0925477399001161.

Wray, G.A. The evolutionary significance of *cis*-regulatory mutations. *Nature Reviews Genetics*, 8(3):206–216, March 2007. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg2063. URL http://www.nature.com/doifinder/10.1038/nrg2063.

Yang, J., Ramos, E. and Corces, V.G. The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species. *Genome Research*, 22(11):2199–2207, November 2012. ISSN 1088-9051. doi: 10.1101/gr.142125.112. URL http://genome.cshlp.org/cgi/doi/10.1101/gr.142125.112.

Yu, G. *ChIPseeker: ChIPseeker for ChIP peak Annotation, Comparison, and Visualization*, 2014. R package version 1.0.8.

Yu, H.H., Araj, H.H., Ralls, S.A. and Kolodkin, A.L. The transmembrane semaphorin Sema I is required in *Drosophila* for embryonic motor and CNS axon guidance. *Neuron*, 20(2):207–220, 1998. doi: 10.1016/S0896-6273(00)80450-X. URL http://www.cell.com/neuron/abstract/S0896-6273(00)80450-X.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S. *et al.* Model-

based analysis of ChIP-seq (MACS). *Genome Biol*, 9(9):R137, 2008. URL http://www.biomedcentral.com/content/pdf/gb-2008-9-9-r137.pdf.

Zhao, G. and Skeath, J.B. The Sox-domain containing gene *Dichaete/fish-hook* acts in concert with *vnd* and *ind* to regulate cell fate in the *Drosophila* neuroectoderm. *Development*, 129(5):1165–1174, 2002. URL http://dev.biologists.org/content/129/5/1165.short.

Zhao, G., Wheeler, S.R. and Skeath, J.B. Genetic control of dorsoventral patterning and neuroblast specification in the *Drosophila* central nervous system. *The International Journal of Developmental Biology*, 51(2):107–115, 2007. ISSN 0214-6282. doi: 10.1387/ijdb.062188gz. URL http://www.intjdevbiol.com/paper.php?doi=062188gz.

Zhen, Y. and Andolfatto, P. Methods to detect selection on noncoding DNA. In Anisimova, M., editor, *Evolutionary Genomics*, volume 856, pages 141–159. Humana Press, Totowa, NJ, 2012. ISBN 978-1-61779-584-8, 978-1-61779-585-5. URL http://www.springerlink.com/index/10.1007/978-1-61779-585-5_6.

Zhong, L., Wang, D., Gan, X., Yang, T. and He, S. Parallel expansions of Sox transcription factor group B predating the diversifications of the arthropods and jawed vertebrates. *PLoS ONE*, 6(1):e16570, January 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0016570. URL http://dx.plos.org/10.1371/journal.pone.0016570.

Zhu, L.J., Gazin, C., Lawson, N.D., Pags, H., Lin, S.M. *et al.* ChIPpeakAnno: a bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC bioinformatics*, 11(1):237, 2010. URL http://www.biomedcentral.com/1471-2105/11/237/.

Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M. and Furlong, E.E.M. Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, 462(7269):65–70, November 2009. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature08531. URL http://www.nature.com/doifinder/10.1038/nature08531.