UNIVERSITY OF CAMBRIDGE

Sarah Hamilton Carl

Darwin College

A thesis submitted on , 2014 for the Degree of Doctor of Philosophy

Abstract

Drug repositioning is the discovery of new indications for approved or failed drugs. This practice is commonly done within the drug discovery process in order to adjust or expand the application line of an active molecule. Nowadays, an increasing number of computational methodologies aim at predicting repositioning opportunities in an automated fashion. Some approaches rely on the direct physical interaction between molecules and protein targets (docking) and some methods consider more abstract descriptors, such as a gene expression signature, in order to characterise the potential pharmacological action of a drug (Chapter 1).

On a fundamental level, repositioning opportunities exist because drugs perturb multiple biological entities, (on and off-targets) themselves involved in multiple biological processes. Therefore, a drug can play multiple roles or exhibit various mode of actions responsible for its pharmacology. The work done for my thesis aims at characterising these various modes and mechanisms of action for approved drugs, using a mathematical framework called description logics.

In this regard, I first specify how living organisms can be compared to complex black box machines and how this analogy can help to capture biomedical knowledge using description logics (Chapter 2). Secondly, the theory is implemented in the Functional Therapeutic Chemical Classification System (FTC - `https://www.ebi.ac.uk/chembl/ftc/`), a resource defining over 20,000 new categories representing the modes and mechanisms of action of approved drugs. The FTC also indexes over 1,000 approved drugs, which have been classified into the mode of action categories using automated reasoning. The FTC is evaluated against a gold standard, the Anatomical Therapeutic Chemical Classification System (ATC), in order to characterise its quality and content (Chapter 3).

Finally, from the information available in the FTC, a series of drug repositioning hypotheses were generated and made publicly available via a web application (`https://www.ebi.ac.uk/chembl/research/ftc-hypotheses`). A sub-

set of the hypotheses related to the cardiovascular hypertension as well as for Alzheimer's disease are further discussed in more details, as an example of an application (Chapter 4).

The work performed illustrates how new valuable biomedical knowledge can be automatically generated by integrating and leveraging the content of publicly available resources using description logics and automated reasoning. The newly created classification (FTC) is a first attempt to formally and systematically characterise the function or role of approved drugs using the concept of mode of action. The open hypotheses derived from the resource are available to the community to analyse and design further experiments.

Declaration

This thesis:

- is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specified in the text;
- is not substantially the same as any that I have submitted for a degree or diploma or other qualification at any other university; and
- does not exceed the prescribed limit of 60,000 words (approximately 39,853 words).

Sarah Hamilton Carl
, 2014

Acknowledgements

Writing a thesis is a challenging enterprise, and I would like to thank all the people who provided support along the journey. First to both my supervisors, Dietrich Rebholz-Schuhmann (DRS) and John Overington (JPO) for giving me the opportunity and freedom of doing my own science. Then to all the folks who provided me valuable advice and enlightening discussion, in particular Sarah Carl, Robert Hoehndorf, Anika Oellrich, Felix Krueger, Rita Santos, Christoph Grabmuller, Benjamin Stauch, John May, Gerard van Westen, George Papadatos, Grace Mugambate, Syed Asad Rahman and Pedro Ballester. I have interacted mostly with two research groups during my PhD time at the EMBL-EBI, the text-mining and the ChEMBL group; I would like to show my gratitude to all the members of both of these teams. I would also like to include my family in the acknowledgements, in particular my mother Evelyne, my father Serge and my little brother Elliott, for providing mental support and cheese. All my friends should be present in this list, both from here such as the predocs or out from of the country such as the Haute-Savoie folks. A special thanks to Sarah Carl (<3) and Jean-Baptiste Pettit, as well as to Jacek for the spiritual support. Finally I would like to sincerely thank the reader, for taking the time and energy to go through this document. I deeply apologize in advance for the numerous people I did not mention here, and I will offer a free pub lunch to thank anyone reading this manuscript - just send me an email (samuel.croset@gmail.com) with the subject ”THESIS ACKNOWLEDGEMENT” to claim your meal.

Contents

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 through 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., 2000b; 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., 2000b; 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., 1995b; Wood and Episkopou, 1999a), 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.), and group F genes are expressed in the developing vascular and lymphatic systems (Downes and Koopman; 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 particular lineages (Figure

1.1) (Bowles et al., 2000b; 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., 1995a; Wood and Episkopou, 1999a; Wegner and Stolt, 2005). Group B *Sox* genes have been analyzed on both a sequence and expression level in several species of invertebrates as well, 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., 2000a; 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 (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; 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 both group B1 and B2 *Sox* genes in vertebrates and cannot be neatly divided into activator and repressor categories (Ferrero et al., 2014b). 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 (Crmazy et al., 2000; Soriano and Russell, 1998; Uwanogho et al., 1995a; Wood and Episkopou, 1999b; 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, while also 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 fly embryos observed upon expression of Sox2 protein (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 neuroblast 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., 2014a).

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; Crmazy et al., 2000; Girard et al., 2006; Snchez-Soriano and Russell, 2000; Shen et al., 2013). *Dichaete* mutant embryos show axonal and midline defects, which can be rescued by expressing Dichaete in the midline (Snchez-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 1.2A). *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 1.2B) (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). Such apparent redundancy is also present among 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). These results strongly suggest a partial functional compensation between group B1 *Sox* genes in vertebrates and between *SoxN* and *Dichaete* in *D. melanogaster*, the evolutionary driver for which 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., 2014a). 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, the TFs Dr and vnd, which are involved in dorso-ventral patterning in the CNS (Zhao et al., 2007), and the neuroblast temporal identity genes svp, hb, Kr and pdm2 (Ferrero et al., 2014a; 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 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., 2014a; 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 1.3). 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., 2014a). Considering 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 occurs is that the new paralogous gene 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 either lead to loss of function and transformation of the new paralogue into a pseudogene or, if favorable mutations occur, either subfunctionalization, in which the role of the original gene is divided amongst the new paralogues either by functional domain or by spatial or temporal expression pattern, or neofunctionalization, in which the new copy acquires functions that did not belong to the original gene(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 of the fly embryo and provide spatial information to downstream genes in order to direct the development of appropriate segmental morphology. Although they sometimes work in a combinatorial manner, their functions are largely non-redundant, with some exceptions, and individual deletions generally show strong mutant phenotypes in both flies and vertebrates (Foronda et al., 2009; Maconochie et al., 1996). 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 (Vavouri et al.). In contrast to the *Hox* genes, functional redundancy in *Sox* genes 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; Overton et al., 2002; Rizzoti et al., 2004; Uchikawa et al., 2011; Uwanogho et al., 1995b; Wegner and Stolt, 2005; Wood and Episkopou, 1999a).

One possible explanation for the compensatory action of *Dichaete* and *SoxN* is to provide greater robustness to the developing CNS; it has been argued that functional redundancy may be a general mechanism for promoting robustness in genetic regulatory networks (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., 2014a). 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 (Snchez-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 elected 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 underlying regulatory networks (Biggin, 2011; Fisher et al., 2012; MacArthur et al., 2009). Comparative studies of transcription factor binding in multiple *Drosophila* species facilitates 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 exist for directly or indirectly studying genome-wide transcription factor binding patterns in *Drosophila*. Two of the primary *in vivo* techniques are ChIP (chromatin immunoprecipitation) and DamID, which is based on DNA methylation by a tethered DNA adenine methyltransferase (dam) (Greil et al., 2006) (Figure 1.4). Each 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 via 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., 2011; Paris et al., 2013; Villar et al., 2014), representing an evolutionary span of approximately 40 million years.

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 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. One of these 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., 2011). Each of these studies considered both presence/absence of peaks in each species as well as quantitative changes in binding strength. 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., 2011). 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., 2011).

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., 2011; 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 Twists 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., 2011). 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., 2011; 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 to identify 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 coregulators 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., 2011).

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). 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). 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?

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 to use for 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 a Dichaete-Dam, SoxN-Dam and Dam-only construct 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, that of 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 generally, 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., 2011; 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

CHAPTER 3

Exploration of Dichaete and SoxNeuro in Four Species of *Drosophila*

CHAPTER 4

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

CHAPTER 5

Chromatin Accessibility During Development in *Drosophila pseudoobscura*

5.1 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 quite 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. 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) . I sequenced three biological replicates from each stage. A detailed description of the methods used in the FAIRE protocol and for processing the sequencing data can be found in Chapter 2. 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.

5.2 FAIRE-seq results

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

Bibliography

Haruhiko Akiyama, Marie-Christine Chaboissier, James F. Martin, Andreas Schedl, and Benoit de Crombrugghe. 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`.

Jelena Aleksic and Steven Russell. 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`.

Jelena Aleksic, Enrico Ferrero, Bettina Fischer, Shih Pei Shen, and Steven Russell. 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`.

Davide-Carlo Ambrosetti, Claudio Basilico, and Lisa Dailey. 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`.

Tenley C. Archer, Jing Jin, and Elena S. Casey. 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`.

M. Bergsland, D. Ramskold, C. Zaouter, S. Klum, R. Sandberg, and J. Muhr. 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`.

A. Bery, B. Martynoga, F. Guillemot, J.-S. Joly, and S. Retaux. 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`.

Pallavi Bhattaram, Alfredo Penzo-Mndez, Elisabeth Sock, Clemencia Colmenares, Kotaro J. Kaneko, Alex Vassilev, Melvin L. DePamphilis, Michael Wegner, and Vronique Lefebvre. 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`.

MarkD. Biggin. 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`.

J Bowles, G Schepers, and P Koopman. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental Biology*, 227(2):239–255, November 2000a.

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

Robert K. Bradley, Xiao-Yong Li, Cole Trapnell, Stuart Davidson, Lior Pachter, Hou Cheng Chu, Leath A. Tonkin, Mark D. Biggin, and Michael B. Eisen. 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.

Marita Buescher, Fook Sion Hing, and William Chia. 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.

Andrew G. Clark, Michael B. Eisen, Douglas R. Smith, Casey M. Bergman, Brian Oliver, Therese A. Markow, Thomas C. Kaufman, Manolis Kellis, William Gelbart, Venky N. Iyer, Daniel A. Pollard, Timothy B. Sackton, Amanda M. Larracuente, Nadia D. Singh, Jose P. Abad, Dawn N. Abt, Boris Adryan, Montserrat Aguade, Hiroshi Akashi, Wyatt W. Anderson, Charles F. Aquadro, David H. Ardell, Roman Arguello, Carlo G. Artieri, Daniel A. Barbash, Daniel Barker, Paolo Barsanti, Phil Batterham, Serafim Batzoglou, Dave Begun, Arjun Bhutkar, Enrico Blanco, Stephanie A. Bosak, Robert K. Bradley, Adrianne D. Brand, Michael R. Brent, Angela N. Brooks, Randall H. Brown, Roger K. Butlin, Corrado Caggese, Brian R. Calvi, A. Bernardo de Carvalho, Anat Caspi, Sergio Castrezana, Susan E. Celniker, Jean L. Chang, Charles Chapple, Sourav Chatterji, Asif Chinwalla, Alberto Civetta, Sandra W. Clifton, Josep M. Comeron, James C. Costello, Jerry A. Coyne, Jennifer Daub, Robert G. David, Arthur L. Delcher, Kim Delehaunty, Chuong B. Do, Heather Ebling, Kevin Edwards, Thomas Eickbush, Jay D. Evans, Alan Filipski, Sven Findei, Eva Freyhult, Lucinda Fulton, Robert Fulton, Ana C. L. Garcia, Anastasia Gardiner, David A. Garfield, Barry E. Garvin, Greg Gibson, Don Gilbert, Sante Gnerre, Jennifer Godfrey, Robert Good, Valer Gotea, Brenton Gravely, Anthony J. Greenberg, Sam Griffiths-Jones, Samuel Gross, Roderic Guigo, Erik A. Gustafson, Wilfried Haerty, Matthew W. Hahn, Daniel L. Halligan, Aaron L. Halpern, Gillian M. Halter, Mira V. Han, Andreas Heger, LaDeana Hillier, Angie S. Hinrichs, Ian Holmes, Roger A. Hoskins, Melissa J. Hubisz, Dan Hultmark, Melanie A. Huntley, David B. Jaffe, Santosh Jagadeeshan, William R. Jeck, Justin Johnson, Corbin D. Jones, William C. Jordan, Gary H. Karpen, Eiko Kataoka, Peter D. Keightley, Pouya Kheradpour, Ewen F. Kirkness, Leonardo B. Koerich, Karsten Kristiansen, Dave Kudrna, Rob J. Kulathinal, Sudhir Kumar, Roberta Kwok, Eric Lander, Charles H. Langley, Richard Lapoint, Brian P. Lazzaro, So-Jeong Lee, Lisa

Levesque, Ruiqiang Li, Chiao-Feng Lin, Michael F. Lin, Kerstin Lindblad-Toh, Ana Llopart, Manyuan Long, Lloyd Low, Elena Lozovsky, Jian Lu, Meizhong Luo, Carlos A. Machado, Wojciech Makalowski, Mar Marzo, Muneo Matsuda, Luciano Matzkin, Bryant McAllister, Carolyn S. McBride, Brendan McKernan, Kevin McKernan, Maria Mendez-Lago, Patrick Minx, Michael U. Mollenhauer, Kristi Montooth, Stephen M. Mount, Xu Mu, Eugene Myers, Barbara Negre, Stuart Newfeld, Rasmus Nielsen, Mohamed A. F. Noor, Patrick O'Grady, Lior Pachter, Montserrat Papaceit, Matthew J. Parisi, Michael Parisi, Leopold Parts, Jakob S. Pedersen, Graziano Pesole, Adam M. Phillippy, Chris P. Ponting, Mihai Pop, Damiano Porcelli, Jeffrey R. Powell, Sonja Prohaska, Kim Pruitt, Marta Puig, Hadi Quesneville, Kristipati Ravi Ram, David Rand, Matthew D. Rasmussen, Laura K. Reed, Robert Reenan, Amy Reily, Karin A. Remington, Tania T. Rieger, Michael G. Ritchie, Charles Robin, Yu-Hui Rogers, Claudia Rohde, Julio Rozas, Marc J. Rubenfield, Alfredo Ruiz, Susan Russo, Steven L. Salzberg, Alejandro Sanchez-Gracia, David J. Saranga, Hajime Sato, Stephen W. Schaeffer, Michael C. Schatz, Todd Schlenke, Russell Schwartz, Carmen Segarra, Rama S. Singh, Laura Sirot, Marina Sirota, Nicholas B. Sisneros, Chris D. Smith, Temple F. Smith, John Spieth, Deborah E. Stage, Alexander Stark, Wolfgang Stephan, Robert L. Strausberg, Sebastian Strempel, David Sturgill, Granger Sutton, Granger G. Sutton, Wei Tao, Sarah Teichmann, Yoshiko N. Tobari, Yoshihiko Tomimura, Jason M. Tsolas, Vera L. S. Valente, Eli Venter, J. Craig Venter, Saverio Vicario, Filipe G. Vieira, Albert J. Vilella, Alfredo Villasante, Brian Walenz, Jun Wang, Marvin Wasserman, Thomas Watts, Derek Wilson, Richard K. Wilson, Rod A. Wing, Mariana F. Wolfner, Alex Wong, Gane Ka-Shu Wong, Chung-I Wu, Gabriel Wu, Daisuke Yamamoto, Hsiao-Pei Yang, Shiaw-Pyng Yang, James A. Yorke, Kiyohito Yoshida, Evgeny Zdobnov, Peili Zhang, Yu Zhang, Aleksey V. Zimin, Jennifer Baldwin, Amr Abdouelleil, Jamal Abdulkadir, Adal Abebe, Brikti Abera, Justin Abreu, St Christophe Acer, Lynne Aftuck, Allen Alexander, Peter An, Erica Anderson, Scott Anderson, Harindra Arachi, Marc Azer, Pasang Bachantsang, Andrew Barry, Tashi Bayul, Aaron Berlin, Daniel Bessette, Toby Bloom, Jason Blye, Leonid Boguslavskiy, Claude Bonnet, Boris Boukhgalter, Imane Bourzgui, Adam Brown, Patrick Cahill, Sheridon Channer, Yama Cheshatsang, Lisa Chuda, Mieke Citroen, Alville Collymore, Patrick Cooke, Maura Costello, Katie D'Aco, Riza Daza,

Georgius De Haan, Stuart DeGray, Christina DeMaso, Norbu Dhargay, Kimberly Dooley, Erin Dooley, Missole Doricent, Passang Dorje, Kunsang Dorjee, Alan Dupes, Richard Elong, Jill Falk, Abderrahim Farina, Susan Faro, Diallo Ferguson, Sheila Fisher, Chelsea D. Foley, Alicia Franke, Dennis Friedrich, Loryn Gadbois, Gary Gearin, Christina R. Gearin, Georgia Giannoukos, Tina Goode, Joseph Graham, Edward Grandbois, Sharleen Grewal, Kunsang Gyaltsen, Nabil Hafez, Birhane Hagos, Jennifer Hall, Charlotte Henson, Andrew Hollinger, Tracey Honan, Monika D. Huard, Leanne Hughes, Brian Hurhula, M Erii Husby, Asha Kamat, Ben Kanga, Seva Kashin, Dmitry Khazanovich, Peter Kisner, Krista Lance, Marcia Lara, William Lee, Niall Lennon, Frances Letendre, Rosie LeVine, Alex Lipovsky, Xiaohong Liu, Jinlei Liu, Shangtao Liu, Tashi Lokyitsang, Yeshi Lokyitsang, Rakela Lubonja, Annie Lui, Pen MacDonald, Vasilia Magnisalis, Kebede Maru, Charles Matthews, William McCusker, Susan McDonough, Teena Mehta, James Meldrim, Louis Meneus, Oana Mihai, Atanas Mihalev, Tanya Mihova, Rachel Mittelman, Valentine Mlenga, Anna Montmayeur, Leonidas Mulrain, Adam Navidi, Jerome Naylor, Tamrat Negash, Thu Nguyen, Nga Nguyen, Robert Nicol, Choe Norbu, Nyima Norbu, Nathaniel Novod, Barry O'Neill, Sahal Osman, Eva Markiewicz, Otero L. Oyono, Christopher Patti, Pema Phunkhang, Fritz Pierre, Margaret Priest, Sujaa Raghuraman, Filip Rege, Rebecca Reyes, Cecil Rise, Peter Rogov, Keenan Ross, Elizabeth Ryan, Sampath Settipalli, Terry Shea, Ngawang Sherpa, Lu Shi, Diana Shih, Todd Sparrow, Jessica Spaulding, John Stalker, Nicole Stange-Thomann, Sharon Stavropoulos, Catherine Stone, Christopher Strader, Senait Tesfaye, Talene Thomson, Yama Thoulutsang, Dawa Thoulutsang, Kerri Topham, Ira Topping, Tsamla Tsamla, Helen Vassiliev, Andy Vo, Tsering Wangchuk, Tsering Wangdi, Michael Weiand, Jane Wilkinson, Adam Wilson, Shailendra Yadav, Geneva Young, Qing Yu, Lisa Zembek, Danni Zhong, Andrew Zimmer, Zac Zwirko, David B. Jaffe, Pablo Alvarez, Will Brockman, Jonathan Butler, CheeWhye Chin, Sante Gnerre, Manfred Grabherr, Michael Kleber, Evan Mauceli, and Iain MacCallum. 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.

Jrme Collignon, Shanthini Sockanathan, Adam Hacker, Michel Cohen-Tannoudji, Dominic Norris, Sohaila Rastan, Milena Stevanovic, Peter N. Goodfellow, and Robin Lovell-Badge. 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.

S. Contrino, R. N. Smith, D. Butano, A. Carr, F. Hu, R. Lyne, K. Rutherford, A. Kalderimis, J. Sullivan, S. Carbon, E. T. Kephart, P. Lloyd, E. O. Stinson, N. L. Washington, M. D. Perry, P. Ruzanov, Z. Zha, S. E. Lewis, L. D. Stein, and G. Micklem. 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.

Frdric Crmazy, Philippe Berta, and Franck Girard. < i> sox neuro</i>, 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.

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

Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein,

Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guig, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, Zhiping Weng, Kevin P. White, Barbara Wold, Jainab Khatun, Yanbao Yu, John Wrobel, Brian A. Risk, Harsha P. Gunawardena, Heather C. Kuiper, Christopher W. Maier, Ling Xie, Xian Chen, Morgan C. Giddings, Bradley E. Bernstein, Charles B. Epstein, Noam Shoresh, Jason Ernst, Pouya Kheradpour, Tarjei S. Mikkelsen, Shawn Gillespie, Alon Goren, Oren Ram, Xiaolan Zhang, Li Wang, Robbyn Issner, Michael J. Coyne, Timothy Durham, Manching Ku, Thanh Truong, Lucas D. Ward, Robert C. Altshuler, Matthew L. Eaton, Manolis Kellis, Sarah Djebali, Carrie A. Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, Chenghai Xue, Georgi K. Marinov, Jainab Khatun, Brian A. Williams, Chris Zaleski, Joel Rozowsky, Maik Rder, Felix Kokocinski, Rehab F. Abdelhamid, Tyler Alioto, Igor Antoshechkin, Michael T. Baer, Philippe Batut, Ian Bell, Kimberly Bell, Sudipto Chakrabortty, Xian Chen, Jacqueline Chrast, Joao Curado, Thomas Derrien, Jorg Drenkow, Erica Dumais, Jackie Dumais, Radha Duttagupta, Megan Fastuca, Kata Fejes-Toth, Pedro Ferreira, Sylvain Foissac, Melissa J. Fullwood, Hui Gao, David Gonzalez, Assaf Gordon, Harsha P. Gunawardena, Cdric Howald, Sonali Jha, Rory Johnson, Philipp Kapranov, Brandon King, Colin Kingswood, Guoliang Li, Oscar J. Luo, Eddie Park, Jonathan B. Preall, Kimberly Presaud, Paolo Ribeca, Brian A. Risk, Daniel Robyr, Xiaoan Ruan, Michael Sammeth, Kuljeet Singh Sandhu, Lorain Schaeffer, Lei-Hoon See, Atif Shahab, Jorgen Skancke, Ana Maria Suzuki, Hazuki Takahashi, Hagen Tilgner, Diane Trout, Nathalie Walters, Huaien Wang, John Wrobel, Yanbao Yu, Yoshihide Hayashizaki, Jennifer Harrow, Mark Gerstein, Timothy J. Hubbard, Alexandre Reymond, Stylianos E. Antonarakis, Gregory J. Hannon, Morgan C. Giddings, Yijun Ruan, Barbara Wold, Piero Carninci, Roderic Guig, Thomas R. Gingeras, Kate R. Rosenbloom, Cricket A.

Sloan, Katrina Learned, Venkat S. Malladi, Matthew C. Wong, Galt P. Barber, Melissa S. Cline, Timothy R. Dreszer, Steven G. Heitner, Donna Karolchik, W. James Kent, Vanessa M. Kirkup, Laurence R. Meyer, Jeffrey C. Long, Morgan Maddren, Brian J. Raney, Terrence S. Furey, Lingyun Song, Linda L. Grasfeder, Paul G. Giresi, Bum-Kyu Lee, Anna Battenhouse, Nathan C. Sheffield, Jeremy M. Simon, Kimberly A. Showers, Alexias Safi, Darin London, Akshay A. Bhinge, Christopher Shestak, Matthew R. Schaner, Seul Ki Kim, Zhuzhu Z. Zhang, Piotr A. Mieczkowski, Joanna O. Mieczkowska, Zheng Liu, Ryan M. McDaniell, Yunyun Ni, Naim U. Rashid, Min Jae Kim, Sheera Adar, Zhancheng Zhang, Tianyuan Wang, Deborah Winter, Damian Keefe, Ewan Birney, Vishwanath R. Iyer, Jason D. Lieb, Gregory E. Crawford, Guoliang Li, Kuljeet Singh Sandhu, Meizhen Zheng, Ping Wang, Oscar J. Luo, Atif Shahab, Melissa J. Fullwood, Xiaoan Ruan, Yijun Ruan, Richard M. Myers, Florencia Pauli, Brian A. Williams, Jason Gertz, Georgi K. Marinov, Timothy E. Reddy, Jost Vielmetter, E. Partridge, Diane Trout, Katherine E. Varley, Clarke Gasper, Anita Bansal, Shirley Pepke, Preti Jain, Henry Amrhein, Kevin M. Bowling, Michael Anaya, Marie K. Cross, Brandon King, Michael A. Muratet, Igor Antoshechkin, Kimberly M. Newberry, Kenneth McCue, Amy S. Nesmith, Katherine I. Fisher-Aylor, Barbara Pusey, Gilberto DeSalvo, Stephanie L. Parker, Sreeram Balasubramanian, Nicholas S. Davis, Sarah K. Meadows, Tracy Eggleston, Chris Gunter, J. Scott Newberry, Shawn E. Levy, Devin M. Absher, Ali Mortazavi, Wing H. Wong, Barbara Wold, Matthew J. Blow, Axel Visel, Len A. Pennachio, Laura Elnitski, Elliott H. Margulies, Stephen C. J. Parker, Hanna M. Petrykowska, Alexej Abyzov, Bronwen Aken, Daniel Barrell, Gemma Barson, Andrew Berry, Alexandra Bignell, Veronika Boychenko, Giovanni Bussotti, Jacqueline Chrast, Claire Davidson, Thomas Derrien, Gloria Despacio-Reyes, Mark Diekhans, Iakes Ezkurdia, Adam Frankish, James Gilbert, Jose Manuel Gonzalez, Ed Griffiths, Rachel Harte, David A. Hendrix, Cdric Howald, Toby Hunt, Irwin Jungreis, Mike Kay, Ekta Khurana, Felix Kokocinski, Jing Leng, Michael F. Lin, Jane Loveland, Zhi Lu, Deepa Manthravadi, Marco Mariotti, Jonathan Mudge, Gaurab Mukherjee, Cedric Notredame, Baikang Pei, Jose Manuel Rodriguez, Gary Saunders, Andrea Sboner, Stephen Searle, Cristina Sisu, Catherine Snow, Charlie Steward, Andrea Tanzer, Electra Tapanari, Michael L. Tress, Marijke J. van Baren, Nathalie Walters, Stefan Washietl, Laurens Wilming, Amonida Zadissa, Zhengdong Zhang, Michael Brent, David Haussler, Manolis Kellis, Alfonso Valencia, Mark Gerstein, Alexandre Reymond, Roderic Guig, Jennifer Harrow, Timothy J. Hubbard, Stephen G. Landt, Seth Frietze, Alexej Abyzov, Nick Addleman, Roger P. Alexander, Raymond K. Auerbach, Suganthi Balasubramanian, Keith Bettinger, Nitin Bhardwaj, Alan P. Boyle, Alina R. Cao, Philip Cayting, Alexandra Charos, Yong Cheng, Chao Cheng, Catharine Eastman, Ghia Euskirchen, Joseph D. Fleming, Fabian Grubert, Lukas Habegger, Manoj Hariharan, Arif Harmanci, Sushma Iyengar, Victor X. Jin, Konrad J. Karczewski, Maya Kasowski, Phil Lacroute, Hugo Lam, Nathan Lamarre-Vincent, Jing Leng, Jin Lian, Marianne Lindahl-Allen, Renqiang Min, Benoit Miotto, Hannah Monahan, Zarmik Moqtaderi, Xinmeng J. Mu, Henriette OGeen, Zhengqing Ouyang, Dorrelyn Patacsil, Baikang Pei, Debasish Raha, Lucia Ramirez, Brian Reed, Joel Rozowsky, Andrea Sboner, Minyi Shi, Cristina Sisu, Teri Slifer, Heather Witt, Linfeng Wu, Xiaoqin Xu, Koon-Kiu Yan, Xinqiong Yang, Kevin Y. Yip, Zhengdong Zhang, Kevin Struhl, Sherman M. Weissman, Mark Gerstein, Peggy J. Farnham, Michael Snyder, Scott A. Tenenbaum, Luiz O. Penalva, Francis Doyle, Subhradip Karmakar, Stephen G. Landt, Raj R. Bhanvadia, Alina Choudhury, Marc Domanus, Lijia Ma, Jennifer Moran, Dorrelyn Patacsil, Teri Slifer, Alec Victorsen, Xinqiong Yang, Michael Snyder, Kevin P. White, Thomas Auer, Lazaro Centanin, Michael Eichenlaub, Franziska Gruhl, Stephan Heermann, Burkhard Hoeckendorf, Daigo Inoue, Tanja Kellner, Stephan Kirchmaier, Claudia Mueller, Robert Reinhardt, Lea Schertel, Stephanie Schneider, Rebecca Sinn, Beate Wittbrodt, Jochen Wittbrodt, Zhiping Weng, Troy W. Whitfield, Jie Wang, Patrick J. Collins, Shelley F. Aldred, Nathan D. Trinklein, E. Christopher Partridge, Richard M. Myers, Job Dekker, Gaurav Jain, Bryan R. Lajoie, Amartya Sanyal, Gayathri Balasundaram, Daniel L. Bates, Rachel Byron, Theresa K. Canfield, Morgan J. Diegel, Douglas Dunn, Abigail K. Ebersol, Tristan Frum, Kavita Garg, Erica Gist, R. Scott Hansen, Lisa Boatman, Eric Haugen, Richard Humbert, Gaurav Jain, Audra K. Johnson, Ericka M. Johnson, Tattyana V. Kutyavin, Bryan R. Lajoie, Kristen Lee, Dimitra Lotakis, Matthew T. Maurano, Shane J. Neph, Fiedencio V. Neri, Eric D. Nguyen, Hongzhu Qu, Alex P. Reynolds, Vaughn Roach, Eric Rynes, Peter Sabo, Minerva E. Sanchez, Richard S. Sandstrom, Amartya Sanyal, Anthony O. Shafer, Andrew B. Stergachis, Sean Thomas, Robert E. Thurman, Benjamin Vernot, Jeff Vierstra, Shinny Vong, Hao Wang, Molly A. Weaver, Yongqi Yan, Miaohua Zhang, Joshua M. Akey, Michael Bender, Michael O. Dorschner, Mark Groudine, Michael J. MacCoss, Patrick Navas, George Stamatoyannopoulos, Rajinder Kaul, Job Dekker, John A. Stamatoyannopoulos, Ian Dunham, Kathryn Beal, Alvis Brazma, Paul Flicek, Javier Herrero, Nathan Johnson, Damian Keefe, Margus Lukk, Nicholas M. Luscombe, Daniel Sobral, Juan M. Vaquerizas, Steven P. Wilder, Serafim Batzoglou, Arend Sidow, Nadine Hussami, Sofia Kyriazopoulou-Panagiotopoulou, Max W. Libbrecht, Marc A. Schaub, Anshul Kundaje, Ross C. Hardison, Webb Miller, Belinda Giardine, Robert S. Harris, Weisheng Wu, Peter J. Bickel, Balazs Banfai, Nathan P. Boley, James B. Brown, Haiyan Huang, Qunhua Li, Jingyi Jessica Li, William Stafford Noble, Jeffrey A. Bilmes, Orion J. Buske, Michael M. Hoffman, Avinash D. Sahu, Peter V. Kharchenko, Peter J. Park, Dannon Baker, James Taylor, Zhiping Weng, Sowmya Iyer, Xianjun Dong, Melissa Greven, Xinying Lin, Jie Wang, Hualin S. Xi, Jiali Zhuang, Mark Gerstein, Roger P. Alexander, Suganthi Balasubramanian, Chao Cheng, Arif Harmanci, Lucas Lochovsky, Renqiang Min, Xinmeng J. Mu, Joel Rozowsky, Koon-Kiu Yan, Kevin Y. Yip, and Ewan Birney. 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`.

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

S. Ferrari, V. R. Harley, A. Pontiggia, P. N. Goodfellow, R. Lovell-Badge, and M. E. Bianchi. 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/`.

Enrico Ferrero, Bettina Fischer, and Steven Russell. SoxNeuro orchestrates central nervous system specification and differentiation in drosophila and is only partially redundant with dichaete. *Genome biology*, 15(5), May 2014a. doi: 10.1186/gb-2014-15-5-r74.

Enrico Ferrero, Bettina Fischer, and Steven Russell. SoxNeuro orchestrates central nervous system specification and differentiation in drosophila and is only partially redundant with dichaete. *Genome Biology*, 15(5):R74, 2014b. ISSN 1465-6906. doi: 10.1186/gb-2014-15-5-r74. URL http://genomebiology.com/2014/15/5/R74.

A. L. M. Ferri. 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.

W. W. Fisher, J. J. Li, A. S. Hammonds, J. B. Brown, B. D. Pfeiffer, R. Weiszmann, S. MacArthur, S. Thomas, J. A. Stamatoyannopoulos, M. B. Eisen, P. J. Bickel, M. D. Biggin, and S. E. Celniker. 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.

Allan Force, Michael Lynch, F. Bryan Pickett, Angel Amores, Yi-lin Yan, and John Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999. URL http://www.genetics.org/content/151/4/1531.short.

David Foronda, Luis F. de Navas, Daniel L. Garaulet, and Ernesto Sanchez-Herrero. 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.

M. B. Gerstein, C. Bruce, J. S. Rozowsky, D. Zheng, J. Du, J. O. Korbel, O. Emanuelsson, Z. D. Zhang, S. Weissman, and M. Snyder. 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.

Klaus Giese, Jeffery Cox, and Rudolf Grosschedl. 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.

Franck Girard, Willy Joly, Jean Savare, Nathalie Bonneaud, Conchita Ferraz, and Florence Maschat. 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.

Paul G. Giresi and Jason D. Lieb. 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.

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

D. Graur, Y. Zheng, N. Price, R. B. R. Azevedo, R. A. Zufall, and E. Elhaik. 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.

Frauke Greil, Celine Moorman, and Bas van Steensel. [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.

S. I. E. Guth and M. Wegner. 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.

Melissa M. Harrison, Xiao-Yong Li, Tommy Kaplan, Michael R. Botchan, and Michael B. Eisen. 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`.

Qiye He, Anas F Bardet, Brianne Patton, Jennifer Purvis, Jeff Johnston, Ariel Paulson, Madelaine Gogol, Alexander Stark, and Julia Zeitlinger. High conservation of transcription factor binding and evidence for combinatorial regulation across six drosophila species. *Nature Genetics*, 43(5):414–420, April 2011. ISSN 1061-4036, 1546-1718. doi: 10.1038/ng.808. URL `http://www.nature.com/doifinder/10.1038/ng.808`.

Jez Huang, Michel Arsenault, Martin Kann, Carlos Lopez-Mendez, Monique Saleh, Dorota Wadowska, Mary Taglienti, Jacqueline Ho, Yuan Miao, David Sims, Jonathan Spears, Alfonso Lopez, Glenda Wright, and Sunny Hartwig. 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`.

Takako Isshiki, Bret Pearson, Scott Holbrook, and Chris Q. Doe. 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`.

Muriel Jager, Eric Quinnec, Evelyn Houliston, and Michal Manuel. 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`.

Muriel Jager, Eric Quinnec, Roxane Chiori, Herv Le Guyader, and Michal Manuel. 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.

Sam John, Peter J Sabo, Robert E Thurman, Myong-Hee Sung, Simon C Biddie, Thomas A Johnson, Gordon L Hager, and John A Stamatoyannopoulos. 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.

Yusuke Kamachi, Masanori Uchikawa, Jrme Collignon, Robin Lovell-Badge, and Hisato Kondoh. 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.

Tommy Kaplan, Xiao-Yong Li, Peter J. Sabo, Sean Thomas, John A. Stamatoyannopoulos, Mark D. Biggin, and Michael B. Eisen. 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.

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

Hashem Koohy, Thomas A. Down, and Tim J. Hubbard. 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.

E. Z. Kvon, G. Stampfel, J. O. Yanez-Cuna, B. J. Dickson, and A. Stark. 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.

S. G. Landt, G. K. Marinov, A. Kundaje, P. Kheradpour, F. Pauli, S. Batzoglou, B. E. Bernstein, P. Bickel, J. B. Brown, P. Cayting, Y. Chen, G. DeSalvo, C. Epstein, K. I. Fisher-Aylor, G. Euskirchen, M. Gerstein, J. Gertz, A. J. Hartemink, M. M. Hoffman, V. R. Iyer, Y. L. Jung, S. Karmakar, M. Kellis, P. V. Kharchenko, Q. Li, T. Liu, X. S. Liu, L. Ma, A. Milosavljevic, R. M. Myers, P. J. Park, M. J. Pazin, M. D. Perry, D. Raha, T. E. Reddy, J. Rozowsky, N. Shoresh, A. Sidow, M. Slattery, J. A. Stamatoyannopoulos, M. Y. Tolstorukov, K. P. White, S. Xi, P. J. Farnham, J. D. Lieb, B. J. Wold, and M. Snyder. 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`.

C. Larroux, G. N. Luke, P. Koopman, D. S. Rokhsar, S. M. Shimeld, and B. M. Degnan. 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`.

Claire Larroux, Bryony Fahey, Danielle Liubicich, Veronica F. Hinman, Marie Gauthier, Milena Gongora, Kathryn Green, Gert Wrheide, Sally P. Leys, and Bernard M. Degnan. 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`.

Vronique Lefebvre, Bogdan Dumitriu, Alfredo Penzo-Mndez, Yu Han, and Bhattaram Pallavi. 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`.

Xiao-Yong Li, Sean Thomas, Peter J. Sabo, Michael B. Eisen, John A. Stamatoyannopoulos, and Mark D. Biggin. 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.

M. Lynch. 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.

Yue Ma, Kaan Certel, Yanping Gao, Emily Niemitz, Jack Mosher, Ashim Mukherjee, Mousumi Mutsuddi, Neda Huseinovic, Stephen T. Crews, Wayne A. Johnson, and others. Functional interactions between DrosophilabHLH/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.

Stewart MacArthur, Xiao-Yong Li, Jingyi Li, James B. Brown, Hou Cheng Chu, Lucy Zeng, Brandi P. Grondona, Aaron Hechmer, Lisa Simirenko, and S. V. Keranen. 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.

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

S. Malas, S. Duthie, P. Deloukas, and V. Episkopou. 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.

Shinji Masui, Yuhki Nakatake, Yayoi Toyooka, Daisuke Shimosato, Rika Yagi, Kazue Takahashi, Hitoshi Okochi, Akihiko Okuda, Ryo Matoba, Alexei A. Sharov, Minoru S. H. Ko, and Hitoshi Niwa. 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.

T. Matsui. 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.

Cdric Maurange and Alex P. Gould. 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.

DanielJ. McKay and JasonD. Lieb. 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.

Carol McKimmie, Gertrud Woerfel, and Steven Russell. 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/.

Alan M. Moses, Daniel A. Pollard, David A. Nix, Venky N. Iyer, Xiao-Yong Li, Mark D. Biggin, and Michael B. Eisen. 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.

Shane Neph, Jeff Vierstra, Andrew B. Stergachis, Alex P. Reynolds, Eric Haugen, Benjamin Vernot, Robert E. Thurman, Sam John, Richard Sandstrom, Audra K. Johnson, Matthew T. Maurano, Richard Humbert, Eric Rynes, Hao Wang, Shinny Vong, Kristen Lee, Daniel Bates, Morgan Diegel, Vaughn Roach, Douglas Dunn, Jun Neri, Anthony Schafer, R. Scott Hansen, Tanya Kutyavin, Erika Giste, Molly Weaver, Theresa Canfield, Peter Sabo, Miaohua Zhang, Gayathri Balasundaram, Rachel Byron, Michael J. MacCoss, Joshua M. Akey, M. A. Bender, Mark Groudine, Rajinder Kaul, and John A. Stamatoyannopoulos. 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.

Xiaochun Ni, Yong E. Zhang, Nicolas Ngre, Sidi Chen, Manyuan Long, and Kevin P. White. 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.

Seiji Nishiguchi, Heather Wood, Hisato Kondoh, Robin Lovell-Badge, and Vasso Episkopou. 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.

Duncan T Odom, Robin D Dowell, Elizabeth S Jacobsen, William Gordon, Timothy W Danford, Kenzie D MacIsaac, P Alexander Rolfe, Caitlin M Conboy, David K Gifford, and Ernest Fraenkel. 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.

P. M. Overton, W. Chia, and M. Buescher. 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.

Paul M. Overton, Lisa A. Meadows, Joachim Urban, and Steven Russell. 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.

Mathilde Paris, Tommy Kaplan, Xiao Yong Li, Jacqueline E. Villalta, Susan E. Lott, and Michael B. Eisen. 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.

Nichanun Phochanukul and Steven Russell. 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`.

Hilary L. Pioro and Angelika Stollewerk. 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`.

Karine Rizzoti, Silvia Brunelli, Danielle Carmignac, Paul Q Thomas, Iain C Robinson, and Robin Lovell-Badge. 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`.

S. R. Russell, Natalia Sanchez-Soriano, Charles R. Wright, and Michael Ashburner. 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`.

C. A. Russo, Naoko Takezaki, and Masatoshi Nei. 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`.

R. Satija and R. K. Bradley. 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`.

Goslik E Schepers, Rohan D Teasdale, and Peter Koopman. 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.

D. Schmidt, M. D. Wilson, B. Ballester, P. C. Schwalie, G. D. Brown, A. Marshall, C. Kutter, S. Watt, C. P. Martinez-Jimenez, S. Mackay, I. Talianidis,

P. Flicek, and D. T. Odom. 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.

David B. Searls. 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.

David B. Searls. 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.

David B. Searls. 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.

Shih Pei Shen, Jelena Aleksic, and Steven Russell. 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/.

Jeremy M Simon, Paul G Giresi, Ian J Davis, and Jason D Lieb. 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.

Andrew H. Sinclair, Philippe Berta, Mark S. Palmer, J. Ross Hawkins, Beatrice L. Griffiths, Matthijs J. Smith, Jamie W. Foster, Anna-Maria Frischauf, Robin Lovell-Badge, and Peter N. Goodfellow. 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.

Patrick Smits, Ping Li, Jennifer Mandel, Zhaoping Zhang, Jian Ming Deng, Richard R Behringer, Benoit de Crombrugghe, and Vronique Lefebvre.

The transcription factors l-sox5 and sox6 are essential for cartilage formation. *Developmental Cell*, 1(2):277–290. doi: 10.1016/S1534-5807(01)00003-X. URL http://www.cell.com/developmental-cell/abstract/S1534-5807(01)00003-X.

E. Sock, S. D. Rettig, J. Enderich, M. R. Bosl, E. R. Tamm, and M. Wegner. 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.

Natalia Snchez Soriano and Steven Russell. 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.

Mansi Srivastava, Emina Begovic, Jarrod Chapman, Nicholas H. Putnam, Uffe Hellsten, Takeshi Kawashima, Alan Kuo, Therese Mitros, Asaf Salamov, Meredith L. Carpenter, Ana Y. Signorovitch, Maria A. Moreno, Kai Kamm, Jane Grimwood, Jeremy Schmutz, Harris Shapiro, Igor V. Grigoriev, Leo W. Buss, Bernd Schierwater, Stephen L. Dellaporta, and Daniel S. Rokhsar. 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.

Klara Stefflova, David Thybert, MichaelD. Wilson, Ian Streeter, Jelena Aleksic, Panagiota Karagianni, Alvis Brazma, DavidJ. Adams, Iannis Talianidis, JohnC. Marioni, Paul Flicek, and DuncanT. Odom. 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.

Natalia Snchez-Soriano and Steven Russell. 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.

S. Tanaka, Y. Kamachi, A. Tanouchi, H. Hamada, N. Jing, and H. Kondoh. 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.

The modENCODE Consortium, S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, S. Washietl, B. I. Arshinoff, F. Ay, P. E. Meyer, N. Robine, N. L. Washington, L. Di Stefano, E. Berezikov, C. D. Brown, R. Candeias, J. W. Carlson, A. Carr, I. Jungreis, D. Marbach, R. Sealfon, M. Y. Tolstorukov, S. Will, A. A. Alekseyenko, C. Artieri, B. W. Booth, A. N. Brooks, Q. Dai, C. A. Davis, M. O. Duff, X. Feng, A. A. Gorchakov, T. Gu, J. G. Henikoff, P. Kapranov, R. Li, H. K. MacAlpine, J. Malone, A. Minoda, J. Nordman, K. Okamura, M. Perry, S. K. Powell, N. C. Riddle, A. Sakai, A. Samsonova, J. E. Sandler, Y. B. Schwartz, N. Sher, R. Spokony, D. Sturgill, M. van Baren, K. H. Wan, L. Yang, C. Yu, E. Feingold, P. Good, M. Guyer, R. Lowdon, K. Ahmad, J. Andrews, B. Berger, S. E. Brenner, M. R. Brent, L. Cherbas, S. C. R. Elgin, T. R. Gingeras, R. Grossman, R. A. Hoskins, T. C. Kaufman, W. Kent, M. I. Kuroda, T. Orr-Weaver, N. Perrimon, V. Pirrotta, J. W. Posakony, B. Ren, S. Russell, P. Cherbas, B. R. Graveley, S. Lewis, G. Micklem, B. Oliver, P. J. Park, S. E. Celniker, S. Henikoff, G. H. Karpen, E. C. Lai, D. M. MacAlpine, L. D. Stein, K. P. White, M. Kellis, D. Acevedo, R. Auburn, G. Barber, H. J. Bellen, E. P. Bishop, T. D. Bryson, A. Chateigner, J. Chen, H. Clawson, C. L. G. Comstock, S. Contrino, L. C. DeNapoli, Q. Ding, A. Dobin, M. H. Domanus, J. Drenkow, S. Dudoit, J. Dumais, T. Eng, D. Fagegaltier, S. E. Gadel, S. Ghosh, F. Guillier, D. Hanley, G. J. Hannon, K. D. Hansen, E. Heinz, A. S. Hinrichs, M. Hirst, S. Jha, L. Jiang, Y. L. Jung, H. Kashevsky, C. D. Kennedy, E. T. Kephart, L. Langton, O.-K. Lee, S. Li, Z. Li, W. Lin, D. Linder-Basso, P. Lloyd, R. Lyne, S. E. Marchetti, M. Marra, N. R. Mattiuzzo, S. McKay, F. Meyer, D. Miller, S. W. Miller, R. A. Moore, C. A. Morrison, J. A. Prinz, M. Rooks, R. Moore, K. M. Rutherford, P. Ruzanov, D. A. Scheftner, L. Senderowicz, P. K. Shah, G. Shanower,

R. Smith, E. O. Stinson, S. Suchy, A. E. Tenney, F. Tian, K. J. T. Venken, H. Wang, R. White, J. Wilkening, A. T. Willingham, C. Zaleski, Z. Zha, D. Zhang, Y. Zhao, and J. Zieba. Identification of functional elements and regulatory circuits by drosophila modENCODE. *Science*, 330(6012):1787–1797, December 2010. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1198374. URL http://www.sciencemag.org/cgi/doi/10.1126/science.1198374.

Sean Thomas, Xiao-Yong Li, Peter J. Sabo, Richard Sandstrom, Robert E. Thurman, Theresa K. Canfield, Erika Giste, William Fisher, Ann Hammonds, and Susan E. Celniker. 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.

Quang M. Trinh, Fei-Yang A. Jen, Ziru Zhou, Kar M. Chu, Marc D. Perry, Ellen T. Kephart, Sergio Contrino, Peter Ruzanov, and Lincoln D. Stein. 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).

Masanori Uchikawa, Yusuke Kamachi, and Hisato Kondoh. 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.

Masanori Uchikawa, Megumi Yoshida, Makiko Iwafuchi-Doi, Kazunari Matsuda, Yoshiko Ishida, Tatsuya Takemoto, and Hisato Kondoh. 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.

D Uwanogho, M Rex, E J Cartwright, G Pearl, C Healy, P J Scotting, and P T Sharpe. Embryonic expression of the chicken sox2, sox3 and sox11 genes suggests an interactive role in neuronal development. *Mechanisms of development*, 49(1-2):23–36, January 1995a.

Dafe Uwanogho, Maria Rex, Elizabeth J. Cartwright, Gina Pearl, Chris Healy, Paul J. Scotting, and Paul T. Sharpe. Embryonic expression of the chicken< i> sox2, sox3</i> and< i> sox11</i> genes suggests an interactive role in neuronal development. *Mechanisms of development*, 49(1):23–36, 1995b. URL `http://www.sciencedirect.com/science/article/pii/0925477394002993`.

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

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

Diego Villar, Paul Flicek, and Duncan T. Odom. 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`.

Andreas Wagner. 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`.

Andreas Wagner. 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`.

M. Wegner. 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`.

Michael Wegner and C. Claus Stolt. 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.

Ling Wei, Daojun Cheng, Dong Li, Meng Meng, Lina Peng, Lin Tang, Minhui Pan, Zhonghuai Xiang, Qingyou Xia, and Cheng Lu. 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.

WarrenA. Whyte, DavidA. Orlando, Denes Hnisz, BrianJ. Abraham, CharlesY. Lin, MichaelH. Kagey, PeterB. Rahl, TongIhn Lee, and RichardA. Young. 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.

Maria E. Wilson, Katherine Y. Yang, Anna Kalousova, Janet Lau, Yasuhiro Kosaka, Francis C. Lynn, Juehu Wang, Caroline Mrejen, Vasso Episkopou, Hans C. Clevers, and Michael S. German. 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.

Megan J Wilson and Peter K Dearden. 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.

H B Wood and V Episkopou. Comparative expression of the mouse sox1, sox2 and sox3 genes from pre-gastrulation to early somite stages. *Mechanisms of development*, 86(1-2):197–201, August 1999a.

Heather B. Wood and Vasso Episkopou. Comparative expression of the mouse< i> sox1</i>,< i> sox2</i> and< i> sox3</i> genes from pre-gastrulation to early somite stages. *Mechanisms of development*, 86(1):197–

201, 1999b. URL http://www.sciencedirect.com/science/article/pii/S0925477399001161.

Gregory A. Wray. 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.

J. Yang, E. Ramos, and V. G. Corces. 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.

Guoyan Zhao, Scott R. Wheeler, and James B. Skeath. 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.

Lei Zhong, Dengqiang Wang, Xiaoni Gan, Tong Yang, and Shunping He. 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.

Robert P. Zinzen, Charles Girardot, Julien Gagneur, Martina Braun, and Eileen E. M. Furlong. 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.

APPENDIX A

Glossary

appendix B

Abbreviations

Appendix C

Peer-reviewed publications