UNIVERSITY OF
CAMBRIDGE

Sox Gene Redundancy in the *Drosophila* Nervous System

Edridge Kevin D'Souza

Churchill College

This dissertation is submitted for the degree of Master of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. It is not substantially the same as any that I have submitted, or am concurrently submitting, for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or is being concurrently submitted, for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the prescribed limit of 20 000 words.

Edridge Kevin D'Souza
September 2020

Abstract

Sox Gene Redundancy in the *Drosophila* Nervous System

The early stages of Central Nervous System (CNS) development in the fruit fly *Drosophila melanogaster* are influenced by the activity of transcription factors belonging to the SRY-related HMG-box (Sox) gene family. This family, which shares a conserved High Mobility Group (HMG) domain, includes the sex determination factor SRY and is present in all metazoans characterised to date. With respect to CNS formation in fruit flies, the specification and differentiation of the neuroectoderm is mediated by gene regulatory networks regulated by the two SoxB family genes *SoxNeuro* (*SoxN*) and *Dichaete* (*D*), which are known to exhibit partial redundancy.

Genetic analysis reveals that while single mutants of either gene exhibit abnormal CNS phenotypes in regions where they are uniquely expressed, there is substantial functional compensation in the regions where they are normally co-expressed. Strikingly, double mutants exhibit severely defective CNS phenotypes. This functional compensation is a hallmark of the Sox family, and SoxB proteins are functionally conserved across species. It is therefore of interest to explore the degree to which Sox proteins are capable of compensating for each other, as well as to identify differences in their regulatory activities.

I aimed to test this by replacing the endogenous *SoxN* and *D* loci with the homologous mouse SoxB gene *Sox2*, which *SoxN* has been shown to functionally replace in mouse ES cells. CRISPR constructs were created to generate fly lines with either $SoxN^{Sox2}$ or D^{Sox2} replacement alleles. These lines were to be used to assess phenotypic rescue and to characterise any CNS phenotypes by immunohistochemistry followed by genomic analysis of Sox2 binding via ChIP-seq. In parallel, I also performed a computational analysis of the differences and similarities between *SoxN* and *D* expression at the single cell level, utilising published scRNA-seq datasets from the *Drosophila* embryo, larval brain, and adult ventral nerve cord (VNC) to examine expression in cell clusters that are SoxB-positive.

These analyses uncovered clusters of cells in the embryonic neuroectoderm consistent with the known expression of *SoxN* and *D*. In the larval brain, *SoxN* expression is higher than that of *D* and identifies cell clusters likely to represent neuroblasts or neural progenitors. Analyis of the much larger adult VNC dataset uncovered groups of cells exhibiting expression profiles indicative of differentiating GABAergic or cholinergic neurons, consistent with a role for *SoxN* in neuronal differentiation. Taken together, this thesis provides the foundation for generating transgenic fly lines to assess Sox protein functional compensation and evolutionary conservation *in vivo* and also identifies specific SoxB expressing cell populations that can be a focus for future analysis of gene regulatory networks regulated by SoxB.

Acknowledgements

I would like to thank Dr. Steve Russell for taking me into his lab, for providing me with guidance and feedback during for my experiments and writing, and for going above and beyond in helping me complete my thesis project. Special thanks to the other members of the Russell lab who helped me acclimate to the wet lab—Dr. Dagmara Korona, Bettina Fischer, Sabila Chilaeva, Maria Ouvarova, and especially Barbara Joo-Lara, who gave me personal guidance for all my wet lab experiments. I would also like to thank Dr. Anne Ferguson-Smith for being my departmental advisor, and Dr. Jennifer Nichols for providing the *mSox2* sequence that was critical for this project.

I additionally would like extend my thanks to the numerous anonymous voices of the bioinformatics communities in sites like BioStars, Stack Exchange, GitHub, and Reddit for helping me find the right direction whenever I felt truly lost.

I am thankful to my parents, who supported me and had the patience to watch me complete the latter half of this degree while quarantined at home. I am forever grateful to the Winston Churchill Foundation, Churchill College, and the University of Cambridge for providing me with the opportunity that made this MPhil research possible. Finally, I would like to thank the COVID-19 pandemic for providing me with the time to write this manuscript.

Contents

Glossary

AP Anterior-Posterior.

AS-C The *achaete/scute* gene complex.

bHLH Basic Helix-Loop-Helix.

BP Biological Process.

CDS Coding Sequence.

ChIP-seq Chromatin immunoprecipitation sequencing.

CNS Central Nervous System.

D *Dichaete*.

DamID DNA adenine methyltransferase identification.

DSBs Double-Strand Breaks.

DV Dorsal-Ventral.

Ek Velvet worm *Euperipatoides kanangrensis*.

ES cell Embryonic stem cell.

EtBr Ethidium Bromide.

FISSEQ Fluorescent *in situ* RNA sequencing.

Gm Pillbug *Glomeris marginata*.

GO Gene Ontology.

GRN Gene Regulatory Network.

gRNA Guide RNA.

HDR Homology Directed Repair.

HLTS Hypotrichosis–Lymphedema–Telangiectasia Syndrome.

HMG High Mobility Group.

iNSCs Induced Neural Stem Cells.

iPSCs Induced Pluripotent Stem Cells.

ISCs Intestinal Stem Cells.

IVT *In Vitro* Transcription.

LB Lysogeny broth.

LCA Last Common Ancestor.

MMLV Moloney Murine Leukemia Virus.

NGS Next-Generation Sequencing.

NPCs Neural Progenitor Cells.

ORF Open Reading Frame.

PCA Principal Component Analysis.

PCR Polymerase Chain Reaction.

Pt Spider *Parasteatoda tepidariorum.*

RHA Right Homology Arm.

SAZ Segment Addition Zone.

scRNA-seq Single-cell RNA sequencing.

Sm Spider *Stegodyphus mimosarum.*

SMART Switching mechanism at the 5' end of the RNA transcript.

smFISH Single molecule fluorescence *in situ* hybridisation.

SNN SharedNearest Neighbor.

Sox SRY-related HMG box.

SoxN *SoxNeuro*.

STRT Single-cell Tagged Reverse Transcription.

TAE Tris-Acetate-EDTA.

Tc Beetle *Tribolium castaneum*.

TF Transcription Factor.

UMAP Uniform Manifold Approximation and Projection.

VNC Ventral Nerve Cord.

Chapter 1

Introduction

The fruit fly *Drosophila melanogaster* has long been used as a model organism for biomedical research. Its short lifespan, well-characterised developmental cycle, and wealth of available genetic tools make it an attractive system for studying processes that are conserved in humans and other vertebrates (Tolwinski, 2017). As a result, it was one of the first animals to have a fully sequenced genome (Adams et al., 2000), making it suitable for detailed examination of gene expression and regulation. These characteristics, along with modern advances in technology such as next-generation sequencing (NGS) and CRISPR genome editing, make *Drosophila* a powerful tool for manipulating and studying the role of specific genes and genetic regulation.

An important group of genetic regulators is the Sox family of transcription factor (TF) genes, which encode a group of DNA-binding proteins that are implicated in a wide range of developmental and regulatory processes, and which are conserved across the metazoans, including humans (Kamachi & Kondoh, 2013). The role of Sox family TFs in *Drosophila* can therefore provide insight to the regulatory systems governing human embryonic development. In particular, examining the roles of the *Drosophila* Sox genes *SoxNeuro* (*SoxN*) and *Dichaete* (*D*) can elucidate some of the ways in which homologous mammalian Sox genes can influence neural development.

1.1 Sox Genes

Sox (SRY-related HMG box) genes derive their name from the founding member of the family, the mammalian testis-determining factor SRY (Sarkar & Hochedlinger, 2013). Their protein products are characterised by the presence of a sequence-specific high mobility group (HMG) DNA binding domain that is conserved between members of the Sox family (Kamachi & Kondoh, 2013), with family membership conferred by having at least 50% sequence similarity to the HMG box of SRY (Prior & Walter, 1996). Sox proteins are related to the distinct HMG domain in LEF/TCF transcription factors of the Wnt signalling pathway, and HMG domains from both Sox and LEF/TCF are related to the non-sequence specific DNA-binding HMG domain in structural

proteins including HMG1, HMG2, and the fungal mating-type protein MatA (Czaja et al., 2014; Kormish et al., 2010; Thomas & Travers, 2001). Sox HMG domains share a consensus DNA recognition sequence AACAAT, and all Sox proteins characterised to date are known to bind to this core sequence (Mertin et al., 1999; Rimini et al., 1995).

The HMG domain mediates Sox protein-DNA binding in a sequence-specific manner within the minor groove, bending the DNA approximately 90° to render nucleotides accessible to other components of the transcriptional initiation complex machinery or to partner TFs (Reményi et al., 2003). Sox proteins are often characterised by the presence of a C-terminal transactivation domain, which operates in conjunction with the p300 cofactor to activate transcription (Nowling et al., 2000). However, at least some Sox subgroups also exhibit behavior consistent with transcriptional repression (Uchikawa et al., 1999). Sox HMG domains additionally feature two independent nuclear localisation sequences that are widely conserved among the Sox family (Südbeck & Scherer, 1997), further marking Sox proteins for a variety of transcription-related functions. While SRY functions specifically to control sex differentiation between males and females in eutherian mammals, other Sox genes perform a wide variety of functions in different cell types, and are categorised into subgroups based on functional and sequence-level similarities.

The molecular functions of the Sox gene family are diverse, as Sox genes have evolved numerous roles in stem cell potentiation, developmental regulation, and neural formation. While not every Sox gene has an exact equivalent across different species, some common groupings allow classification of functional differences in Sox genes within an individual species and also across related phyla (Figure 1). Sox genes are therefore categorised into groups (A through H) with members of a given group sharing sequence-level and structural similarities across species (Bowles et al., 2000).

The sole member of the SoxA family is the mammalian testis determining factor *SRY*, which was the first characterised Sox gene. SoxB genes are expressed in neural progenitor cells (NPCs) and are involved in maintaining stem cell pluripotency (Sarkar & Hochedlinger, 2013). They also play a vital role in embryogenesis and the formation of the nervous system (Guth & Wegner, 2008). This subfamily includes both *SoxN* and *D* in *Drosophila*, as well as the homologous *Sox1, 2 and 3* in mice and humans (Schepers et al., 2002). SoxC genes *Sox4, Sox11,* and *Sox12* are expressed in stem cells and contribute to differentiation, giving rise to components of the skeletal, cardiovascular, nervous, and endocrine systems (Lefebvre & Bhattaram, 2016). As with the SoxB family, SoxD is also involved in neurodevelopment and embryogenesis, and its member *Sox6* is a downstream regulatory target of Sox2 (Ji & Kim, 2016). Despite not containing transactivation or transrepression domains, they are known to perform transcriptional activation and repression and play a role in both lymphocyte differentiation and erythropoiesis (Chew & Gallo, 2009; Lefebvre, 2010). SoxE is implicated in a variety of stem cell related functions, and lineage tracing experiments have shown that Sox9 plays a broad role

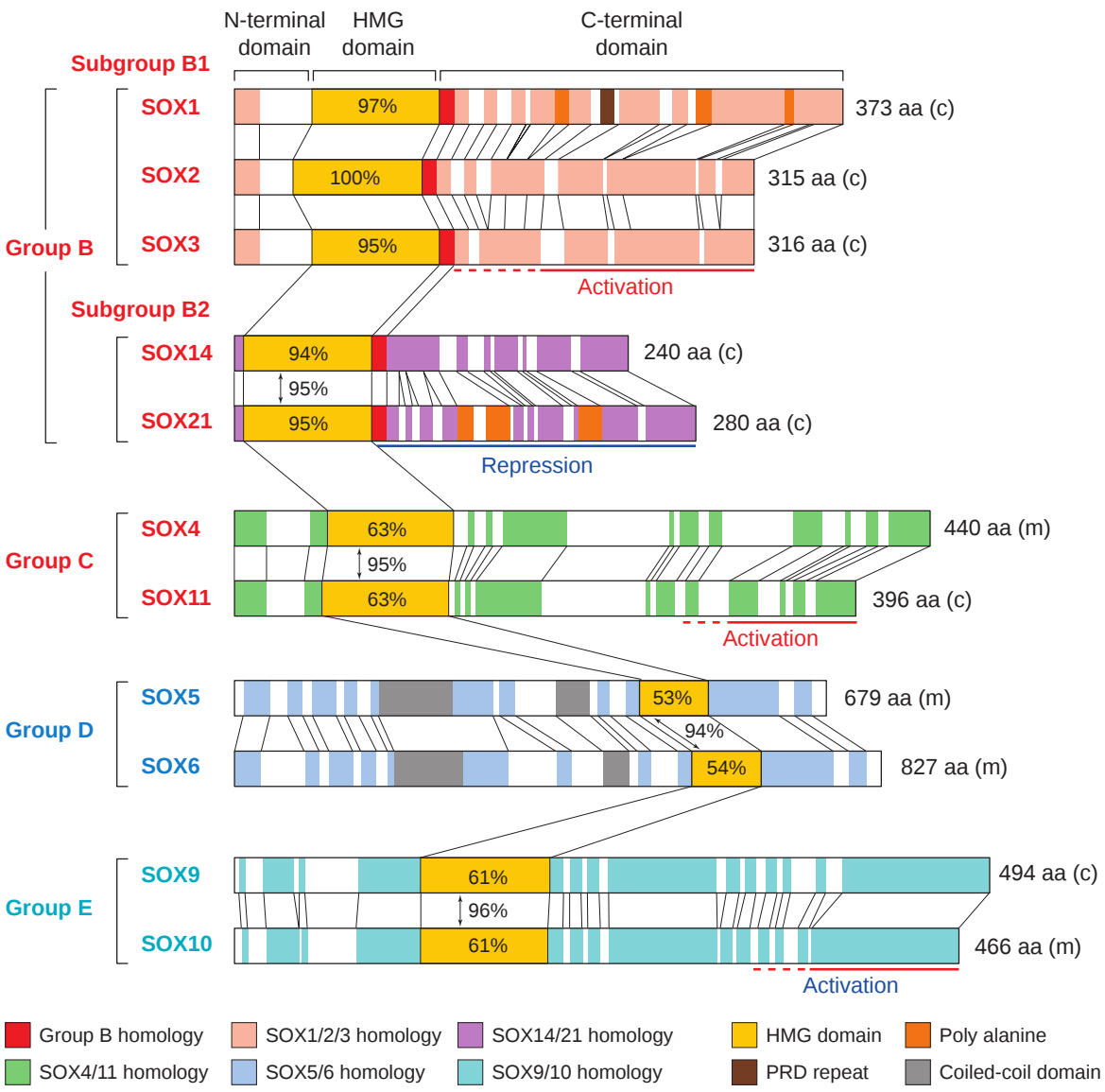

Figure 1: Diversity in Sox protein sequence homology. Reprinted with permission from Kamachi *et al.* (2000). Representative mammalian members of Sox families B through E are shown with sequence homology within the HMG domain quantified in comparison to Sox2. The HMG domain is the only domain that displays sequence homology between different families. However, there are several other domains that provide sequence homology within a given group; for instance, SoxB1 and SoxB2 genes contain a homology domain specific to group B, and additionally contain homology in activation or repression domains specific to their subgroup.

in priming progenitor cells for differentiation in intestinal, hepatic, pancreatic, mammary, and retinal tissues (Furuyama et al., 2011; Sarkar & Hochedlinger, 2013). Sox9 also acts downstream of SRY together with SoxE members Sox8 and Sox10 as well as with SoxB member Sox3 to regulate the development of gonadal cells during sex differentiation (Graves, 1998; She & Yang, 2017). The SoxF family plays a vital role in developing both the cardiovascular and lymphatic systems, and mutations in the SoxF gene *Sox18* are associated with hypotrichosis–lymphedema–telangiectasia syndrome (HLTS), a lymphatic disorder (Francois et al., 2010; Hokari et al., 2008). Evidently, Sox transcription factors provide a diverse arsenal of developmental and regulatory controls that are conserved between animal species. The several subfamilies represent structural similarities that translate into broad functional redundancy and conservation.

Much of the specificity and diversity of Sox genes comes from their partner factors. Early studies into the expression of the Sox2 target gene Fibroblast Growth Factor 4 (*FGF-4*) showed that neither Sox2 nor its cofactor Oct-3 were alone sufficient to induce *FGF-4* expression, but together, the Sox2/Oct-3 complex could interact with the protein Oct-1 to bind the *FGF-4* enhancer and drive expression (Yuan et al., 1995). These Oct proteins bind preferentially to specific DNA octomer motifs, and are a subfamily of POU homeodomain TFs known to be common cofactors of Sox TFs (Tantin, 2013). Because both Sox and Oct TFs have their own sequence-specific binding profiles, specific Sox and Oct gene combinations can result in complexes that can target enhancer sequences of a relatively small subset of genes at a time, lending tissue specificity to the Sox-partner code (Kondoh & Kamachi, 2010).

One of the most well-studied examples of these Sox-partner interactions is the interplay between Sox2 and Oct-3/4. When both are coexpressed, the resulting complex can drive expression not only of target genes that define the embryonic stem cell (ES cell) state, but also for the *Sox2* and *Oct-4* genes themselves, allowing the complex to stabilise its own expression and cell type (Kamachi & Kondoh, 2013; Kondoh et al., 2004). Interestingly, more recent work has shown that in the context of ES and neural progenitor cells, Sox2 acts as a pioneer factor to open chromatin and facilitate POU binding (Mistri et al., 2015). There is evidence that the Sox-POU interaction is conserved since, as described below, Dichaete is known to interact with the POU protein Ventral veins lacking in the CNS midline (Soriano & Russell, 1998).

In addition to stabilising an existing cell type, Sox-partner interactions can also drive the transition to the next stage of development. A Sox-partner complex is capable of driving expression of another TF that can then complex with either the original Sox gene or its cofactor to promote the expression of genes for a new stage of development (Kamachi & Kondoh, 2013). This is the case when Sox10 and its cofactor Pax3 promote expression of the *Mitf* gene, encoding a TF that partners with Sox10 to drive genes that differentiate neural crest cells into melanocytes (Bondurand et al., 2000; Ludwig et al., 2004). Modern sequencing techniques have provided a quantitative look at the interaction profiles of Sox TFs and their POU-domain cofactors, revealing that Sox2 is capable of cooperating not only with Oct-3/4 but also with

several of the class III POU factors that are endogenously coexpressed with Sox2 in neural cells (Chang et al., 2017). Heterodimerisation is not the only mode that Sox TFs can use, as SoxD and SoxE proteins have been found to homodimerise and act as their own cofactors (Lefebvre et al., 2007). Thus, the Sox-partner code can either stabilise a cell type or drive subsequent stages of tissue differentiation. Sequence-level specificity of Sox genes and their binding partners allows them to target specific subsets of genes in different cell types, and the interplay between these factors can both maintain and progress developmental stages. These processes exemplify the regulatory versatility of Sox genes at the molecular level.

1.1.1 SoxB subfamily

Of the Sox subfamilies, the SoxB group is particularly relevant for understanding and modelling animal development. This group contains the mammalian gene *Sox2*, which, as mentioned previously, has diverse roles related to ES cell pluripotency, and neural development. Many of the functions mediated by Sox2 depend on expression of specific cofactors, conferring a tissue specificity to its TF activity. For instance, interaction with Oct-3/4 will drive genes necessary for ES cell pluripotency, while interaction with the TF Pax6 drives the expression of crystallin to promote induction of the eye lens (Bernard & Harley, 2010; Kamachi et al., 2001; Kondoh et al., 2004). Antagonistic interactions between Sox2 and tissue-specific cofactors occur in a dosage-dependent manner to regulate cell fate in the neuroectoderm, digestive system, and peripheral nervous system, among others (Sarkar & Hochedlinger, 2013). While *Sox2* is one of the most well-studied Sox genes in vertebrates, it is far from the only major player in the SoxB subfamily.

The SoxB group is divided into subgroup SoxB1 (Sox1, 2 and 3), which generally comprises transcriptional activators, and SoxB2 (Sox14 and 21), which are generally transcriptional repressors (Uchikawa et al., 1999). These subfamilies interplay in vertebrates to regulate neural formation. This requires a balancing act between Sox1-3, which keep stem cells in a pluripotent state, and Sox21, which commits them to differentiation (Sandberg et al., 2005; Wegner, 2010). In *Drosophila*, sequence based analysis indicates that the SoxB1 group contains *SoxN* (Kamachi & Kondoh, 2013; Phochanukul & Russell, 2010). The SoxB2 subgroup in *Drosophila* contains *Dichaete*, *Sox21a,* and *Sox21b* (Crémazy et al., 2001; Nambu & Nambu, 1996; Russell et al., 1996; Uchikawa et al., 1999). However, sequence homology analysis appears to show that *Sox21a* is closer to vertebrate subgroup B2 while *Sox21b* and *D* are more likely to represent an invertebrate-specific SoxB subgroup (McKimmie et al., 2005).

Compared to *SoxN* and *D*, the fly genes *Sox21a* and *Sox21b* are less well-characterised. However, *Sox21a* has been studied in the context of intestinal stem cells (ISCs). Unlike many other SoxB members, *Sox21a* appears to function primarily in adult flies and its mutants show no defects during development, although it is expressed in the embryo gut anlage (McKimmie et al., 2005; Meng & Biteau, 2015). There is less known about the role of *Sox21b*, but it has been

found to localise to the embryonic hindgut (Crémazy et al., 2001; McKimmie et al., 2005). In wild type flies, *Sox21a* acts as a tumor suppressor and helps ISCs transition from enteroblasts to enterocytes (Meng & Biteau, 2015; Zhai et al., 2015); when faced with cellular damage resulting in oxidative stress, *Sox21a* expression decreases and results in a feedback loop that proliferates differentiation-defective enteroblasts that help with tissue repair (J. Chen et al., 2016). Its proliferative and regenerative properties come as a result of being downstream of the JNK, ERK, and JAK/STAT pathways (Meng & Biteau, 2015; Zhai et al., 2017), and DamID experiments suggest that it may join as a cofactor with the repressive TF Capicua (Cic) in ISCs and enteroblasts to control differentiation downstream of the EGF pathway (Doupé et al., 2018). While less is known about *Sox21b*, sequence analysis has shown that both *Sox21a* and *Sox21b* contain introns in their HMG box, and that a conserved genomic cluster of *D*, *D* 3' regulatory sequences, *Sox21b*, and *Sox21a* (McKimmie et al., 2005; M. J. Wilson & Dearden, 2008) appears in several analysed insect species (Figure 2). This points to a model in which these three genes, as well as *SoxN*, may have evolved as a result of redundant duplication and specialisation of an ancestral SoxB gene.

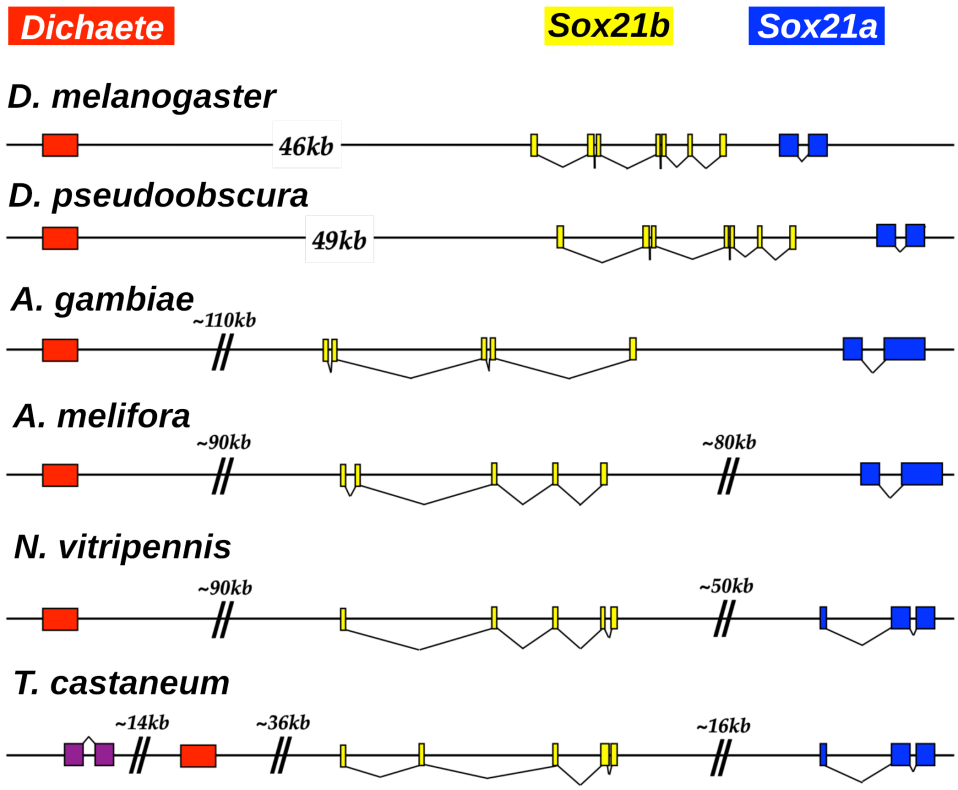

Figure 2: Conserved cluster of *Dichaete*, *Sox21b*, and *Sox21a*. The cluster of *D*, *Sox21b*, and *Sox21a* is conserved in several analysed insect species including the *Drosophila* species *melanogaster* and *pseudoobscura*, the mosquito *A. gambiae*, the honeybee *A. melifora*, the wasp *N. vitripennis*, and the beetle *T. castaneum*. Data adapted from McKimmie *et al.* (2005) and Wilson & Dearden (2008). Modified from figure provided by S. Russell (personal communication, 28 Aug 2020).

While Sox B1 and B2 genes in vertebrates exhibit a relatively pronounced distinction between activation and repression, the respective *Drosophila* genes *SoxN* and *D* exhibit much more overlap in terms of function, localisation, and regulatory control (Neriec & Desplan, 2014). As a result, it is difficult to characterise either of them as directly homologous to any particular vertebrate SoxB gene, although given the functional arguments below, it is likely that SoxN is a true orthologue of the SoxB1 family. In common with the vertebrate SoxB1 family, SoxN and D show a degree of functional redundancy which is of some interest, as identifying the shared and unique properties of the two genes can elucidate much of the behavior of SoxB TFs in animal systems. Interestingly, *SoxN* and *D* are known to be intronless in insects and vertebrate SoxB1 genes have been confirmed to be intronless in humans, mice, and chickens (McKimmie et al., 2005; Uchikawa et al., 1999). Because *Drosophila* genes often contain regulatory sequences in their first intron (Marais et al., 2005), these genes represent a subset of Sox factors that can be swapped into the endogenous *SoxN* or *D* loci without introducing secondary regulatory effects. This primes *Drosophila* as a useful system for studying the effects of introducing exogenous SoxB genes to the fly genome.

Both *SoxN* and *D* are expressed in the neuroectoderm of the developing *Drosophila* embryo, where they overlap in the medial and intermediate columns. In addition, *D* is expressed in the ventral midline from its formation, while *SoxN* shows expression in the lateral column of the neuroectoderm where *D* is absent (Crémazy et al., 2000; Phochanukul & Russell, 2010). This tissue specificity supports functional differences between the two genes during development, as *D* is necessary for proper midline formation (Soriano & Russell, 1998) while SoxN is necessary for lateral neuroblast differentiation (Overton et al., 2002). In the medial and intermediate neuroectoderm where both genes are expressed, single mutants for either gene display only mild Central Nervous System (CNS) phenotypes, but double mutants show severe neural hypoplasia (Buescher et al., 2002; Overton et al., 2002), suggesting that these genes can partially compensate for each other in the tissues where they are coexpressed.

The functional differences between SoxN and D may be better understood through their DNA binding profiles. DNA adenine methyltransferase identification (DamID) and ChIP experiments have shown both TFs binding to 1,890 common genes, with S*oxN* uniquely binding to 1,649 genes and *D* uniquely binding to 1,753 (Ferrero et al., 2014). In mutants that were null for either gene, DamID using the remaining TF showed that both *SoxN* and *D* were able to bind some of the missing TF's unique targets, indicating functional compensation at the genomic level. However, this was not the only behavior exhibited, as some regions demonstrated *de novo* binding of the remaining TF (Ferrero et al., 2014). This illustrates that on the level of molecular binding, the partial redundancy lets either TF rescue some of the other's function, but also causes behavior that deviates from the wild type. Later experiments showed not only that *SoxN* and *D* share common binding targets in *Drosophila melanogaster*, but that genes targeted by both TFs were more likely to display binding conservation in *Drosophila simulans*, suggesting

conservation not only of the TFs themselves but also of their regulatory networks (Carl & Russell, 2015). While *SoxN* and *D* share several similarities, it is also informative to examine the areas in which they are unique. Their differences represent clues to their non-redundant functions in *Drosophila*, and many of their unique functions are preserved in related species.

The *Dichaete* gene was independently identified by two different groups who noted its unique role in embryonic segmentation. Russell *et al.* showed that the HMG gene known as *Sox70D* corresponded to the known mutation *Dichaete* and then demonstrated that it is necessary in early embryogenesis for normal segmentation where it regulates the pair-rule segmentation genes *even-skipped*, *hairy*, and *runt* (Russell et al., 1996). Simultaneously, Nambu & Nambu identified the gene as *fish-hook* (*fish*) based on its expression pattern in the embryo and identified similar segmentation defects in mutants, also suggesting roles in CNS development (Nambu & Nambu, 1996). Assays based on lacZ reporters showed that the sequences flanking *D* contain multiple enhancer elements that determine its complex expression across embryogenesis, and identified functions in hindgut development, in *hedgehog* (*hh*) and *decapentaplegic* (*dpp*) signalling, and in brain development (Sánchez-Soriano & Russell, 2000).

Later work showed that *D* also plays a role in CNS patterning along the dorsoventral (DV) axis and that it works downstream of *Epidermal growth factor receptor* (*Egfr*) and in concert with both *intermediate neuroblasts defective* (i*nd*) and *ventral nerve cord defective* (*vnd*) to determine the cellular fate of neuroblasts at different parts of the DV axis (Zhao & Skeath, 2002). Work in the CNS established not only that *D* was expressed in the CNS midline, but also that it was necessary for proper midline formation (Soriano & Russell, 1998). In this role D forms complexes with Single minded and the POU domain protein Ventral veins lacking to regulate the expression of *Slit*, a midline-expressed gene responsible for normal axon formation at the midline (Ma et al., 2000). Of interest, these studies also showed that Sox2 could functionally substitute for Dichaete in the midline. Screens with dominant-negative variants of *D* identified *commisureless* (*comm*) and *asense* (*ase*) as direct targets, and also found other targets specific to tissue type and developmental stage (Shen et al., 2013). ChIP-array and DamID binding data along with gene expression data identified *D* as a regulatory hub during embryogenesis, with many of its direct targets serving also as conserved targets for the related TF *Sox2* in mammals (Aleksic et al., 2013). These experiments highlight the diversity of the known functions performed by *Dichaete*, ranging from developmental regulation of segmentation to promotion of neurodevelopment. In particular, the unique segmentation and axis patterning phenotypes of *Dichaete* and its close homologues appear to be conserved in other arthropod species, even though they do not share the exact same expression patterns as in *Drosophila* (Clark & Peel, 2018; Phochanukul & Russell, 2010; Paese et al., 2018).

In contrast, the functions of *SoxN* appear much more specific to CNS development. *SoxN* was first identified as part of a PCR screen of embryonic cDNA, and its expression was found to be under the control of zygotic DV patterning genes such as *dpp* and *twi*, linking its expression

to the patterning established by *D* (Crémazy et al., 2000). SoxN expression was found to colocalise with D expression in the CNS and PNS, with unique expression in the lateral column and no expression in the midline (Crémazy et al., 2000; Phochanukul & Russell, 2010). *SoxN* was found to be necessary for neuroblast formation, with mutants defective in the lateral column of the neuroectoderm where *D* is unable to compensate for its loss; this phenotype was more pronounced in mutants also lacking *vnd* or *ind*, suggesting that these TFs interact with both SoxN and D to give rise to ventral and intermediate neuroblast formation (Buescher et al., 2002; Overton et al., 2002; Zhao & Skeath, 2002). The complexities of these interactions were further explored in the context of ubiquitous ectopic *Egfr* expression, which caused *vnd* and *ind* to be expressed more in the lateral neuroectoderm where they could then act as *SoxN* cofactors; this suggests both spatial and temporal regulation of *vnd* and *ind* by *Egfr* in wild-type neuroblast development (Zhao et al., 2007). The neuroblast formation phenotype of *SoxN* was also found to work upstream and in parallel with the proneural *achaete/scute* gene complex (*AS-C*) genes, with both *SoxN* and the *AS-C* members *achaete* (*ac*), *scute* (*sc*), and *lethal of scute* (*l'sc*) contributing to neuroblast formation through separate mechanisms (Buescher et al., 2002; Overton et al., 2002). Thus, several of the major proneural genes involved with segment patterning (Hartenstein & Wodarz, 2013) have interactions with both *SoxN* and *D* during the course of neurogenesis.

Later work highlighted the role of *SoxN* in signalling pathways such as Wg/Wnt, where it was shown to act as a negative regulator of *wingless* (*wg*) activity in flies through interaction with the HMG-containing TF Pangolin (Lef/Tcf) (Chao et al., 2007). *SoxN* was found to work in concert with *D* and antagonistically with *wg* to activate *shavenbaby* (*svb*), leading to the formation of trichome projections in the epidermis (Overton et al., 2007). This indicates the possible relevance of *SoxN* and its homologues to canonical vertebrate and invertebrate signalling pathways. While its functions in early CNS development exhibit conservation in other species, several of its functions later in development appear unique to *Drosophila*, suggesting a model in which *SoxN* evolved from an ancestral metazoan SoxB and kept its basic functions while also undergoing additional functional specialisation specific to flies (Ferrero et al., 2014). Altogether, these studies point to ways in which *SoxN* and *D*, despite their unique expression patterns and functions, help regulate some of the same systems during development.

1.2 Evolution and specification of SoxB genes

Sox genes in different animal species exhibit a high degree of structural and functional similarities, but also exhibit unique species-specific functions. Given the relative functional conservation of its broad subfamilies across species, this raises the question of how Sox genes in their current form originated and how they gained their own specialised functions. Evidence points to an evolutionary history of Sox genes in the animal kingdom in which Sox genes

originated from duplication of an ancestral HMG gene in metazoa or choanozoa (Jager et al., 2006), suggesting that diversification of the Sox gene family occurred prior to the formation of the Bilateria clade that contains both arthropods and chordates. Within the animal kingdom, there has been a stable core set of Sox genes in subfamilies B through F, indicating that the last common ancestor (LCA) of chordates contained this core set in addition to a SoxH gene (Heenan et al., 2016), whereas the arthropods inherited only the core set of genes.

Specifically for SoxB genes, phylogenetic reconstruction based on sequence alignment of HMG domain genes has suggested that the bilaterian LCA contained a SoxB1 gene and a SoxB2 gene, and that both of these genes followed separate evolutionary paths following the divergence of arthropods from chordates (Zhong et al., 2011). Prior models had suggested that SoxB genes diversified through separate genome duplication and tandem duplication events following the deuterostome/protostome split (McKimmie et al., 2005), but the presence of SoxB1 and SoxB2 in the bilaterian LCA suggests that a tandem duplication event occurred prior to the split (Zhong et al., 2011). In humans, it is suggested that *Sox1-3* evolved from genome duplication events affecting the *SoxB1* copy, while *Sox14* and *Sox21* arose from genome duplications of the *SoxB2*. Meanwhile, in insects, the model suggests that *SoxB2* underwent two tandem duplications to give rise to *Sox21a*, *Sox21b*, and *Dichaete*, and that *SoxB1* specialised to give rise to *SoxNeuro* (Figure 3) (Zhong et al., 2011). This explains the presence of the conserved cluster of *D-Sox21a-Sox21b* and their related enhancer sequences present in all sequenced insect species to date (McKimmie et al., 2005; Paese et al., 2018; Phochanukul & Russell, 2010).

The evolutionary history of SoxB in insects raises the question of whether there are also functional commonalities between related genes in different species. Studies comparing Sox proteins in *D. melanogaster*, *A. gambiae*, the honeybee *Apis mellifera*, the wasp *Nasonia vitripennis*, and the beetle *Tribolium castaneum* (Tc) found that each of these insect groups contain eight or nine Sox genes in total, with *SoxB* as the largest subgroup with at least four members per species (M. J. Wilson & Dearden, 2008). DamID binding analysis of *SoxN* and *D* in four *Drosophila* species reveals that while the common targets of both TFs display some binding conservation in other species, suggesting functional importance of the shared targets, binding site turnover as a result of compensatory evolution is proportional to phylogenetic distance between species and may account for changes in gene expression as well as function (Carl & Russell, 2015).

Tandem duplication and genome duplication events explain why there are so many *SoxB* genes in insects, but do not on their own explain their functional similarities and differences. In the case of neofunctionalisation, SoxB genes would be expected to show changes in their coding sequence or their flanking regulatory sequences. However, it is known that the *SoxB2* sequence cluster in *Drosophila* is highly conserved across the insects (McKimmie et al., 2005). The shared regulatory controls therefore point to a model in which SoxB factors share a network of conserved gene targets to maintain the robustness of SoxB downstream regulation

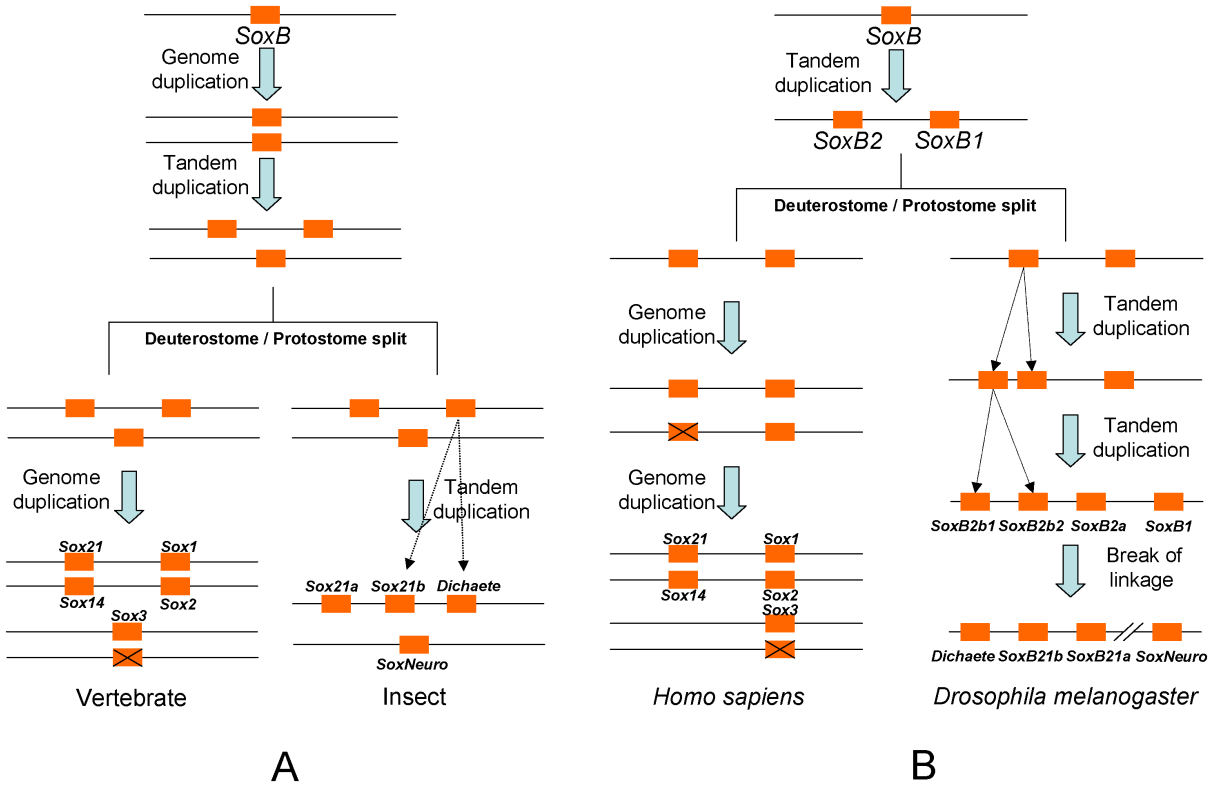

Figure 3: Models for SoxB evolutionary history. The diversity in *SoxB* genes was attributed to both tandem duplications and genome duplications. **(A)** McKimmie *et al.* (2005) proposed that *SoxN* and *D* diverged due to an ancestral genome duplication, followed by a tandem duplication that gave rise to *Sox21a* and an insect-specific tandem duplication that resulted in *Sox21b*. **(B)** However, Zhong *et al.* note that the bilaterian LCA possessed both *SoxB1* and *SoxB2*, suggesting that an ancestral tandem duplication occurred first; further tandem duplications gave rise to the conserved cluster in insects while genome duplications and pseudogenisation contributed to the current vertebrate *SoxB* organisation. Reprinted from Zhong *et al* (2011).

while individual SoxB members can pick up different targets to specialise their expression and targeting spatially and temporally (Carl & Russell, 2015; Neriec & Desplan, 2014). In other words, SoxB genes can maintain a shared regulatory network of conserved targets from the ancestral gene (Aleksic et al., 2013; Ferrero et al., 2014; Neriec & Desplan, 2014) while evolutionary distance from related species can allow them to pick up new expression patterns and functions (Carl & Russell, 2015).

1.2.1 Conservation in invertebrates

Because the evolutionary history of Sox genes caused an expansion of similar SoxB genes in arthropod species, it is important to examine how arthropod SoxB expression is both similar and different to the *Drosophila* model. Namely, *Dichaete* and its close homologues in different species appear to retain functions related to segment patterning during embryonic development,

while *SoxN* is expressed in the neuroectoderm in all characterised arthropod species do date.

Drosophila has historically been one of the most versatile systems for studying segmentation and embryogenesis, but research into other arthropods reveals ways in which its pathways are not necessarily representative of invertebrate development. Specific parts of the developmental cascade differ between species, but the most conserved components are homologues of *Drosophila* segment polarity genes *engrailed* (*en*), *wingless* (*wg*), and *hedgehog* (*hh*) (Peel et al., 2005). Other parts of the *Drosophila* system exhibit some partial conservation—each arthropod contains at least one homologous gene to a *Drosophila* pair-rule gene (Peel et al., 2005), but these homologues does not represent any broadly conserved patterning mechanisms.

Despite significant differences between the *Drosophila* model and other arthropods, the Sox gene family still has conserved roles relating to CNS development and segmentation. Janssen *et al.* examined SoxB-F expression in Tc as well as in the velvet worm *Euperipatoides kanangrensis* (Ek) and the pillbug *Glomeris marginata* (Gm) (Janssen et al., 2018). They found that *Tc-SoxN* is expressed in the developing embryonic nervous system, and *Gm-SoxN* and *Ek-SoxN* are both expressed in the brain and ventral nervous system. *Tc-*, *Gm-*, and *Ek-Dichaete* expression varies by species, but shares a common localisation in the segment addition zone (SAZ), where their respective *SoxN* genes are excluded, and is also expressed in the neuroectoderm. Tc additionally expresses both *Tc-Sox21a* and *Tc-Sox21b*, with *Sox21b* largely recapitulating the localisation patterns of *Dichaete* and *Sox21a* showing much weaker overall expression. Similar experiments in all three organisms reveal different localisation patterns in SoxC-F and suggest conserved functional roles despite their different expression patterns, with SoxC genes taking a greater role in the developing CNS than in *Drosophila.* Saliently, the expression of *Dichaete* is related to segmentation in all of these arthropods, and localisation in the SAZ suggests pair rule gene-like expression that may be present in the LCA but lost in some sister clades such as *Apis* (Janssen et al., 2018).

More practically, *Dichaete* regulation explains how related processes may govern the development of long germ band and short germ band insects. While all adult insects share similar body organisation, the developmental processes they use to pattern the anterior-posterior (AP) axis bear marked differences—long germ band insects like *Drosophila* pattern all segments of the developing embryo simultaneously during the blastoderm stage of development, while intermediate and short germ band insects like *Tribolium* first specify only anterior segments representing the future head and thorax, and then extend patterning to the entire embryo by extending at the posterior segment in a manner akin to vertebrate segmentation (Liu & Kaufman, 2005). Because of the differences in germ band development, the regulation of these two processes was thought to be substantially different until Clark & Peel suggested that a shared cascade of *Caudal*, *Dichaete*, and *Odd-paired* may explain both mechanisms (Clark & Peel, 2018). Spatiotemporal analyses of these three genes in *Drosophila* and *Tribolium* show that while both organisms express these factors to regulate pair-rule genes and influence

germ band patterning, *Drosophila* tends to express these factors as discrete on-off sequential switches while *Tribolium* expresses them simultaneously but varies their spatial organisation over time, suggesting *Dichaete* operates within a conserved developmental framework that differs in regulation rather than substance (Clark & Peel, 2018). This result suggests that while arthropods may share conserved Sox regulatory networks, they can still regulate expression spatiotemporally to influence the behavior of an individual Sox factor.

The role of Sox genes has been examined not only in arthropods of the class Insecta but also in Arachnida. The spiders *Parasteatoda tepidariorum* (Pt) and *Stegodyphus mimosarum* (Sm) have both undergone whole genome duplications during their evolutionary history, and display higher copy number of genes derived from *Sox21a* and *Sox21b*, as well as of genes derived from SoxC-F (Bonatto Paese et al., 2018). A spider gene *Sox21b-1* related to *D* and *Sox21b* is expressed in the SAZ of both Pt and Sm, as was the case with *Tc-*, *Gm-*, and *Ek-Dichaete*, pointing to a conserved role that dates back to some arthropod LCA (Bonatto Paese et al., 2018; Janssen et al., 2018). In Pt, *Dichaete* is not involved with segmentation, but *Sox21b-1* is known to regulate *fkh* and *h* in a gap gene-like fashion in the anterior parts of the embryo as well as *Dl* in the posterior SAZ (Paese et al., 2018). While *Dichaete* is not specifically involved here, it remains apparent that *Dichaete*-like genes play a conserved role in segmentation of arthropods, and that SoxB genes represent a diverse set of genes with shared regulatory networks and conserved functions.

1.2.2 Conservation in vertebrates

The presence of the SoxA gene *SRY* is a unique but not universal feature of vertebrates, present in placental and marsupial mammals but not in their sister monotreme clade, suggesting that *SRY* may originate in the LCA of eutherians and marsupials (Katsura et al., 2018; Wallis et al., 2007). PCR-based assays in mice have revealed that *SRY* is expressed in the urogenital ridge of males but not females, pointing to the role of *SRY* in testis differentiation (Gubbay et al., 1990). This machinery appears conserved in mammals, and even has homologues in yeast mating-type proteins (Sinclair et al., 1990). The HMG box contributes to this sex differentiation phenotype, and studies of XY females with gonadal dysgenesis reveal that all identified mutations to the *SRY* gene were related to an open reading frame (ORF) within the HMG domain (McElreavy et al., 1992), suggesting that the DNA binding behavior of the domain is necessary for proper functioning of the protein. SRY interacts with other Sox proteins containing an HMG box, such as Sox8 and Sox9, as part of its role in differentiating the testis and maintaining fertility in both humans and mice (Jiang et al., 2013).

Vertebrates maintain a large degree of Sox subgroup conservation, with the LCA of the chordate phylum containing SoxB-H and SoxB featuring the largest number of member genes (Heenan et al., 2016). In contrast to the conserved roles of SoxB in arthropod segmentation and neurodevelopment, vertebrate SoxB genes are most associated with neurogenesis and

maintaining pluripotency in ES cells (Karnavas et al., 2013; Sarkar & Hochedlinger, 2013). As with invertebrates, the diversity in SoxB is likely due to a tandem duplication of the ancestral SoxB gene, resulting precursors to the SoxB1 and SoxB2 subgroups, followed by genome duplications that resulted in human and vertebrate lineages containing *Sox1-3* in the B1 group and *Sox14* and *Sox21* in the B2 group (Zhong et al., 2011).

The role of SoxB in neurogenesis has been studied in the vertebrate systems of chickens, mice, and humans, among others. As mentioned previously, the functional split between SoxB1 activation and SoxB2 repression is more clearly defined in vertebrates as compared to invertebrate systems, though there is some evidence of SoxB1 factors also acting as repressors (Karnavas et al., 2013). Studies in chick embryos have revealed that *Sox1-3* maintain pluripotency in NPCs by blocking the downstream effects of proneural basic helix-loop-helix (bHLH) TFs that drive differentiation, and that these bHLH proteins can upregulate *Sox21* to counter the effect of *Sox1-3* and commit cellular fate (Bylund et al., 2003; Sandberg et al., 2005). Thus, SoxB2 repressors above a certain threshold are able to overcome the repressive effect of SoxB1 on proneural genes to drive differentiation (Episkopou, 2005). The balance of activating and repressing SoxB proteins controls development spatially as well as temporally. *In situ* probes of developing chicken embryos show that Sox2 largely defines the CNS expression zones of SoxB1 factors, and that the coexpression of SoxB2 factors occurs in tissues that have undergone terminal differentiation (Uchikawa et al., 1999).

SoxB factors largely exhibit regulatory conservation between species, but comparisons of SoxB studies in chicken and mice have revealed species-specific effects (Uchikawa et al., 2011). Many of the expression patterns of Sox2 shared between chicken and mouse embryos have been attributed to conservation of enhancers within 50 kb of the *Sox2* sequence, with sequential activation of enhancers driving spatial and temporal specificity in different CNS tissues. Interspecies conservation of enhancer sequences believed to be common to all vertebrates also highlights *Sox21*, whereas enhancer specificity of *Sox1* and *Sox3* are more variable and species-dependent (Uchikawa et al., 2011; Woolfe et al., 2005). *Sox2* activation and *Sox21* repression in mice models are therefore relevant to the study of SoxB expression in humans. In mice, *Sox2* is known to be necessary for proper neural formation, as mutants exhibit deformed cerebra and a depletion of NPCs (Ferri et al., 2004). Studies indicate that SoxB1 and Sox11 factors can interact as a cofactor with class III POU TFs and then activate the expression of the neural stem cell marker *Nestin*, which has binding sites for both types of TF in its enhancer (Tanaka et al., 2004). Further examination of these players reveals that the Sox factors act sequentially, with Sox2 in ES cells preselecting genes that are later bound by Sox3 in NPC phase, and Sox11 binding these targets during neuron differentiation (Bergsland et al., 2011). This sequential binding mechanism can provide temporal control for the activation of target genes, as Sox2 primes Sox3 to outcompete Sox11 for binding sites before differentiation is ready. Humans also exhibit similar sequential expression of Sox2 and Sox11 during neurogenesis,

with Sox2 binding data revealing it as a pioneer for eventual Sox11 targets (Dodonova et al., 2020; Mu et al., 2012). Thus, the conservation of SoxB proteins during vertebrate neurogenesis is likely driven by conservation in both enhancers and gene targets.

The role of SoxB1 factors in priming genes for differentiation in the nervous system is underscored by their role in holding off differentiation until the proper time. Indeed, *Sox2* can act as a reprogramming factor to drive the formation of induced pluripotent stem cells (iPSCs) or induced neural stem cells (iNSCs), transforming differentiated fibroblast tissues into newly pluripotent populations with the help of cofactors such as Oct4 and Myc (Sarkar & Hochedlinger, 2013; Takahashi & Yamanaka, 2006). In this capacity, other SoxB1 proteins can act redundantly with Sox2 due to their shared ability to partner with the POU domain protein Oct4, a trait not shared by Sox TFs from other families (Nakagawa et al., 2008). The redundancy of SoxB1 proteins in maintaining ES cell pluripotency means that knocking out any individual SoxB1 factor will not cause major defects due to the ability of the remaining factors to provide functional compensation in the tissues where they are usually coexpressed (Sarkar & Hochedlinger, 2013). The pluripotency phenotype of mammalian SoxB factors can be recapitulated by the endogenous SoxG factor Sox15, as well as by several homologous SoxB1 factors taken from invertebrates (Niwa et al., 2016).

1.2.3 SoxB Rescues

In *Drosophila*, experiments with *SoxN* and *D* mutants have indicated a high degree of functional redundancy, as is the case with mammalian *Sox1-3*, but Gal4-UAS controlled rescue experiments with both fly and mammalian SoxB genes suggest more complex interactions. Soriano and Russell first demonstrated that *Drosophila* midline phenotypes in *Dichaete* mutants could be rescued with the introduction of the mouse *Sox2* gene (Soriano & Russell, 1998). While mouse *Sox2* provides significant rescue of *D* midline phenotypes, *SoxN* and *Sox1* do not (Overton, 2003). Conversely, *Sox1* can rescue some *SoxN* lateral CNS phenotypes while *D* and *Sox2* cannot (Shen et al., 2013; Overton, 2003). Fly SoxB1 genes also appear able to rescue their mammalian homologues. When endogenous *Sox2* is replaced with fly *SoxN* in mouse embryonic stem cells, the fly gene is capable of rescuing and even enhancing the usual stem cell pluripotency phenotype, and also of supporting early embryogenesis (Niwa et al., 2016). Structurally, this was found to be possible because of a conserved K57 residue in both *Sox2* and *SoxN*, as well as the presence of a transactivation domain in *Sox2* (Niwa et al., 2016; Nowling et al., 2000). These interactions reveal a more complex picture of the functional compensation possible for SoxB genes both within the same species and between vertebrates and invertebrates. Notably, these rescues were performed via knockdown and overexpression, highlighting the need for true genomic swaps in future investigation.

1.3 Project Aims

This review provides an examination of the diverse functions performed by the Sox gene family, specifically the SoxB subgroup that is responsible for aspects of neurodevelopment and embryonic segmentation. The SoxB subfamily is so diverse because of the high degree of redundancy between its members due to gene duplication events and shared regulatory networks. Paradoxically, this redundancy is what makes it possible for organisms to specialise these SoxB genes for additional functions.

While much is known about the roles of *SoxNeuro* and *Dichaete* in *Drosophila* as well as more broadly in arthropods and vertebrates, it is still of interest to examine how the function of these genes may relate to their vertebrate homologues. This project aimed to study the functional conservation between these fly SoxB genes and their mammalian homologue *Sox2* by creating knock-in constructs to replace each endogenous gene with *Sox2* and examining the phenotype of the resulting organism. I aimed to create CRISPR-based replacement constructs using PCR cloning and Gibson-like assembly. Conditioned on the success of these constructs and the viability of the knock-in flies, it will then be possible to examine the CNS phenotypes of the resulting flies and to perform ChIP-seq to examine how the binding profile of the exogenous *Sox2* differs from that of the endogenous SoxB genes. This would paint a clearer picture of the ways in which *Sox2* can compensate for *SoxN* and *D*, as well as of how the regulatory pressures on each locus may influence expression and target binding.

Another aim of this project is to examine the regulatory networks that *SoxN* and *D* participate in by examining available single cell RNA sequencing (scRNA-seq) datasets. I analysed publicly available data from the *Drosophila* embryo, larval brain, and adult ventral nerve cord to characterise the cell subgroups expressing either *SoxB* gene. Analysis of other marker genes expressed in these cell subgroups can provide insight to the shared regulatory networks that make *SoxN* and *D* redundancy possible, as well as to their differences. Overall, this project seeks to integrate both molecular and computational approaches to better characterise the field's understanding of the interplay between these two key Sox genes.

Chapter 2

Swapping *Drosophila* and mammalian Sox genes

Previous experiments have established some degree of functional conservation between different members of the insect and mammalian Sox gene family. Overexpression-based assays have shown that *Sox2* can rescue the *D* midline phenotype in the *Drosophila* embryo, while *Sox1* can rescue *SoxN* lateral phenotypes (Shen et al., 2013; Soriano & Russell, 1998; Overton, 2003). While these experiments provided insight to the conservation of Sox factors between different species, they are based on overexpression systems that rescue null mutations, and therefore do not demonstrate the true degree of functional conservation. Furthermore, previous experiments in the Russell lab using such overexpression approaches indicate that phenotypic rescue is variable depending on the transgene and driver (Overton, 2003). An alternative approach is to employ a functional replacement or knock-in system where the coding sequence (CDS) of the endogenous *Drosophila* gene is replaced entirely with a mammalian SoxB CDS. Experiments with the reverse approach have previously demonstrated that replacing *Sox2* with *Drosophila SoxN* in mice allows normal maintenance of stem cell pluripotency (Niwa et al., 2016); I aimed to explore whether the reverse may also be true and whether it may also be the case for *Dichaete*.

Such a system would allow the mammalian gene to be expressed under the control of the endogenous SoxB enhancers, providing better insight into the functional conservation between an exogenous mammalian SoxB gene and the partially redundant fly *SoxN* and *D* genes. The objective of the wet lab portion of this project was therefore to create fly knock-in lines replacing endogenous *SoxN* or *D* with the mouse *Sox2* (*mSox2*) gene. Such fly lines would open the door to further exploration of how *mSox2* is capable of recapitulating the function of the endogenous gene. Differences in $SoxN^{mSox2}$ *and* D^{mSox2} phenotypes could potentially reveal areas in which the two endogenous genes are not redundant, shedding light on their unique functions in *Drosophila* development. Ongoing work in the Russell lab is aimed at swapping *Dichaete* and *SoxN* coding sequences, providing complementary information on the functional equivalence of the fly SoxB genes.

In the past, knock-ins based on homologous recombination were comparatively difficult and time-intensive. Site-specific mutations in the *Drosophila* genome involved using FLP-FRT site-specific recombination in conjunction with the I-*Sce*I site-specific endonuclease to induce double strand breaks (DSBs) (Rong et al., 2002; Rong & Golic, 2000). In contrast, newer CRISPR/Cas9 based methods allow for a more robust and streamlined approach. The crux of this approach involves designing a donor template that contains three major components—an arm homologous to the 5' genomic sequence (left homology arm; LHA), the CDS of the knock-in gene, and an arm homologous to the 3' genomic sequence (right homology arm; RHA). Once this donor template is created, it can be injected into *Drosophila* embryos expressing the endonuclease Cas9 in the germ line, along with two chimeric guide RNA (gRNA) molecules that direct Cas9 to induce targeted DSBs near the junctions between the endogenous gene and its LHA/RHA, excising the genomic CDS. From there, homology directed repair (HDR) will match the genomic LHA/RHA to the LHA/RHA of the donor template, integrating the knock-in CDS to the location where the endogenous gene was excised (Gratz et al., 2014; Gratz, Harrison, et al., 2015; Gratz, Rubinstein, et al., 2015).

Using modern CRISPR techniques, it is therefore possible to create *mSox2* knock-in constructs simply by synthesising two separate donor plasmids: one with *mSox2* flanked by *SoxN* homology arms and one flanked by *D* homology arms. Once these components are created, generation of the corresponding $SoxN^{mSox2}$ and D^{mSox2} flies depends on microinjection of the donor templates and their corresponding gRNAs, followed by screening of the resultant flies for successful integration of the exogenous *mSox2* knock-in. The primary goal for this section of the project was to create plasmids that could serve as donor templates.

The general approach to creating such templates relied on synthesising individual fragments and then annealing them together in order to create the final plasmid. For both "Sox2 in SoxN" and "Sox2 in D" constructs, the target plasmid consisted of four major components: the *SoxN/D* LHA, the *mSox2* CDS, the *SoxN/D* RHA, and a pUC19 backbone with a multiple cloning site that can be cleaved by the enzyme *Sma*I. Aside from the backbone, each of these fragments was synthesised with a minimum 25 nt overlap with the adjacent fragment, allowing for a final ligation of these overlapping ends using a Gibson-like assembly protocol provided by New England Biolabs (R. Chao et al., 2015). Once assembled, the resultant plasmids could be verified with PCR, used to transform *Escherichia coli* cells, and amplified to create a high-concentration stock of the assembled plasmid. The bulk of these wet lab experiments focus on the process of synthesising fragments and ligating them to create donor templates. Figures 4 and 5 depict schematics of the desired target plasmids, along with the PCR primers used to generate each of the fragments. Supplementary Figure 1 shows the features of plasmid 1444, the template DNA that was used to clone the *mSox2* sequence.

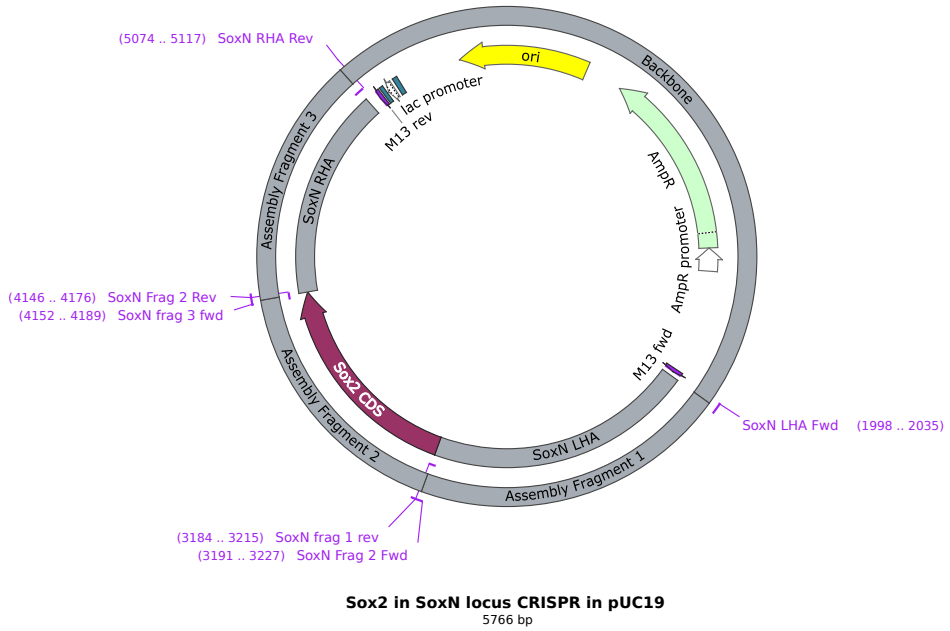

Figure 4: Schematic of "Sox2 in SoxN" plasmid construct. Assembly fragments and primers indicate the intended synthesis plan for NEBuilder Gibson-like assembly. mSox2 CDS is positioned for homology-driven repair with the LHA and RHA of the enodgenous SoxN sequence.

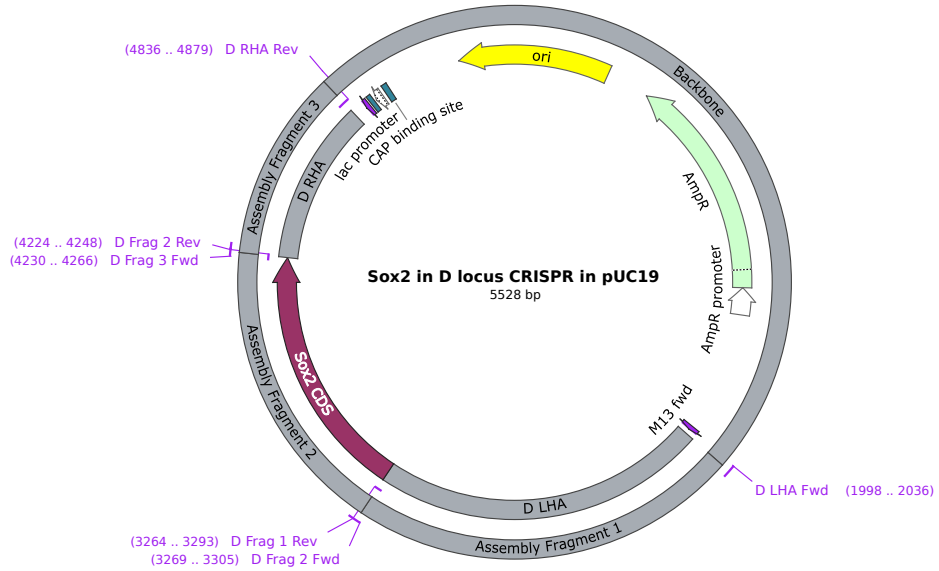

Figure 5: Schematic of "Sox2 in D" plasmid construct. As with the SoxN construct, this places the mSox2 CDS between the LHA and RHA for the endogenous D locus.

2.1 Materials and Methods

Reagents and PCR parameters for these experiments are summarised in Tables 1-5. All materials were from existing stocks or new orders in the Russell lab. Plasmid 1444, a source of the *mSox2* CDS , was provided by the lab of Prof. Jennifer Nichols (Cambridge Stem Cell Inst).

Reagent	Supplier	Purpose
pUC19 + SmaI digest	Russell Lab	Assembly vector
NEB 5-alpha chemically competent E. coli	New England Biolabs	Bacterial transformation
D in SoxN plasmid	Russell Lab	DNA source for PCR
Plasmid 1444	Jennifer Nichols Lab	DNA source for PCR
SoxN in D plasmid	Russell Lab	DNA source for PCR
NEBuilder Assembly Cloning kit	New England Biolabs	Fragment Assembly
6x loading dye	New England Biolabs	Gel electrophoresis
Agarose	Sigma-Aldrich	Gel electrophoresis
Ethidium Bromide (EtBr)	Sigma-Aldrich	Gel electrophoresis
GeneRuler 1kb, 100bp ladders	ThermoFisher	Gel electrophoresis
Lysogeny broth (LB) media and LB agar	Sigma-Aldrich	Growing E. coli
Q5 Hi-Fidelity DNA Polymerase Kit	New England Biolabs	PCR
Ampicillin	Sigma-Aldrich	Resistance selection
CloneJet PCR cloning kit	ThermoFisher	Sequencing
Nuclease-free water	Sigma-Aldrich	Solution preparation

Table 1: Reagents and suppliers used in experiments

2.1.1 Polymerase Chain Reaction

All PCR reactions were performed using the Q5 High-Fidelity Polymerase protocol from New England Biolabs as summarised in Table 2. Bulk mixes of Q5 mix were prepared in 500 µL aliquots containing all reagents except for primers and template DNA. Template DNA and primer combinations as summarised in Table 4 were used to create the 25 µL reaction mix for each fragment in separate PCR tubes. Initial conditions were based on the NEB T_m calculator tool. However, differences in length and GC content between fragments necessitated gradient PCR assays at different elongation times to optimise reaction conditions, with optimal settings for each fragment summarised in Table 5. Each PCR included a negative no-template control. Following PCR, 1 µL of each reaction was visualised using gel electrophoresis, while the remainder was stored at −20 °C for future ligation.

Following bacterial transformation, PCR was used to identify positive colonies with SoxN/D Frag 2 primers used to identify the *mSox2* CDS. Colonies were mixed in 10 µL of nuclease-free water, streaked on a separate LB agar plate for preservation, and heated for 10 minutes at 95 °C to rupture cell walls; 1 µL of this solution was then used in the place of the usual template DNA. Colonies that passed the *mSox2* CDS screen were then PCR tested for the LHA-Sox2

Item	Composition
1x Tris-Acetate-EDTA (TAE) buffer solution	- 2 M Tris base - 1 M EDTA - 1 M Acetic acid - MilliQ water to 1 L total
LB agar	- 10 g LB agar powder - 250 mL MilliQ water - 250 µL of 100 mg/mL ampicillin
LB Media	- 4 µL aliquot LB liquid media - 4 µL of 100 mg/mL ampicillin
Agarose gel	- 0.48 g agarose - 60 mL 1x TAE - 2.5 µL Ethidium Bromide
Q5 mix	- 1X Q5 Reaction buffer - 1X Q5 High GC Enhancer - 200 µM dNTP mix - 0.5 µM forward primer - 0.5 µM reverse primer - 0.02 units/µL Q5 Polymerase - 90 ng template DNA; equivalent volume of nuclease-free water for negative controls - Nuclease-free water to 25 µL total
Alkaline Lysis Solution I + 2% RNAse	- 25 mMTris-Hcl - 10 mM EDTA
Alkaline Lysis Solution II	- 0.2 N NaOH - 1% w/v SDS
Alkaline Lysis Solution III	- 3 M KOAc, pH 5.2

Table 2: Composition of solutions and mixtures used during experimentation

Primer Name	Primer Sequence	Intended Fragment
SoxN LHA Fwd	tgaattcgagctcggtaccc GGATCGATATAGTGACGG	SoxN LHA
SoxN Frag 1 Rev	catgttatacatCTTGCGGG GATTACTTCCAG	SoxN LHA
SoxN Frag 2 Fwd	taatccccgcaagATGTATA ACATGATGGAGACGGAG	SoxN^{mSox2} CDS
SoxN Frag 2 Rev	tattttcaataatTCACATG TGCGACAGGGG	SoxN^{mSox2} CDS
SoxN Frag 3 Fwd	tcgcacatgtgaATTATTGA AAATATTAACTAAAGCCC	SoxN RHA
SoxN RHA Rev	gtcgactctagaggatcccc GAGATGTTTTTGTAAAGATTTTTC	SoxN RHA
D LHA Fwd	tgaattcgagctcggtaccc CGATTGCCCTTGTCCTTTC	D LHA
D Frag 1 Rev	catgttatacatTCCAGCTA TTTTGAACAC	D LHA
D Frag 2 Fwd	caaaatagctggaATGTATA ACATGATGGAGACGGAG	D^{mSox2} CDS
D Frag 2 Rev	ctaaaactcgactTCACATG TGCGACAGGGG	D^{mSox2} CDS
D Frag 3 Fwd	tcgcacatgtgaAGTCGAGT TTTAGGTTAGAGTACAG	D RHA
D RHA Rev	gtcgactctagaggatcccc ATCTTCCAAGCTGTAAATTATTTG	D RHA

Table 3: PCR primers used for fragment cloning. Uppercase nucleotides represent where the primers anneal to the source DNA for their intended fragment, while lowercase nucleotides represent overlap with other fragments during ligation.

Fragment Name	DNA Source	Fwd Primer	Rev Primer	Expected length (bp)
SoxN LHA	D in SoxN	SoxN LHA Fwd	SoxN Frag 1 Rev	1186
SoxN^{mSox2}	Plasmid 1444	SoxN Frag 2 Fwd	SoxN Frag 2 Rev	960
SoxN RHA	D in SoxN	SoxN Frag 3 Fwd	SoxN RHA rev	934
D LHA	SoxN in D	D LHA Fwd	D Frag 1 Rev	1264
D^{mSox2}	Plasmid 1444	D Frag 2 Fwd	D Frag 2 Rev	960
D RHA	SoxN in D	D Frag 3 Fwd	D RHA Rev	618

Table 4: Components for synthesising each amplification fragment.

Fragment	Anneal temperature (°C)	Anneal time (s)	Extension time (s)
SoxN LHA	65	30	90
$SoxN^{mSox2}$	65	30	90
SoxN RHA	58	30	90
D LHA	65	30	90
D^{mSox2}	65	30	90
D RHA	58.2	40	7

Table 5: Summary of PCR parameters for different fragments. All PCR setups start with a single 30-second denaturation step at 98 °C, followed by 32 three-step cycles of melting for 10 seconds at 98 °C, annealing at variable temperatures/times, and extending at 72 °C at variable times. All cycles finish with a final 120 second extension at 72 °C.

and Sox2-RHA junctions. Because colony PCR resulted in many false positives, an alternative method was to first grow each colony in liquid media, miniprep a purified plasmid, and then perform PCR screens for *mSox2* on the resultant plasmid.

2.1.2 Gel electrophoresis and imaging

Following PCR, the product of each reaction was visualised using agarose gel electrophoresis in 1x TAE with Ethidium bromide (EtBr) included for visualisation. The NEB GeneRuler ladder was used as a size marker and gels were visualised under UV light with the Syngene NuGenius imaging platform.

2.1.3 Fragment assembly

Once all fragments for a given target plasmid were synthesised, they were ligated using the NEBuilder HiFI DNA Assembly cloning kit. The three PCR product fragments were combined with SmaI-digested pUC19 in a 1:1 vector:insert ratio, following the guidelines for a four-fragment assembly. The total fragment amount was slightly under 0.5 pmol, and the fragments were combined with HiFi DNA Assembly master mix and nuclease-free water to a total reaction volume of 20 µL. A parallel assembly using the kit's pre-made positive control mix was also performed. The assembly reactions were kept at 50 °C for at least 60 minutes to facilitate ligation. Following the ligation process, 1 µL of the reaction product was used for bacterial transformation, and the remainder was stored at −20 °C.

2.1.4 Bacterial transformation

NEB DH5-alpha *E. coli* cells were transformed with the ligated plasmids. Thawed chemically competent cells from the NEBuilder assembly kit were mix with 1 µL of plasmid product,

chilled on ice for 30 minutes, heat shocked at 42 °C for 30 seconds to induce transformation, and then chilled for another 2 minutes on ice. The transformed *E. coli* cells were added to 300 µL of SOC outgrowth medium and incubated for an hour at 37 °C in a 300 RPM shaker. A 100 µL portion of the culture was plated and spread on a warmed LB agar and ampicillin plate, and then incubated overnight for no longer than 20 hours. For the NEBuilder assembly products, the process was repeated in parallel with the positive control product. Positive colonies were grown in liquid media and combined in a 1:1 ratio with 50% glycerol for long-term storage at −80 °C.

2.1.5 Miniprep plasmid extraction

After overnight growth of *E. coli* in liquid media, alkaline lysis was used to isolate and extract plasmid DNA. A total of 3 mL of *E. coli* culture was pelleted at 8000 RPM and then resuspended in 150 µL of Solution I + 2% RNAse. This was gently mixed with 150 µL of Solution II to lyse cell walls followed by 150 µL of Solution III to neutralise the pH of the mixture. The reaction was chilled at −20 °C for 5 minutes, and the DNA was precipitated with 500 µL of isopropanol, followed by another −20 °C chill for 10 minutes and a 5 minute centrifugation at 13000 RPM. The supernatant was discarded and the remaining pellet was washed with 400 µL of 100% ethanol to remove excess salt. This was followed by centrifuging again for 5 minutes at 13000 RPM and discarding the supernatant. The pellet was completely dried for 10 minutes in an open tube and then resuspended in 20 µL of nuclease-free water to produce the final product.

2.1.6 Sanger sequencing

Two different approaches were used for sequencing, one involving cloning the *mSox2* CDS into a ThermoFisher CloneJET system for sequencing with pJET1.2 universal primers, and one involving new primers specifically for sequencing from the plasmid 1444 template. In both cases, each Sanger sequencing attempt took two separate 10 µL reactions: 6.5 µL of nuclease-free water, 1 µL of template DNA at a starting concentration of 90 ng/µL, and 2.5 µL of either the forward or reverse 10 µM primer.

To clone the *mSox2* DNA into the pJET1.2 vector, 1 µL of raw PCR product from the reaction that used $SoxN^{mSox2}$ primers was combined on ice with 10 µL of reaction buffer, 1 µL of T4 DNA ligase, 1 µL of the pJET1.2 blunt cloning vector, and 7 µL of nuclease-free water. The ligation reaction occurred for 5 minutes at room temperature, and was then used to transform chemically competent DH5-alpha *E. coli* cells. After incubation at 37 °C overnight on an ampicillin selection plate, a colony was selected, grown in liquid media, and plasmid DNA was extracted via miniprep. This was used for sequencing in conjunction with the pJET1.2 forward primer 5'-d(CGACTCACTATAGGGAGAGCGGC)-3' and the reverse primer 5'-d(AAGAACATCGATTTTCCATGGCAG)-3'. The sample was sent to the Genewiz sequencing

service for analysis.

When this approach failed to yield any results, it was repeated with new custom primers designed specifically for the plasmid 1444 template. The computational tool Primer3 was used to optimise a set of primers between 100-300 bp outside of the *mSox2* CDS (Untergasser et al., 2012). The sequences for these primers are 5'-d(GCTTCTGGCGTGTGACCG)-3' for the primer "Sox2 Sanger Fwd" and 5'-d(CACACCGGCCTTATTCCAAG)-3' for "Sox2 Sanger Rev." While this approach did result in a sequence output, the sequence was half as long as expected and low-fidelity. No further sequencing attempts occurred due to the onset of COVID-19 shutdowns the following week.

2.2 Results

Fragments were synthesised from DNA templates, and PCR conditions were optimised so that consistent fragments could be generated from a base plasmid or from colony miniprep templates. The expected length of each fragment is shown in Table 4, and the optimised PCR conditions are summarised in Table 5. For both the "Sox2 in SoxN" and "Sox2 in D" donor template plasmids, synthesis of each individual fragment was successful, but transformation of the ligation product proved more problematic. The SoxN plasmid successfully ligated, generated transformed bacterial colonies, and was analysed for the presence of each component fragment (Figure 6). Unfortunately, contamination in the bacterial stocks meant that this assembled plasmid was lost. It is not certain that the presence of the LHA, *mSox2* CDS, and RHA fragments were sufficient to determine that the assembly was successful, as secondary primer pairs to test the junction between the CDS and each homology arms yielded fragment sizes that were inconsistent; however, it is unclear that this is due to unsuccessful assembly, as significant contamination was observed in these experiments, which were conducted right before the COVID shutdown. Conversely, each of the D fragments were successfully synthesised (Figure 7), but efforts to introduce the ligated products into bacteria and select positive colonies proved unsuccessful. PCR products of each fragment necessary to construct either construct are still available, and can be used to create fresh plasmids. Because the final PCR conditions are now known, the fragments can readily be freshly synthesised and ligated without further optimisation.

While all fragments were eventually optimised, certain fragments took more effort to ensure ideal conditions. Various two-step and three-step synthesis protocols were attempted, with different annealing temperatures and extension times. LHA fragments and *mSox2* CDS fragments for both the SoxN and D constructs were readily synthesised at a 65 °C anneal temperature and 90 second extension time. The RHA fragments were considerably more difficult to produce. The SoxN RHA fragment was not generated with a 65 °C anneal temperature, but was found when the annealing condition was changed to 58 °C. The D RHA fragment

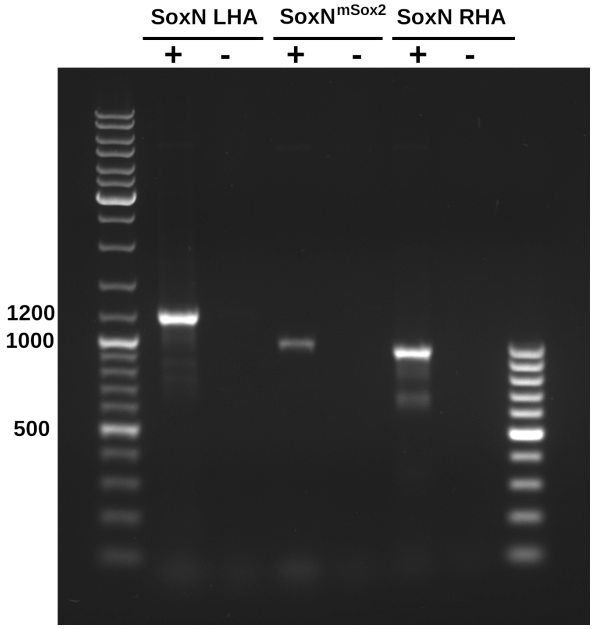

Figure 6: PCR verification of each fragment in assembled "Sox2 in SoxN" plasmid. Lane order: GeneRuler 1kb, SoxN LHA, SoxN LHA n.c., SoxN^{mSox2}, SoxN^{mSox2} n.c., SoxN RHA, SoxN RHA n.c., GeneRuler 100bp.

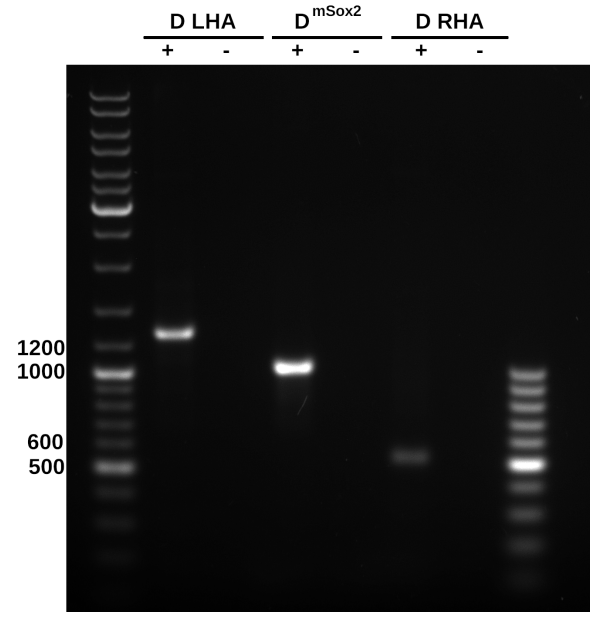

Figure 7: Unassembled fragments for "Sox2 in D" construct. These represent the stable PCR products for each assembly fragment prior to ligation. Assembly into final plasmid was unsuccessful, but fragments prior to ligation were of expected lengths. Lane order: GeneRuler 1kb, D LHA, D LHA n.c., D^{mSox2}, D^{mSox2} n.c., D RHA, D RHA n.c., GeneRuler 100bp.

required multiple combinations of both annealing temperature and extension time before stable conditions were found. Gradient PCR revealed that 58.2 °C annealing temperature combined with an extension time of 19 seconds initially generated a PCR product (Figure S2), however, this was not reproducible and the gradient PCR scheme was modified under different extension times to yield a more reliable product with an extension time of only 7 seconds.

Colony PCR was used to select *E. coli* colonies that contained either the assembled SoxN or D constructs. Primers for the *mSox2* CDS fragment were used as a first-past screen, after which positive colonies were grown and miniprepped to yield a more pure DNA template for a second-pass screen that also included the LHA and RHA fragments. However, the colony PCR method yielded a high rate of false positives. While several colonies would appear to be positive for the *mSox2* CDS fragment (Figure S3), this fragment was not present in a stable form once the same PCR was repeated on purified miniprep DNA (Figure S4). Therefore, it became more reliable during the selection process to first grow out each colony in liquid media, use a miniprep extraction to obtain purified plasmid DNA, and then perform the first-pass *mSox2* screen using the purified DNA. While this technique yielded fewer false positives, a significant drawback was that it was more time-intensive. Differences in transformation efficiency were observed depending on the *E. coli* cells being used, as transformation with cells that were made chemically competent in the lab yielded no colonies on an ampicillin selection plate, while the high-efficiency DH5-alpha competent cells from the NEBuilder kit were consistently able to produce colonies on selection plating.

Attempts to sequence the *mSox2* CDS from plasmid 1444 proved ultimately unsuccessful. As described previously, Genewiz Sanger sequencing attempts failed due to lack of primer annealing for the pJET1.2 universal primer construct. Secondary sequencing attempts with the custom primers optimised specifically for plasmid 1444 were able to bind, but produced an essentially meaningless output, with 282 of the 435 bases marked as N despite the fact that the primers were over 1300 bp apart from each other (Figures S1, S5). It is unclear why these primers failed to result in a high-quality Sanger sequence, but verification of the identity and integrity of the *mSox2* CDS sequence from plasmid 1444 will be necessary before the ligated donor templates are ultimately used to create CRISPR knock-ins.

2.3 Discussion

Difficulties during PCR synthesis arose due to a combination of the properties of the PCR primers and of the intended target fragments. Because the assembly primers created by the NEBuilder assembly tool contained a minimum overlap of 25 nt, each primer had a relatively large estimated estimate from the NEB T_m calculator, and therefore the first attempt for most fragments involved a two-step synthesis that combined annealing and extension at 72 °C. Through trial and error, it became evident that these T_m values were actually lower in practice

as compared to the calculated theoretical values.

In addition to difficulties due to the T_m, the next biggest difficulty with PCR was due to the sequence of the fragments themselves. For both the SoxN and D RHA fragments, the lower annealing temperature of near 58 °C was likely related to the low GC content of the target fragment, as the SoxN RHA fragment was 34% GC and the D RHA was 33% GC while other fragments were closer to 50%. This finding is consistent with previous observations that high GC content of the desired PCR product correlates with higher annealing temperature (Mammedov et al., 2008). Gradient PCR (Figure S2) was instrumental in rapidly varying the annealing temperature condition to identify viable conditions; if a band of the proper size was present but faint, slightly altering the reaction conditions could produce a more consistent product.

While PCR yielded reliable results when the DNA template was a purified plasmid, false positives were frequent when attempting colony PCR screens. When performing a screen for *mSox2* CDS in transformed colonies, it was common to see positive colonies (Figure S3), but when the PCR was repeated for the purified plasmids for each colony, there was a high degree of nonspecific bands, as well as a low intensity for bands of the desired size (Figure S4), indicating that many of the supposedly Sox2-positive colonies did not actually incorporate the CDS successfully. One likely explanation for the high degree of false positives is that unassembled fragments from the ligation step may have been present in the outgrowth media that was used to plate the transformed cells. This is consistent with the fact that many of these positives disappeared after purifying the internal plasmid DNA.

Because of difficulties with Sanger sequencing, it was not possible to verify conclusively that the *mSox2* CDS contained the true *Sox2* sequence, but the proper PCR fragment length was produced consistently whether using the SoxN or D assembly primers, indicating some degree of consistency in the product under equivalent sets of primers. Possible explanations for the lack of Sanger sequence using pJET universal primers, and for the low-quality results when the process was repeated with custom primers for plasmid 1444 (Figure S5), involve the idea that Sanger sequencing simply has a lower tolerance for primer defects as compared to PCR. With PCR, reaction conditions can be modified in terms of buffer components and cycle parameters, but Sanger sequencing is a purely linear process that proceeds uniformly. Defects in primer binding will result in a lower quality product under Sanger sequencing as compared to PCR, which can better tolerate sub-optimal binding and reaction conditions. A priority for future assembly attempts will be verifying that the sequence is indeed identical to the known *mSox2* nucleotide sequence. If problems persist, it may be necessary to obtain another *mSox2* CDS from a source plasmid that is more amenable to sequencing.

2.3.1 Future Work

Assuming that it is possible to verify the *mSox2* sequence and to eventually create the desired donor plasmids, the next step would involve providing both plasmids and their associated gRNAs to the Genetics Department microinjection service for insertion into embryos expressing Cas9. Crossing these to relevant balancer stocks (*Sco/SM6a* for *SoxN* and *TM2/TM6c* for *D*), F1 progeny would be screened for the presence of the *mSox2* sequence via PCR. There are two possible outcomes for each of the replacements. One is that they will be homozygous viable and fertile, indicating that Sox2 is able to provide all wild type *Drosophila* SoxB functions. Alternatively, there may be observable phenotypes, most likely lethality, that indicate that Sox2 cannot fully fulfill the role of the endogenous genes.

Whether the knock-in mutations are viable or lethal, it would first be necessary to perform a phenotypic assessment of homozygous knock-in embryos. Immunostaining with the BP102 antibody will allow for an initial assessment of any CNS phenotypes, since these are well described for *Dichaete* and *SoxN* null mutants. Further investigation avenues, such as examining neuroblast markers including *Eagle* and *Worinu* in both knock-ins (Buescher et al., 2002; Overton et al., 2002), expression of segmentation genes like *eve* in the $Dichaete^{Sox2}$ lines (Russell et al., 1996), or neuronal markers including *Nerfin-1* or *Sema-1a* for late *SoxN* functions (Ferrero et al., 2014), will help identify processes where Sox2 cannot adequately provide normal SoxB function. Because the aforementioned genes have been previously characterised in the context of SoxB binding and mutant phenotypes, it may be possible to identify sequence contexts at mapped enhancers to understand any lack of functional rescue.

To further explore the relationship between the mammalian and fly SoxB proteins at the level of the genome, it was my plan to perform ChIP-seq analysis on the knock-in stocks to determine how the binding profile of mSox2 compares to the endogenous profiles of SoxN and D as described previously by the lab (Aleksic et al., 2013; Ferrero et al., 2014). For example, a cross of $SoxN^{mSox2}/SM6a$ x $SoxN^{GFP}/SoxN^{GFP}$ would result in progeny that are 50% $SoxN^{Sox2}/SoxN^{GFP}$ and 50% $SoxN^{GFP}/SM6a$. For stage 9-10 embryos, it will not be possible to differentiate the progeny by phenotype in sufficient quantities for ChIP-seq, so an appropriate experimental setup would be to perform ChIP with both α-mSox2 and α-GFP antibodies using the same chromatin preparations. ChIP validated antibodies are available for both Sox2 and GFP (Lodato et al., 2013; Porcelli et al., 2019). While the level of $SoxN^{mSox2}$ would only be half that of the endogenous *SoxN*, only one copy of *SoxN*—the one with the GFP tag—would be immunopurified. Additionally, because the antibodies have different binding efficiency, it will not be possible to perform a true normalisation between their binding profiles, but a qualitative comparison of binding profiles would allow for comparison between SoxN and Sox2 binding in the same cells, revealing the extent of functional equivalence at the level of genome binding. This approach would be repeated for D^{mSox2} and D^{GFP}, with similar caveats.

Similar questions have already been asked in other organisms. To study whether the

genomic sequence or the nuclear environment were more important in determining transcriptional regulation, Wilson *et al.* added an aneuploid human chromosome 21 to Tc1 mouse hepatocytes, finding that TF binding and transcription initiation along this aneuploid chromosome recapitulated the patterns present in chromosome 21 of human hepatocytes (M. D. Wilson et al., 2008). Qiu *et al.* similarly explored the levels of human chromosome 21 expression as compared to the expression of endogenous orthologues in Tc1 mouse neurons, finding a significant correlation that they attributed to sequence-level similarity between orthologous human and mouse genes; the differences in expression were attributed to sequence divergence (Qiu et al., 2016).

This implies a mode of regulation in which differences in cellular environments and nuclear transcription factors between species matters less than the primary sequence of the exogenous genes or chromosomes, at least in similar tissues. From this, we might expect that the sequence differences between fly SoxB genes and *mSox2* may make it unlikely for *mSox2* to fully rescue the endogenous phenotype, especially since there is evidence that TF binding divergence between closely related species is a function of both factor-independent variables like chromatin state and factor-dependent differences in consensus site recognition (Bradley et al., 2010). Ostensibly, any failure for *mSox2* to recapitulate endogenous function may relate to the factor's inability to recognise the proper binding sequence. Alternatively, regulatory sequence divergence for orthologous genes appears to impact TF binding affinity without qualitatively changing which loci are bound (Bradley et al., 2010; Wittkopp, 2010), so under the regulation of the fly Sox loci, *mSox2* may target and be targeted by the same network members as *SoxN* and *D*. Depending on whether or not *mSox2* recapitulates the SoxB phenotypes, it may be possible either to identify the sequence-level differences that prevent it from binding dSoxB recognition sequences or to find the identities of the regulatory targets and cofactors that it shares with the dSoxB genes.

Taken together, these proposed analyses would help elucidate the extent to which SoxB genomic binding is determined by the amino acid sequences of the proteins or by chromatin context. If this successful, it may even be worthwhile to perform single-cell RNA-sequencing to judge whether changes in binding targets for the exogenous *mSox2* also correlate with changes in expression patterns of downstream target genes and whether there are pleiotropic effects of the knock-in for tissues such as the eyes or intestine. Such differences may provide a new focus for other aspects of mutant phenotypic screening. In turn, examining mutant phenotype differences between the $SoxN^{mSox2}$ and D^{mSox2} flies can reveal how the regulatory pressures on the SoxB loci differ and can also provide a starting point for further exploration of the tissues and cofactors associated with each endogenous gene.

Chapter 3

Computational Experiments

To provide a perspective complementary to that of the proposed wet lab experiments, we decided to explore the expression of *SoxN* and *D in vivo*. Previously published single cell RNA-seq (scRNA-seq) datasets (Allen et al., 2020; Brunet Avalos et al., 2019; Karaiskos et al., 2017) provide a wealth of information on the expression of genes in the *Drosophila* nervous system and mining these datasets may provide insights into the cell types that express *SoxN* and *D,* suggest biological functions, and identify other factors that these cells express. Compared to methods such as immunostaining or the use of lacZ reporters that focus only on a gene of interest, genomic approaches allows us to define groups of cells that are positive for a gene of interest and then examine the contribution of other genes in those same cell groups. By comparing transcriptomic data across different stages of fly development, it is possible correlate expression groups with known cell types, and then to examine the other commonalities of those expression groups.

Transcriptome analysis via scRNA-seq offers a relatively unbiased way to profile the transcriptional output in a group of cells. Protocols for scRNA-seq generally involve isolating mRNA via its poly-A tail, fragmenting transcripts, reverse transcribing the fragments, and amplifying the cDNA fragments before sequencing (Hebenstreit, 2012). Fragmentation helps keep read length short, reducing 5' biases caused by differences in the polymerase progression (Mortazavi et al., 2008) but comes at the expense of transcript integrity.

In addition to standard scRNA-seq methods, other protocols help optimise characteristics such as transcript integrity or spatial/temporal specificity. Switching mechanism at the 5' end of the RNA transcript (SMART) protocols use the Moloney murine leukemia virus (MMLV) reverse transcriptase, which provides an "anchor" for polymerase template switching, allowing for production and amplification of cDNA for the full length of the transcript, averaging 2 kb at a time (Zhu et al., 2001). SMART-based scRNA-seq protocols like Single-cell tagged reverse transcription (STRT) add a barcode for each sample to allow for parallel multiplexed sequencing, and contain a one-to-one relation between sequencing reads and original mRNAs (Hebenstreit, 2012; Islam et al., 2011). Similar methods like Smart-Seq trade off strand specificity for increased

coverage depth and isoform specificity (Hebenstreit, 2012; Ramsköld et al., 2012). The CEL-seq protocol attempts to avoid the pitfalls of data loss during exponential PCR amplification by instead using the linear progression of *in vitro* transcription (IVT), resulting in less variation and greater reproducibility as compared to STRT (Hashimshony et al., 2012). The Drop-Seq protocol separates each cell into nanoliter droplets and assigns a unique barcode to each cell's mRNA output, allowing for rapid profiling of the entire transcriptome of a group of cells (Macosko et al., 2015).

The aforementioned protocols all use next-generation sequencing (NGS) to sequence amplified fragments. Other protocols allow for single cell resolution transcriptomic profiling with a more limited scope. Fluorescent *in situ* RNA sequencing (FISSEQ) relies on *in situ* reverse transcription to produce cDNA that is then crosslinked and hybridised to fluorescent probes which are imaged with confocal microscopy (J. H. Lee et al., 2014). This offers a high degree of spatial resolution for transcripts, but fails to profile the entire transcriptome and involves shorter (< 30 nt) sequencing reads. Single molecule fluorescence *in situ* hybridisation (smFISH) similarly uses fluorescently tagged DNA oligo probes to find nascent transcripts, and imaging can provide a detailed view of the spatial distribution of a given mRNA species as well as real-time information on its transcriptional rate (G. Li & Neuert, 2019). However, these methods fail to provide the unbiased whole-transcriptome profiling that NGS-based scRNA-seq methods do.

This analysis looks at three public scRNA-seq datasets to examine SoxB expression. The embryonic dataset from Karaiskos *et al.* uses the 10X Chromium protocol to examine approximately 8000 cells from stage 6 embryos; the dataset from Avalos *et al.* uses Drop-Seq to sequence approximately 5000 cells from the first instar larval brain separated from the ventral nerve cord (VNC); and the adult VNC dataset from Allen *et al.* contains approximately 26000 cells sequenced via 10X Chromium. Both *SoxN* and *D* are known to be expressed in the neuroectoderm of stage 6 embryos (Nambu & Nambu, 1996; Russell et al., 1996; Overton et al., 2002). While neither gene has been characterised in first instar larvae, both are highly expressed in the embryonic brain. *D* is expressed in larval optic lobes (Melnattur et al., 2013) and late larval brain (Suzuki et al., 2013). *SoxN* is more widely expressed in the larval brain and VNC (Crémazy et al., 2000), and is present in various neurons of the adult nervous system (Schilling et al., 2019). Together, these datasets provide a way to independently validate existing expression patterns of *SoxN* and *D* while uncovering nuances in their regulation.

3.1 Materials and Methods

To begin, processed read matrices for each of the three datasets were downloaded from the NCBI Gene Expression Omnibus. While different data processing steps were detailed in each of the original papers, for my analysis I elected to employ the same pipeline for all datasets to

ensure a degree of consistency, unless otherwise noted. The majority of computational analyses used a local Ubuntu 18.04 installation running bash version 4.4.20 and R version 3.4.3 on a machine with 8 GB of RAM. For tasks that required higher computational power, the Rustbucket server in the University of Cambridge network provided a Debian 9 environment with 32 GB of RAM, running bash version 4.4.12 and R version 4.4.0. A complete list of software packages and versions used in these analyses is provided in Table 6. For each dataset, processing scripts and output files are provided at *https://github.com/edridgedsouza/mphil-thesis*.

Attached Package	Ubuntu Version	Rustbucket Version
ALL	1.28.0	-
Seurat	3.1.5	3.1.5
purrr	0.3.4	-
UpSetR	1.4.0	-
ggplot2	3.3.0	3.3.0
rvest	0.3.5	-
xml2	1.3.0	-
org.Dm.eg.db	3.10.0	-
topGO	2.38.1	-
SparseM	1.78	-
GO.db	3.10.0	-
AnnotationDbi	1.48.0	-
IRanges	2.20.2	-
S4Vectors	0.24.3	-
Biobase	2.46.0	-
graph	1.64.0	-
BiocGenerics	0.32.0	-
tibble	3.0.1	3.0.1
dplyr	0.8.5	0.8.5
ClusterMap	-	0.1.0

Table 6: R software packages and corresponding version numbers. Packages shown are non-base libraries listed under the "other attached packages" output of the `sessionInfo()` function. Versions are shown for the local Ubuntu installation running R v3.4.5 and for the Debian installation on the Rustbucket server running R v4.4.0. A complete list of all packages loaded via namespace but not attached to the R session can be found on GitHub

3.1.1 Data Preprocessing

Data for each analysis was obtained from the NCBI GEO database using accession numbers GSE95025, GSE134722, and GSE141807 for the embryo, larval brain, and adult VNC datasets, respectively. The original methods used for processing for these datasets involved some filtering to eliminate cells or genes with low expression. However, because this project seeks to explore

the nuances of *SoxN* and *D* expression, even in cells with low transcript counts, I performed all analyses using unfiltered data to capture the total variance present in the original samples.

The embryo study included replicates from both *Drosophila melanogaster* and *Drosophila virilis,* the larval brain study included replicates from starved and unstarved individuals, and the adult VNC study examined both male and female replicates separately. In order to maintain consistency, the final datasets used for analysis used pooled expression data across multiple replicates, with no replicate-specific normalisation. I restricted the embryo analysis to *D. melanogaster* expression for ease of comparison with other data, while only unstarved flies were considered for the analysis of the larval brain dataset. Each replicate of the adult VNC dataset contained approximately 30,000 sequencing barcodes, an order of magnitude larger than the other datasets that used approximately 1,000 to 2,000 barcodes per replicate. While pooling male and female replicates would have been more biologically comparable to the other datasets, it would also require much greater computational power. Consequently, only male replicates from the adult VNC dataset were pooled for analysis. However, the scripts used to process these data can be easily adapted to consider only females or both males and females.

The R package Seurat (Satija et al., 2015) provides a complete set of methods for processing and analysing scRNA-seq data, and was the main package used for this project. Gene identifiers unique to *Drosophila* were resolved via org.Dm.eg.db from the Bioconductor repository (Gentleman et al., 2004; Carlson, 2019). To prepare each of the three datasets, each replicate under consideration was independently loaded with the `CreateSeuratObject()` function and then pooled together using `merge.Seurat()`.

Exploratory analyses revealed the degree to which each cell expressed transcripts corresponding to noncoding RNAs and mitochondrial genes, and the `FindVariableFeatures()` function with a variance-stabilised transform selection method helped identify the most variable genes in each condition. `ScaleData()` centered and normalised expression counts relative to each gene, and Principal Component Analysis (PCA) used the variable features of each dataset to calculate significant dimensions. The `JackStraw()` function identified which of the PCA dimensions were most significant (Chung & Storey, 2015); after confirming that the first 20 dimensions were significant for all three datasets, `FindNeighbors()` used dimensions 1 through 20 to to construct a Shared Nearest Neighbor (SNN) Graph for each dataset. The larger adult VNC dataset was also rerun with the first 45 dimensions to provide a resolution more in line with the findings in the original paper. `FindClusters()` with a resolution parameter of 0.5 used the Louvain algorithm to automatically partition each dataset into clusters of cells by identifying local "communities" within the larger network of cells (Blondel et al., 2008; Waltman & van Eck, 2013). Because these clusters are based on a standardised processing pipeline, there is not a one-to-one correspondence between these clusters and clusters identified in each original paper.

After processing and clustering, the data was reduced to two dimensions via the Uniform

Manifold Approximation and Projection (UMAP) algorithm (McInnes et al., 2018). Visualisation of this projection allowed each cluster identified by the Louvain algorithm to appear as distinct groupings of cells; by querying this processed dataset for expression of *SoxN*, *D*, and other neural markers, it was possible to identify which clusters were positive for each marker, as well as the distribution of expression levels within each cluster. `FindAllMarkers()` with a minimum percent threshold of 0.25 and other parameters set to their default values used a Wilcoxon rank sum test to identify genes significantly enriched as markers for each cluster. These marker lists provided a rich source of information for Gene Ontology (GO) analysis, subset exploration, and cross-dataset comparisons.

3.1.2 GO Analysis

After producing a list of cell clusters and their corresponding marker genes for each dataset, the next step was to examine the clusters that expressed *SoxN* or *D* as one of their markers. The two major approaches used were Gene Ontology (GO) analysis and subset analysis, with the former providing a more high-level view of the biological processes associated with each cluster. The topGO package from Bioconductor offered a stable set of GO annotations for each gene as well as statistical tests to identify the significance of different GO terms within each subset of genes (Alexa & Rahnenfuhrer, 2020). To increase the depth of information available, these analyses also used previously established datasets to identify genes annotated as transcription factors, SoxN bound targets, and D bound targets. TF annotation was scraped from the first version of the FlyTF database (Adryan & Teichmann, 2006), while binding information for SoxN and D came from DamID and ChIP-seq experiments by Ferrero *et al.* and Aleksic *et al.*, respectively (Aleksic et al., 2013; Ferrero et al., 2014).

For each individual GO analysis, a list of "foreground" genes was tested against a "background" list consisting of all the genes identified during sequencing. For the purpose of this investigation, only Biological Process (BP) annotations were considered. When calculating significant enrichment of GO terms, both weighted and unweighted versions of the Fisher's exact test were performed to identify enriched BP terms with and without consideration for annotation hierarchy (Alexa et al., 2006). The results from both types of tests were combined, and the p-values of the weighted test results were adjusted to account for multiple testing using the Benjamini-Hochberg correction (Benjamini & Hochberg, 1995; S.-Y. Chen et al., 2017). Using the ggplot2/tidyverse environment (Wickham et al., 2019), GO terms with adjusted p-values of less than 0.1 were plotted against their significance levels to identify which terms were most significantly enriched in each cluster. Each plot represents the culmination of one GO gene set analysis, and this process was repeated systematically for multiple gene lists. Because not all genes contain GO annotations, the number of genes in a given GO analysis may be less than the total number of genes used as input.

For each dataset, individual clusters expressing either SoxB gene were analysed for GO

enrichment, first considering all marker genes in the cluster and then only the genes that are known to be TFs. After examining individual clusters, the script looked at the supersets and intersection sets for all the clusters expressing *SoxN*, all the clusters expressing *D*, and all the clusters expressing either gene. For each of these conditions, the foreground gene list consisted of the full gene list for the condition, the list restricted to only TFs, the list restricted to SoxN bound targets, and the list restricted to D bound targets. In this way, it was possible to perform an unbiased review of the biological processes that are active in cells expressing either *SoxN* or *D* in any of the datasets.

3.1.3 Subset Analysis

While the GO analysis provided a high-level view of the major processes occurring in each cell cluster, further examination of each subset provided a more nuanced view of the genes expressed in each cluster. The GO plots offer a starting point for understanding expression in individual clusters and in groups or subsets of clusters. The BP annotations reveal the major processes occurring in each group, while the annotations for the TF-specific lists show a more focused view of the major regulatory players within these groups. Supersets and intersection sets for clusters expressing *SoxN*, *D*, or both reveal shared players and targets involved in Gene Regulatory Networks (GRNs) that involve *SoxN* and *D*. By incorporating annotations of SoxN and D bound targets, it is possible to identify the parts of each GRN that SoxN and D directly impact. GRNs were visualised by adding gene lists to the Cytoscape v3.8.0 software suite (Shannon et al., 2003). Edge weights for each graph were calculated based on experimental evidence annotations provided by the STRING database of protein-protein interactions (Szklarczyk et al., 2019). By highlighting subsets of the GRN based on degree of separation from *SoxN* and *D*, it was possible to further narrow each cluster to genes in the near regulatory vicinity of SoxB factors. Upon identifying these genes, there was enough information to assign putative cell types to clusters depending on markers and expression patterns previously identified in *Drosophila* neural cells.

The analysis looked to identify candidate genes that may have biologically significant interactions with *SoxN* and *D*, so it was therefore important to set parameters permissively to produce larger gene lists. This approach was utilised in the GO analysis, using Benjamini-Hochberg correction rather than Bonferroni to lower the false negative rate and cluster supersets to widen the spectrum of genes under consideration. For a full subset analysis, it is necessary to examine both permissive and restrictive gene lists. Using the UpSetR package, it was possible to examine how many genes lie at the intersection of multiple SoxB-positive clusters (Conway et al., 2017); genes that were present in the intersection of multiple clusters are likely candidates for future exploration in conjunction with Sox genes.

3.1.4 Cross-Dataset Comparison

Besides identifying cell cluster identities within individual datasets, it was of interest to explore whether clusters from one dataset could be mapped to clusters from another based on similarities in gene expression patterns. Doing so would provide an avenue to explore changes in Sox-associated GRNs for individual cell types throughout the course of development. The ClusterMap package provides an interface to compare two different Seurat datasets based on their expression counts as well as the marker genes identified in each cluster (Gao et al., 2019). Comparisons were attempted between the embryo and larval datasets, as well as between the larval and adult VNC data. Because this method failed to yield any comparisons with a high degree of correlation, the next best way to compare the development of cell types over time was to use subset analysis to identify putative cell identities within each of the datasets.

3.2 Results

Analysis of each dataset resulted in a gene list for each identified cluster, as well as GO enrichment analysis for each cluster associated with *SoxN* or *D*, along with their respective subsets and supersets. To identify features of each cluster, the GO enrichment scores provided a first-pass filter for understanding the major cellular processes occurring in each cluster. Because all three datasets focus on different parts of the *Drosophila* CNS, several of the top GO terms are unsurprisingly associated with processes related to neurodevelopment. Interestingly, other common GO terms related to processes involving transcriptional regulation and ribosomal assembly, indicating that those particular clusters may correspond to undifferentiated cells that have a high baseline level of transcriptional output before differentiation decreases global expression in favor of more specialised functions (Efroni et al., 2008). Because both *SoxN* and *D* are known to play roles in CNS differentiation, clusters that express either gene and that also have high transcriptional output may represent neural precursors to later differentiated tissues.

3.2.1 Embryonic Dataset

The embryonic dataset by Karaiskos *et al.* provides a detailed look at the expression of SoxB in the *Drosophila* stage 6 embryo at the onset of gastrulation. While this dataset covers the entire embryo and is therefore not CNS-specific, expression of both SoxB genes is restricted to the neuroectoderm at this stage. We therefore expect that clusters that are positive for *SoxN* or *D* are likely to represent the broad range of neural precursor and neuroectodermal cells in the early embryo.

The UMAP projection and ridge plots for this dataset clusters into 14 total groups of cells, with clusters numbered according to the number of cells (Figure 8a; highest cluster number has fewest cells). The ridge plots reveal that *SoxN* expression is generally higher than that of *D*

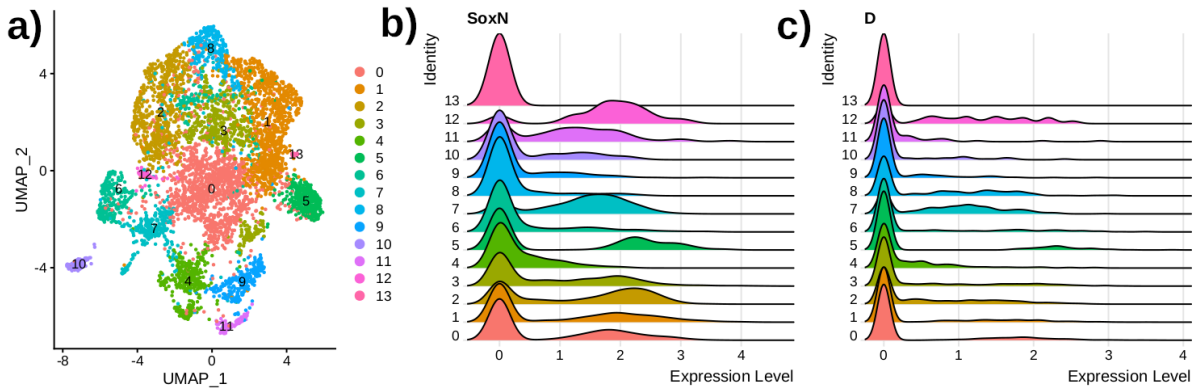

Figure 8: Cell populations within the stage 6 embryo. Data from Karaiskos *et al.* was pooled to consider only the replicates from *D. melanogaster*. **a)** UMAP projection was performed using the first 20 dimensions of data. Louvain clustering with 20 dimensions and a resolution parameter of 0.5 reveals 14 distinct clusters of cells, with cluster 0 containing the largest number of cells. **b-c)** Ridge plots reveal log-normalised and variance-stabilised expression of both *SoxN* and *D* within several clusters. *SoxN* exhibits generally higher expression levels than *D*, though *D* shows higher expression in cluster 8. Significance testing shows that *SoxN* is significantly enriched in clusters 2 and 12 while *D* is enriched in clusters 5 and 12.

(Figure 8b,c). While both genes show expression in multiple clusters, *SoxN* is only significantly enriched as a marker gene in clusters 2 and 12 while *D* is expressed weakly in cluster 5 and significantly in cluster 12. Because both genes are coexpressed in cluster 12, this represents an area in which both factors may share some targets in their GRNs.

Both genes are also expressed in clusters where they are not significantly enriched as a marker. In particular, *SoxN* is moderately expressed in clusters 0-3, 7, and 11 and weakly expressed in clusters 4, 6, and 8-10; *D* is weakly expressed in clusters 0-4, 7, 8, 9, and 11. This analysis focuses on the clusters that specifically express *SoxN* or *D* as a significant marker, but interactions are also likely in these other clusters as well. For instance, cluster 7 contains moderate expression of both factors and its 111 genes are enriched for terms related to regulation of embryonic development ($\sim$ 3E-5); cluster 8 contains higher expression of *D* than *SoxN* and contains 137 genes, of which 83 (60.6%, 8.5E-23) relate to anatomic structure development, 13 (9.5%, 3.7E-6) relate to blastoderm segmentation, and 31 (22.6%, 1.4E-31) show publication enrichment related to AP and DV patterning (Saunders et al., 2013). This is consistent with a model of *D* coordinating a network of developmental processes in the tissues where it is uniquely expressed, despite being present in low absolute quantities.

The superset of clusters 2, 5, and 12 contains 494 genes that are significantly enriched for several GO terms, the most significant of which involve axon guidance, VNC development, regulation of transcription and signalling, and neurogenesis (Figure 9a). Within this superset, 69 of the genes represent transcription factors. The GO terms most associated with these

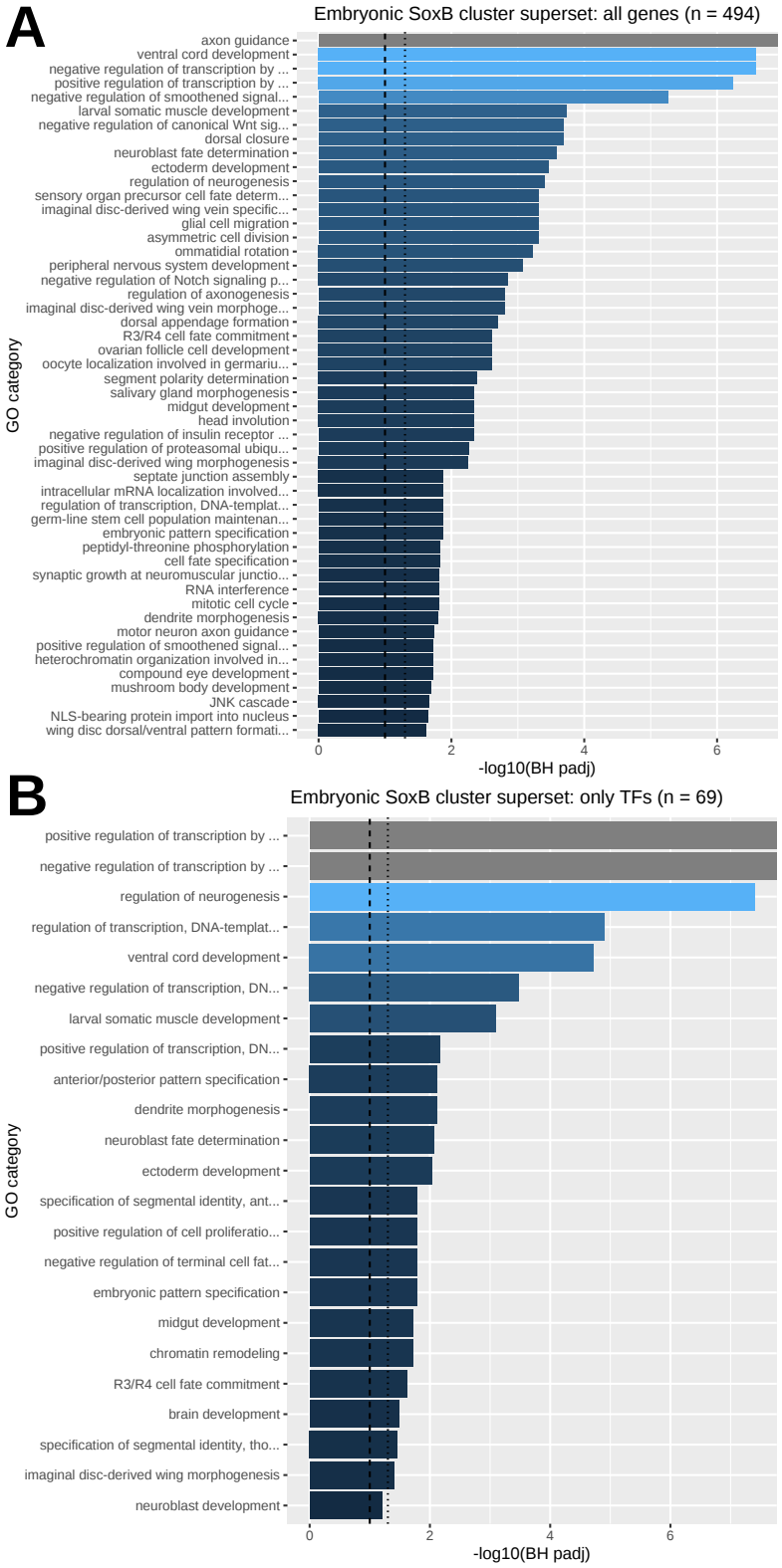

Figure 9: Gene Ontology enrichment for superset of embryonic SoxB-marked clusters. Enriched GO terms are shown along with the negative $\log_{10}$ of their Benjamini-Hochberg adjusted p-values. Dashed and dotted lines signify 0.1 and 0.05 significance thresholds, respectively. Grey bars represent significant GO terms whose adjusted p-values were corrected to 0. **A)** Enrichment of all 494 annotated genes within the superset of clusters 2, 5, and 12 of the embryonic dataset. **B)** Enrichment of only the 69 annotated TFs.

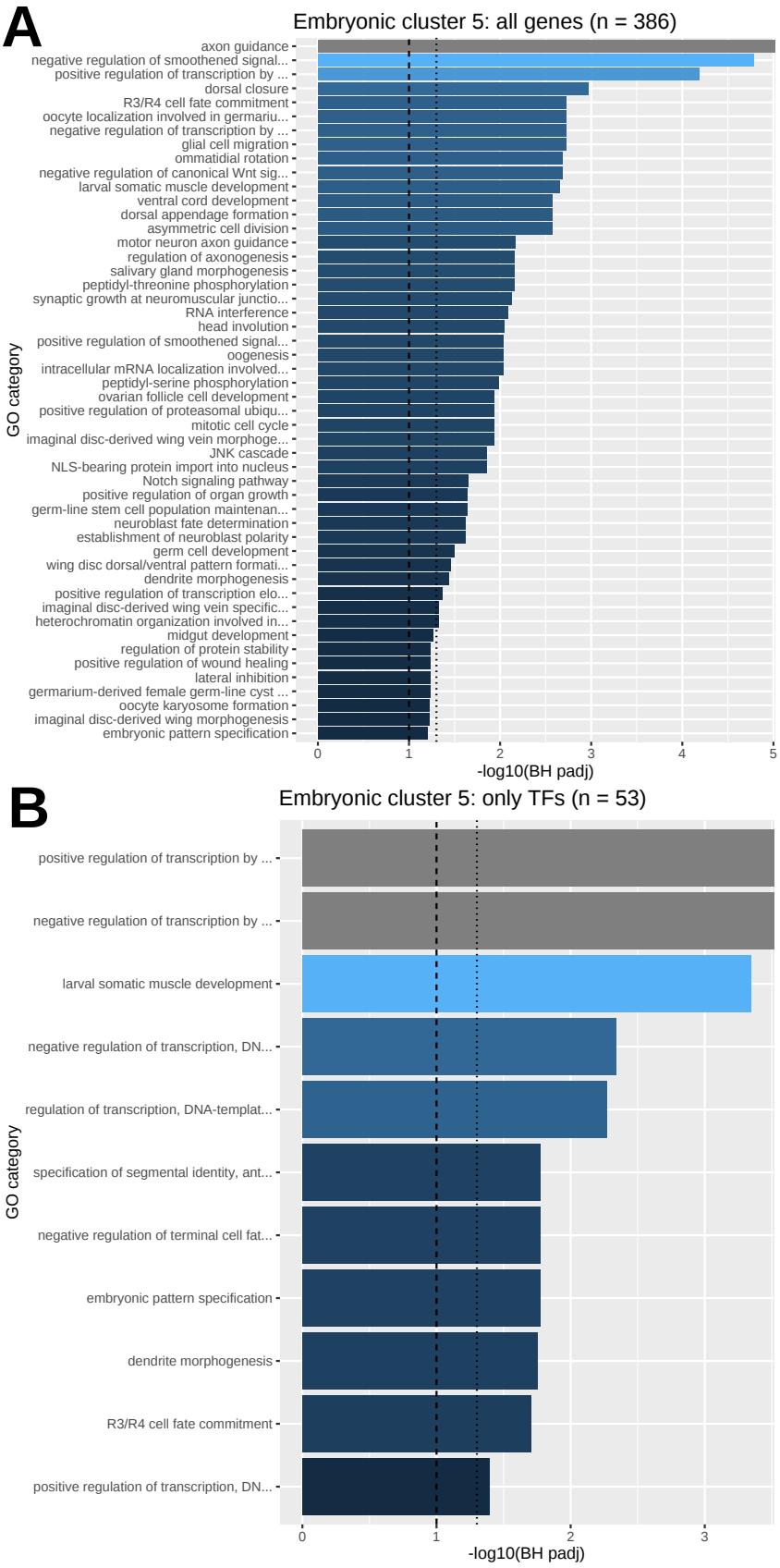

Figure 10: Gene Ontology enrichment for embryonic cluster 5. *D* is a significant marker for cluster 5 of the embryonic dataset. **A)** Enrichment plot of all 386 annotated genes within cluster 5. **B)** Enrichment of the 53 annotated TFs in the cluster.

are unsurprisingly related to regulation of transcription, since it is a set of TFs, but more importantly regulation of neurogenesis, anterior/posterior patterning, and development of VNC are all enriched terms (Figure 9b). This is consistent with the known functions of *D* and *SoxN* in CNS development. Furthermore, within the superset, 167 and 372 genes are known binding targets of *SoxN* and *D* respectively (Aleksic et al., 2013; Ferrero et al., 2014), with 134 bound by both, reinforcing the view these are neural cells with SoxB activity. The transcription factors in this superset therefore appear to play more of a regulatory role than the other genes, indicating that the biological processes driven by the full superset gene list may largely be a result of genes that are targets of these TFs.

Within cluster 2, there are 44 markers and these show weak enrichment for genes involved in negative regulation of Notch signalling and myoblast migration . This cluster contains 5 genes annotated as TFs, including *SoxN*, *opa*, *odd*, *tsh*, *pnr*, and *gsb*. Analysis of these genes in FlyMine (Lyne et al., 2007) reveals that these genes are primarily related to patterning and segmentation. Indeed, *opa* is known to show a conserved interaction with with *D* during insect segmentation (Clark & Peel, 2018) and later in the larval CNS it functions as a negative regulator of *D* to push neural stem cells toward their later stages of differentiation (Abdusselamoglu et al., 2019). This may imply that lack of significant *D* expression in this cluster is consistent with a mechanism in which *SoxN* but not *D* helps cells progress to a post-neuroblast state (Ferrero et al., 2014). Downregulation of Notch signalling is also known to correspond to transition from progenitor cells to neuroblasts (Contreras et al., 2018), further supporting the idea that cluster 2 represents a relatively large group of cells that includes neuroblasts.

Cluster 5 contains 423 genes, which show a significant enrichment for neural functions and various terms related to developmental processes, including Smo and Wnt signalling pathways. Almost 70% of the genes in this cluster are annotated with the high level GO term "regulation of biological processes", suggesting actively differentiating cells. Analysis of BDGP expression via FlyMine (Hammonds et al., 2013; Lyne et al., 2007; Tomancak et al., 2002, 2007) shows strong enrichment for expression in the ventral ectoderm (5.0E-17), procephalic ectoderm (1.7E-12), and VNC primordium (2.8E-12), consistent with SoxB expression in the neuroectoderm. In this cluster 53 genes (12.5%) are transcription factors with weak enrichment for embryonic patterning and dendrite morphogenesis (Figure 10).

While *Dichaete* is only weakly enriched in this cluster, 277 (65.5%) of these genes are known to be bound by *D*, indicating that even weak *D* presence may indicate a central coordinating role within the GRN. Additionally, while the cluster 5 TFs were enriched for a few distinct processes such as segment specification and muscle development, the full cluster 5 gene list is enriched for various other functions such as R3/R4 photoreceptor fate commitment, glial cell migration, neuromuscular junction growth, axon guidance, JNK signalling, and oogenesis. This points to a role in which *Dichaete* serves as a regulatory hub (Aleksic et al., 2013) that coordinates a wider set of genes during the process of segmentation, CNS specification and

differentiation. Simultaneously, FlyMine localisation analysis shows that many of the TFs in this category, such as the aforementioned *opa,* are known to be active in the developing VNC and ventral ectoderm. This suggests that *D* may be acting in conjunction with other TFs in this cluster to coordinate a wide range of developmental processes related to patterning and differentiation.

Cluster 12 is much smaller than the other clusters, but is the only one that exhibits strong *SoxN* and *D* expression. The cluster contains 92 genes enriched for expression in the procephalic and ventral ectoderm ($\sim$ 2E-14 each), of which 13 (14.1%) are annotated as TFs (Figure 11). Of particular interest is the fact that several of these genes are E(spl) helix-loop-helix (HLH) and Bearded family members, indicating a connection between genes in this cluster and the Notch signalling pathway (Dearden, 2015; Lai et al., 2000). The interaction between Bearded members and the E3 ligase *Neuralised* (*Neur*), which was also present in the cluster, is known to spatially regulate *Delta* signalling to promote neurogenesis (Bardin & Schweisguth, 2006). It is therefore plausible that *SoxN* and *D* coordinate with other neurodevelopmental genes in the cluster, such as *vnd* and *l'sc*, to regulate the signalling pathways that contribute to neuroblast delamination and specification into neurons.

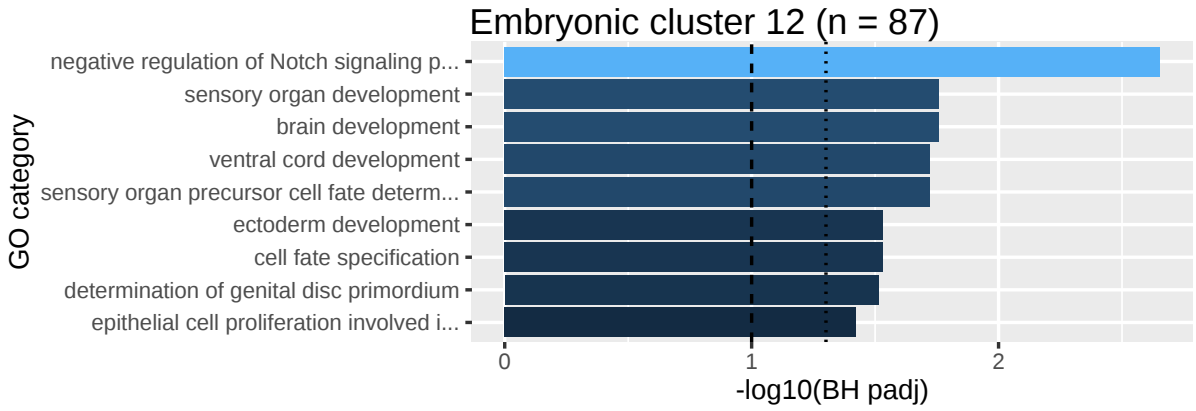

Figure 11: Gene Ontology enrichment for embryonic cluster 12. *SoxN* and *D* are both significant markers for cluster 12 of the embryonic dataset, which contains GO annotations for 87 genes.

In summary, the SoxB-focused analysis of scRNA-seq data from the blastoderm/early gastrulating embryo reveals clusters of cells involved in early aspects of neural development as well as some more specific annotations that may indicate *D* or *SoxN* exclusive functions. Since the expression of *D* and *SoxN* in the early embryo have been well characterised, the analysis of this dataset offers confidence in extending this approach to the larval brain and adult VNC datasets.

3.2.2 Larval Brain Dataset

The dataset from Avalos *et al.* examines the first instar larval brain dissected away from the ventral nerve cord and is therefore expected to contain more cells and genes that relate specifically to neurodevelopment. The cells cluster into 19 distinct groups (Figure 12a), with *SoxN* highly expressed in several groups and *D* moderately expressed across most clusters (Figure 12b). *SoxN* is significantly expressed as a marker gene in clusters 6, 7, 8, and 10, while *D* is not significantly enriched as a marker for any individual cluster, although some expression of both *D* and *SoxN* is seen in cluster 13. Therefore, the bulk of analysis in this dataset focuses on the clusters where *SoxN* is identified as a marker. Notably, clusters 6, 7, 8, 10, and 13 are closely related on the UMAP projection.

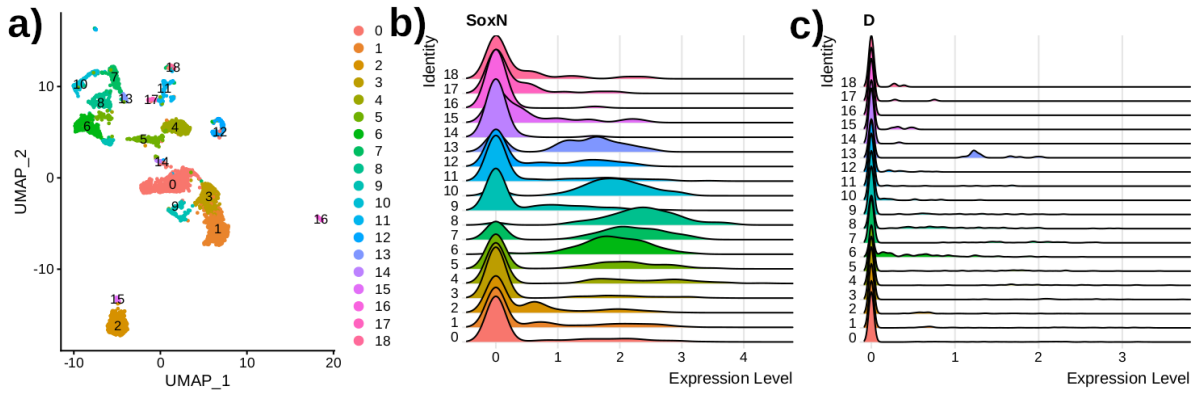

Figure 12: Cell populations within the first instar larval brain. Data from Avalos *et al.* was filtered to keep only unstarved wild-type replicates. **a)** UMAP projection used the first 20 dimensions of data. Louvain clustering with the first 20 dimensions and a resolution parameter of 0.5 results in 19 distinct clusters of cells. **b-c)** Ridge plots show log-normalised expression of *SoxN* and *D*. *SoxN* shows moderate expression in several clusters, and is significantly enriched in clusters 6, 7, 8, and 10. Notably, cluster 13 appears related to these clusters in this projection. While *D* displays a lower baseline level of expression and is not significantly enriched within any cluster, there is some expression in clusters 6 and 13.

The superset of clusters 6, 7, 8, and 10 contains a large number ($n = 972$) of translation-related gene ontologies, as well as annotations for mRNA processing, ubiquitin proteasome activity, and cell proliferation (Figure 13a). In contrast, restricting the list to only the 52 TFs within the superset reveals that the primary annotations are for regulation of neurogenesis, anterior/posterior patterning, neuroblast development, and other nervous system related terms (Figure 13b). Examining only the SoxN bound targets within the superset shows strong enrichment for cytoplasmic translation ($p < 3\text{E-}5$), neural precursor development ($p < 1.8\text{E-}4$), and regulation of Notch signalling ($p < 1.8\text{E-}4$) as well as eye-antennal disc morphogenesis ($p < 3.2\text{E-}3$) that is influenced by anterior/posterior patterning (Morata & Lawrence, 1979) and is a site of *SoxN* expression (Crémazy et al., 2001). While *Dichaete* but not *SoxN* is associated

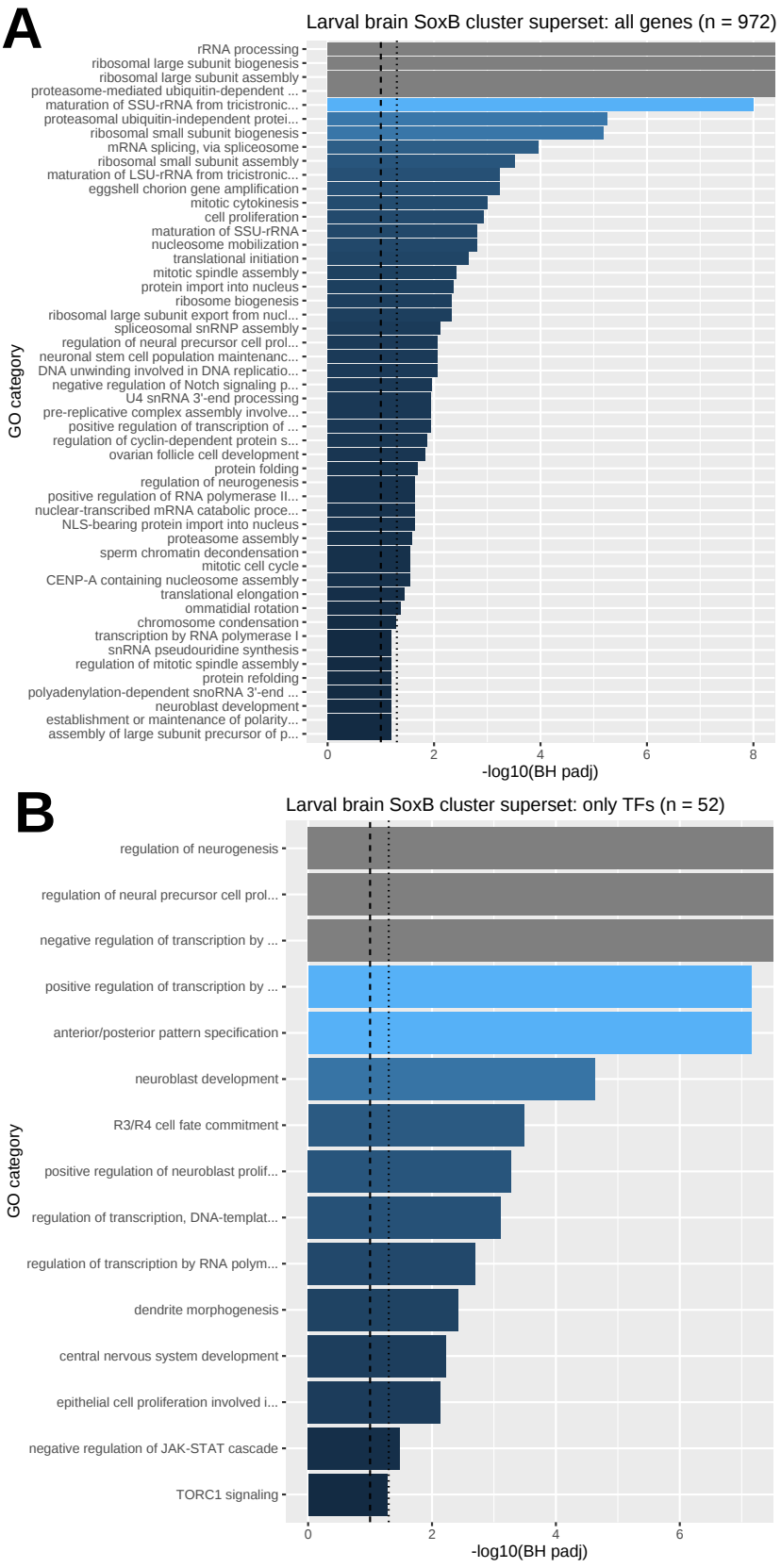

Figure 13: Gene Ontology enrichment for superset of larval brain SoxB-marked clusters. This superset includes clusters 6, 7, 8, and 10. All of these clusters feature *SoxN* as a marker while none include *D.* **A)** Enrichment plot of all 972 annotated genes within this superset. **B)** Enrichment plot of the 52 annotated TFs in the superset.

with anterior/posterior patterning in the embryo, many of the proteins involved in early segmentation are subsequently used in CNS patterning. This may suggest that patterning genes can be regulated by both the SoxB proteins, as might be expected from their extensive binding overlap in the embryo (Ferrero et al., 2014), and that neural-specific regulation is simply a reflection of restricted SoxB expression. In addition, it is possible that these data are also consistent with a model in which *D* works with *SoxN* early in development to regulate transcriptional and translational processes for neurogenesis. It is also possible that the lack of significant *D* expression in the first instar is consistent with data that show *D* expression highly restricted in larval neuroblasts as part of a temporal cascade activated in third instar larva that controls neural identify (Apitz & Salecker, 2015; Suzuki et al., 2013).

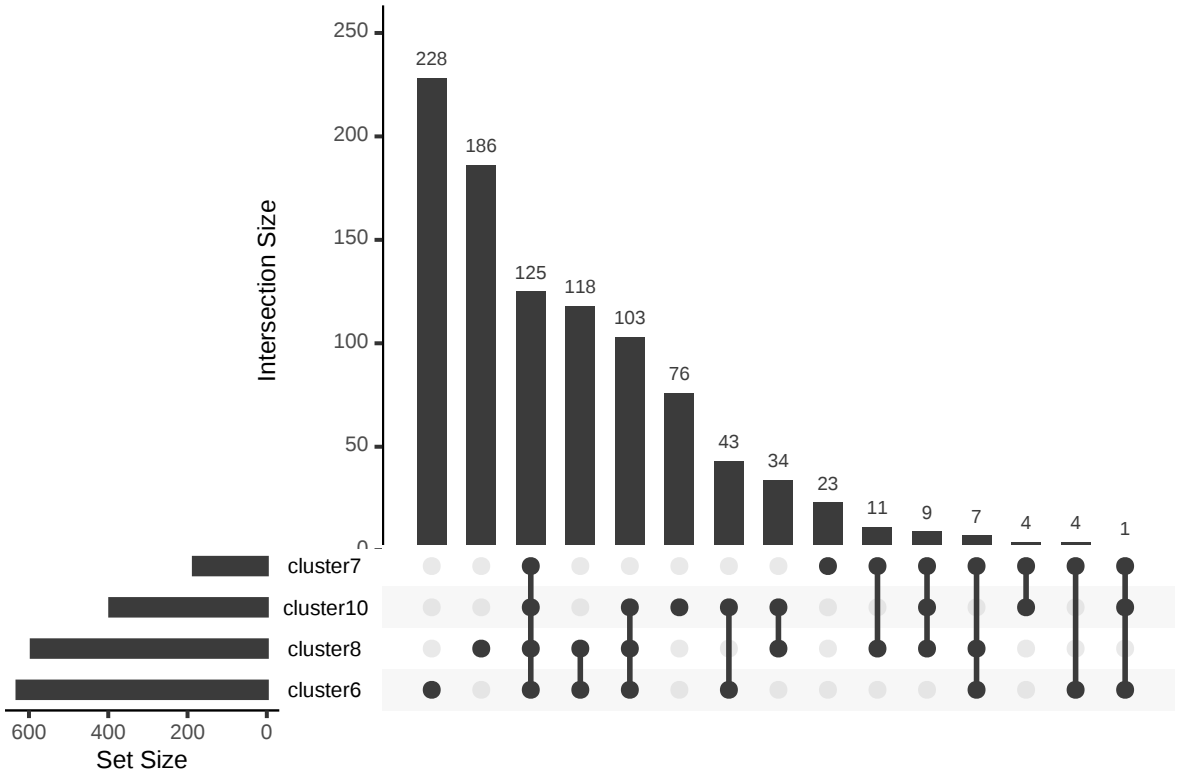

Figure 14: UpSet plot of larval brain dataset clusters. *SoxN*-positive clusters share many of the same marker genes, and UpSet representation can visualise the size of each intersection set. Clusters 6, 7, 8, and 10 bear a high degree of resemblance in terms of their expressed markers: 125 genes are shared by all four clusters, while 118 are shared by clusters 6 and 8. Overlap in these gene lists may indicate the presence of shared network elements that relate to *SoxN* and *D* regulation.

The original paper by Avalos *et al.* used an alternative computational approach to generate 29 different cell clusters, each corresponding to particular cell type annotations (Brunet Avalos et al., 2019). Of these clusters, only one expresses *SoxN* as a marker gene and is annotated

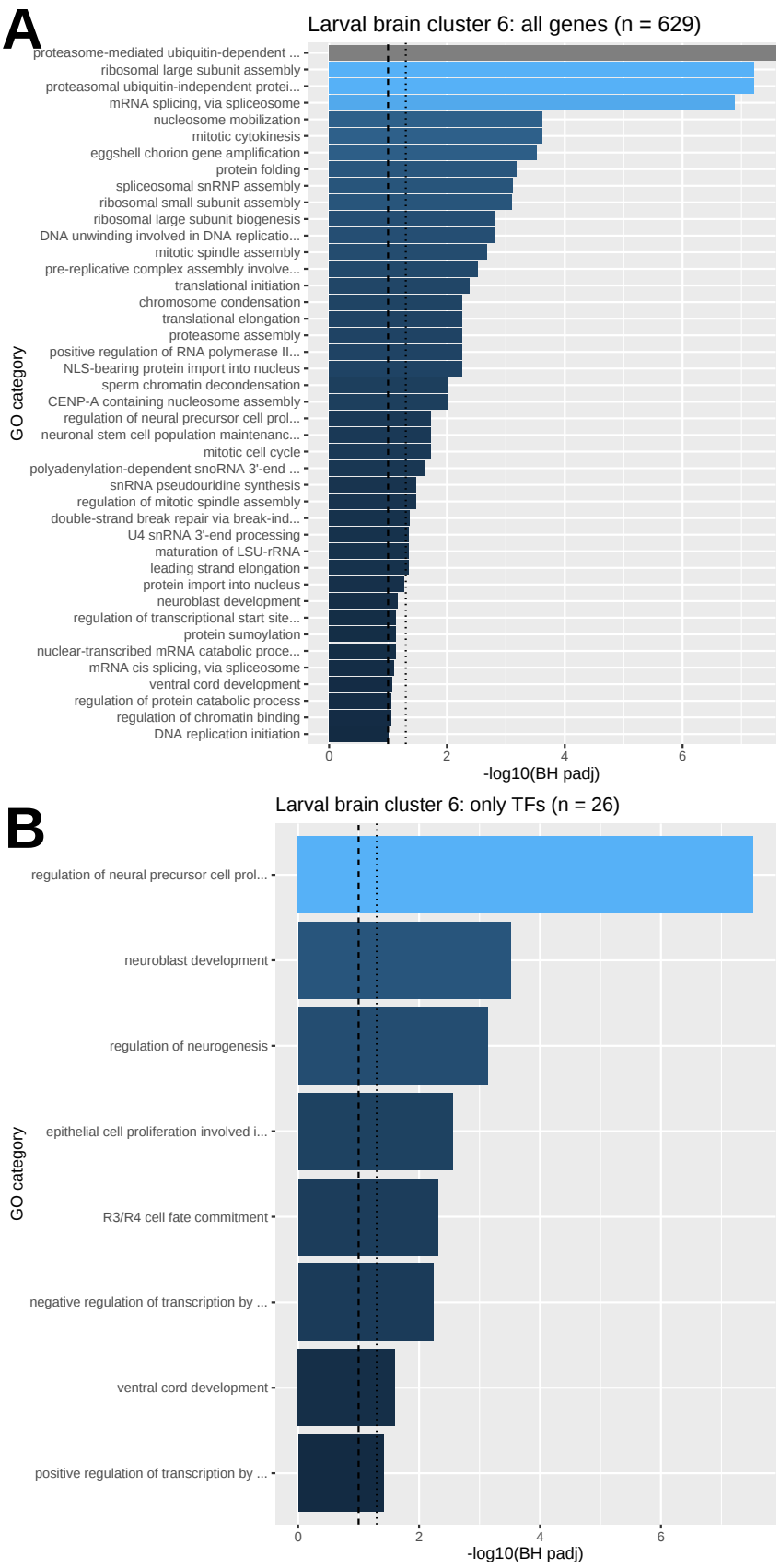

Figure 15: Gene Ontology enrichment for larval brain cluster 6. *SoxN* is a significant marker for cluster 6 of the larval brain dataset. **A)** Enrichment plot of all 629 annotated genes within cluster 6. **B)** Enrichment of the 26 annotated TFs in the cluster.

as “neural progenitor cells 2.” A total of 6 clusters from the original paper are annotated as corresponding to NPCs, all of which received their categorisation based on literature-based annotations. Within this cluster, other marker genes include *sna*, *klu*, *fru*, and *cas*; all four of which are co-expressed with *SoxN* in clusters 6, 7, and 8 in my analysis, indicating that the three clusters in this categorisation may be related to “neural progenitor cells 2” from the original paper. However, while all of these factors are markers for clusters 6 through 8, all of them except *sna* and *SoxN* are also expressed in cluster 13, and each gene is also uniquely expressed in some other clusters. While there is not an exact correspondence between clusters from the two analyses, this provides a starting assumption that *SoxN*-positive cells in clusters 6, 7, and 8 will generally behave similarly to NPCs. While none of these factors appear to be shared in cluster 10, UpSet analysis of the *SoxN*-positive clusters reveals that while only 7 genes are shared uniquely between clusters 6 through 8, 125 genes are shared between clusters 6, 7, 8, and 10, suggesting that cluster 10 of this analysis may also be NPC-related (Figure 14).

Despite the similarities between these clusters, there still appear to be marked differences. The neural marker *Dpn* is present in clusters 6 and 8 but not 7 or 10. Within cluster 6, only 26 (4.1%) of the 629 identified markers are annotated as transcription factors. Within the full set, the most enriched GO terms involve mRNA processing, proteasome degradation, and translation, but within only the TFs, neuroblast development and regulation of neurogenesis are the most significant processes (Figure 15). This, along with general CNS development, is also true of cluster 8 (Figure 16), which shares a full 16 of the 26 TFs and 353 of the same cluster markers as cluster 6. Cluster 6 possesses marker genes such as *sna*, *wor*, and *ase* that suggest neuroblast selection and delamination (Arefin et al., 2019). Both clusters share a large number of ribosomal genes related to translation, and interestingly, both contain over 200 genes annotated as *D* bound targets. Despite not expressing *D* as a marker, this is consistent with an interpretation of clusters 6 and 8 as neuroblasts that have not yet initiated *D* expression but that have a Dichaete-regulated GRN ready for temporal activation.

Cluster 7 is similar to clusters 6 and 8, but is more highly dominated by ribosomal and translational genes; considering only the 10 TFs in this cluster yields high GO enrichment for CNS development and NPC proliferation (Figure 17). Because these TFs include *SoxN* and the shared *sna*, *klu*, *fru*, and *cas*, it is possible that cluster 7 represents a subgroup of neuroblasts that are actively promoting translation as they commit to becoming neurons. While fewer of the genes in cluster 7 are binding targets for *D* as compared to clusters 6 and 8, 36 of the 65 targets come from ribosomal genes; indeed, both the superset and intersection set of clusters 6, 7, 8, and 10 are enriched for ribosomal assembly processes when considering only *D* binding targets (Figure 18). While *D* is not a marker in any of these putative neuroblast clusters, it still has moderate baseline expression within these cells and may direct networks related to translation and neuroblast commitment once the temporal cascade activates its expression.

While the UpSet analysis (Figure 14) shows that cluster 10 shares similarities with clusters

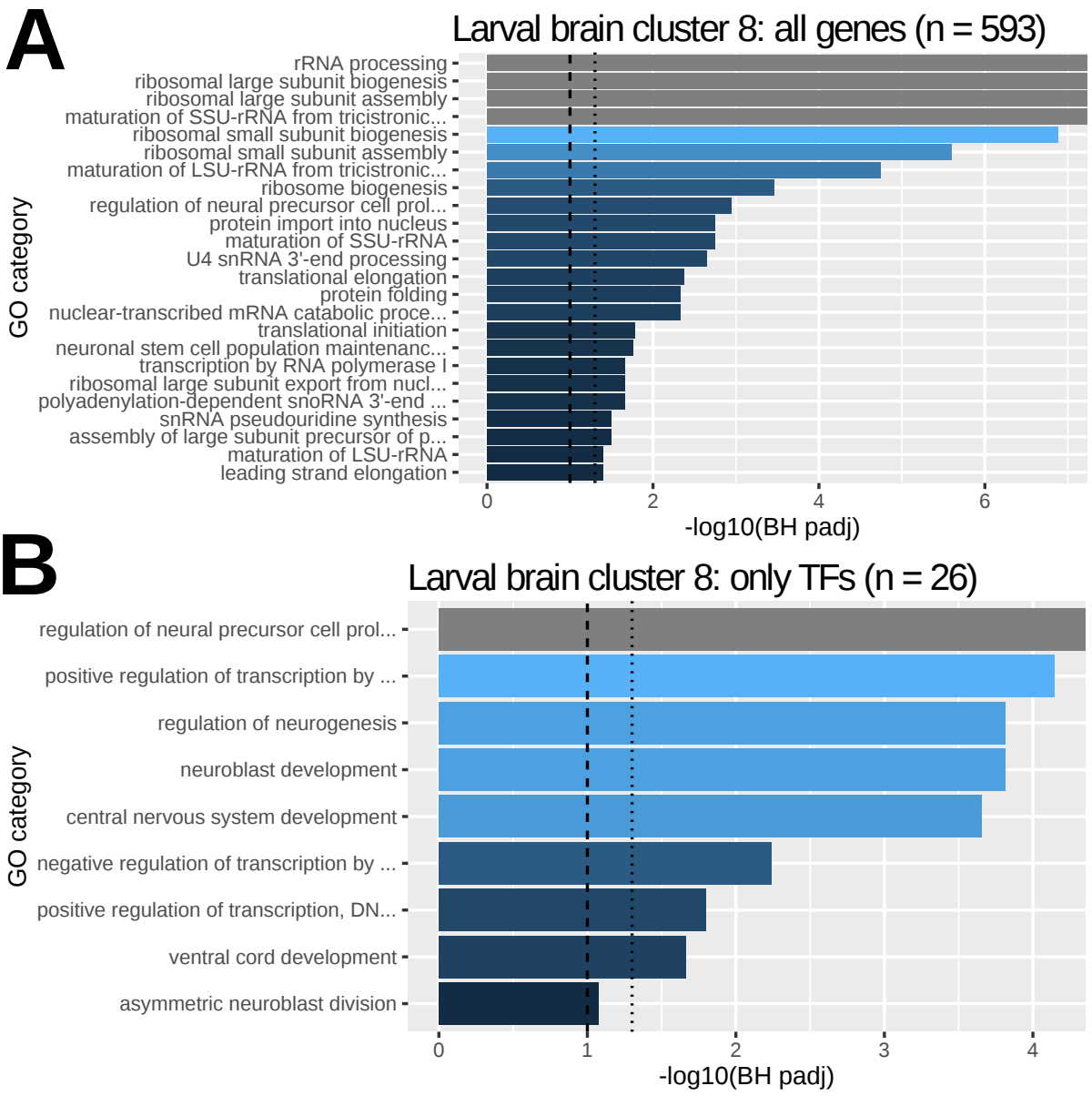

Figure 16: Gene Ontology enrichment for larval brain cluster 8. *SoxN* is a significant marker for cluster 8 of the larval brain dataset. **A)** Enrichment plot of all 593 annotated genes within cluster 8. **B)** Enrichment of the 26 annotated TFs in the cluster.

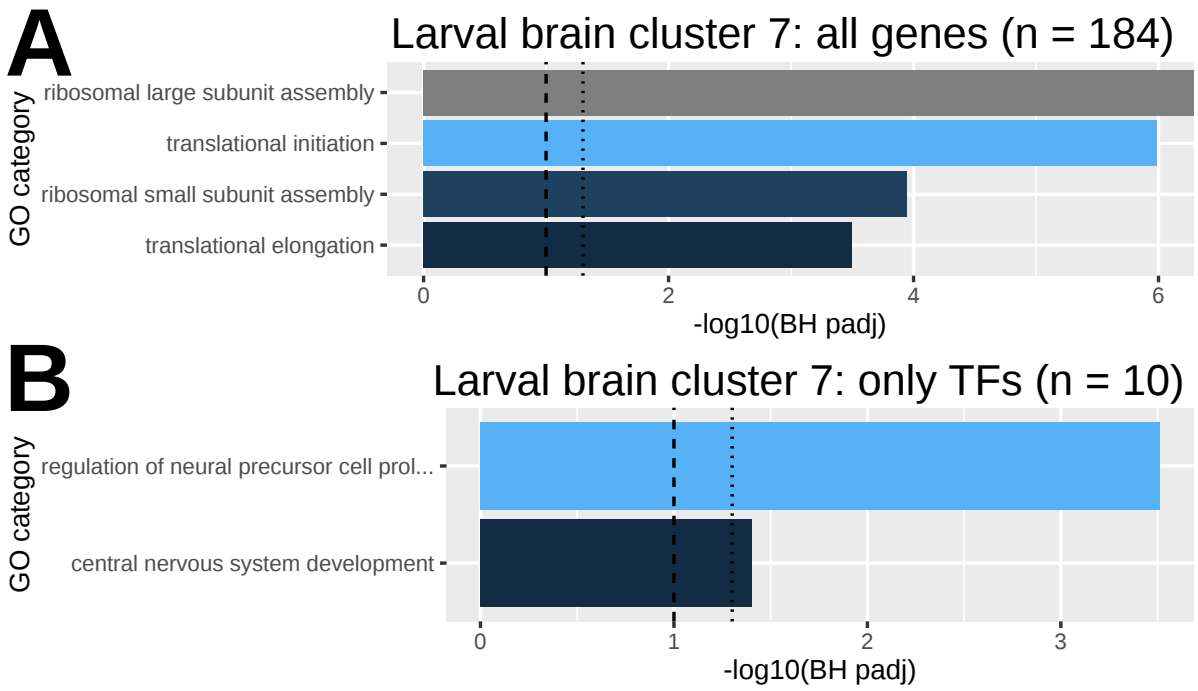

Figure 17: Gene Ontology enrichment for larval brain cluster 7. *SoxN* is a significant marker for cluster 7 of the larval brain dataset. **A)** Enrichment plot of all 184 annotated genes within cluster 7. **B)** Enrichment of the 26 annotated TFs in the cluster.

6 through 8, it does not share the core markers defined by Avalos *et al.* Though its full GO enrichment is dominated by processes related to ribosomal assembly, its 20 transcription factors are enriched for negative regulation of transcription, anterior/posterior patterning, and neurogenesis (Figure 19). This cluster expresses *Tll*, the final element after *D* in the temporal cascade that determines the neural progeny of neuroblasts (X. Li et al., 2013). It is therefore possible that cluster 10 represents a version of the cells in clusters 6 through 8 that are in the process of committing to neuronal fate or have already committed.

Taken together, my analysis of the larval brain scRNA-seq data reveals clusters of neural cells that appear to be highly translationally active. The absence of genes encoding neurotransmitters, which are present in several of the clusters that do not express SoxB, indicates that SoxB expressing cells are all neuroblasts or progenitors, a suggestion supported by enrichment for cell-cycle related genes. In keeping with the mapped expression in the larval CNS, *SoxN* shows much greater enrichment than *Dichaete*, the latter being expressed in a highly restricted pattern in the developing optic lobes.

3.2.3 Adult Ventral Nerve Cord Dataset

The dataset from Allen *et al.* is derived from the dissected adult ventral nerve cord, and sampled by far the greatest number of cells ($\sim$ 26,000) of the 3 analyses. My UMAP analysis identified

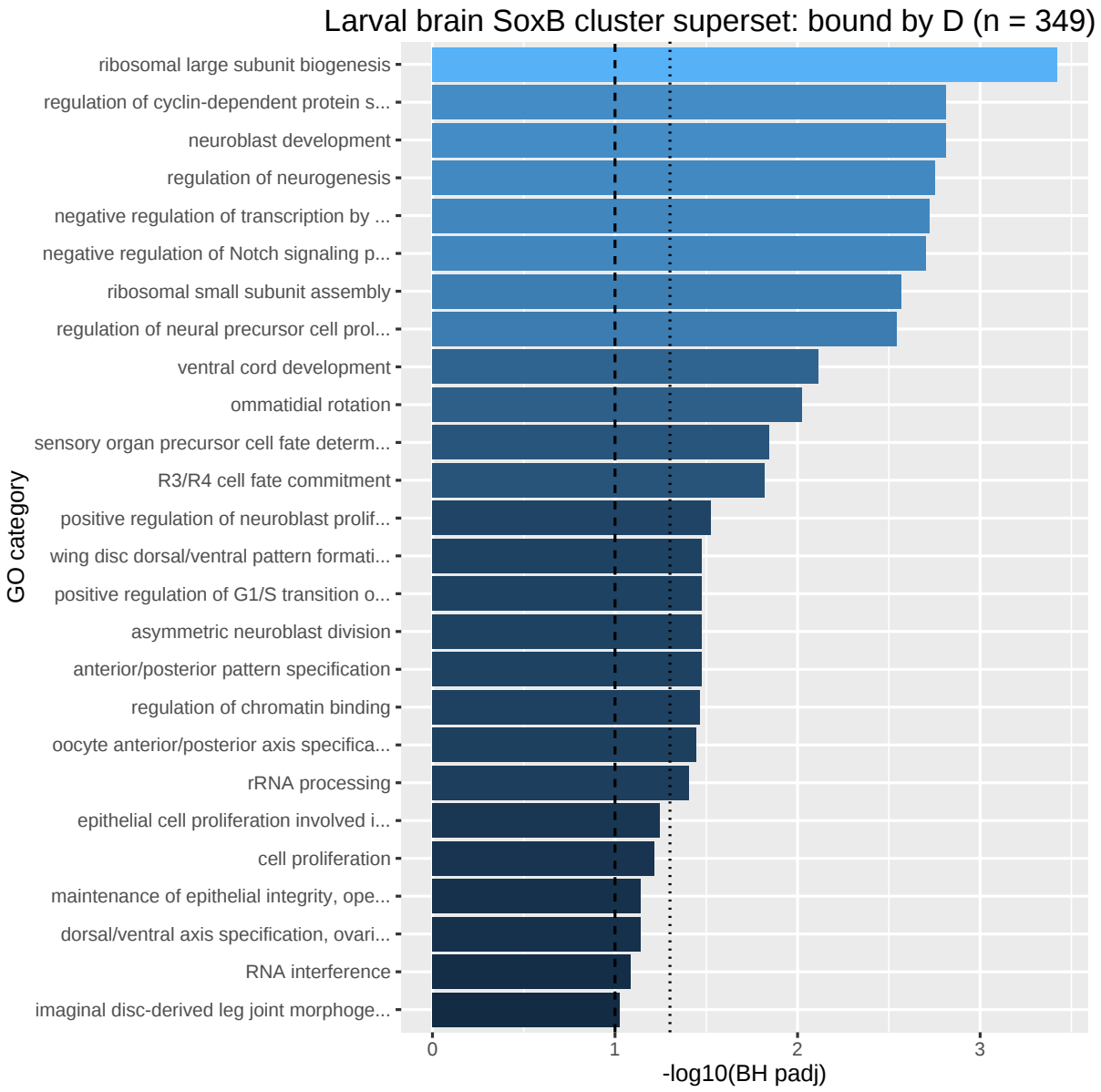

Figure 18: Gene Ontology enrichment for D-bound genes within superset of larval brain SoxB-marked clusters. 35.9% of genes in the superset of markers for clusters 6, 7, 8, and 10 (Figure 13a) are bound by Dichaete in the embryo (Aleksic et al., 2013).

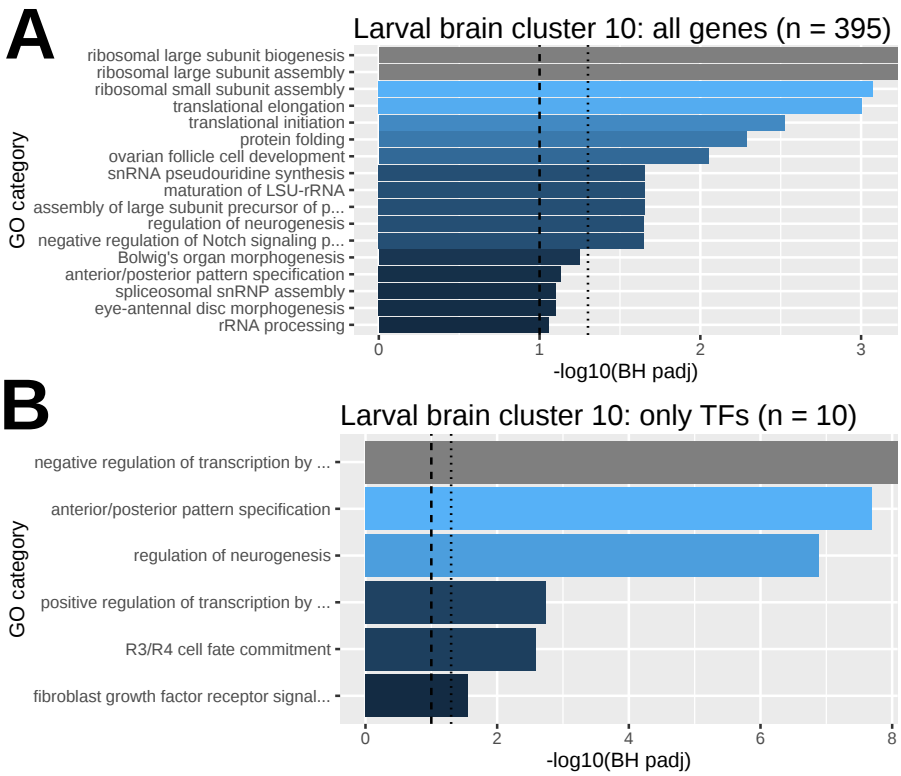

Figure 19: Gene Ontology enrichment for larval brain cluster 10. *SoxN* is a significant marker for cluster 10 of the larval brain dataset. **A)** Enrichment plot of all 395 annotated genes within cluster 10. **B)** Enrichment of the 20 annotated TFs in the cluster.

174 distinct cell clusters (Figure 20a). This was the only dataset processed with dimensions 1 through 45 and with a cluster resolution of 12, as the original paper reveals that there is significant information far beyond the first 20 dimensions (Allen et al., 2020). Projecting *SoxN*, *D* and the neuroblast/GMC marker *Prospero* (L. Li & Vaessin, 2000) into these UMAP plots revealed widespread *pros* expression in many clusters, reflecting a high proportion of neuroblast and post-neuroblast lineages (Figure 20b). In contrast, *SoxN* and *D* are far more restricted in their expression, with *Dichaete* in particular limited to very few clusters. (Figure 20c-d). Within these defined clusters, *SoxN* is a marker for clusters 53, 108, 126, 135, 140, 151, 159, and 167, while *D* is a marker only in cluster 108. Overall, *SoxN* is expressed in 2.5% of the cells in the dataset while *D* is expressed in only 0.4%.

In cluster 108, *D* is present in 63.7% of cells and *SoxN* in 40.9%. The cluster is enriched for neuronal-related GO terms, in particular synaptic functions, suggesting these are post-mitotic cells. The cluster is enriched for genes expressed in the embryonic brain (140 genes, 4.6E-16) with 40% (220 genes) bound by Dichaete in the embryo. Of the 34 TFs in this cluster, 24 are expressed in the embryonic brain and 11 of these have annotated roles in neuron development (3.5E-5), reinforcing the view that this cluster represents differentiating neurons. (Figure 21). Interestingly, this cluster also expresses *Sox21b*, which exhibits CNS expression in the late embryo and larva and is likely co-regulated with *Dichaete* (McKimmie et al., 2005).

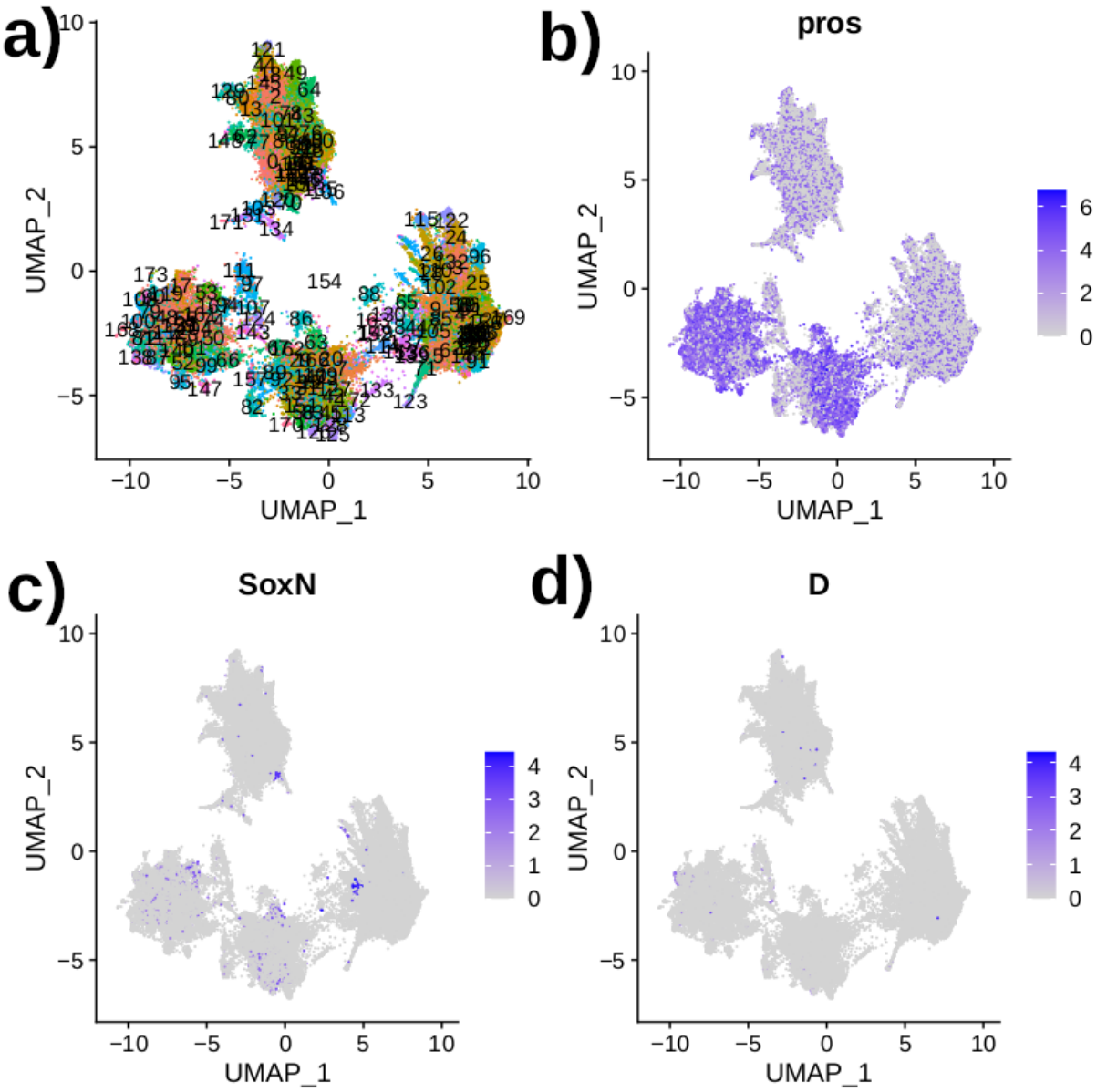

Figure 20: Cell populations within the adult ventral nerve cord. Male replicates from Allen *et al.* were pooled and processed on the Rustbucket server. **a)** UMAP projection used the first 45 dimensions, and Louvain clustering used the first 45 dimensions with a resolution parameter of 12 to produce 174 total cell clusters. **b-d)** Relative log-normalised expression levels of the neuroblast/GMC marker *pros* as well as of *SoxN* and *D* are shown on the original projection. While *pros* exhibits expression over a wide range of cell clusters, *SoxB* expression is more limited, with *SoxN* more highly expressed than *D*. Significance testing shows *SoxN* enriched in clusters 53, 108, 126, 135, 140, 151, 159, and 167, with *D* enriched in cluster 108 as well.

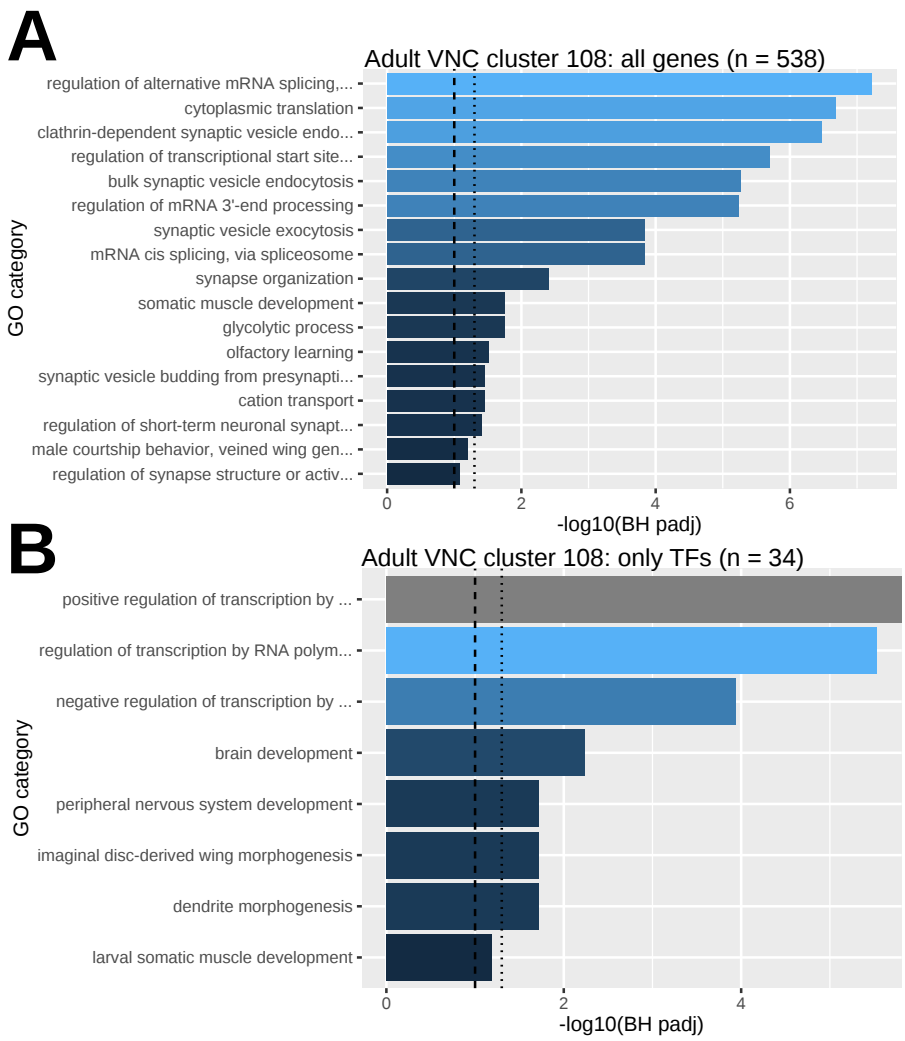

Figure 21: Gene Ontology enrichment for adult ventral nerve cord cluster 108. *SoxN* and *D* are significant markers for cluster 108 of the adult VNC dataset. **A)** Enrichment plot of all 538 annotated genes within cluster 108. **B)** Enrichment of the 34 annotated TFs in the cluster.

Within the superset of the *SoxN*-positive clusters, the most enriched GO terms for these 1694 genes involve translation, mRNA splicing, signal transduction, and several functions related to the formation of axons and synapses (Figure 22a). Focusing only on the 111 TFs within this *SoxN* superset, the most significant terms include dendrite morphogenesis, PNS development, and a variety of other specific terms related to developmental processes (Figure 22b). Additionally, other terms specify differentiation and development for body parts including the brain, eyes, midgut, and muscles. Estacio-Gómez *et al.* recently described a DamID based analysis of factors associated with particular neuronal lineages (Estacio-Gómez et al., 2020). They report several TFs associated with the specification of cholinergic, GABAergic and glutamergic neurons, including *Ets65A*, a repressor of cholinergic lineages. This TF is enriched in clusters 53, 135, and 167, indicating these are unlikely to be cholinergic (however, see below). In contrast, *knot* is reported to be a cholinergic marker and is identified in cluster 140. Finally, *Ptx1* is

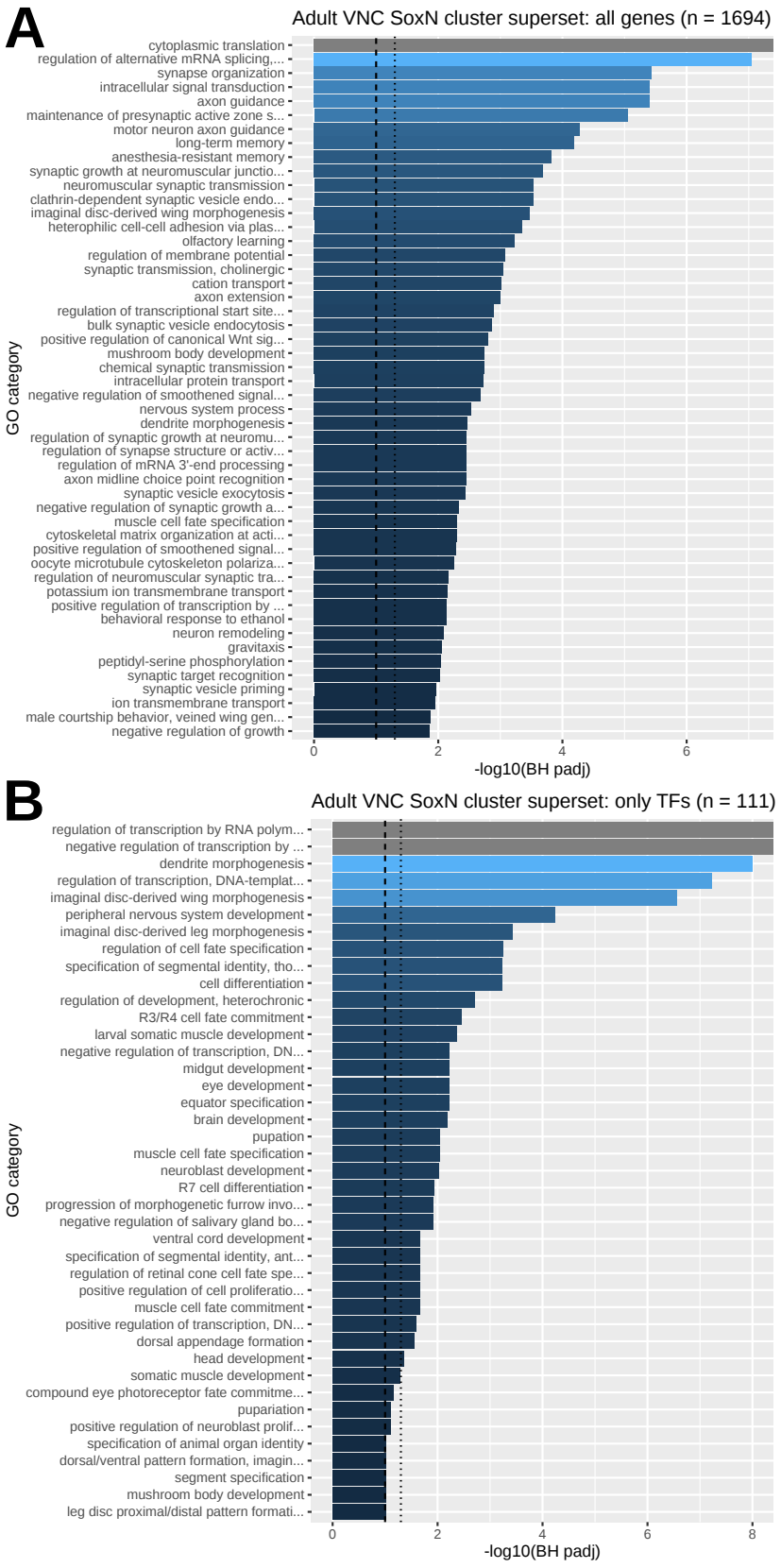

Figure 22: Gene Ontology enrichment for superset of adult ventral nerve cord *SoxN*-marked clusters. This superset of a clusters that feature *SoxN* as a marker includes clusters 53, 108, 126, 135, 140, 151, 159, and 167. **A)** Enrichment plot of all 1694 annotated genes within this superset. **B)** Enrichment plot of the 111 annotated TFs in the superset.

enriched in GABAergic lineages and we found this only in cluster 159. Restricting the superset to *SoxN* bound targets, 414 genes are enriched for functions mainly related to axon ($p < 3.2\text{E-}5$) and dendrite ($p < 3.2\text{E-}5$) development as well as several behavioral traits related to learning, memory, and response to ethanol. This paints a picture of *SoxN* in the adult VNC being involved in the networks that regulate the specification and development of particular neuronal lineage, with most of the clusters under consideration likely to be neurons. This is in line with previous work showing *SoxN* involvement in late neuronal differentiation in the embryo (Ferrero et al., 2014).

Turning to each cluster in turn: 53 contains 578 genes and is enriched for synaptic transmission and organisation terms as well as olfactory learning (Figure 23). As noted above, expression of *Ets65A* would suggest this cluster contains non-cholinergic cells; however, the cluster contains 6 nicotinic acetylcholine receptor subunit genes which are indicative of cholinergic functions, as well as the chloride channel *Rdl* which is associated with GABAergic activity (McGonigle & Lummis, 2009). This may suggest the cluster contains cells of mixed lineage. The 45 TFs in this cluster show weak enrichment ($\sim$ 3E-3) for processes related to eye and PNS development, suggesting that it either comprises nerves that handle sensory input or is simply a more general reflection of neurogenesis, with at least 8 of the TFs known to have roles in axonogenesis.

Cluster 126 contains 484 genes that are highly enriched for ion transport and cholinergic synaptic transmission, as well as more general terms related to axon/synapse development (Figure 24). As with cluster 53, it contains several nAChR subunit genes, suggesting that these cells are likely cholinergic. Enrichment for Wnt and Hh (via Smo) signalling is associated with neural development and processes such as planar polarity (McFarland et al., 2008) as well as axon guidance (Charron & Tessier-Lavigne, 2007), and may indicate that these cells are differentiating. This cluster is also enriched for the functional terms proboscis extension and olfactory learning, perhaps suggesting that they are differentiating progenitors that have initiated specific neural expression programmes.

Cluster 135 only contains 9 genes, 4 of which encode mitochondrial proteins related to electron transport and ATP production. It contains *SoxN* and the cholinergic repressor *Ets65A*. It is possible that this small cluster is simply an artifact of the partitioning approach.

Clusters 140 and 151 are the most similar of any two clusters, with 121 genes in common that are not shared by any of the other SoxB clusters (Figure 25); both are enriched for synapse formation and development, suggesting their GRNs play a role in organising the growing CNS, and in the case of cluster 151, also the PNS (Figure 26a,c). It is worth noting that cluster 140 contains the *knot* TF, a marker for cholinergic neurons, along with the fly Acetylcholine esterase and Choline acetyltransferase genes (*Ace* and *ChAT*), reinforcing the view that these are cholinergic neurons. In contrast, cluster 151 expresses Tropomyosin 1 (*Tm1*), a gene enriched in L3 larval GABAergic neurons (Estacio-Gómez et al., 2020). While the 44 TFs within cluster

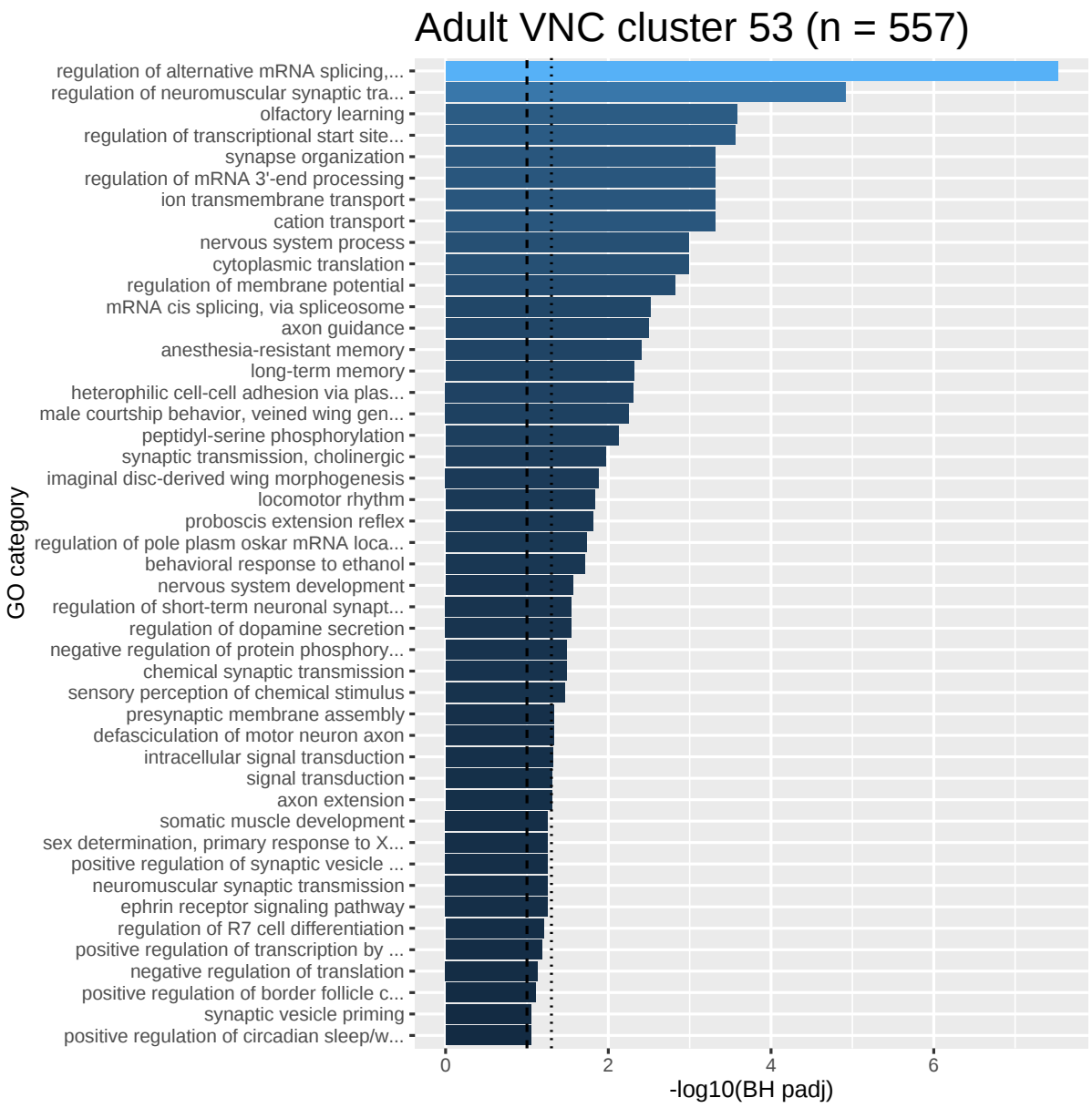

Figure 23: Gene Ontology enrichment for adult ventral nerve cord cluster 53. *SoxN* is a significant marker for cluster 53 of the adult VNC dataset. GO plot shows enrichment for the 557 annotated genes in this cluster.

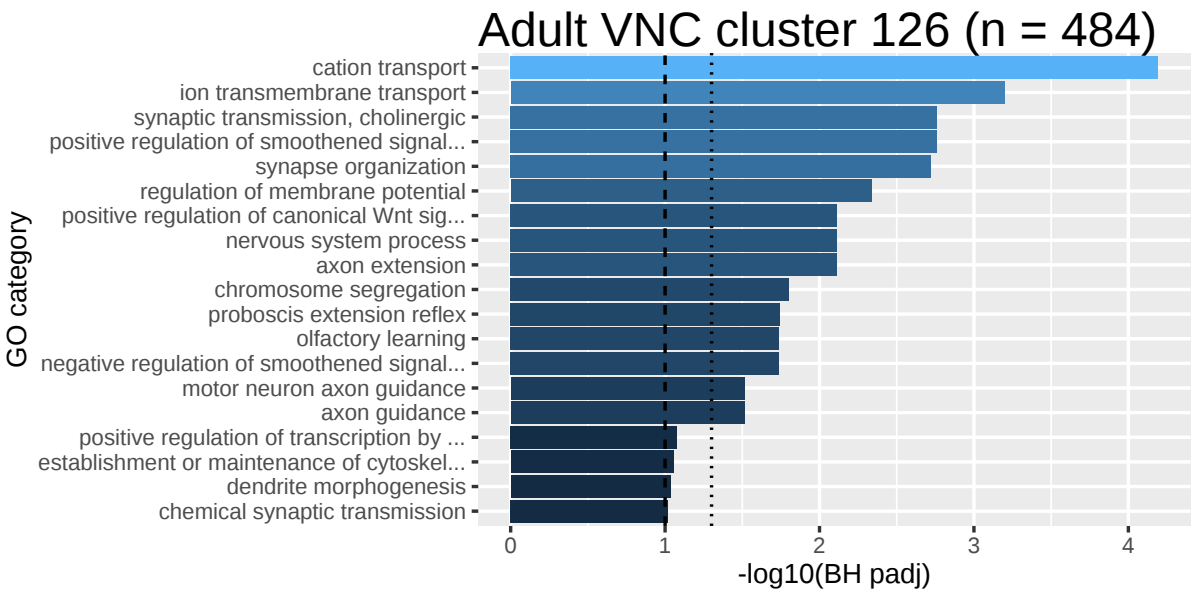

Figure 24: Gene Ontology enrichment for adult ventral nerve cord cluster 126. *SoxN* is a significant marker for cluster 126 of the adult VNC dataset. GO plot shows enrichment for the 484 annotated genes in this cluster.

140 show relatively few significant ontology enrichments, the 44 TFs in cluster 151 are enriched for regulators implicated in PNS development (Figure 26b,d), perhaps suggesting that, given their similarity, 151 cells are progenitors of cluster 140 neurons.

Though cluster 159 appears similar to other clusters, with enrichment for genes involved with synaptic formation and organisation, it also appears to contain TFs involved with morphogenesis and PNS formation, with weak enrichment for TFs implicated in the regulation of cell proliferation. The cluster contains the *Ptx1* TF and other markers found in GABAergic neurons, but also features two nAChR subunits and the muscarinic Acetylcholine receptor (*mAChR-A*), which are more associated with cholinergic neurons. The presence of mixed neurotransmitter types and positive regulators of cell proliferation shuch as *Tsh* and *Hth*, may suggest this cluster represents progenitors.

Cluster 167 genes appear enriched for mRNA regulation and ion transport, with a number of ontology terms pointing to aspects of neuronal differentiation (synaptic functions, axon differentiation/guidance) while its TFs show some enrichment for eye development terms indicative of general neuronal functions.

3.2.3.1 *SoxN* and *D* Compensation

In cluster 108, both *SoxN* and *D* are expressed, signifying that genes in this cluster may provide insight to the shared parts of the GRN that allow one factor to partially compensate in the absence of the other. Using Cytoscape with the edge-weighted spring embedded layout, it was

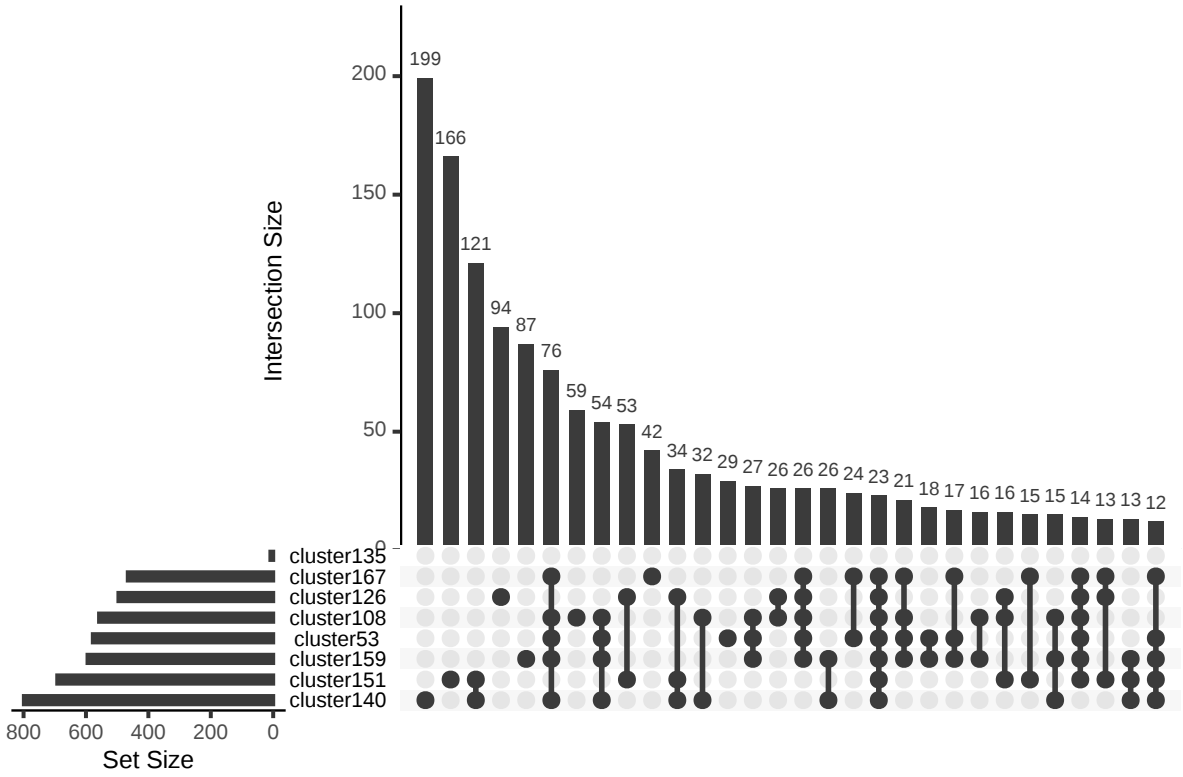

Figure 25: UpSet plot of adult ventral nerve cord dataset clusters. The 30 largest intersection sets for marker genes in the *SoxB*-positive clusters 53, 108, 126, 135, 140, 151, 159, and 167 are shown. Cluster 135 did not contain any marker genes that were in these largest intersection sets. Clusters 141 and 150 show a high degree of overlap in their marker genes, with a total of 121 genes uniquely shared by these two clusters.

possible to visualise the GRN of cluster 108, as well as the degrees of separation between *SoxN*/*D* and the rest of the network (Figure 27). Within two degrees of separation, there are 92 other genes in the vicinity of the two SoxB factors, and several of these genes appear closely linked to dense clusters within the network. Cluster A appears on the top of the network and contains 47 genes whose GO terms are highly enriched for synapse formation, neurotransmission, and sensation. Cluster B is on the right of the network and contains nodes close to cluster A; its 35 genes are enriched for functions related to the regulation of mRNA slicing and processing. Cluster C on the bottom of the GRN is the furthest separated from the other clusters and contains a dense network of 73 genes related to translation as well as metabolic processes such as electron transport and ATP production.

Each of these three clusters represents a different subset of genes that *SoxN* and *D* may be able to regulate through various pathways. Supplementary Table 1 provides a full list of genes in each cluster, as well as the genes in the regulatory vicinity of the Sox genes. Within clusters A-C, 24.7% and 34.7% of the 170 genes are bound by SoxN or D in the embryo respectively, in line with the average of 23.3% and 39.8% for the 558 genes throughout cluster 108. This rises to

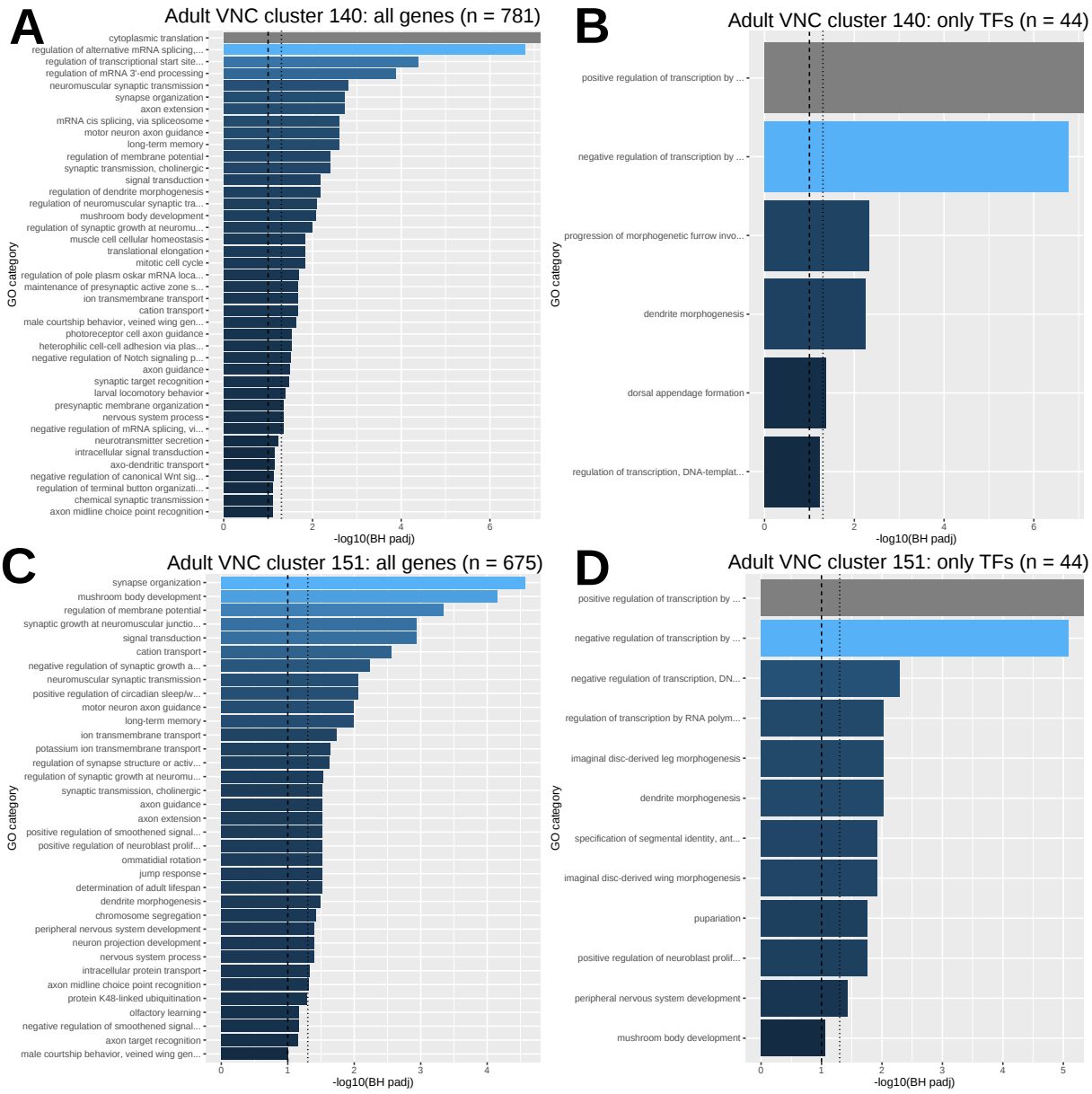

Figure 26: Gene Ontology enrichment for adult ventral nerve cord clusters 140 and 151. *SoxN* is a significant marker for these clusters the adult VNC dataset. **A,C)** Enrichment plot of all 781 and 675 annotated genes in clusters 140 and 151, respectively. **B,D)** Enrichment of the 44 annotated TFs in each cluster.

35.5% and 60.1% ($z = 2.511$ and 4.684, two population Z test) when considering the genes in the vicinity of *SoxN* and *D* within the network.

Examining these genes within two degrees of separation of *SoxN* and *D*, the most enriched terms involve transcriptional regulation, mRNA processing, and processes related to cellular signalling and stem cell maintenance. These genes include *Sox21b* as well as neural markers like *pros* and *elav*. Additionally, this group contains 15 of the 34 TFs in the entire cluster 108. Interestingly, several of the genes within two degrees of the Sox genes provide entry points to

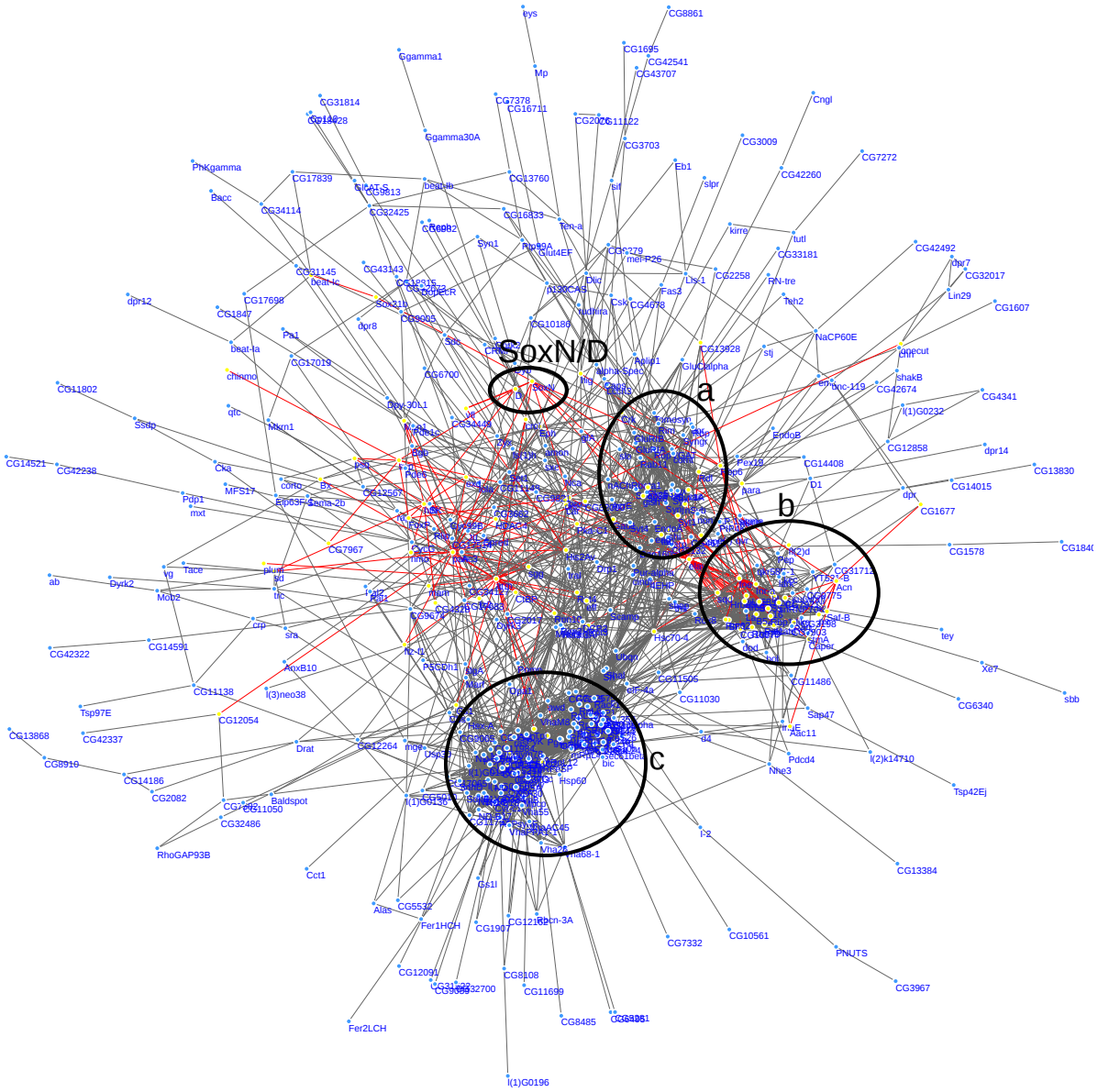

Figure 27: Gene Regulatory Network for adult ventral nerve cord cluster 108. Visualised in Cytoscape, this GRN uses an edge-weighted spring embedded layout to yield an undirected graph of interactions in the sole cluster expressing both *SoxN* and *D* as marker genes. Line weights are proportional to STRING database experimental evidence annotations for protein-protein interaction. Nodes within two degrees of separation from *SoxN* and *D* are highlighted in yellow and are connected by red lines. Three main clusters of genes deal with functions related to synapse neurotransmission, transcriptional regulation, and metabolic maintenance, respectively. Genes within two degrees of *SoxN* and *D* appear to provide regulatory entry points into these clusters, as well as to other genes not part of these clusters.

the other regulatory clusters. Clusters A and B appear to contain several genes that interact with the differentiated neuron marker *elav* (Yankura et al., 2013), while cluster C metabolic genes *Tpi* and *Pgk* are linked to the beta-catenin homologue *arm* that is involved in Wnt signalling and segment polarity (Loureiro & Peifer, 1998). A central group of the genes *pros*, *dnc*, and *Lar* extends its connections to other genes in the network that are not parts of the main three clusters. While these connections require empirical testing for further confirmation, the parts of this GRN within two degrees of separation from *SoxN* and *D* represent potential regulatory entry points for Sox genes to regulate transcription, metabolism, and neurotransmission within differentiated nerve cells. Overlaps in these networks may reveal the shared targets that allow either gene to partially compensate for the other.

The analysis of this large single cell dataset uncovered some interesting insights but also some confusing observations. While in general *SoxN* is associated with the expression of TFs implicated in aspects of neuron development and differentiation, the clustering approach combined cells expressing markers indicative of different neuronal functions (cholinergic and GABAergic). The consistent enrichment for genes associated with neuronal activity, such as axons, synapse functions and, in particular, neurotransmitters, strongly suggests these are not neuroblast lineages. Thus the analysis either indicates that *SoxN* is expressed in progenitors that give rise to neurons of different fate, or the clustering method is combining cells of different lineage. Nevertheless, the analysis has uncovered sets of genes expressed in particular cell types that are likely to be direct SoxB targets in the adult CNS.

3.2.4 Comparing Datasets

After assigning cell identities to clusters in each dataset, it was of interest to explore whether any of these clusters could correspond to others at different developmental stages. Using the ClusterMap software, the embryo and larval brain as well as the larval brain and adult VNC datasets were compared. While the software was able to identify correlations between clusters, none of these correlations were significant enough to conclude a relation between any two datasets. Within the embryonic-larval comparison, the *SoxN*-positive larval clusters 6, 7, 8, and 10 exhibited a similarity score of only 0.04 to embryonic cluster 10; for the larval-adult VNC comparison, they did not map onto any of the VNC clusters. It was not possible to perform this automated correlation between datasets, so drawing conclusions from the data required exploring the literature to infer how clusters from one dataset could relate to another.

3.2.5 Data Reliability

While the methods used for this analysis aimed to provide a consistent process for evaluating multiple datasets systematically, the drawback of this approach is that it produces results that differ from the originally published analyses for these datasets. The lack of data filtering and

normalisation was intended to avoid discarding signals that are potentially biologically relevant. Indeed, as expression of *D* was more restricted than that of *SoxN*, this approach helped ensure that even lowly-expressed genes could be captured by this analysis.

However, there are potential pitfalls of this approach. Lack of filtering has the potential to decrease the signal-to-noise ratio within the data, and instead favors false positives over false negatives. This leaves the analysis susceptible to artifacts of bias from the sequencing methodology, which is amplified by the fact that scRNA-seq data is inherently more noisy than bulk RNA-seq data (G. Chen et al., 2019). Particularly, while the 10X Chromium method used for the adult VNC dataset allows for cheap profiling of a high number of cells compared to methods like Smart-Seq2, the number of detected genes per cell is considerably lower (Baran-Gale et al., 2018). While the log transformation used in this analysis accounts for some of the variability in low-count genes, other normalisation methods like the sctransform variance-stabilised transformation make fewer assumptions about the underlying structure of the data (Hafemeister & Satija, 2019). Given the potentially biased structure of the data and the goal of identifying as many leads as possible, the subjective determination of this analysis was to decrease false negatives even at the expense of increasing false positives. Therefore, the lack of filtering and normalisation beyond the standard log-transformation had the possibility to incorporate artifacts of bias into the results relative to the original analysis by Allen *et al.*

Another factor to consider is that the UMAP method used to categorise clusters of cells is based on a stochastic process and is therefore less reproducible than deterministic dimensionality reduction algorithms like PCA. The consequence of this is that performing the same UMAP reduction with different seed values has the potential to a produce a different clustering of cells each time. For this analysis, the default seed value of 1 was used each time for consistency. However, because of the bias added by omitting the filtration and extra normalisation steps, it was necessary to ensure that the clusters produced by these analyses had some biological validity based on correlation to the clusters determined by Allen *et al.*

Marker gene lists for *SoxN*-positive and *D*-positive clusters were obtained from the SCope database (Davie et al., 2018) and compared to the corresponding clusters from this unfiltered analysis. Allen *et al.* identified 10 *SoxN*-positive clusters, 2 *D*-positive clusters, and one cluster positive for both. The marker gene lists for the superset of the *SoxN* clusters was compared to the corresponding list for the *SoxN* clusters in the unfiltered analysis; likewise, the process was repeated for the *D* clusters and for the clusters expressing both.

The Jaccard similarity of the corresponding gene lists from each analysis were relatively low (0.128 comparing *SoxN* clusters, 0.065 comparing *D* clusters, and 0.037 comparing *SoxN-D* clusters), suggesting that the gene lists that these analyses produced were highly dissimilar. However, upon closer inspection, these low values are explained by the fact that the gene lists in the unfiltered analysis were significantly more expansive than those identified by Allen *et al.*; this is expected, as the goal of the unfiltered analysis was to be unrestrictive in order to reveal

more candidate genes. When considering the percentage of genes identified by Allen *et al.* that were also detected in this analysis, there is considerably more overlap. For the *SoxN*-positive clusters, 78.8% of marker genes overlapped with the gene lists produced in the unfiltered analysis. For *D* clusters and *SoxN-D* clusters, these values were similarly high at 80.4% and 80.8%, respectively. This signifies that while the two methods produced different cell clusters, the clusters that were identified as positive for either gene were robust between analyses, lending validity to the idea that this consistent unfiltered approach produces results that are biologically similar to the original filtered analyses while also identifying more candidate genes.

Because these methods produced different results, the intersection sets of their gene lists are of particular interest because they represent the most high-confidence marker genes for *SoxN* and *D* clusters. The consensus marker list for *SoxN*-positive clusters includes 231 genes enriched for nervous system development ($p = 1.65\text{E-}13$) and neuron differentiation ($p = 2.08\text{E-}13$), consistent with known roles of *SoxN* in later neuron differentiation. The 31 genes in the *D* consensus list were enriched for localisation in the embryonic brain ($p = 1.59\text{E-}6$) and ventral midline ($8.94\text{E-}6$), consistent with known roles of *D* regulating earlier brain development and later localisation in the midline (Aleksic et al., 2013). The 21 consensus genes from the *SoxN-D* clusters exhibit low expression during early embryogenesis (1 gene) with increasing expression until peak expression at stages 13-16 (12 genes), consistent with the idea that *SoxN* and *D* pathways are considerably active after stage 11 of embryonic development (Ferrero et al., 2014).

While these consensus sets give insight into the most reproducible marker genes, it is also of interest to characterise the $\sim 20\%$ of genes in each category identified by Allen *et al.* which were not identified by the unfiltered analysis. For all three categories of gene lists, there were too few genes to yield significant GO enrichment when compared to a background gene set of all *Drosophila* genes. The GO analysis was repeated to consider the background set as only the genes in the union of the gene lists from the two analyses, with similar inconclusive results. However, when looking only at the enrichment for the consensus set of *SoxN*-positive clusters, there was a significant enrichment for motor neuron axon guidance ($p < 10\text{E-}2$). None of the other intersection sets exhibited significant enrichments when compared to their corresponding restricted background gene sets. Taken together, though, these results show that the gene lists produced by the unfiltered analysis produced largely similar cell type clustering and marker gene lists as compared to the original analysis. This lack of difference suggests that the algorithmic deviations still preserved significant consensus genes while also allowing for more genes to be considered in each gene set.

3.3 Discussion

These computational experiments represent a more nuanced understanding of the expression of *SoxN* and *D in vivo* based on a systematic scRNA-seq approach. Most of the Sox-positive

clusters in the embryo and larval brain appear to be NPCs and maturing neuroblasts/GMCs, while in the adult VNC, Sox genes are expressed in differentiated neurons. Examining the marker genes in each cluster provided an idea of the factors that are regularly coexpressed with *SoxN* and *D*, and through GO analysis it was possible to examine the different cellular functions carried out in each tissue. Annotations for TF status as well as *SoxN* and *D* binding gave a focused view at the factors that most likely work with Sox genes directly to influence the expression of other structural genes in each cluster.

While this approach approximates a systematic and unbiased approach, there are definite pitfalls of the methods used. Within the embryonic dataset, for instance, cluster 5 showed *D* as a marker gene, but after Benjamini-Hochberg correction, it was no longer a significant marker for that cluster. The aim of this project was to be permissive in identifying potential candidates in the Sox interactome, and therefore, *D* was still considered as a marker for this cluster despite low significance. Additionally, identifying marker genes within each cluster uses the Wilcoxon rank sum test, a nonparametric measure that tests whether the prevalence and intensity of a given gene's expression level is higher within a cluster as compared to cells outside the cluster. Because of how significance is calculated, it is often the case that certain genes will not be counted as markers if they are ubiquitous and not uniquely expressed. Despite the fact that the larval brain dataset exhibited some *D* expression in multiple clusters (Figure 12c), *D* was not a significant marker gene for any cluster. However, as stated before, several of the gene targets within clusters 6, 7, and 8 are *D* binding targets in the embryo, and even a baseline level of presence may be possible for *D* to regulate expression in those cells. For this reason, the list of genes in each marker list is a suggestive but non-exhaustive look at the unique expression within different cell groups.

Many of the cell type inferences depended upon performing Gene Ontology enrichment analyses for each set of genes. This was a useful first-pass filter for understanding the main biological processes occurring in each cell cluster. However, GO analysis is a flawed metric notorious for yielding false positives (Pavlidis et al., 2012), so any GO enrichment scores must be viewed with caution. Because of how enrichment is calculated, the validity of results can vary depending on the number of genes being examined and the degree to which they are annotated with GO functions. Ostensibly, future analysis with more complete gene annotations may result in different GO terms appearing as enriched. For this reason, the GO analysis was used primarily as a guide rather than as a standard. By examining common GO terms within each cluster or group of clusters, it was possible to see how GO terms changed when focusing on subsets such as TFs and bound genes; after that, manual examination of each list was necessary for drawing meaningful conclusions from the data. Future analyses may seek to provide a finer resolution of inference by restricting the background gene sets—for instance, when evaluating only the TFs within a larger gene list, it may be beneficial to use the list of annotated TFs as a background set rather than the full *Drosophila* genome so that significant

terms are less heavily biased towards GO terms involving transcriptional regulation.

The attempted unbiased approach for processing each dataset allowed the same pipeline to apply to multiple unrelated sequencing projects. While this provided consistency between datasets, this came at the expense of congruity with the original results. In the larval brain dataset, clusters 6, 7, 8, and 10 appear to be related to one single cluster from the original analysis by Avalos *et al.*; however, because the cluster grouping in this analysis was different, it was possible to see how cluster 10 may be a more mature version of the cells in the other clusters. However, it is still possible that some of the incongruity may also be because the lack of low-count filtering in this analysis introduced statistical noise that biased the data. Despite this, comparison of clusters in the unfiltered adult VNC analysis against clusters produced originally by Allen *et al.* indicate a relatively high degree ($\sim$ 80%) of replicability along with sufficient permissiveness to identify more candidate genes than the original analysis. This indicates that despite deviations from the original analyses caused by lack of filtration or normalisation, and despite the stochastic nature of the UMAP algorithm, the identified clusters were still robust and can meaningfully be used to draw inference about the underlying regulatory biology. This reflects the fact that UMAP tends to favor high reproducibility and biological validity when compared to other nondeterministic algorithms like t-SNE and scvis (Becht et al., 2018). The unfiltered approach contained the added benefit of systematic consistency of analysis among different datasets, and largely preserved gene list structure while also identifying more potential hits.

While comparisons between datasets attempted to hold all factors constant, this was not always possible. For instance, the adult VNC dataset was an order of magnitude larger than the others, and was run on the Rustbucket server because analysis was not possible on a local machine. However, for the sake of computational efficiency, only male flies were pooled for the final dataset. This decision was arbitrary and can be trivially modified to include female flies instead; however, the failure to include female samples in this analysis can be seen as an example of sex bias in scientific research (S. K. Lee, 2018). With access to higher computing power, it would be possible to repeat the analysis with two male and two female replicates in order to lend higher validity to these findings. However, the approaches used to study the GRNs in these male flies is versatile and can be applied to a network that includes both males and females.

Overall, these computational experiments detail a process for extracting information about Sox genes from preexisting public datasets. Using GO analysis to identify significant processes, subset analysis to highlight genes of interest, and network analysis to uncover the structure of shared regulatory networks, it was possible to identify putative cell identities as well as candidate genes for further exploration within Sox-related GRNs. In the future, it would be worthwhile to extend this analysis to other scRNA-seq datasets that may include more *D* expression. Because *Dichaete* plays a conserved role in embryonic patterning, it may be

worthwhile to isolate *D*-positive cells with flow cytometry and then identify features of its interactome that may be shared between species. Additionally, wet lab experiments will be necessary to validate each of the putative leads identified in this analysis. Methods such as co-IP and pulldown assays will be necessary to validate whether each Sox factor physically interacts with these identified targets. Additionally, once the $SoxN^{mSox2}$ and D^{mSox2} constructs are ready, it will be of interest to explore whether exogenous *mSox2* interacts with target genes in the same way due to either structural similarities or regulatory forces. These data have identified roles in which *SoxN* and *D* contribute to the function and regulation of cells throughout the development of the fly CNS.

Chapter 4

Conclusion

Sox genes play a vital role in regulating developmental processes in flies and in animals. Understanding the SoxB genes *SoxN* and *D* is a valuable step in characterising the impact that homologous Sox genes may have in mice and in humans. This project represents an attempt to characterise the known functional redundancies between SoxB genes and to explore what makes them different.

Much of the wet lab portion of this thesis relies on future experimentation to validate or reject any proposed hypotheses about how *SoxN* and *D* can functionally compensated by an exogenous homologue. The procedures detailed in the wet lab portion of this thesis will provide sufficient detail for future experimenters to assemble the donor templates and then use CRISPR/Cas9 to produce knock-in flies. Whether *mSox2* is indeed capable of functionally compensating for fly Sox genes will depend on whether any phenotype is observable in the resulting progeny. This may be in the form of macroscopic developmental defects, CNS hypoplasia identifiable via BP102 antibody staining, or even molecular differences in binding targets. Assuming that the knock-in flies are viable, comparing the differences between $SoxN^{mSox2}$ and D^{mSox2} phenotypes will provide insight to how the regulatory features of the *SoxN* and *D* loci affect their expression and behavior in the cell. If successful, this will elucidate some of the mechanisms that allow the two genes to functionally compensate for each other, as well as the unique structural and regulatory factors that cause them to have different roles in development.

The computational experiments in this project represent an attempt to leverage existing data to find new insights. The known evolutionary history of Sox genes points to the idea that functional redundancy is supported by conservation in gene targets and GRNs. By systematically examining all genes at different developmental stages, it was possible to explore the expression of *SoxN* and *D in vivo* and infer not only functions specific to each cell type but also the other elements with which Sox genes share their GRNs. From this data, it was possible to identify putative neuroblast and NPC cell groups at the onset of embryonic gastrulation and in early larval development. Additionally, expression of *SoxN* in the adult ventral nerve cord reveals groups of Sox-expressing nerve cells that specialise in different functions. Examination of the

other genes that are coexpressed in these clusters has provided insight to possible regulatory networks that impact shared *SoxN* and *D* functions, and also demonstrate how these factors are able to impact a variety of cellular functions in different tissues.

While these clusters have been given putative identities, there is still more that can be learned from them. This project has looked at the markers in cells that are positive for *SoxN* and *D*. This has identified other players in these GRNs that may merit further exploration in the lab. If any of these leads are successful, the next step would be to use a similar process to explore regulatory networks in the clusters positive for those validated leads. This, in turn, represents an iterative methodology for identifying potential hits and then exploring their expression further. Potentially, it would be of interest to incorporate even more datasets that include greater spatial and temporal specificity so that it may be possible to examine Sox genes without being influenced by background expression from cells that do not express these factors.

Sox genes are important in development throughout the animal kingdom, and SoxB genes in particular play a vital role in segmentation and neurodevelopment. The experiments and analyses conducted as part of this thesis project provide a starting point for future investigators to look at Sox expression and regulatory networks. Discoveries regarding *SoxN* and *D* expression may immediately translate to a better understanding of arthropod developmental regulation, but homology to mammalian group B Sox genes paves a way for these insights to potentially impact our understanding of human stem cell maintenance and neurodevelopment.

References

Abdusselamoglu, M. D., Eroglu, E., Burkard, T. R., & Knoblich, J. A. (2019). The transcription factor odd-paired regulates temporal identity in transit-amplifying neural progenitors via an incoherent feed-forward loop. *eLife*, *8*. doi: 10.7554/eLife.46566

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195. doi: 10.1126/science.287.5461.2185

Adryan, B., & Teichmann, S. A. (2006). FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. *Bioinformatics*, *22*(12), 1532–1533. Retrieved from `https://www.mrc-lmb.cam.ac.uk/genomes/FlyTF/old_index.html`

Aleksic, J., Ferrero, E., Fischer, B., Shen, S. P., & Russell, S. (2013). The role of Dichaete in transcriptional regulation during Drosophila embryonic development. *BMC Genomics*, *14*, 861. doi: 10.1186/1471-2164-14-861

Alexa, A., & Rahnenfuhrer, J. (2020). *topGO: Enrichment Analysis for Gene Ontology*. Bioconductor version: Release (3.11). doi: 10.18129/B9.bioc.topGO

Alexa, A., Rahnenführer, J., & Lengauer, T. (2006). Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, *22*(13), 1600–1607. doi: 10.1093/bioinformatics/btl140

Allen, A. M., Neville, M. C., Birtles, S., Croset, V., Treiber, C. D., Waddell, S., & Goodwin, S. F. (2020). A single-cell transcriptomic atlas of the adult Drosophila ventral nerve cord. *eLife*, *9*, e54074. doi: 10.7554/eLife.54074

Apitz, H., & Salecker, I. (2015). A region-specific neurogenesis mode requires migratory progenitors in the Drosophila visual system. *Nature Neuroscience*, *18*(1), 46–55. doi: 10.1038/nn.3896

Arefin, B., Parvin, F., Bahrampour, S., Stadler, C. B., & Thor, S. (2019). Drosophila Neuroblast Selection Is Gated by Notch, Snail, SoxB, and EMT Gene Interplay. *Cell Reports*, *29*(11), 3636–3651.e3. doi: 10.1016/j.celrep.2019.11.038

Baran-Gale, J., Chandra, T., & Kirschner, K. (2018). Experimental design for single-cell RNA sequencing. *Brief Funct Genomics*, *17*(4), 233–239. doi: 10.1093/bfgp/elx035

Bardin, A. J., & Schweisguth, F. (2006). Bearded family members inhibit Neuralized-mediated

endocytosis and signaling activity of Delta in Drosophila. *Dev. Cell*, *10*(2), 245–255. doi: 10.1016/j.devcel.2005.12.017

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., … Newell, E. W. (2018, December). Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol*. doi: 10.1038/nbt.4314

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bergsland, M., Ramsköld, D., Zaouter, C., Klum, S., Sandberg, R., & Muhr, J. (2011). Sequentially acting Sox transcription factors in neural lineage development. *Genes Dev.*, *25*(23), 2453–2464. doi: 10.1101/gad.176008.111

Bernard, P., & Harley, V. R. (2010). Acquisition of SOX transcription factor specificity through protein-protein interaction, modulation of Wnt signalling and post-translational modification. *Int. J. Biochem. Cell Biol.*, *42*(3), 400–410. doi: 10.1016/j.biocel.2009.10.017

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech.*, *2008*(10), P10008. doi: 10.1088/1742-5468/2008/10/P10008

Bonatto Paese, C. L., Leite, D. J., Schönauer, A., McGregor, A. P., & Russell, S. (2018). Duplication and expression of Sox genes in spiders. *BMC Evolutionary Biology*, *18*(1), 205. doi: 10.1186/s12862-018-1337-4

Bondurand, N., Pingault, V., Goerich, D. E., Lemort, N., Sock, E., Caignec, C. L., … Goossens, M. (2000). Interaction among SOX10, PAX3 and MITF, three genes altered in Waardenburg syndrome. *Hum Mol Genet*, *9*(13), 1907–1917. doi: 10.1093/hmg/9.13.1907

Bowles, J., Schepers, G., & Koopman, P. (2000). Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev. Biol.*, *227*(2), 239–255. doi: 10.1006/dbio.2000.9883

Bradley, R. K., Li, X.-Y., Trapnell, C., Davidson, S., Pachter, L., Chu, H. C., … Eisen, M. B. (2010). Binding Site Turnover Produces Pervasive Quantitative Changes in Transcription Factor Binding between Closely Related Drosophila Species. *PLOS Biology*, *8*(3), e1000343. doi: 10.1371/journal.pbio.1000343

Brunet Avalos, C., Maier, G. L., Bruggmann, R., & Sprecher, S. G. (2019). Single cell transcriptome atlas of the Drosophila larval brain. *eLife*, *8*, e50354. doi: 10.7554/eLife.50354

Buescher, M., Hing, F. S., & Chia, W. (2002). Formation of neuroblasts in the embryonic central nervous system of Drosophila melanogaster is controlled by SoxNeuro. *Development*, *129*(18), 4193–4203.

Bylund, M., Andersson, E., Novitch, B. G., & Muhr, J. (2003). Vertebrate neurogenesis is counteracted by Sox1–3 activity. *Nature Neuroscience*, *6*(11), 1162–1168. doi: 10.1038/nn1131

Carl, S. H., & Russell, S. (2015). Common binding by redundant group B Sox proteins is evolutionarily conserved in Drosophila. *BMC Genomics*, *16*(1), 292. doi: 10.1186/s12864-015-1495-3

Carlson, M. (2019). *org.Dm.eg.db: Genome wide annotation for fly.* Retrieved from `https://bioconductor.org/packages/org.Dm.eg.db/`

Chang, Y. K., Srivastava, Y., Hu, C., Joyce, A., Yang, X., Zuo, Z., … Jauch, R. (2017). Quantitative profiling of selective Sox/POU pairing on hundreds of sequences in parallel by Coop-seq. *Nucleic Acids Res*, *45*(2), 832–845. doi: 10.1093/nar/gkw1198

Chao, A. T., Jones, W. M., & Bejsovec, A. (2007). The HMG-box transcription factor SoxNeuro acts with Tcf to control Wg/Wnt signaling activity. *Development*, *134*(5), 989–997. doi: 10.1242/dev.02796

Charron, F., & Tessier-Lavigne, M. (2007). The Hedgehog, TGF-beta/BMP and Wnt Families of Morphogens in Axon Guidance. In D. Bagnard (Ed.), *Axon Growth and Guidance* (pp. 116–133). New York, NY: Springer. doi: 10.1007/978-0-387-76715-4_9

Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Front. Genet.*, *10*. doi: 10.3389/fgene.2019.00317

Chen, J., Xu, N., Huang, H., Cai, T., & Xi, R. (2016). A feedback amplification loop between stem cells and their progeny promotes tissue regeneration and tumorigenesis. *eLife*, *5*, e14330. doi: 10.7554/eLife.14330

Chen, S.-Y., Feng, Z., & Yi, X. (2017). A general introduction to adjustment for multiple comparisons. *J Thorac Dis*, *9*(6), 1725–1729. doi: 10.21037/jtd.2017.05.34

Chew, L.-J., & Gallo, V. (2009). The Yin and yang of Sox proteins: activation and repression in development and disease. *J Neurosci Res*, *87*(15), 3277–3287. doi: 10.1002/jnr.22128

Chung, N. C., & Storey, J. D. (2015). Statistical significance of variables driving systematic variation in high-dimensional data. *Bioinformatics*, *31*(4), 545–554. doi: 10.1093/bioinformatics/btu674

Clark, E., & Peel, A. D. (2018). Evidence for the temporal regulation of insect segmentation by a conserved sequence of transcription factors. *Development*, *145*(10). doi: 10.1242/dev.155580

Contreras, E. G., Egger, B., Gold, K. S., & Brand, A. H. (2018). Dynamic Notch signalling regulates neural stem cell state progression in the Drosophila optic lobe. *Neural Development*, *13*(1), 25. doi: 10.1186/s13064-018-0123-8

Conway, J. R., Lex, A., & Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, *33*(18), 2938–2940. doi: 10.1093/bioinformatics/btx364

Crémazy, F., Berta, P., & Girard, F. (2000). Sox neuro, a new Drosophila Sox gene expressed in the developing central nervous system. *Mech. Dev.*, *93*(1-2), 215–219. doi: 10.1016/s0925-4773(00)00268-9

Crémazy, F., Berta, P., & Girard, F. (2001). Genome-wide analysis of Sox genes in Drosophila melanogaster. *Mech. Dev.*, *109*(2), 371–375. doi: 10.1016/s0925-4773(01)00529-9

Czaja, W., Miller, K. Y., Skinner, M. K., & Miller, B. L. (2014). Structural and functional conservation of fungal MatA and human SRY sex-determining proteins. *Nature Communications*, *5*(1), 5434. doi: 10.1038/ncomms6434

Davie, K., Janssens, J., Koldere, D., Waegeneer, M. D., Pech, U., Kreft, L., … Aerts, S. (2018). A Single-Cell Transcriptome Atlas of the Aging Drosophila Brain. *Cell*, *174*(4), 982–998.e20. doi: 10.1016/j.cell.2018.05.057

Dearden, P. K. (2015). Origin and evolution of the enhancer of split complex. *BMC Genomics*, *16*(1). doi: 10.1186/s12864-015-1926-1

Dodonova, S. O., Zhu, F., Dienemann, C., Taipale, J., & Cramer, P. (2020). Nucleosome-bound SOX2 and SOX11 structures elucidate pioneer factor function. *Nature*, *580*(7805), 669–672. doi: 10.1038/s41586-020-2195-y

Doupé, D. P., Marshall, O. J., Dayton, H., Brand, A. H., & Perrimon, N. (2018). Drosophila intestinal stem and progenitor cells are major sources and regulators of homeostatic niche signals. *PNAS*, *115*(48), 12218–12223. doi: 10.1073/pnas.1719169115

Efroni, S., Duttagupta, R., Cheng, J., Dehghani, H., Hoeppner, D. J., Dash, C., … Meshorer, E. (2008). Global transcription in pluripotent embryonic stem cells. *Cell Stem Cell*, *2*(5), 437–447. doi: 10.1016/j.stem.2008.03.021

Episkopou, V. (2005). SOX2 functions in adult neural stem cells. *Trends Neurosci.*, *28*(5), 219–221. doi: 10.1016/j.tins.2005.03.003

Estacio-Gómez, A., Hassan, A., Walmsley, E., Le, L. W., & Southall, T. D. (2020). Dynamic neurotransmitter specific transcription factor expression profiles during Drosophila development. *Biology Open*, *9*(5). doi: 10.1242/bio.052928

Ferrero, E., Fischer, B., & Russell, S. (2014). SoxNeuro orchestrates central nervous system specification and differentiation in Drosophila and is only partially redundant with Dichaete. *Genome Biol.*, *15*(5), R74. doi: 10.1186/gb-2014-15-5-r74

Ferri, A. L. M., Cavallaro, M., Braida, D., Cristofano, A. D., Canta, A., Vezzani, A., … Nicolis, S. K. (2004). Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development*, *131*(15), 3805–3819. doi: 10.1242/dev.01204

Francois, M., Koopman, P., & Beltrame, M. (2010). SoxF genes: Key players in the development of the cardio-vascular system. *The International Journal of Biochemistry & Cell Biology*, *42*(3), 445–448. doi: 10.1016/j.biocel.2009.08.017

Furuyama, K., Kawaguchi, Y., Akiyama, H., Horiguchi, M., Kodama, S., Kuhara, T., … Uemoto, S. (2011). Continuous cell supply from a Sox9-expressing progenitor zone in adult liver, exocrine pancreas and intestine. *Nat. Genet.*, *43*(1), 34–41. doi: 10.1038/ng.722

Gao, X., Hu, D., Gogol, M., & Li, H. (2019). ClusterMap: compare multiple single cell RNA-Seq datasets across different experimental conditions. *Bioinformatics*, *35*(17), 3038–3045. doi:

10.1093/bioinformatics/btz024

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., … Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi: 10.1186/gb-2004-5-10-r80

Graves, J. A. M. (1998). Interactions between SRY and SOX genes in mammalian sex determination. *BioEssays*, *20*(3), 264–269. doi: 10.1002/(SICI)1521-1878(199803)20:3⟨264::AID-BIES10⟩3.0.CO;2-1

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou, A., Münsterberg, A., … Lovell-Badge, R. (1990). A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*, *346*(6281), 245–250. doi: 10.1038/346245a0

Guth, S. I. E., & Wegner, M. (2008). Having it both ways: Sox protein function between conservation and innovation. *Cell. Mol. Life Sci.*, *65*(19), 3000–3018. doi: 10.1007/s00018-008-8138-7

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, *20*(1), 296. doi: 10.1186/s13059-019-1874-1

Hammonds, A. S., Bristow, C. A., Fisher, W. W., Weiszmann, R., Wu, S., Hartenstein, V., … Celniker, S. E. (2013). Spatial expression of transcription factors in Drosophila embryonic organ development. *Genome Biol.*, *14*(12), R140. doi: 10.1186/gb-2013-14-12-r140

Hartenstein, V., & Wodarz, A. (2013). Initial neurogenesis in Drosophila. *Wiley Interdiscip Rev Dev Biol*, *2*(5), 701–721. doi: 10.1002/wdev.111

Hashimshony, T., Wagner, F., Sher, N., & Yanai, I. (2012). CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. *Cell Rep*, *2*(3), 666–673. doi: 10.1016/j.celrep.2012.08.003

Hebenstreit, D. (2012). Methods, Challenges and Potentials of Single Cell RNA-seq. *Biology (Basel)*, *1*(3), 658–667. doi: 10.3390/biology1030658

Heenan, P., Zondag, L., & Wilson, M. J. (2016). Evolution of the Sox gene family within the chordate phylum. *Gene*, *575*(2, Part 2), 385–392. doi: 10.1016/j.gene.2015.09.013

Hokari, R., Kitagawa, N., Watanabe, C., Komoto, S., Kurihara, C., Okada, Y., … Miura, S. (2008). Changes in regulatory molecules for lymphangiogenesis in intestinal lymphangiectasia with enteric protein loss. *Journal of Gastroenterology and Hepatology*, *23*(7pt2), e88–e95. doi: 10.1111/j.1440-1746.2007.05225.x

Islam, S., Kjällquist, U., Moliner, A., Zajac, P., Fan, J.-B., Lönnerberg, P., & Linnarsson, S. (2011). Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.*, *21*(7), 1160–1167. doi: 10.1101/gr.110882.110

Jager, M., Quéinnec, E., Houliston, E., & Manuel, M. (2006). Expansion of the SOX gene family predated the emergence of the Bilateria. *Molecular Phylogenetics and Evolution*, *39*(2), 468–477. doi: 10.1016/j.ympev.2005.12.005

Janssen, R., Andersson, E., Betnér, E., Bijl, S., Fowler, W., Höök, L., … Tiemann, S. (2018). Embryonic expression patterns and phylogenetic analysis of panarthropod sox genes: insight into nervous system development, segmentation and gonadogenesis. *BMC Evol. Biol.*, *18*(1), 88. doi: 10.1186/s12862-018-1196-z

Ji, E. H., & Kim, J. (2016). SoxD Transcription Factors: Multifaceted Players of Neural Development. *Int J Stem Cells*, *9*(1), 3–8. doi: 10.15283/ijsc.2016.9.1.3

Jiang, T., Hou, C.-C., She, Z.-Y., & Yang, W.-X. (2013). The SOX gene family: function and regulation in testis determination and male fertility maintenance. *Mol. Biol. Rep.*, *40*(3), 2187–2194. doi: 10.1007/s11033-012-2279-3

Kamachi, Y., & Kondoh, H. (2013). Sox proteins: regulators of cell fate specification and differentiation. *Development*, *140*(20), 4129–4144. doi: 10.1242/dev.091793

Kamachi, Y., Uchikawa, M., & Kondoh, H. (2000). Pairing SOX off: with partners in the regulation of embryonic development. *Trends Genet.*, *16*(4), 182–187. doi: 10.1016/s0168-9525(99)01955-1

Kamachi, Y., Uchikawa, M., Tanouchi, A., Sekido, R., & Kondoh, H. (2001). Pax6 and SOX2 form a co-DNA-binding partner complex that regulates initiation of lens development. *Genes Dev*, *15*(10), 1272–1286. doi: 10.1101/gad.887101

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., … Zinzen, R. P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science*, *358*(6360), 194–199. doi: 10.1126/science.aan3235

Karnavas, T., Mandalos, N., Malas, S., & Remboutsika, E. (2013). SoxB, cell cycle and neurogenesis. *Front Physiol*, *4*. doi: 10.3389/fphys.2013.00298

Katsura, Y., Kondo, H. X., Ryan, J., Harley, V., & Satta, Y. (2018). The evolutionary process of mammalian sex determination genes focusing on marsupial SRYs. *BMC Evol Biol*, *18*. doi: 10.1186/s12862-018-1119-z

Kondoh, H., & Kamachi, Y. (2010). SOX–partner code for cell specification: Regulatory target selection and underlying molecular mechanisms. *The International Journal of Biochemistry & Cell Biology*, *42*(3), 391–399. doi: 10.1016/j.biocel.2009.09.003

Kondoh, H., Uchikawa, M., & Kamachi, Y. (2004). Interplay of Pax6 and SOX2 in lens development as a paradigm of genetic switch mechanisms for cell differentiation. *Int. J. Dev. Biol.*, *48*(8-9), 819–827. doi: 10.1387/ijdb.041868hk

Kormish, J. D., Sinner, D., & Zorn, A. M. (2010). Interactions between SOX factors and Wnt/beta-catenin signaling in development and disease. *Dev Dyn*, *239*(1), 56–68. doi: 10.1002/dvdy.22046

Lai, E. C., Bodner, R., & Posakony, J. W. (2000). The enhancer of split complex of Drosophila includes four Notch-regulated members of the bearded gene family. *Development*, *127*(16), 3441–3455.

Lee, J. H., Daugharthy, E. R., Scheiman, J., Kalhor, R., Ferrante, T. C., Yang, J. L., … Church,

G. M. (2014). Highly multiplexed subcellular RNA sequencing in situ. *Science*, *343*(6177), 1360–1363. doi: 10.1126/science.1250212

Lee, S. K. (2018). Sex as an important biological variable in biomedical research. *BMB Rep*, *51*(4), 167–173. doi: 10.5483/BMBRep.2018.51.4.034

Lefebvre, V. (2010). The SoxD transcription factors – Sox5, Sox6, and Sox13 – are key cell fate modulators. *Int J Biochem Cell Biol*, *42*(3), 429–432. doi: 10.1016/j.biocel.2009.07.016

Lefebvre, V., & Bhattaram, P. (2016). SOXC genes and the control of skeletogenesis. *Curr Osteoporos Rep*, *14*(1), 32–38. doi: 10.1007/s11914-016-0296-1

Lefebvre, V., Dumitriu, B., Penzo-Méndez, A., Han, Y., & Pallavi, B. (2007). Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *Int. J. Biochem. Cell Biol.*, *39*(12), 2195–2214. doi: 10.1016/j.biocel.2007.05.019

Li, G., & Neuert, G. (2019). Multiplex RNA single molecule FISH of inducible mRNAs in single yeast cells. *Scientific Data*, *6*(1), 94. doi: 10.1038/s41597-019-0106-6

Li, L., & Vaessin, H. (2000). Pan-neural Prospero terminates cell proliferation during Drosophila neurogenesis. *Genes Dev*, *14*(2), 147–151.

Li, X., Erclik, T., Bertet, C., Chen, Z., Voutev, R., Venkatesh, S., … Desplan, C. (2013). Temporal patterning of Drosophila medulla neuroblasts controls neural fates. *Nature*, *498*(7455), 456–462. doi: 10.1038/nature12319

Liu, P. Z., & Kaufman, T. C. (2005). Short and long germ segmentation: unanswered questions in the evolution of a developmental mode. *Evolution & Development*, *7*(6), 629–646. doi: 10.1111/j.1525-142X.2005.05066.x

Lodato, M. A., Ng, C. W., Wamstad, J. A., Cheng, A. W., Thai, K. K., Fraenkel, E., … Boyer, L. A. (2013). SOX2 Co-Occupies Distal Enhancer Elements with Distinct POU Factors in ESCs and NPCs to Specify Cell State. *PLOS Genetics*, *9*(2), e1003288. doi: 10.1371/journal.pgen.1003288

Loureiro, J., & Peifer, M. (1998). Roles of Armadillo, a Drosophila catenin, during central nervous system development. *Curr. Biol.*, *8*(11), 622–632. doi: 10.1016/s0960-9822(98)70249-0

Ludwig, A., Rehberg, S., & Wegner, M. (2004). Melanocyte-specific expression of dopachrome tautomerase is dependent on synergistic gene activation by the Sox10 and Mitf transcription factors. *FEBS Lett.*, *556*(1-3), 236–244. doi: 10.1016/s0014-5793(03)01446-7

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., … Micklem, G. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. *Genome Biol.*, *8*(7), R129. doi: 10.1186/gb-2007-8-7-r129

Ma, Y., Certel, K., Gao, Y., Niemitz, E., Mosher, J., Mukherjee, A., … Nambu, J. R. (2000). Functional Interactions between Drosophila bHLH/PAS, Sox, and POU Transcription Factors Regulate CNS Midline Expression of the slit Gene. *J Neurosci*, *20*(12), 4596–4605. doi: 10.1523/JNEUROSCI.20-12-04596.2000

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll,

S. A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, *161*(5), 1202–1214. doi: 10.1016/j.cell.2015.05.002

Marais, G., Nouvellet, P., Keightley, P. D., & Charlesworth, B. (2005). Intron Size and Exon Evolution in Drosophila. *Genetics*, *170*(1), 481–485. doi: 10.1534/genetics.104.037333

McElreavy, K., Vilain, E., Abbas, N., Costa, J. M., Souleyreau, N., Kucheria, K., … Flamant, F. (1992). XY sex reversal associated with a deletion 5' to the SRY "HMG box" in the testis-determining region. *PNAS*, *89*(22), 11016–11020. doi: 10.1073/pnas.89.22.11016

McFarland, K. A., Topczewska, J. M., Weidinger, G., Dorsky, R. I., & Appel, B. (2008). Hh and Wnt signaling regulate formation of olig2+ neurons in the zebrafish cerebellum. *Developmental Biology*, *318*(1), 162–171. doi: 10.1016/j.ydbio.2008.03.016

McGonigle, I., & Lummis, S. C. R. (2009). RDL receptors. *Biochem. Soc. Trans.*, *37*(Pt 6), 1404–1406. doi: 10.1042/BST0371404

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861. doi: 10.21105/joss.00861

McKimmie, C., Woerfel, G., & Russell, S. (2005). Conserved genomic organisation of Group B Sox genes in insects. *BMC Genet.*, *6*, 26. doi: 10.1186/1471-2156-6-26

Melnattur, K. V., Berdnik, D., Rusan, Z., Ferreira, C. J., & Nambu, J. R. (2013). The Sox gene Dichaete is expressed in local interneurons and functions in development of the Drosophila adult olfactory circuit. *Dev Neurobiol*, *73*(2), 107–126. doi: 10.1002/dneu.22038

Meng, F., & Biteau, B. (2015). A Sox Transcription Factor Is a Critical Regulator of Adult Stem Cell Proliferation in the Drosophila Intestine. *Cell Reports*, *13*(5), 906–914. doi: 10.1016/j.celrep.2015.09.061

Mertin, S., McDowall, S. G., & Harley, V. R. (1999). The DNA-binding specificity of SOX9 and other SOX proteins. *Nucleic Acids Res.*, *27*(5), 1359–1364. doi: 10.1093/nar/27.5.1359

Mistri, T. K., Devasia, A. G., Chu, L. T., Ng, W. P., Halbritter, F., Colby, D., … Wohland, T. (2015). Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells. *EMBO reports*, *16*(9), 1177–1191. doi: 10.15252/embr.201540467

Morata, G., & Lawrence, P. A. (1979). Development of the eye-antenna imaginal disc of Drosophila. *Developmental Biology*, *70*(2), 355–371. doi: 10.1016/0012-1606(79)90033-2

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628. doi: 10.1038/nmeth.1226

Mu, L., Berti, L., Masserdotti, G., Covic, M., Michaelidis, T. M., Doberauer, K., … Lie, D. C. (2012). SoxC Transcription Factors Are Required for Neuronal Differentiation in Adult Hippocampal Neurogenesis. *J. Neurosci.*, *32*(9), 3067–3080. doi: 10.1523/JNEUROSCI.4679-11.2012

Nakagawa, M., Koyanagi, M., Tanabe, K., Takahashi, K., Ichisaka, T., Aoi, T., … Yamanaka, S. (2008). Generation of induced pluripotent stem cells without Myc from mouse and

human fibroblasts. *Nat. Biotechnol.*, *26*(1), 101–106. doi: 10.1038/nbt1374

Nambu, P. A., & Nambu, J. R. (1996). The Drosophila fish-hook gene encodes a HMG domain protein essential for segmentation and CNS development. *Development*, *122*(11), 3467–3475.

Neriec, N., & Desplan, C. (2014). Different ways to make neurons: parallel evolution in the SoxB family. *Genome Biol*, *15*(5), 116. doi: 10.1186/gb4177

Niwa, H., Nakamura, A., Urata, M., Shirae-Kurabayashi, M., Kuraku, S., Russell, S., & Ohtsuka, S. (2016). The evolutionally-conserved function of group B1 Sox family members confers the unique role of Sox2 in mouse ES cells. *BMC Evolutionary Biology*, *16*(1), 173. doi: 10.1186/s12862-016-0755-4

Nowling, T. K., Johnson, L. R., Wiebe, M. S., & Rizzino, A. (2000). Identification of the transactivation domain of the transcription factor Sox-2 and an associated co-activator. *J. Biol. Chem.*, *275*(6), 3810–3818. doi: 10.1074/jbc.275.6.3810

Overton, P. M. (2003). *The role of sox genes in the development of* Drosophila melanogaster (Unpublished doctoral dissertation). University of Cambridge.

Overton, P. M., Chia, W., & Buescher, M. (2007). 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. doi: 10.1242/dev.02878

Overton, P. M., Meadows, L. A., Urban, J., & Russell, S. (2002). Evidence for differential and redundant function of the Sox genes Dichaete and SoxN during CNS development in Drosophila. *Development*, *129*(18), 4219–4228.

Paese, C. L. B., Schoenauer, A., Leite, D. J., Russell, S., & McGregor, A. P. (2018). A SoxB gene acts as an anterior gap gene and regulates posterior segment addition in a spider. *Elife*, *7*. doi: 10.7554/eLife.37567

Pavlidis, P., Jensen, J. D., Stephan, W., & Stamatakis, A. (2012). A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. *Mol Biol Evol*, *29*(10), 3237–3248. doi: 10.1093/molbev/mss136

Peel, A. D., Chipman, A. D., & Akam, M. (2005). Arthropod segmentation: beyond the Drosophila paradigm. *Nat. Rev. Genet.*, *6*(12), 905–916. doi: 10.1038/nrg1724

Phochanukul, N., & Russell, S. (2010). No backbone but lots of Sox: Invertebrate Sox genes. *Int. J. Biochem. Cell Biol.*, *42*(3), 453–464. doi: 10.1016/j.biocel.2009.06.013

Porcelli, D., Fischer, B., Russell, S., & White, R. (2019). Chromatin accessibility plays a key role in selective targeting of Hox proteins. *Genome Biology*, *20*(1), 115. doi: 10.1186/s13059-019-1721-4

Prior, H. M., & Walter, M. A. (1996). SOX genes: architects of development. *Mol. Med.*, *2*(4), 405–412.

Qiu, J., McQueen, J., Bilican, B., Dando, O., Magnani, D., Punovuori, K., … Hardingham, G. E.

(2016). Evidence for evolutionary divergence of activity-dependent gene expression in developing neurons. *eLife*, *5*, e20337. doi: 10.7554/eLife.20337

Ramsköld, D., Luo, S., Wang, Y.-C., Li, R., Deng, Q., Faridani, O. R., … Sandberg, R. (2012). Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.*, *30*(8), 777–782. doi: 10.1038/nbt.2282

Reményi, A., Lins, K., Nissen, L. J., Reinbold, R., Schöler, H. R., & Wilmanns, M. (2003). Crystal structure of a POU/HMG/DNA ternary complex suggests differential assembly of Oct4 and Sox2 on two enhancers. *Genes Dev*, *17*(16), 2048–2059. doi: 10.1101/gad.269303

Rimini, R., Pontiggia, A., Spada, F., Ferrari, S., Harley, V. R., Goodfellow, P. N., & Bianchi, M. E. (1995). Interaction of normal and mutant SRY proteins with DNA. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, *350*(1333), 215–220. doi: 10.1098/rstb.1995.0154

Russell, S. R., Sanchez-Soriano, N., Wright, C. R., & Ashburner, M. (1996). The Dichaete gene of Drosophila melanogaster encodes a SOX-domain protein required for embryonic segmentation. *Development*, *122*(11), 3669–3676.

Sandberg, M., Källström, M., & Muhr, J. (2005). Sox21 promotes the progression of vertebrate neurogenesis. *Nature Neuroscience*, *8*(8), 995–1001. doi: 10.1038/nn1493

Sarkar, A., & Hochedlinger, K. (2013). The Sox Family of Transcription Factors: Versatile Regulators of Stem and Progenitor Cell Fate. *Cell Stem Cell*, *12*(1), 15–30. doi: 10.1016/j.stem.2012.12.007

Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), 495–502. doi: 10.1038/nbt.3192

Saunders, A., Core, L. J., Sutcliffe, C., Lis, J. T., & Ashe, H. L. (2013). Extensive polymerase pausing during Drosophila axis patterning enables high-level and pliable transcription. *Genes Dev.*, *27*(10), 1146–1158. doi: 10.1101/gad.215459.113

Schepers, G. E., Teasdale, R. D., & Koopman, P. (2002). Twenty Pairs of Sox: Extent, Homology, and Nomenclature of the Mouse and Human Sox Transcription Factor Gene Families. *Developmental Cell*, *3*(2), 167–170. doi: 10.1016/S1534-5807(02)00223-X

Schilling, T., Ali, A. H., Leonhardt, A., Borst, A., & Pujol-Martí, J. (2019). Transcriptional control of morphological properties of direction-selective T4/T5 neurons in Drosophila. *Development*, *146*(2). doi: 10.1242/dev.169763

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., … Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research*, *13*(11), 2498. doi: 10.1101/gr.1239303

She, Z.-Y., & Yang, W.-X. (2017). Sry and SoxE genes: How they participate in mammalian sex determination and gonadal development? *Seminars in Cell & Developmental Biology*, *63*, 13–22. doi: 10.1016/j.semcdb.2016.07.032

Shen, S. P., Aleksic, J., & Russell, S. (2013). Identifying targets of the Sox domain protein

Dichaete in the Drosophila CNS via targeted expression of dominant negative proteins. *BMC Dev. Biol.*, *13*, 1. doi: 10.1186/1471-213X-13-1

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

Soriano, N. S., & Russell, S. (1998). The Drosophila SOX-domain protein Dichaete is required for the development of the central nervous system midline. *Development*, *125*(20), 3989–3996.

Suzuki, T., Kaido, M., Takayama, R., & Sato, M. (2013). A temporal mechanism that produces neuronal diversity in the Drosophila visual center. *Developmental Biology*, *380*(1), 12–24. doi: 10.1016/j.ydbio.2013.05.002

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., … Mering, C. v. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, *47*(D1), D607–D613. doi: 10.1093/nar/gky1131

Sánchez-Soriano, N., & Russell, S. (2000). Regulatory mutations of the Drosophila Sox gene Dichaete reveal new functions in embryonic brain and hindgut development. *Dev. Biol.*, *220*(2), 307–321. doi: 10.1006/dbio.2000.9648

Südbeck, P., & Scherer, G. (1997). Two Independent Nuclear Localization Signals Are Present in the DNA-binding High-mobility Group Domains of SRY and SOX9. *J. Biol. Chem.*, *272*(44), 27848–27852. doi: 10.1074/jbc.272.44.27848

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676. doi: 10.1016/j.cell.2006.07.024

Tanaka, S., Kamachi, Y., Tanouchi, A., Hamada, H., Jing, N., & Kondoh, H. (2004). Interplay of SOX and POU factors in regulation of the Nestin gene in neural primordial cells. *Mol. Cell. Biol.*, *24*(20), 8834–8846. doi: 10.1128/MCB.24.20.8834-8846.2004

Tantin, D. (2013). Oct transcription factors in development and stem cells: insights and mechanisms. *Development*, *140*(14), 2857–2866. doi: 10.1242/dev.095927

Thomas, J. O., & Travers, A. A. (2001). HMG1 and 2, and related 'architectural' DNA-binding proteins. *Trends in Biochemical Sciences*, *26*(3), 167–174. doi: 10.1016/S0968-0004(01)01801-1

Tolwinski, N. S. (2017). Introduction: Drosophila—A Model System for Developmental Biology. *J Dev Biol*, *5*(3). doi: 10.3390/jdb5030009

Tomancak, P., Beaton, A., Weiszmann, R., Kwan, E., Shu, S., Lewis, S. E., … Rubin, G. M. (2002). Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, *3*(12), RESEARCH0088. doi: 10.1186/gb-2002-3-12-research0088

Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., … Ru-

bin, G. M. (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, *8*(7), R145. doi: 10.1186/gb-2007-8-7-r145

Uchikawa, M., Kamachi, Y., & Kondoh, H. (1999). Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken. *Mech. Dev.*, *84*(1-2), 103–120. doi: 10.1016/s0925-4773(99)00083-0

Uchikawa, M., Yoshida, M., Iwafuchi-Doi, M., Matsuda, K., Ishida, Y., Takemoto, T., & Kondoh, H. (2011). B1 and B2 Sox gene expression during neural plate development in chicken and mouse embryos: universal versus species-dependent features. *Dev. Growth Differ.*, *53*(6), 761–771. doi: 10.1111/j.1440-169X.2011.01286.x

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3—new capabilities and interfaces. *Nucleic Acids Res*, *40*(15), e115. doi: 10.1093/nar/gks596

Wallis, M. C., Waters, P. D., Delbridge, M. L., Kirby, P. J., Pask, A. J., Grützner, F., … Graves, J. a. M. (2007). Sex determination in platypus and echidna: autosomal location of SOX3 confirms the absence of SRY from monotremes. *Chromosome Res.*, *15*(8), 949–959. doi: 10.1007/s10577-007-1185-3

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B*, *86*(11), 471. doi: 10.1140/epjb/e2013-40829-0

Wegner, M. (2010). All purpose Sox: The many roles of Sox proteins in gene expression. *Int. J. Biochem. Cell Biol.*, *42*(3), 381–390. doi: 10.1016/j.biocel.2009.07.006

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L. J., … Odom, D. T. (2008). Species-Specific Transcription in Mice Carrying Human Chromosome 21. *Science*, *322*(5900), 434–438. doi: 10.1126/science.1160930

Wilson, M. J., & Dearden, P. K. (2008). Evolution of the insect Sox genes. *BMC Evol Biol*, *8*, 120. doi: 10.1186/1471-2148-8-120

Wittkopp, P. J. (2010). Variable Transcription Factor Binding: A Mechanism of Evolutionary Change. *PLOS Biology*, *8*(3), e1000342. doi: 10.1371/journal.pbio.1000342

Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., … Elgar, G. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, *3*(1), e7. doi: 10.1371/journal.pbio.0030007

Yankura, K. A., Koechlein, C. S., Cryan, A. F., Cheatle, A., & Hinman, V. F. (2013). Gene regulatory network for neurogenesis in a sea star embryo connects broad neural specification and localized patterning. *PNAS*, *110*(21), 8591–8596. doi: 10.1073/pnas.1220903110

Yuan, H., Corbi, N., Basilico, C., & Dailey, L. (1995). Developmental-specific activity of the

FGF-4 enhancer requires the synergistic action of Sox2 and Oct-3. *Genes Dev.*, *9*(21), 2635–2645. doi: 10.1101/gad.9.21.2635

Zhai, Z., Boquete, J.-P., & Lemaitre, B. (2017). A genetic framework controlling the differentiation of intestinal stem cells during regeneration in Drosophila. *PLOS Genetics*, *13*(6), e1006854. doi: 10.1371/journal.pgen.1006854

Zhai, Z., Kondo, S., Ha, N., Boquete, J.-P., Brunner, M., Ueda, R., & Lemaitre, B. (2015). Accumulation of differentiating intestinal stem cell progenies drives tumorigenesis. *Nature Communications*, *6*(1), 10219. doi: 10.1038/ncomms10219

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

Zhao, G., Wheeler, S. R., & Skeath, J. B. (2007). Genetic control of dorsoventral patterning and neuroblast specification in the Drosophila Central Nervous System. *Int. J. Dev. Biol.*, *51*(2), 107–115. doi: 10.1387/ijdb.062188gz

Zhong, L., Wang, D., Gan, X., Yang, T., & He, S. (2011). Parallel Expansions of Sox Transcription Factor Group B Predating the Diversifications of the Arthropods and Jawed Vertebrates. *PLOS ONE*, *6*(1), e16570. doi: 10.1371/journal.pone.0016570

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R., & Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *BioTechniques*, *30*(4), 892–897. doi: 10.2144/01304pf02

Appendix A

Supplementary Figures and Tables

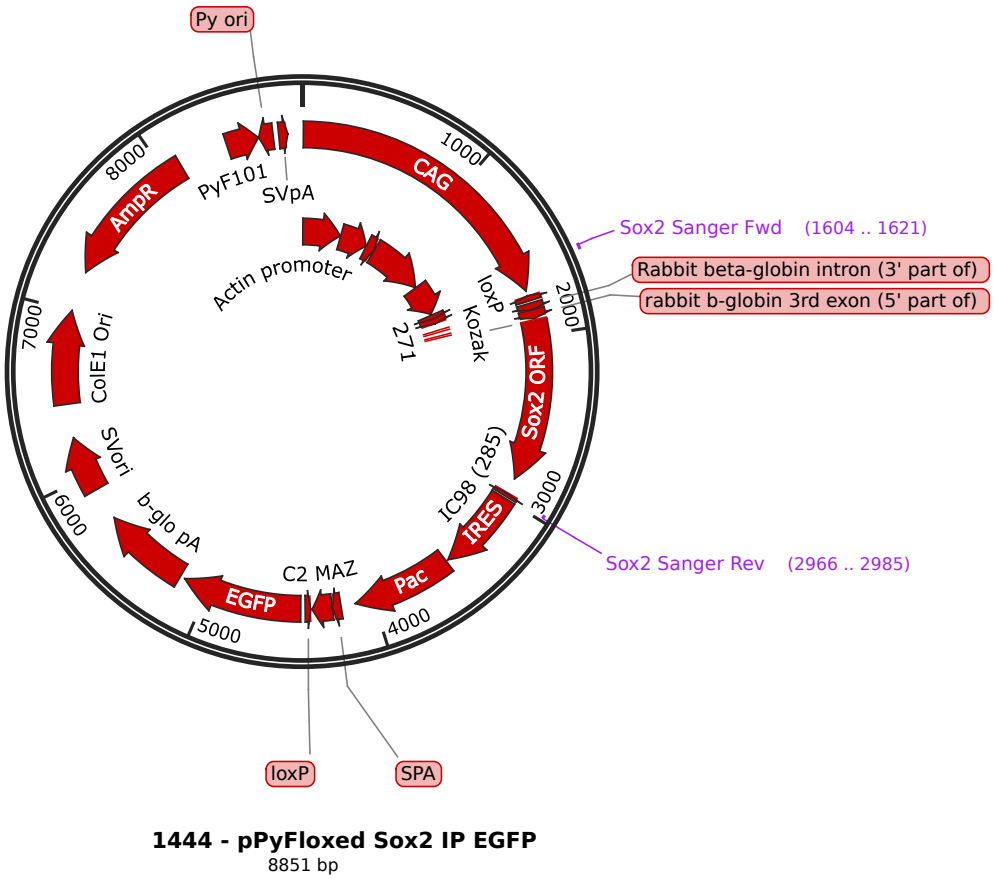

Figure S1: Schematic of plasmid 1444 containing the mSox2 CDS. Custom primers for Sanger sequencing are indicated, as are selection markers and other features native to the plasmid. Plasmid and annotations provided by the lab of Prof. Jennifer Nichols (Cambridge Stem Cell Institute).

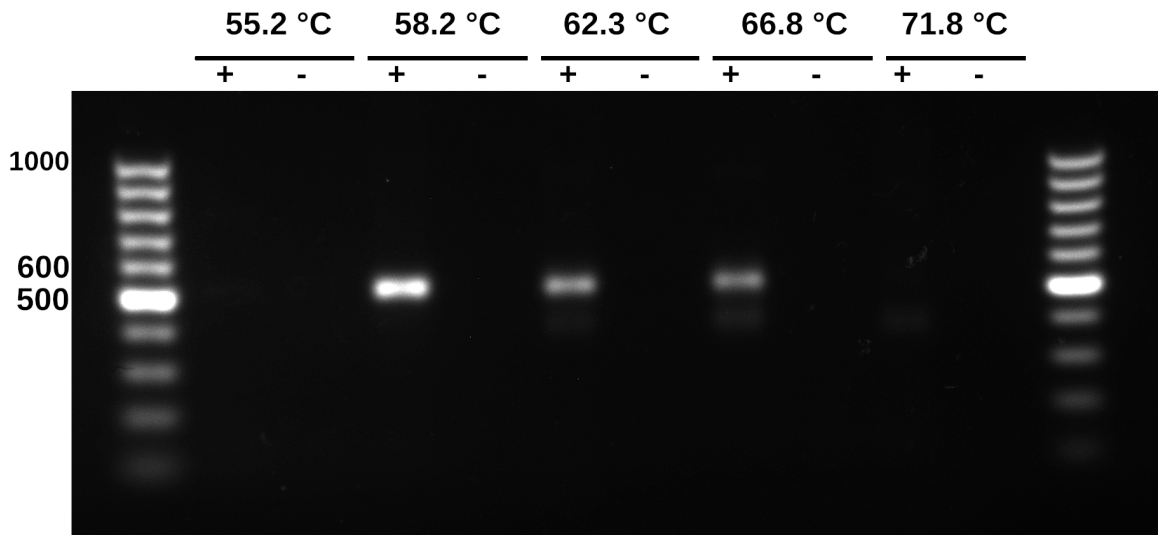

Figure S2: Gradient PCR of D RHA fragment under different annealing temperature conditions shows optimal PCR parameters under a 19 second extension. Lane order: GeneRuler 100bp, 55.2 °C, 55.2 °C n.c., 58.2 °C, 58.2 °C n.c., 62.3 °C, 62.3 °C n.c., 66.8 °C, 66.8 °C n.c., 71.8 °C, 71.8 °C n.c., GeneRuler 100bp. PCR products were visible for the middle three temperature settings, but were most pronounced at an annealing temperature of 58.2 °C.

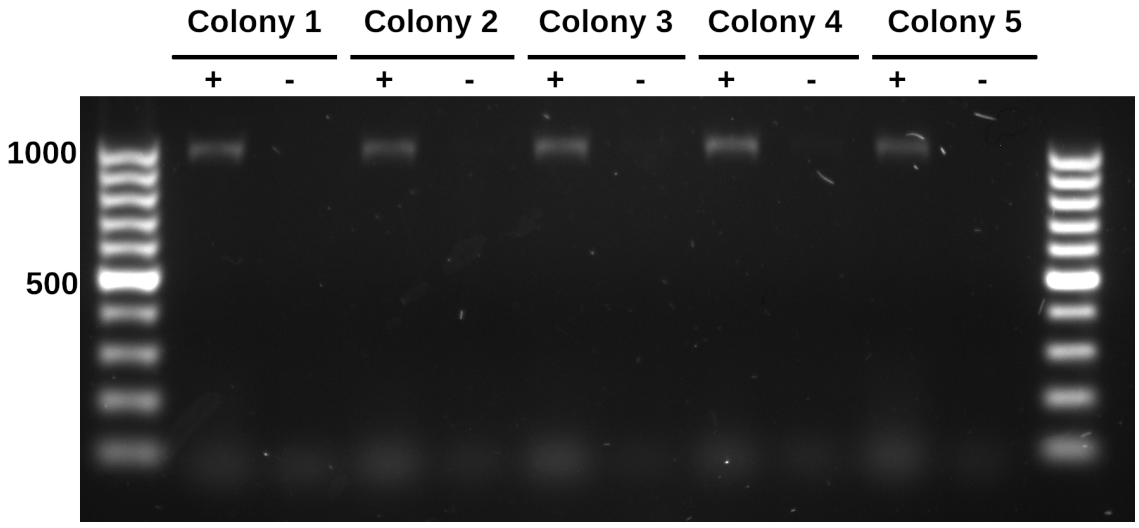

Figure S3: Colony PCR first-pass screen for mSox2 CDS fragment in "Sox2 in D" transformed colonies. All five colonies appear to show a positive result for the presence of mSox2. Lane order: GeneRuler 100bp, Colony 1, Colony 1 n.c., Colony 2, Colony 2 n.c., Colony 3, Colony 3 n.c., Colony 4, Colony 4 n.c., Colony 5, Colony 5 n.c., GeneRuler 100bp

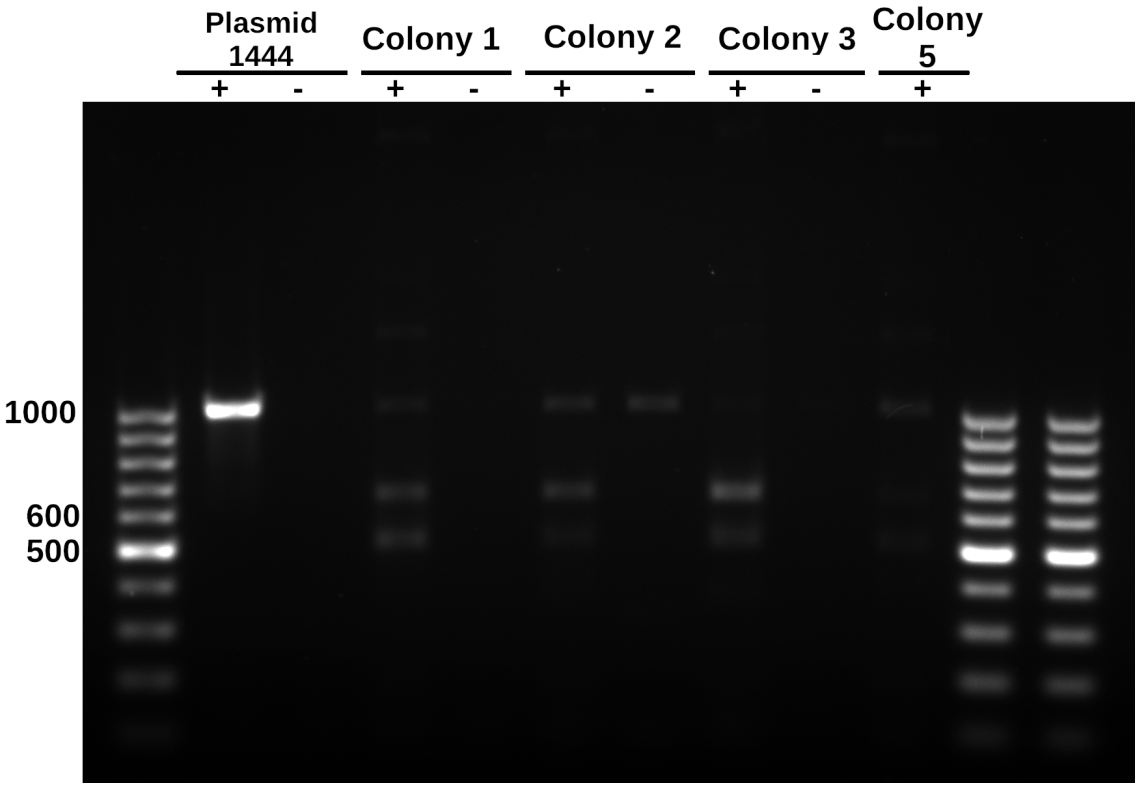

Figure S4: PCR for mSox2 CDS fragment in purified DNA taken from "Sox2 in D" transformation colonies indicates high degree of false positives. PCR of mSox2 sequence from plasmid 1444 is used as a positive control, with miniprepped plasmid DNA from colonies in Figure S3 as comparison. In all cases, there is a significant amount of nonspecific bands, faint specific bands, as well as some contamination in the negative control lane for colony 2. Lane order: GeneRuler 100bp, plasmid 1444 p.c., plasmid 1444 n.c., Colony 1, Colony 1 n.c., Colony 2, Colony 2 n.c., Colony 3, Colony 3 n.c., Colony 5, GeneRuler 100bp, GeneRuler 100bp

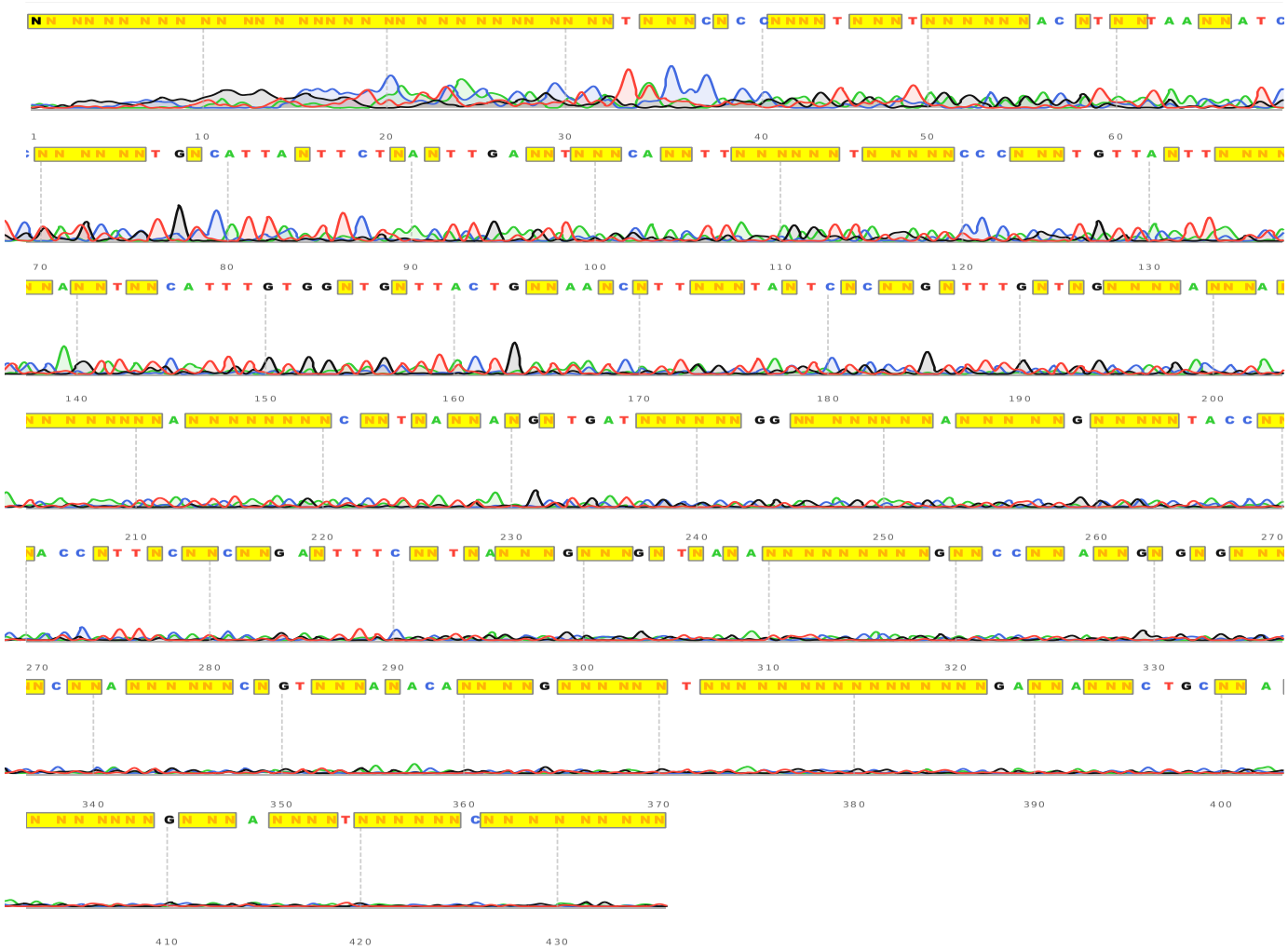

Figure S5: Sanger sequencing attempt with custom PCR primers for plasmid 1444 yields low-quality results. Only 435 bases were reported despite the primers being 1375 bp apart, of which 282 bases were reported simply as “N”.

SoxN-D Two Degrees of Separation				Cluster A		Cluster B		Cluster C			
Aac11	dnc	nmo	Saf-B	AcrE	PIP4K	Acn	Rbp1	Ald	l(2)37Cc	RpL18	Vha26
Acn	eff	nSyb	SC35	AP-1sigma	Pkc53E	B52	Rbp1-like	ATPsynB	levy	RpL23A	Vha55
arm	elav	onecut	scaf6	cac	Pur-alpha	bl	Ref1	ATPsynF	Mdh2	RpL24	Vha68-1
B52	Eph	pan	sd	CG4612	qvr	bol	Rm62	awd	Mpcp	RpL35A	VhaAC45
beat-Ic	exd	para	SF2	CG9132	Rab11	Caper	Rsf1	bic	mRpL12	RpL37a	VhaM8.9
bl	fl(2)d	Pgk	sgg	comt	Rbp	CG10077	Saf-B	blw	mRpL2	RpL4	VhaPPA1-1
Bx	fne	Pka-C1	shi	cpx	Rbp6	CG31712	SC35	CG11752	mRpL22	RpL6	vig
cac	FoxP	Pkc53E	snRNP70K	Crk	Rdl	CG3198	scaf6	CG11876	mRpL4	RpL8	
Caper	ftz-f1	plum	Sox21b	Csp	Rim	CG7903	SF2	CG11984	Nacalpha	RpLP0	
CG10077	fz2	pros	SoxN	Dap160	Rop	CG7971	snRNP70K	CG13220	ND-18	RpLP1	
CG12054	Gad1	ps	sqd	dor	shi	CG9775	sqd	CG17065	ND-30	RpS10b	
CG13220	Gapdh1	psq	svp	e(r)	slo	dod	stmA	CG42575	ND-39	RpS14a	
CG13928	gish	pUf68	Syn	eag	Snap25	fl(2)d	su(w[a])	CG4300	ND-75	RpS14b	
CG1677	Gs1	Rb97D	Syt1	elav	Syn	fne	x16	CG5214	ND-B14.5A	RpS19a	
CG32000	HDAC4	Rbp1	tau	EndoA	Syngr	Hrb87F	YT521-B	CG5903	ND-B17	RpS2	
CG4328	hig	Rbp1-like	Tpi	Gad1	Synj	Hrb98DE		CG7215	ND-B18	RpS26	
CG5214	His2Av	Rbp6	vfl	gish	Syt1	Inr-a		COX4	Pgk	RpS8	
CG7967	Hrb87F	Rdl	x16	GluRIA	Syt4	kcc		COX5A	Pglym78	SdhA	
CG7971	Hrb98DE	Ref1		GluRIB	Syx1A	La		Cyt-c1	PyK	SdhB	
chinmo	Hsc70-4	RFeSP		lap	Syx6	lark		Eno	Rack1	SdhD	
Crk	hth	Rm62		Lerp	tau	Pep		Gapdh1	ras	Sec61beta	
crol	Inr-a	Rpn10		nAChRbeta1	Tomosyn	ps		Hsp60	RFeSP	sta	
Csp	Lar	Rpt3		nonA	VGAT	pUf68		l(1)G0156	RpL13	Tpi	
CtBP	mam	Rpt4		nSyb		qkr58E-1		l(1)G0255	RpL13A	UQCR-C1	
D	Mnn1	Rsf1		para		Rb97D		l(1)G0334	RpL17	UQCR-C2	

Table S1: Full gene list for sub-clusters within adult ventral nerve cord cluster 108. Clusters A, B, and C represent the core genes for the three high-density areas within the GRN of cluster 108 (Figure 27). *SoxN*, *D*, and all genes within two degrees of separation are also included. Gene Ontology analysis reveals that these gene clusters are enriched for different biological processes. Cluster A is enriched for neurotransmission, Cluster B for mRNA processing, and Cluster C for translation and metabolism. The genes associated with *SoxN* and *D* are enriched for transcriptional regulation and functions related to clusters A-C.