Investigating the functional and evolutionary significance of Group B Sox genes in arthropods

Joshua Paul Maher

Department of Genetics

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Selwyn College November 2017

This thesis is dedicated to my Oma. You have been an inspiration to so many people, and I miss you every day. I hope that I have made you proud.

Declaration

I hereby declare that 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.
- Is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as 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; and
- Does not exceed the prescribed word of 60,000 words limit for the Faculty of Biology's Degree Committee.

Joshua Paul Maher

November 2017

Acknowledgements

First, I would like to thank my supervisor, Steve Russell, for providing me the opportunity to carry out research in your lab. You gave me the freedom to explore a topic of my choosing and provided a lot of support along the way, including financial assistance at the very end. You have encouraged and challenged me, and I am grateful to have had such a learned and knowledgeable mentor throughout. Thanks to my funders, the Medical Research Council, for making this research possible. I would also like to thank Anne Ferguson-Smith, my Graduate Advisor, and Gavin Jarvis, my College Graduate Tutor, for all of your advice and help over the course of my studies, and John Welch for your guidance on evolutionary analysis. A special thanks goes to Hilary Jeanes who has provided invaluable mentorship over the last year: your support has been instrumental to my PhD.

I would like to thank other members of the Russell Lab too. Sarah Carl, you were the first lab member I met – thank you for sharing your love and passion for all things Sox with me! Alex Chen, we began our journey together and grew together as researchers. Thank you for all your amazing dinner parties and great company! Dagmara Korona, thanks for all of your help and guidance on cloning. Bettina Fisher, you have provided indispensable support – you helped me find my feet, taught me all that I know about R, and were incredibly patient whenever I'd approach you having attempted something in R that I couldn't get quite right. Thanks for being such a wonderful coding guru and mentor. Finally, Stefan Köstler and Damiano Porcelli, thanks for all the great lab and pub banter, chats about science, and your friendship and support.

Thanks to all my friends in the department. I won't name everybody, as I'm sure I'll leave somebody out. But a special mention and thanks goes to John McGonigle and Alex Patto. You've been the greatest of friends throughout. Thanks to all my College friends, and indeed Selwyn College itself: you've made my time here at Cambridge truly memorable and I'm so grateful for having been part of such a warm and friendly community over the last few years. A special mention goes to Nat, Isa, Max, and Nigel. You've all provided amazing company and friendship.

And thanks to my friends outside the Cambridge bubble: Austin Wellbelove, Bysshe Coffey, Danny Weston, Matthew Robinson, and Dominic Swain. You've helped keep me connected to the 'real world' beyond the ivory towers.

Amy – you’ve been my rock through all the ups and downs. Your patience, care, and love have helped me through the most trying of times, and you have broadened my world. We have had so much fun together, and I can’t wait to begin the next chapter of our lives.

Finally, I would like to thank my family. Dad, thanks for all your love and support, and for reading through this entire thesis despite not having had any scientific training. Thank you, Mum, for your unwavering belief in me. Thanks also to my brother, Jacob, and my grandparents. I hope I have made all of you proud.

Abstract

Group B Sox genes play a critical developmental role in both vertebrates and insects. Within the model species *Drosophila melanogaster*, two SoxB genes, *Dichaete* and *SoxNeuro*, have been shown to act as 'master regulators' in the early development of the central nervous system. Genetic studies have demonstrated the intimate level at which each gene establishes neural stem cell (neuroblast) development, and the redundant properties they share. SoxB genes have only been characterised in a handful of arthropod species thus far, with most work to date focusing on drosophilids. Recent studies have investigated the functional role of the Dichaete and SoxNeuro proteins at the genomic level, establishing thousands of loci within the *Drosophila* genome which each protein binds to and interacts with. These investigations have demonstrated that this genomic binding is highly conserved, even across the 25 million year evolutionary divergence of different drosophilid species. Moreover, these investigations show a striking overlap of bound targets of Dichaete and SoxNeuro within *Drosophila* genomes, providing further evidence for the redundant role these two proteins play within the fruit fly.

The purpose of this investigation was twofold. First, I set out to resolve the phylogenetic origins of arthropod SoxB genes, as mutually exclusive models explaining their emergence are still contested. Using the highly conserved signature region of Sox genes, the high mobility group-box (HMG) encoding domain, I have identified and annotated the SoxB of several invertebrate taxa. In total, my investigation includes 24 different metazoan taxa, which represents the largest investigation of arthropod SoxB phylogeny to date. In light of this research, I have proposed a new model of SoxB evolution which resolves the conflicting elements of the two primary competing models.

Second, to study the evolution of SoxB in terms of functional conservation/divergence, I selected the emerging model organism *Tribolium castaneum*, a Coleopteran species with a reasonably well assembled and annotated genome, as a model in which to draw a comparative analysis with *Drosophila melanogaster*. I first began by characterising the spatiotemporal expression patterns of *SoxNeuro* mRNA in early *Tribolium* embryos using whole mount *in situ* hybridisation, and examined published *Dichaete* expression patterns in the context of central nervous system development in *T. castaneum*. Using these data, I draw a comparison to the

expression profiles of *Dichaete* and *SoxNeuro* orthologues in *Drosophila melanogaster* and other species. I have found that both *Dichaete* and *SoxNeuro* expression patterns in the developing central nervous system are remarkably well-conserved across species. Secondly, I attempted to characterise the genome-wide binding profiles of both Dichaete and SoxNeuro proteins in *Tribolium* in what would have represented the first genomic investigation of its kind in this emerging species.

Using a tethered DNA adenine methyltransferase (Dam) enzyme for both SoxNeuro and Dichaete, I hoped to characterise the genomic loci with which each protein interacts within the beetle genome (a technique known as DamID). It was my aim to use these data to generate a consensus binding-recognition motif for each transcription factor, and compare these to the orthologous motifs identified in *Drosophila*, to investigate their functional divergence/conservation across a 350 million year timescale. Furthermore, I wished to identify the genomic regions most strongly bound by each transcription factor to determine if, as in *Drosophila*, these were genes most closely associated with central nervous system development, and to compare these *Tribolium* target genes with those in the fruit fly. Finally, I hoped to investigate whether there was a significant overlap in the binding targets of both Dichaete and SoxNeuro in order to help determine whether functional redundancy plays as important a role in *Tribolium* development as it does in *Drosophila*.

Unfortunately, these last set of experiments have proved unsuccessful, despite several attempts which have made use of different promoters, different DNA enrichment methodologies, and tackling unforeseen DNA contamination issues. Nevertheless, the troubleshooting experiments that I have carried out will pave the way for further genomic experiments in *Tribolium*, easing the establishment of genomic research in this emerging organism so that we can better understand arthropod development beyond the *Drosophila melanogaster* paradigm.

Table of Contents

List of Figures

List of Tables

Abbreviations

Abb.	**Word**	**Abb.**	**Word**
Acep	*Atta cepholates*	Hduj	*Hypsibius dujardini*
Agam	*Anopheles gambiae*	Hmel	*Heliconius melpomene*
Amel	*Apis mellifera*	HMG	High Mobility Group
AP	Anterior-posterior	Hsap	*Homo sapiens*
Apis	*Acyrthosiphon pisum*	Isc	*Ixodes scapularis*
Bmor	*Bombyx mori*	ML	Maximum Likelihood
bp	Base pair	Mmus	*Mus musculus*
Cele	*Caenorhabditis elegans*	NB	Neuroblast
ChIP	Chromatin immunoprecipitation	Nvit	*Nasonia vitripennis*
CNS	Central nervous system	Phum	*Pediculus humanus*
DamID	DNA adenine methyltranseferase identification	Ptep	*Parasteatoda tepidariorium*
Dmel	*Drosophila melanogaster*	Rpro	*Rhodnius prolixus*
Dpon	*Dendroctonus ponderosae*	Smar	*Strigamia maritima*
Dps	*Drosophila psueodobscura*	Tcas	*Tribolium castaneum*
Dpu	*Daphnia pulex*	TF	Transcription factor
DV	Dorsal-ventral	Turt	*Tetranychus urticae*
Ggal	*Gallus gallus*	VNC	Ventral nerve cord
GMC	Ganglion Mother Cell	Znev	*Zootermopsis nevadensis*

Chapter 1

Introduction

1.1 Introduction to Group B Sox genes

Sox are an ancient and ubiquitous family of metazoan genes. The Sox family encode master regulators (Prior & Walter, 1996; Chan & Kyba, 2013; Whyte *et al*., 2013) involved in a plethora of biological processes (Prior & Walter, 1996; Wegner, 1999; Bowles *et al.*, 2000; Guth & Wegner, 2008). All animals studied thus far possess multiple Sox genes, from the most basal animals such as sponges and cnidarians (Shinzato *et al.*, 2008; Fortunato *et al.*, 2012) to more complex metazoans including vertebrates and insects (*e.g.* see Bowles *et al.*, 2000; McKimmie *et al.*, 2005; Wilson & Dearden, 2008). Sox genes encode proteins which are characterised by a highly conserved amino acid region, the high mobility group (HMG)-box domain, which is implicated in sequence-specific DNA binding in the minor groove, DNA bending, protein interactions, and nuclear transport (Ferrari *et al.*, 1992; Lefebvre *et al.*, 2007). Sox genes were first identified in mammals based on homology with the eutherian mammal testis-determining factor *Sry*, and are defined as sharing ≥50% sequence similarity with the HMG domain of SRY (Laudet *et al.*, 1993; Soullier *et al.*, 1999; Bowles *et al.*, 2000). The Sox name itself comes from "**S**RY-related HMG b**ox** containing gene", chosen to evoke parallels with the developmentally important Hox gene family (Lovell-Badge, 2010). The HMG superfamily of proteins includes not just Sox, but also the non-sequence-specific HMG1 and HMG2 proteins, the nucleolar transcription factor UBF, the sequence-specific TCF-1 and LEF-1 proteins involved in Wnt signalling, and fungal TFs such as mat-Mc and MATA1 (Laudet *et al.*, 1993). It quickly became apparent that this family of proteins is highly diverse, and within mammalian genomes, over 30 Sox have been identified and are implicated in numerous functions (Pevny & Lovell-Badge, 1997; Wegner, 1999).

In 1993, a comparative study of partial HMG domain protein sequences was conducted by Wright *et al.*, which included 15 known Sox genes from the mouse. This analysis identified 6 provisional groups within the Sox family: A: *Sry*; B: *Sox1*, *Sox2*, *Sox3*, and *Sox14*; C: *Sox4*, *Sox11*, and *Sox12*; D: *Sox5*, *Sox6*, and *Sox13*; E: *Sox8*, *Sox9*, and *Sox10*; and F: *Sox7* (Wright *et al.*, 1993). This was later expanded to 7 groups upon the discovery of *Sox15* and *Sox20*, which were assigned to group G, and *Sox21*, which was assigned to group B (van de Wetering & Clevers, 1993; Meyer *et al.*, 1996), before finally an 8th group was added, H, following the discovery of *Sox30* (Osaki *et al.*, 1999). Bowles *et al.* (2000) investigated these groupings further through comparisons with orthologues in other metazoans and concluded that the HMG domain sequence alone can be used to accurately identify relatedness, being congruent with relatedness as suggested by overall gene and protein structure. The protein structure of most members of the Sox family possess the HMG domain close to the N-terminus, and a transactivation/repression domain towards the C-terminus, separated by a hinge region. The transactivation domains tend to be serine-rich, in common with many transactivator proteins, and are essential for the transactivation activity of some Sox proteins (Wright *et al.*, 1995; van de Wetering *et al.*, 1993; Nowling *et al.*, 2000). The transrepression domains are not rich in particular amino acids, yet do tend to be conserved in closely related Sox proteins (Uchikawa *et al.*, 1999; Kamachi *et al.*, 2000; Kamachi *et al.*, 2009). The HMG domain comprises three alpha helices governing DNA binding and bending, and interactions with other proteins (Reményi *et al.*, 2003; Chakravarthy & Rizzino, 2009). The HMG domain of mouse Sox2, for example, cooperates with partner proteins such as Oct-1 and Oct-3/4 by interacting with their POU domains (Reményi *et al.*, 2003), whereas the N-terminal of the Sox9 HMG domain interacts with the C-terminal Zn fingers of the Snail2 protein in the chick (Sakai *et al.*, 2006).

Vertebrate SoxB

Within vertebrates, Sox are dispersed across multiple chromosomes throughout the genome, which contradicts a model of divergence based solely on tandem duplications (Wegner, 1999). Expansion in the vertebrates is therefore believed to have primarily arisen through whole genome duplications (WGDs): *i.e.* in chordates, two rounds of WGDs occurred ~520-550 million years ago (mya) (Meyer & Van de Peer, 2005; Blomme *et al.*, 2006). Subsequent lineage-specific tandem duplications have given rise to the expansive diversity of Sox seen today in the mammals (Pevny & Lovell-Badge, 1997; Wegner, 1999; Bowles *et al.*, 2000).

In terms of their expression and function, Sox are implicated in the regulation of myriad developmental and reparative processes (Lefebvre *et al*., 2007; Lovell-Badge, 2010; Kamachi & Kondoh, 2013). *Sry* is involved in specifying sex in eutherian mammals through its function in testis specification, mentioned above (Whitfield *et al.*, 1993); Group B genes are expressed in the CNS and eye (Uwanogho *et al.*, 1995; Uchikawa *et al.*, 1999; Kamachi *et al.*, 2000; Wood & Episkopou, 1999; Bergsland *et al.*, 2011); Group C genes in the pancreas and kidney (Sock *et al.*, 2004; Wilson *et al.*, 2005; Huang *et al.*, 2013), groups C, D, and E genes in cartilage and skeleton (Smits *et al.*, 2001; Akiyama *et al.*, 2002); the Group E gene *Sox9* is required for condensation and growth of cartilage (Wright *et al.*, 1995; Yan *et al.*, 2002; Yan *et al.*, 2005) and both *Sox9* and *Sox10* pattern neural crest cells and proliferating crest progenitors that have been newly-induced (Pevny & Placzek, 2005); and expression of group F genes is observed in the lymphatic system and vascular structures (Downes & Koopman, 2001, Matsui *et al*., 2006). However, a critical role for Sox genes appears to be within the central nervous system: at least 12 members of the Sox family are expressed in the CNS at some stage of development (Wegner, 1999; Kamachi *et al.*, 2000; Kim *et al.*, 2003; Kamachi & Kondo, 2013). A key feature of SoxB and SoxE proteins, for example, appears to be their ability to maintain neural progenitor or stem cell identity (Pevny & Placzek, 2005).

In vertebrates, Group B Sox proteins can be clearly classified into two distinct subgroups in terms of function and orthology: Groups B1 and B2 (Bowles *et al.*, 2000; Lefebvre *et al.*, 2007; Guth & Wegner, 2008) (Figure 1.2.1). In the chicken, the Group B1 genes (*Sox1*, *Sox2*, and *Sox3*) act as transcriptional activators via transactivation domains located towards the C-terminus, with all three co-expressed in both adult and embryonic neural progenitor cells (Kamachi *et al.*, 2000; Pevny & Placzek, 2005; Kamachi *et al.*, 2009). In contrast, the B2 genes, *Sox14* and *Sox21*, act as transcriptional repressors with the C-termini regions possessing transrepression domains (Uchikawa *et al.*, 1999; Kamachi *et al.*, 2000). *Sox21* expression is observed throughout the developing CNS, while *Sox14* expression is more limited, only observable in a small subset of interneurons (Uchikawa *et al.*, 1999; Pevny & Placzek, 2005). However, the B2s share highly similar HMG domains with the B1s, and can bind to identified Sox2 targets (Pevny & Placzek, 2005). Moreover, in mouse and HeLa cells *Sox14* has been shown to act as a transcriptional activator (Popovic *et al.*, 2014), similar to the B1 subgroup. Throughout vertebrates, expression of the B1 subgroup correlates with ectodermal cells destined to acquire neural fates, and subsequently with their commitment to this fate (Pevny & Placzek, 2005). The B1 genes also exhibit significant redundancy. In zebrafish, for example, there are 6 SoxB1 genes present, and severe defects in CNS development are only visible in quadruple knockdowns of *Sox2*, *Sox3*, *Sox19a*, and *Sox19b*, suggesting a compensatory mechanism between these genes (Okuda *et al.*, 2010).

Redundancy is observed within other groups too, including groups C, E, and F genes (Reiprich & Wegner, 2015). For example, Bhattaram *et al.* (2010) show evidence of redundancy between the SoxC genes *Sox4*, *Sox11*, and *Sox12* in the fate of neural and mesenchymal progenitor cells, with triple-mutant mice exhibiting the strongest phenotypes. The SoxE genes, *Sox8*, *Sox9*, and *Sox10*, act redundantly in the formation and maintenance of oligodendrocyte precursor cells in mice (Stolt *et al.*, 2003; Stolt *et al.*, 2004; Stolt *et al.*, 2005). Matsui *et al.* (2006) demonstrate how the Group F genes *Sox17* and *Sox18* exhibit redundancy in postnatal vascularization in mice in tissues where both of these genes are co-expressed. Redundancy therefore appears to be a characteristic feature of many Sox genes across the various groups.

Invertebrate SoxB

Within the invertebrates, Sox genes are just as diverse in their functions, although are generally less numerous than their vertebrate homologues (Phochanukul & Russell, 2010). The most basal metazoans such as sponges and placozoans possess only a handful of Sox genes (3-4) (Larroux *et al.*, 2008), although in the calcareous sponge *Sycon ciliatum*, seven Sox genes have been identified in groups B, C, E, and F (Fortunato *et al.*, 2012). The four core groups identified in sponges, B, C, E, and F, are also found in ctenophores (Jager *et al.*, 2008; Schnitzler *et al.*, 2014). The Sox repertoire is greatly expanded in the cnidarians, which possess 10-14 genes in groups B-F (Jager *et al.*, 2006; Shinzato *et al.*, 2008; Phochanukul & Russell, 2010).

Protostomes tend to possess fewer Sox than the Radiata, with <10 genes present in all species examined to date (KcKimmie *et al.*, 2005; Wilson & Dearden, 2008; Phochanukul & Russell, 2010), although at least a single representative of groups B-F are present in most protostomes. Recent work in the molluscs has identified Sox members identical to the groups observed in chordates, with genes in groups B1, B2, C, D, E, F, and H (Yu *et al.*, 2017). Within the deuterostomes, non-vertebrate chordates possess a variable number of Sox genes (Phochanukul & Russell, 2010), with a repertoire being more similar to that of vertebrates; the last common ancestor of the chordates likely possessed at least 7 Sox genes across groups B-F, and H (Heenan *et al.*, 2016).

Within the insects, the fruit fly has 8 Sox genes – four Group B, and 1 in each group C-F. These 8 Sox genes are common to 11 drosophilid species and other Diptera such as *Anopheles gambiae* (Wei *et* al., 2011; Phochanukul & Russell 2010). Two Hymenopteran species, *Apis mellifera* and *Nasonia vitripennis*, have an additional Group E gene. The Coleopteran *Tribolium castaneum* and Lepidopteran *Bombyx mori* have an additional group B gene, possessing 5 in total (Wei *et al.*, 2011). Within the Hemimetabola, *Acyrthosiphon pisum* possesses as few as 6 Sox genes. (See Phochanukul & Russell (2010) for an excellent review of invertebrate Sox evolution).

There is also debate over whether the subgroups B1 and B2 found in vertebrates can be applied to invertebrates. B1 and B2 subgroups had initially been assigned to the cnidarian and sponge Group B genes, however work by Shinzato *et al.* (2008) demonstrates that these cluster outside the Bilaterian B1 and B2 subgroups, suggesting that the subgroups are likely to be restricted to the Bileteria only (Shinzato *et al.*, 2008). Bowles *et al.* (2000) suggest that the B1 and B2 subgroups apply to the 4 SoxB genes of *Drosophila melanogaster*. However, work by McKimmie *et al.* (2005) suggests that this is not the case; while *SoxNeuro* of *D. melanogaster* groups unambiguously with *Sox1*, *Sox2*, and *Sox3* (Group B1), the other 3 SoxB genes of *D. melanogaster*, *Dichaete*, *Sox21a*, and *Sox21b* are instead suggested to be lineage specific and their relationship less clear. Nonetheless, phylogenetic work by Zhong *et al.* (2011) unambiguously clusters *SoxNeuro* with vertebrate B1 genes, and *Dichaete*, *Sox21a*, and *Sox21b* with vertebrate B2 genes.

There thus appears to be a core group of Sox genes that emerged prior to the emergence of the Bilateria (Bowles *et al.*, 2000; Jager *et al.*, 2006; Larroux *et al.*, 2008; Heenan *et al.*, 2016). The core groups B, C, E, and F (van de Wetering *et al.*, 1993; Wright *et al.*, 1993; Meyer *et al.*, 1996) are present in most basal animals, including sponges and ctenophores (Shinzato *et al.*, 2008; Fortunato *et al.*, 2012; Schnitzler *et al.*, 2014). Groups B through to F are found in all higher metazoans, especially the Bilateria, however groups G-J are restricted to particular lineages (Bowles *et al.*, 2000; Larroux *et al.*, 2008; Heenan *et al.*, 2016; Yu *et al.*, 2017).

There is also some debate regarding whether Sox are unique to metazoans: the closest relatives to multicellular eukaryotes, the unicellular choanoflagellates, may possess Sox-like sequences. King *et al.* (2008) has identified two Sox-like sequences in *Monosiga bevicolis*, which suggests that the origin of Sox predates multicellularity (Guth & Wegner, 2008). However, Zhong *et al.* (2011) maintain that these are not true Sox genes, as they share relatively low identities with Sox (<40%), which is significantly lower than the identities shared by metazoans (>50% (Bowles *et al.*, 2000) or >46% (Lefebvre *et al.*, 2007)). Moreover, the Sox-like proteins of the choanoflagellate do not cluster with any identified group of metazoan Sox in phylogenetic analysis (Zhong *et al.*, 2011). This suggests that even if the Sox-like proteins of the choanoflagellates are true Sox orthologues, they are perhaps not orthologous to a specific group of the metazoan Sox family, with the groups arising uniquely in the animal kingdom.

The diverse functions and expression patterns of invertebrate Sox groups is similar to the diversity seen in vertebrates. For example, expression patterns of Sox genes in the ctenophore *Mnemiopsis leidyi* are consistent with the well-described role of Sox genes in stem cell maintenance, with strong expression patterns in proliferating cell zones (Schnitzler *et al.*, 2014), and qRT-PCR data in the scallop *Patinopecten yessoensis* has revealed Sox expression in cells responsible for neurogenesis, haematopoiesis, myogenesis, and gametogenesis (Yu *et al.*, 2017).

For Group C orthologues, honeybee *Am-SoxC* is expressed ubiquitously in late embryos and the adult brain (Wilson & Dearden, 2008). In *C. elegans*, *sem-2* is involved in specifying the cell-fate of sex myoblasts, embryonic muscle development, and egg laying (Broitman-Maduro *et al.*, 2005; Minor *et al.*, 2013). In more basally branching metazoans, such as sponges, *SoxC* is expressed in the ectodermal region within a population of cells that are suspected to become sensory neurons (Shinzato *et al.*, 2008). In the oyster *Crassostrea gigas*, the expression of a SoxC gene in the larval mantle implies a novel function in larval shell formation and biomineralization (Liu *et al.*, 2017). The SoxC gene of *D. melanogaster*, *Sox14*, is expressed in the anterior and posterior endoderm, the anterior mesoderm, and midgut anlage (Fisher *et al.*, 2012, FlyBase report), and during larval and pupal stages it is prominent in the digestive system (Cremazy *et al.*, 2001; Chintapalli *et al.*, 2007; Fisher *et al.*, 2012, FlyBase report).

For Group D genes, *egl-13* mutants cause sterility in female *C. elegans* worms, and *egl-13* has been shown to be required for aspects of vulval development by being necessary for cell fusion between the vulva and uterus (Hanna-Rose & Han, 1999; Oommen & Newman, 2007). Meanwhile, expression data for *Sox102F* in *D. melanogaster* shows that this SoxD gene is expressed in neurons of the ventral nerve cord and embryonic brain, particularly in the mushroom body anlage (Fisher *et al.*, 2012: FlyBase Report). In adults, phenotypes for *Sox102F* mutants include severely impacting cardiac function and disruption of the Wnt signalling pathway (Li *et al.*, 2013).

Group E genes show evidence of conserved function in the insects: in *D. melanogaster*, *Sox100B* is required for correct testes development (Nanda *et al.*, 2009), and in *Apis mellifera* the two SoxE orthologues are both expressed in the testes of male worker drones (Wilson & Dearden, 2008). In the cephalopod *Sepia officinalis*, expression patterns of SoxE suggest a role in vascular genesis (Focareta & Cole, 2016).

Little information exists for Group F Sox genes outside of *D. melanogaster*, although various studies show expression of SoxF homologues in the endoderm of different invertebrates, including the ctenophore *Pleurobrachia pileus* (Jager *et al.*, 2008), the sea anemone *Nematostella vectensis* (Magie *et al.*, 2005), and in the coral *Acropora millepora* (Shinzato *et al.*, 2008). In the fruit fly, *Sox15* is expressed in the embryonic PNS (Cremazy *et al.*, 2001) and sensory primordium (Fisher *et al.*, 2012; FlyBase report). During metamorphosis, *Sox15* is expressed in the anlage of the external sensory organ socket cells, and is necessary for chaeta development (Miller *et al.*, 2009). It is also associated with the *Drosophila* Wnt pathway. Repressing *wg* in the wing imaginal disc, it has a similar phenotype to the dominant *Dichaete* mutation (Russell, 2000; Dichtel-Danjoy *et al.*, 2009; Miller *et al.*, 2009).

Invertebrate Group B Sox genes have also received much attention. In all Bilateria examined thus far, at least one SoxB gene is expressed in the neurogenic region of the developing embryo, suggesting deep conservation of this group's function. For example, in the cnidarian *Nematostella vectansis*, a SoxB orthologue regulates neural progenitor cell behaviour and interacts with the Notch signalling pathway and bHLH genes (Richards & Rentzsch, 2015). In the protostomes, the SoxB expression patterns in the cephalopod *Sepia officinalis* suggest a role in neural specification and development of sensory epithelium. In another protostome, the SoxB expression patterns of the platyhelminthe *Dugesia japonica* imply a conserved role in neural development (Dong *et al.*, 2014). SoxB expression is also strong in the CNS of invertebrate chordates, such as the sea pineapple *Halocynthia roretzi* (Miya & Nishida, 2003) and the sea squirt *Ciona robusta* (Imai *et al.*, 2017). In *C. robusta*, the function of SoxB has been shown to be required for neural development, whereby it regulates genes required for the patterning and specification of posterior neural lineages (Imai *et al.*, 2017). In the chordate *Branchiostoma floridae*, there are 4 SoxB genes; three B1 genes and one B2 gene. B1 genes are expressed in the early neuroectoderm and later in the CNS, and the B2 gene is expressed in later-stage neural cells only (Meulemans & Bronner-Fraser, 2007). In the hemichordates, group B genes are also expressed in the neurogenic ectoderm in the species studied thus far (Taguchi *et al.*, 2002; Lowe *et al.*, 2003).

Within the insects, there are 4 SoxB genes that have been characterised thus far: *Dichaete*, *Sox21a*, *Sox21b*, and *SoxNeuro*, although others have been identified (McKimmie *et al.*, 2005 Wilson & Dearden, 2008, Wei *et al.*, 2011). In the honeybee, *Am-Sox21a* is expressed in the Malpighian tubules, and *Am-Sox21b* is expressed late in embryonic CNS and brain tissue and during oogenesis in adults. *Am-SoxNeuro* is expressed along the ventral gastrulation folds and in the pro-cephalic neurogenic region of gastrulating embryos, and is observed throughout the neuroectoderm and in the neurons of the cephalic lobes post-gastrulation. However, no *Am-Dichaete* expression is detected via *in situ* hybridisation or RT-PCR experiments (Wilson & Dearden, 2008).

Invertebrate SoxB have been most extensively studied within *D. melanogaster*: *Sox21b* expression is observed partially overlapping with *Dichaete* in the hindgut, and is expressed within the ventral epidermis and large intestine (Cremazy *et al.*, 2001; McKimmie *et al.*, 2005). *Sox21a* is expressed in both the hindgut and foregut, and later in development, in unidentified midline cells (McKimmie *et al.*, 2005). Meng & Bitaeu (2015) show that *Sox21a* is expressed in adult intestinal stem cells (ISC), and is necessary for cell proliferation during normal epithelial mitosis, and during gut repair. However, *Sox21a* mutant flies show no developmental defects, implying that this TF is a regulator of adult SCs only (although mutant adults do not live as long as wild type adults) (Meng & Bitaeu, 2015). Moreover, all embryonic *D. melanogaster* SoxB expression is conserved with *D. pseudoobscura* (McKimmie *et al.*, 2005).

The two other SoxB genes of *D. melanogaster*, *Dichaete* and *SoxNeuro*, have been comprehensively studied. *Dichaete* expression initially appears in a broad domain enveloping the entire trunk anlage, then resolving into seven transverse stripes in the blastoderm (Nambu & Nambu, 1996; Russell *et al.*, 1996). *Dichaete* expression can be seen in later stages in the midline glia, and the medial and intermediate columns of the ventral neuroectoderm throughout all waves of NB development. *Dichaete* has been shown to be necessary for correct differentiation of glial lineages within the midline (Sánchez-Soriano & Russell, 1998), however neural phenotypes are weak in the medial and intermediate columns (Nambu & Nambu, 1996; Overton *et al.*, 2002). *Dichaete* has also been shown to be active during *Drosophila* segmentation, with primary pair-rule genes *even-skipped*, *hairy*, and *runt* dependent on *Dichaete* activity (Clark & Peel, 2017); in the hindgut (Sánchez-Soriano & Russell, 2000); and in the ovary during oogenesis (Mukherjee *et al.*, 2006). *Dichaete* is also expressed in the protocerebrum, deuterocerebrum, and tritocerebrum of the embryonic brain (Sánchez-Soriano & Russell, 2000). *SoxNeuro* is expressed in a pan-neuroectodermal manner throughout neurogenesis (Cremazy *et al.*, 2000) across the medial, intermediate, and lateral columns of the neuroectoderm (Buescher *et al.*, 2002; Overton *et al.*, 2002). *SoxNeuro* mutants exhibit severe defects in the head, and in the intermediate and lateral columns of the CNS, however, the medial column forms almost normally (Buescher *et al.*, 2002; Overton *et al.*, 2002).

There is therefore overlapping expression between *Dichaete* and *SoxNeuro* in the medial and intermediate columns (Figure 1.1.2). Perhaps most interesting, however, is the fact that *Dichaete* and *SoxNeuro* double mutants exhibit more severe defects than either single mutant (Buescher *et al.*, 2002; Overton *et al.*, 2002); severe hypoplasia is visible throughout the CNS in double mutants, with the longitudinal axons almost missing entirely. This strongly suggests that similar to the vertebrate SoxB genes mentioned above, these two SoxB genes act in a partially redundant manner (Buescher *et al.*, 2002; Overton *et al.*, 2002). Recent genomic approaches provide substantial evidence supporting this hypothesis of redundancy (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Carl & Russell, 2015). Genome-wide binding studies of Dichaete and SoxNeuro reveal a striking overlap in bound targets of these two TFs in not only *D. melanogaster* (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014), but also in other drosophilids, separated by ~25 million years of divergence (Carl & Russell, 2015) (Figure 1.1.3A-B). Moreover, Dichaete and SoxNeuro exhibit an intricate compensatory binding pattern in the absence of each other; genome-wide binding studies in *Dichaete* mutants and *SoxNeuro* mutants reveal *de novo* binding events occurring in one another's absence (Ferrero *et al.*, 2014) (Figure 1.1.3C). This compensation activity elucidates the redundant role these two genes play in the *D. melanogaster* genome, and hints at a mechanism of neo- and subfunctionalization in the evolutionary history of these homologous genes, whereby paralogues have acquired a subdivision of ancestral function (Lynch & Force, 2000; Larroux *et al.*, 2008; Qian *et al.*, 2010).

However, the fruit fly may not be entirely representative of other insects, and is likely less so for the wider arthropod phylum, as it is a highly specialised species (Hughes & Kaufman, 2000). Indeed, the fact that *Drosophila* larvae do not possess legs or eyes (Kingler, 2004) is highly atypical of insects; thus drawing functional inferences to other species may not always be wise. It is therefore important to widen the scope of analyses to include other invertebrate species, in order to elucidate a more holistic account of Sox function in the animal kingdom (Phochanukul & Russell, 2010).

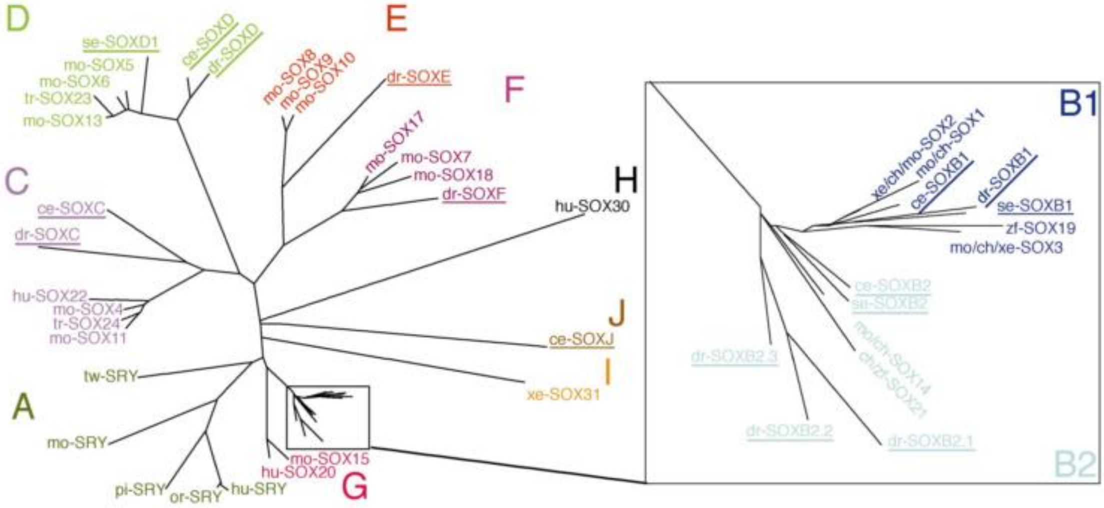


Figure 1.1.1. Unrooted phylogenetic tree of Sox groups A-J. Group B genes are further subdivided into B1 and B2 genes. dr: *Drosophila melanogaster*, ce: *Caenorhabditis elegans*, hu: humans, mo: mouse, or: orangutan, pi: pig, se: sea urchin, rw: rainbow trout, tw: tammar wallaby, xe: *Xenopus laevvis*. Reproduced from Bowles et al. (2000).

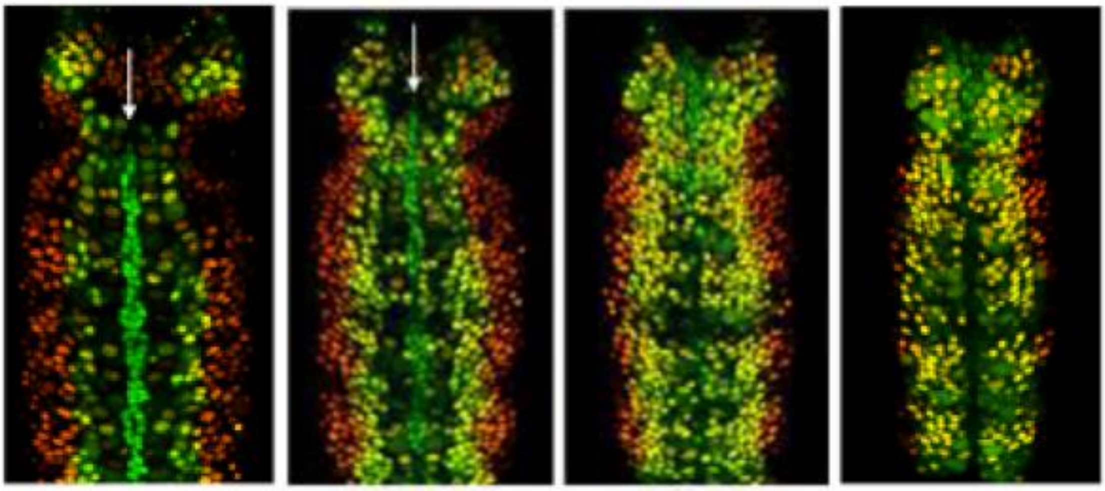


Figure 1.1.2. The neuroectoderm of stage 10 *D. melanogaster* embryos labelled for Dichaete and SoxN expression. Focus is shifted across different planes. *Dichaete* expressed is visible in the glial cells of the ventral midline (green cells, white arrows) as well as the medial and intermediate columns of neuroblasts. In the medial and intermediate columns *SoxNeuro* expression can be seen to overlap with the expression of *Dichaete* (yellow), but not in the lateral column of neuroblasts (red). Figure reproduced from Overton (2003).

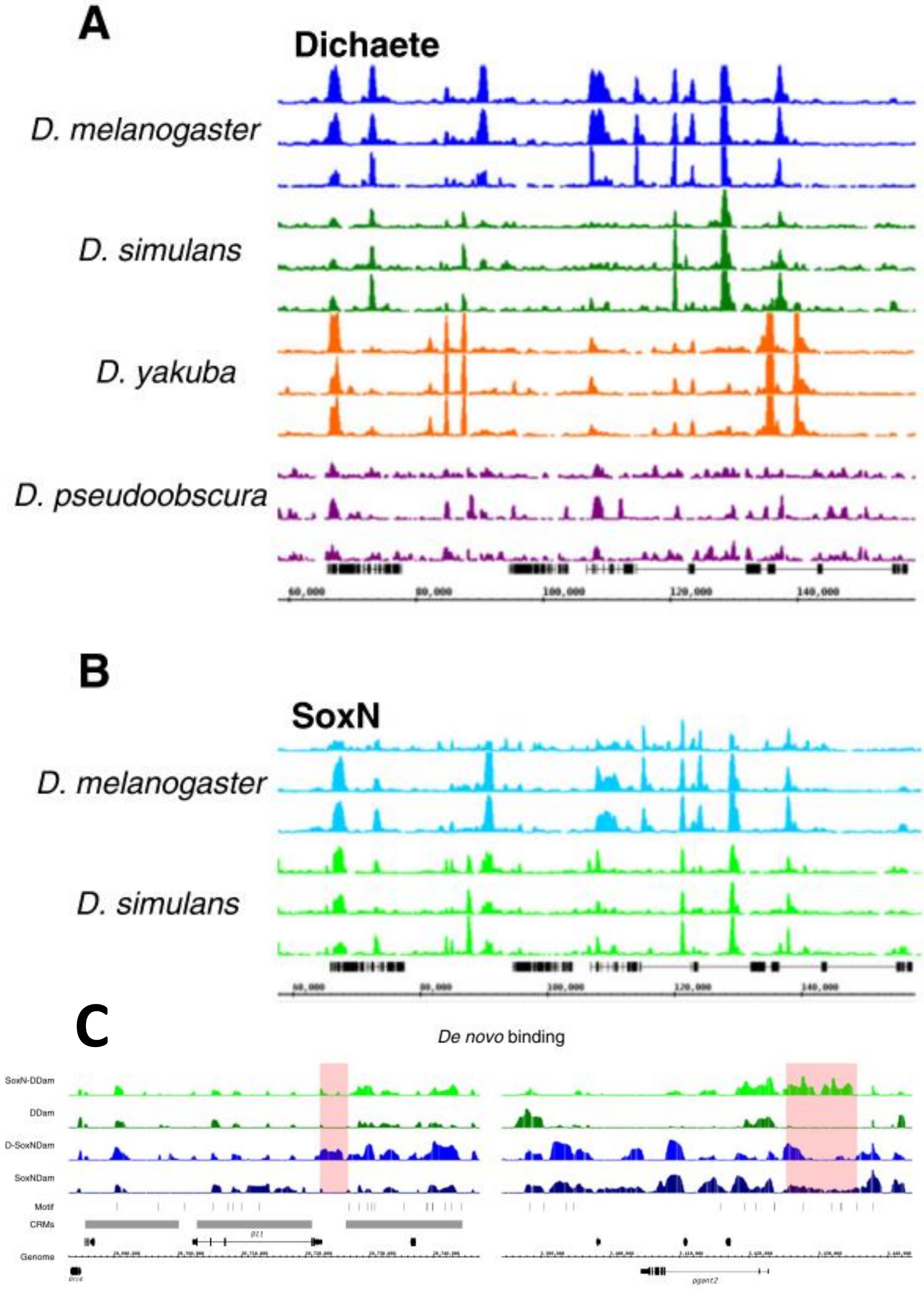


Figure 1.1.3. Common binding of Dichaete (A) and SoxNeuro (B) TFs across drosophilid species, and examples of de novo binding in *D. melanogaster* mutants (C). The same ~120kb region from chromosome 2L is shown for all species in A and B, and the same locus is shown in C. In C, examples of *de novo* binding are highlighted in red. A and B reproduced from Carl & Russell (2015) and C from Ferrero *et al.* (2014).

1.2 The central nervous system of arthropods

Across arthropods, there is remarkable conservation in the development and structure of the central nervous systems (CNS) (Stollewerk & Simpson, 2005; Biffar & Stollewerk, 2014; Hartenstein & Stollewerk, 2015; Stollewerk, 2016). The CNS of arthropods comprises the brain and ventral nerve cord (VNC) (Bhat, 1999; Skeath & Thor, 2003; Doeffinger *et al.*, 2010), and much of the research focus has been on the development of the VNC, which is made up of 14 segmented components called neuromeres (Bhat, 1999; Harzsch, 2003; Boyan & Williams, 2011). These segmental ganglia have the characteristic appearance of a "rope ladder" (Figure 1.2.1), which is conserved across invertebrate phyla (Harzsch, 2003; Ungerer *et al.*, 2011; Biffar & Stollewerk, 2014; Biffar & Stollewerk, 2015). Within insects, there are three gnathal, three thoracic, and eight abdominal neuromeres, which are concomitant with the insect segmental body plan (Bhat, 1999). Neuromeres are divided along the anterior-posterior (AP) axis into symmetrical hemineuromeres, separated by the ventral midline (Bhat, 1999; Harzsch, 2001; Harzsch, 2003; Biffar & Stollewerk, 2014). This highly organised system develops from the neurogenic region, or neuroectoderm, which is a region in the ectodermal layer from which neural stem cells (called neuroblasts (NBs)) delaminate (Hartenstein & Campos-Ortega, 1984; Stollewerk & Simpson, 2005; Hartenstein & Stollewerk, 2015).

The vast majority of research into VNC development to date has been on the fruit fly *Drosophila melanogaster* (Bate & Martinez-Arias, 1993; Skeath, 1999; Bhat, 1999; Hartenstein & Stollewerk, 2015). In *Drosophila*, an invariable number of neural stem cells (neuroblasts, or NBs) delaminate from within the neuroectoderm across five sequential waves of development (S1-S5) during embryonic stages 8-11 (Hartenstein & Campos-Ortega, 1984; Hartenstein *et al.*, 1985). NBs are arranged along seven transverse rows (reviewed in: Hartenstein & Campos-Ortega, 1984; Bhat, 1999), and three longitudinal columns (Weiss *et al.*, 1998; Skeath, 1999; Stollewerk & Simpson, 2005). NBs are identified using a numbering system: the first number identifies the row along the AP axis (1-7; anterior to posterior), and the second along the dorsal-ventral (DV) axis (1-6; medial to lateral) (Bhat, 1999; Skeath, 1999). This numbering system originated in the grasshopper species *Schistocerca americana* (Doe, 1992), and the NB homologues of *Drosophila* are identified accordingly (Figure 1.2.2).

Each NB acquires a distinct identity via positional patterning mechanisms. Along the AP axis, segment polarity genes are responsible for the patterning of NBs, and along the DV axis, columnar genes are responsible. Within *Drosophila*, the segment polarity genes are *wingless* (*wg*), *hedgehog* (*hh*), *patched* (*ptc*), *gooseberry* (*gsb*), *engrailed* (*en*), and *invected* (*inv*) (reviewed by Bhat, 1999), and the columnar genes are *Epidermal growth factor receptor* (*Egfr*), *ventral nerve cord defective* (*vnd*), *intermediate nerve cord defective* (*ind*), and *muscle segment homeodomain* (*msh*) (reviewed by Skeath, 1999), along with *Dichaete* (*D*) and *SoxNeuro* (*SoxN*) (Zhao & Skeath, 2002; Zhao *et al.*, 2007). Expression patterns for many of these genes are found to be similar across arthropods (Wheeler *et al.*, 2005; Doeffinger & Stollewerk, 2010); for example, *en* expression at the segment boundaries is highly conserved across all arthropods examined (Patel, 1994; Patel *et al.*, 1989; Duman-Scheel & Patel, 1999; Chipman & Stollewerk, 2006; Fabritius-Vilpoux *et al.*, 2008), and columnar genes are also similarly expressed in three longitudinal columns in chelicerates and myriapods (Dove, 2003; Dove & Stollewerk, 2003).

While NB fates are determined via segment polarity and columnar gene expression, the neural differentiation of cells is controlled by the proneural genes of the Achaete-Scute Complex (AS-C), *achaete*, *scute*, and *lethal of scute* (Jimenez & Campos-Ortega, 1990; Skeath & Carroll, 1992). These genes are expressed at the onset of neurogenesis in proneural cell clusters in the neuroectoderm. An additional AS-C gene, *asense*, is expressed only in cells destined to become neural precursors (Brand *et al.* 1993). From each of these proneural cell clusters, only single cells differentiate into NBs, with the remainder going on to form epidermal progenitor cells. This is achieved by lateral inhibition of cells via the activity of the neurogenic genes *Notch* and *Delta* (Skeath & Carroll, 1992; Heitzler *et al.*, 1996). The differentiated NB then segregates between the ectodermal and mesodermal layers (Hartenstein & Campos-Ortega, 1984; Skeath & Carroll, 1992), before enlarging and undergoing asymmetrical division (Stent & Weisblat, 1985). This asymmetrical division gives rise to a ganglion mother cell (GMC) and maintains the NB. Following this, the GMC can generate two different neural cell types via its division: neurons or glia (Campos-Ortega & Hartnstein, 1985; Goodman & Doe, 1993).

Temporal changes in gene expression also contribute to the neural fate of individual NBs. Four genes are expressed in a temporal cascade: *hunchback* (*hb*), *Krüppel* (*Kr*), *nubbin* (*nub*), and *castor* (*cas*), are expressed, in that order, in *Drosophila* (reviewed by Brody & Odenwald, 2005). *Hb* is expressed in NBs as they delaminate and during the first round of division; *Hb* is then down-regulated and *Kr* up-regulated for the second round of division*,* before *Kr* is down-regulated and *nub* up-regulated… and so on (Skeath & Thor, 2003; Brody & Odenwald, 2005). This results in a layered pattern of gene expression in the neurones and glia produced by each

NB, with older neurons more lateral to the VNC having delaminated sooner, and retaining the respective temporal gene expression of the GMC and NB (reviewed by Skeath & Thor, 2003) (Figure 1.2.3). Not all NBs undergo this temporal cascade, however (Isshiki *et al.*, 2001), and many NBs express additional genes, including *grainy head* (*gh*) and *Dichaete* (Brody & Odenwald, 2000; Maurange *et al.*, 2008).

The 30 NBs of *Drosophila* generate ~370 neural cells per hemisegment, most of which comprise interneurons (~300), with glial cells and motor neurons (30 and 30, respectively) and neurosecretory cells (7) comprising the rest (comprehensively studied in Schmid *et al.*, 1999). The intermediate and medial column NBs primarily generate neurons, and generate just 3 glial cells; in contrast lateral NBs produce 27 glial cells and 120 neurons (Bossing *et al.*, 1996; Schmid *et al.*, 1999; Landgraf *et al.*, 1997; Granderath & Klämbt, 1999). Nearly all GMCs are believed to acquire a unique fate throughout development (Schmid *et al.*, 1999; Skeath & Thor, 1999). Neurons can be further divided into interneurons which connect to other neurons, motor neurons which innervate muscle tissue, and pioneer neurons which develop into the axonal scaffold observed in the developed CNS and establish the primary axonal tracts (Bate, 1976; Thomas *et al.*, 1984; Landgraf & Thor, 2006). Commissures connect each hemineuromere transversely across the midline, and neuromeres are connected longitudinally by connectives. Collectively, these structures give rise to the “rope ladder” appearance of the CNS in insects (see Figure 1.2.1) (Harzsch, 2003; Ungerer *et al.*, 2011; Biffar & Stollewerk, 2014; Biffar & Stollewerk, 2015).

While many of the genes involved in CNS development and the general structure of the VNC are conserved between insects and other arthropods, significant differences in the developmental mechanisms do exist (reviewed in Biffar, 2013; and Stollewerk, 2016). For example, within crustaceans, NBs can generate both GMCs and precursor cells of the epidermis (Scholtz, 1990), and NBs do not segregate into the embryo, as is observed in insects, but instead remain in the neuroectoderm (Scholtz, 1990; Scholtz, 1992; Harzsch, 2001; Ungerer *et al.*, 2011). In *Daphnia magna*, no proneural clusters are observed, and the first neural gene to be expressed in the CNS is *snail*, prior to the expression of a single AS-C homologue; this is the reverse of the process in *D. melanogaster* (Ungerer & Scholtz, 2008; Ungerer *et al.*, 2011). However, there does appear to be some degree of conservation in that NBs are arranged along invariable rows and columns within hemisegments throughout the developing CNS (Scholtz, 1992; Ungerer & Scholtz, 2008; Ungerer *et al.*, 2012).

In contrast, chelicerates and myriapods lack NBs altogether – instead, neural precursor cell clusters delaminate together in these arthropods (as opposed to a single cell), and acquire their respective fates without further divisions (Stollewerk *et al.*, 2001; Stollewerk & Simpson, 2005; Hartenstein & Stollewerk, 2015). Nonetheless, the proneural genes are conserved and initiate differentiation (Stollewerk *et al.*, 2001; Dove & Stollewerk, 2003). Moreover, the arrangement of these neural precursor clusters is remarkably similar to the arrangement of *Drosophila* NBs, along seven transverse rows and a variable number of columns, suggesting evolutionary conservation (Stollewerk *et al.*, 2001; Dove & Stollewerk, 2003; Hartenstein & Stollewerk, 2015).

The themes explored in this section give rise to many new questions regarding CNS development in arthropods. NB position and gene networks in insects are well-conserved, however, spatiotemporal gene expression is less so (Biffar & Stollewerk, 2014). This is likely to affect neural identity, and neural lineages subsequently need to be examined further in different insect species. Examining the neural lineages of arthropod species that do not possess NBs will also aid in understanding how the "rope ladder" structure of the CNS is so well-conserved. Moreover, the genomes of many arthropods have now been sequenced, enabling genomic techniques that have been used in *Drosophila* to be used in any species which has a published genome. This will help elucidate the regulatory properties of many of the genes outlined in this section across arthropods, so that evolutionary comparisons can be made at the genomic level.

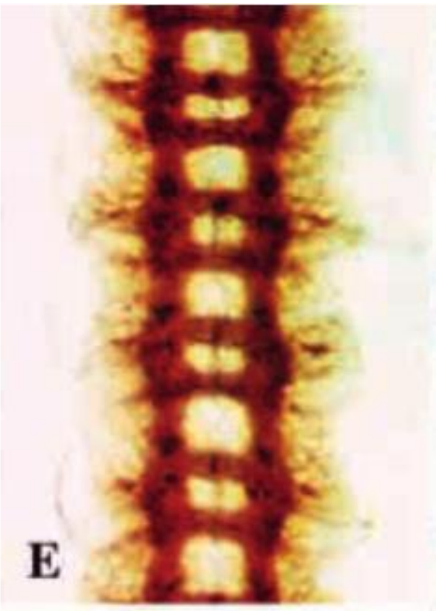


Figure 1.2.1. Flat preparations of a wildtype *D. melanogaster* embryo showing the "rope-ladder" of the arthropod central nervous system. Stage 16 *D. melanogaster* embryo stained with the monoclonal antibody mAbBP102. Neuromeres are repeated in segmental units, connected longitudinally by connectives, and transversally by commissures. Anterior up; figure reproduced from Overton *et al.* (2002).

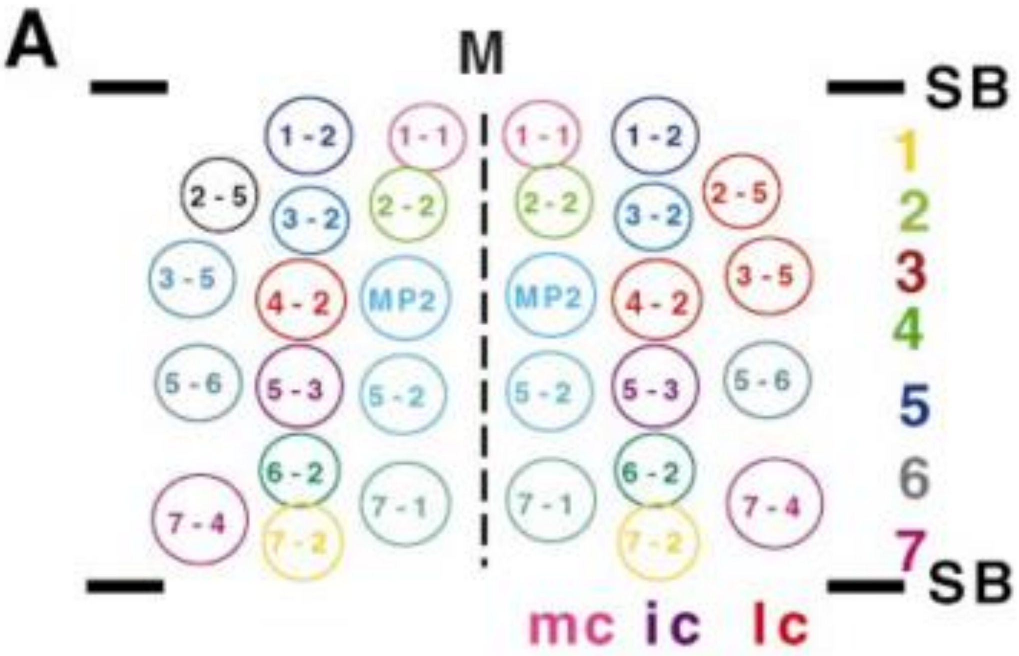


Figure 1.2.2. Schematic diagram of the NBs of a neuromere from an early stage 9 *D. melanogaster* embryo. The segmental boundary (SB) of the neuromeres is shown, and the ventral midline (M) is represented by a dashed line in the centre. Each segment repeats this pattern of NB formation iteratively, producing ~30 NBs in total (however in the stage 9 embryo represented here, only around half of those NBs have delaminated). The NBs are arranged in 7 rows along the AP axis and 3 columns along the DV axis, the medial column (mc), the intermediate column (ic), and the lateral column (lc). Each NB is numbered and colour-coded according to its identity. This system is based on the grasshopper *Schistocerca americana*, however, *D. melanogaster* is less orthogonal in comparison. Figure reproduced from Bhat, 1999.

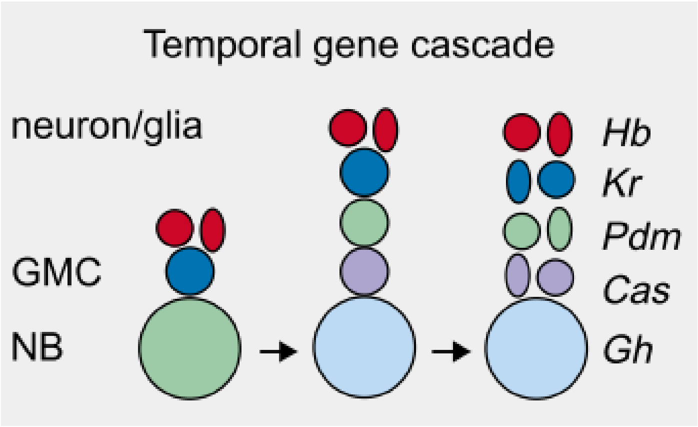


Figure 1.2.3. The temporal gene expression cascade in neural cells of *D. melanogaster*. The gene *hunchback* (*hb*, red) is expressed first, followed by *Krüppel* (*Kr*, blue), *nubbin* (*nub*, or *Pdm* here, green), and *Castor* (*Cas*, purple). *grainy head* (*Gh*) is also expressed in some neural cells (*Gh*, light blue). Most GMC divisions are alternatively asymmetric, represented here by ovals and circles. Figure reproduced from Skeath & Thor, 2003.

1.3 Introduction to genomic approaches

Much of the research discussed above has made use of classical genetic studies. However, techniques have been developed to investigate the interaction of genes and proteins at the genomic level, addressing questions such as how TFs regulate biological processes (reviewed in Latchman, 1997). Much of the early work on DNA regulation was performed in prokaryotes (Jacob & Monod, 1960; Englesberg *et al.*, 1965), and these investigations found that TFs interact with genomic loci by physically binding to the DNA (Karin, 1990; Latchman, 1997) in order to regulate the expression of other genes. Many of these *cis*-regulatory modules (CRMs) (Davidson & Erwin, 2006; Levine, 2010) have now been identified in species such as *Drosophila melanogaster*, giving rise to maps detailing an abundance of regulatory elements (Celniker *et al.*, 2009).

Enhancers are an example of a CRM, and comprise a short sequence of DNA that, when bound by a TF, influences the transcription of an associated gene (Khoury & Gruss, 1983; Serfling *et al.*, 1986; Pennacchio *et al.*, 2013). An enhancer can be up to 1 million base pairs upstream or downstream from its associated gene (Lettice *et al.*, 2003; Pennacchio *et al.*, 2013), yet be spatially adjacent due to the 3D structure of DNA (Maston *et al.*, 2006; Pennacchio *et al.*, 2013). Enhancers have traditionally been identified using enhancer trap protocols, which make use of a reporter gene, *e.g. lacZ*. Randomly inserting the *lacZ* locus into the genome using P-elements (O'Kane & Gehring, 1987) can reveal nearby enhancers; and monitoring the expression of *lacZ* transcripts will elucidate the regulatory effects of the associated enhancer (O'Kane & Gehring, 1987; Hartenstein & Jan, 1992). More modern identification methods make use of a combination of molecular and computational techniques to identify regions commonly bound by TFs (Visel *et al.*, 2007).

The Berkeley Drosophila Transcription Network Project have characterised the enhancers of many genes, with a notable focus on enhancers governing *Drosophila* segmentation (Li *et al.*, 2008). A well-characterised enhancer, for example, is a 480bp region driving the expression of the pair rule gene *even-skipped*, which contains 12 binding sites for different gap gene TFs (Borok *et al.*, 2010). Some enhancers operate through 'enhancer synergy', whereby two or more enhancers work together to produce spatially and temporally regulated gene expression patterns (Perry *et al.*, 2011). One such example of this is the two enhancers that regulate the expression of three gap genes: *hunchback*, *Krüppel*, and *knirps* (Perry *et al.*, 2011).

Another notable project of high-throughput enhancer analysis and characterisation is the FlyLight project, which functionally mapped regulatory elements using the expression of GAL4 (Brand *et al.*, 1993) driven by thousands of different genomic fragments to identify regulatory elements active in the *Drosophila* nervous system (Jenett *et al.*, 2012; Li *et al.*, 2014).

The majority of the studies investigating TF binding utilise chromatin immunoprecipitation (ChIP) to identify regions bound by different TFs (O'Neill & Turner, 1996; Visel *et al.*, 2007; Visel *et al.*, 2009; Collas, 2010), although other techniques, such as DamID, can also be used (van Steensel & Henikoff, 2000; Vogel *et al.*, 2007; Aughey & Southall, 2015; Marshall *et al.*, 2016). Immunoprecipitation (IP) isolates a known protein from biological material (usually a lysate of a biological sample) using an antibody specific to the protein of interest (Rosenberg, 2005). Chromatin IP (ChIP) therefore utilises this technique to examine the interactions of known chromatin-associating proteins and DNA sequences (Gilmour & Lis, 1984; Gilmour & Lis, 1985). Native ChIP can be used to investigate the targets of histone modifiers, identifying nucleosomal fragments to which the histone binds; cross-linked ChIP, in contrast, is more widely used to identify the DNA targets of proteins associated with chromatin, such as TFs (Collas, 2010). In cross-linked ChIP, proteins are temporarily cross-linked with DNA using formaldehyde (Jackson, 1978), or less commonly, UV light (Gilmour & Lis, 1985). The lysate is then sonicated to shear the chromatin, although nuclease digestion may also be performed to fragment chromatin (Jackson & Chalkley, 1981). Fragment sizes of 400-500bp in length are preferred, covering 2-3 nucleosomes (Kornberg, 1974). Protein-DNA complexes are then precipitated using an antibody specific to the protein and washed to remove non-specifically bound chromatin. The cross-linking is reversed, and proteins are removed by digestion with Proteinase K and the isolated DNA is purified. This DNA can then be identified by PCR, hybridisation to a microarray (ChIP-chip), or high-throughput sequencing (ChIP-seq) (Collas, 2010) (Figure 1.3.1A).

In contrast, DamID generates a similar type of data albeit via a very different method. DamID can also be used to map the DNA targets of TFs of interest, however the DNA binding events are captured *post hoc*, as opposed to the snapshot of TF binding achieved in ChIP experiments (van Steensel & Henikoff, 2000; Greil *et al.*, 2006; Vogel *et al.*, 2007; Aughey & Southall, 2015; Marshall *et al.*, 2016). DamID, or **D**NA **a**denine **m**ethyltransferase **id**entification (van Steensel & Henikoff, 2000), utilises the Dam protein, an enzyme endogenous to *Escherichia coli* which methylates adenine nucleotides in the context of GATC sequences (Brooks & Roberts, 1982).

This enzyme does not occur in eukaryotes *in natura*, and thus can be used in transgenic animals or transfected cells to identify protein-DNA interactions. This is achieved by creating a fusion protein between the TF of interest and the Dam protein, inserting this transgene into the host's genome, and ectopically expressing the protein. Subsequently, everywhere the TF binds in the genome, the Dam fusion will methylate nearby adenine regions up to 2.5kb from the TF binding site (van Steensel & Henikoff, 2000). However, adenomethylation is poorly tolerated in eukaryotes, and consequently, low level, 'leaky' expression is required to avoid methylation saturation (van Steensel & Henikoff, 2000; Vogel *et al.*, 2006; Southall *et al.*, 2013). This expression is so low that the protein is undetectable by Western blotting or immunofluorescence (Vogel *et al.*, 2007). Moreover, given the high affinity of the Dam protein for DNA, a Dam-only control is necessary; the binding events of this Dam-only control are subsequently 'subtracted' from the TF-Dam fusion binding events, and only these differential binding events are considered *bona fide* TF binding activity. Methylated DNA is then isolated and enriched using methylation-sensitive nucleases and PCR. First, the *Dpn*I restriction enzyme is used to cleave methylated GATC sites, fragmenting the DNA. Cut DNA is then passed through a size-selecting column, which removes any uncut genomic DNA. An adapter sequence is then ligated to the 5' and 3' ends of fragments, and DNA is digested with *Dpn*II: this second digestion cuts non-methylated GATC sites, and serves as a secondary selection step to cleave any non-methylated fragments that may have passed through the column. DNA is then amplified by PCR using a primer complementary to the adapter sequence; unmethylated DNA that has been cleaved with *Dpn*II thus will not be amplified at this stage (van Steensel & Henikoff, 2000; Vogel *et al.*, 2006; Marshall *et al.*, 2016). Most commonly, sequences are identified via hybridization to a microarray (DamID-chip) or sequenced using high-throughput platforms (DamID-seq) (Figure 1.3.1B).

High-throughput sequencing and microarray analysis for DamID and ChIP each have the advantage of providing whole-genome coverage, and enriched sequences can be mapped to a reference genome. This analysis yields 'peaks', or stacks of binding events, which can be visualised and potentially provide an indication of binding strength (affinity) to individual loci. These peaks can be mapped to nearby genomic features, such as transcription start sites, introns, exons, promoters, and enhancer regions of the genome (Zhu, 2010; Rashid *et al.*, 2011; Yu, 2014). Common patterns in the bound sequences identified, or motifs, can also be identified, and comparisons may be drawn between the motifs of different TFs, or motifs of orthologous TFs in different species (Borneman *et al.*, 2007; Odom *et al.*, 2007; Carl & Russell, 2015). Moreover, associated gene regions can be queried for gene ontology to identify which biological functions TF binding correlates with (Johnson *et al.*, 2007; MacArthur *et al.*, 2009; Zhu *et al.*, 2010; Bailey *et al.*, 2013; Carl & Russell, 2015;).

There are advantages and disadvantages of DamID and ChIP, relative to one another. For example, ChIP is superior to Dam in terms of resolution; since Dam only methylates DNA in the context of GATC sequences (Brooks & Roberts, 1982; van Steensel & Henikoff, 2000), the resolution is limited to how frequently GATC sites occur in the genome and the average distance between them. In contrast, ChIP identifies binding at the true source, independent of non-TF motifs (Collas, 2010), however, both techniques typically enrich fragments 400-500bp in size. Resolution can be further enhanced using exonucleases in parallel with ChIP (ChIP-exo), whereby an exonuclease is introduced to cleave DNA to within a few bp at the protein binding site (Rhee & Pugh, 2011). ChIP can also be less technically challenging than DamID, as ChIP does not require introducing and driving transgene expression with a suitable promoter. Moreover, in ChIP experiments, it is the endogenous protein binding *in situ*, as opposed to the modified trans-protein used in DamID; this may better reflect *in vivo* binding events, especially as the shape of the protein is modified with Dam-fusions which may influence binding events. There may also be post-translational modifications made to TFs that are absent in TF-Dam fusions. Nonetheless, DamID binding data correlates well with binding data generated using ChIP experiments (Aleksic *et al.*, 2013), implying that DNA binding is not distorted in the presence of the Dam-fusion.

Indeed, there are many advantages of DamID in comparison to ChIP. For example, DamID is not reliant on a highly specific antibody, and therefore represents an attractive alternative when antibodies are unavailable for the TF of interest. ChIP also cannot be used to discriminate between different TF isoforms as antibodies are often indiscriminate; DamID can achieve this by engineering the transgene in such a way that different isoforms are expressed. Furthermore, ChIP experiments provide a mere ‘snapshot’ of DNA binding at the time of cross-linking the chromatin; DamID, in contrast, provides a historical ‘signature’ of binding in the genome. (This can also be an advantage of ChIP, however, as it is more readily used to develop a time series of binding events which can be coupled with expression analyses (Sanguinetti *et al.*, 2006; Asif *et al.*, 2010). DamID has also been recently utilised in tissue-specific experiments in *Drosophila*, under the control of the GAL4 system, generating both temporally and spatially specific data (Southall *et al.*, 2013; Marshall *et al.*, 2016).

The advantages and limitations of each technique are therefore context-dependent, meaning researchers possess a degree of flexibility in the techniques available for genomic experiments. Indeed, there are a large number of TF binding experiments utilizing either ChIP or DamID in *D. melanogaster*, for example, investigating the DNA-binding patterns of proteins involved in embryonic AP and DV patterning (MacArthur *et al.*, 2009; Paris *et al.*, 2013), wing patterning (Prasad *et al.*, 2016), cellular transcription machinery (Southall *et al.*, 2013), and CNS development (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014). Studies have also drawn evolutionary comparisons of TF binding between different drosophilid species (Bradley *et al.*, 2010; He *et al.*, 2011; Paris *et al.*, 2013; Carl & Russell, 2015; Prasad *et al.*, 2016). Beyond the *Drosophila* model, both ChIP and DamID have been used in mammals (Vogel *et al.*, 2007; Odom *et al.*, 2007), *Caenorhabditis elegans* (Schuster *et al.*, 2010), *Arabidopsis thaliana* (Germann & Gaudin, 2011), and *Saccharomyces cerevisiae* (Borneman *et al.*, 2007).

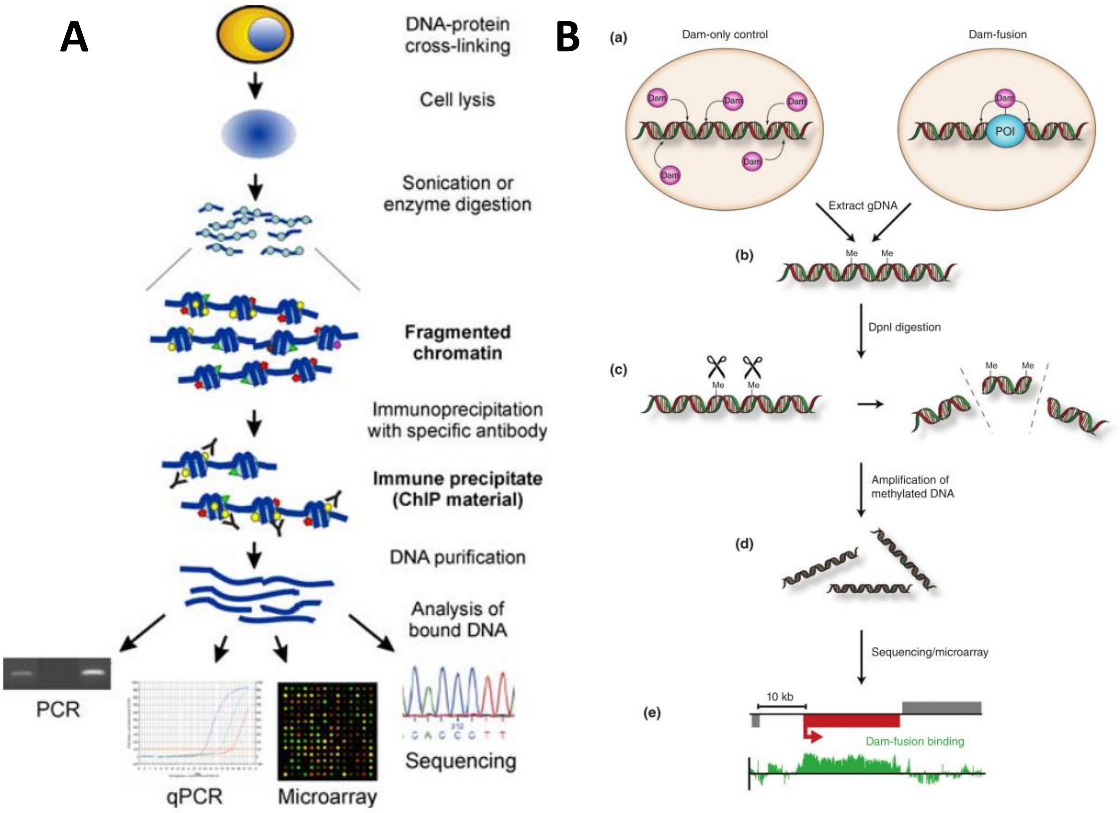


Figure 1.3.1. Schematic diagram of ChIP vs DamID DNA enrichment protocols. (A) ChIP protocol: DNA and proteins are covalently cross-linked (typically achieved using formaldehyde), and biological material is lysed. The genomic DNA of the lysate is fragmented, and the protein of interest (and bound DNA) is immunoprecipitated using a specific antibody. DNA is then purified and amplified. (Figure reproduced from Collas, 2010). (B) DamID protocol: the protein-Dam fusion (and a Dam only control) are inserted via transgenesis, so they bind to DNA *in vivo* and methylate nearby adenine regions. Genomic DNA is extracted, and cleaved with adenomethylation sensitive enzymes. Isolated DNA is then amplified. (Figure reproduced from Aughey & Southall, 2015). The DNA from each method can be analysed via PCR, microarray hybridisation, or high-throughput sequencing.

1.4 *Tribolium castaneum* as an emerging model organism

Beetles are arguably the most successful order of not just insects, but the entire animal kingdom (Hunt *et al.*, 2007). This success arises not only through numerical domination, but also through a diversity which exceeds any other known order of animals (Stork, 1988; Farrell, 1998). It is perhaps surprising then that it is only relatively recently (in the last couple of decades) that they have received much attention, chiefly through the study of the red flour beetle *Tribolium castaneum*; first as an ecological model, and later as a model for studying evolutionary developmental (evo-devo) biology (Brown *et al.*, 2009). The red flour beetle has co-evolved with humankind since the advent of farming practices, acting as pests in cultivated grains and dry foods for millennia (Klingler, 2004; Sallam, 2008). *T. castaneum* has been described as "probably the most common secondary pest of all plant commodities in the world" (Sallam, 2008), and the remains of *T. castaneum* have been discovered in ancient Egyptian tombs (Klingler, 2004), signifying their historical blight on ancient civilisations. Their ability to live in arid environments arises from a remarkable adaptation which enables water recovery from the rectum, and an elongated hindgut that doubles back on itself to further facilitate water reabsorption (King & Denholm, 2014).

T. castaneum, while not nearly as well established as the fruit fly model *Drosophila melanogaster*, has steadily become a widely used model to investigate modes of insect development (Brown *et al.*, 2009). Their short life cycles (~30 days from egg to adult with 6-9 larval instar stages), extended longevity in comparison with other insects (up to 3 years), and high fecundity makes these animals highly amenable to scientific study in the laboratory (Sokoloff, 1972). Females will lay up to 6 eggs at a time, and their polyandrous behaviour (they will mate with multiple males in a single copulation session) ensures favourable genetic diversity in populations (Pai *et al.*, 2005). The genome of *T. castaneum* has also been sequenced and published, revealing a homology with *D. melanogaster* which includes a similarly-sized genome and many orthologous gene regions (Richards *et al.*, 2008).

T. castaneum are also much more representative of insect species than the widely-studied *D. melanogaster*; their larvae possess 3 pairs of thoracic legs, and their fully formed heads include eyes (Klingler, 2004; Bucher & Wimmer, 2005; Brown *et al.*, 2009). This is in stark contrast to *Drosophila*, whose larvae are eyeless and legless, and their brains much less developed (Brown *et al.*, 2009). Indeed, anterior development in *D. melanogaster* is highly derived; for example, the gene *bicoid* (*bcd*) regulates anterior development in *Drosophila* embryos (St Johnston &

Nusslein-Volhard, 1992) and is unique to drosophilids, absent from even closely related Diptera (Stauber *et al.*, 1999). Larval *bcd* mutants develop with missing heads and thoraxes, and *bcd* has been shown to transcriptionally activate the conserved homeobox gene *orthodenticle* (*otd*) and the gap gene *hunchback* (*hb*), both of which are involved in anterior patterning, whilst repressing *caudal*, a gene involved in posterior patterning (Schröder, 2003; Schröder *et al.*, 2008). In the *Tribolium* egg, however, *otd* is maternally contributed, and no orthologue exists for *bcd*. Knock-down of *otd* in beetles results in the absence of larval heads, evocative of *Drosophila bcd* mutants. Double knock-downs of *otd* and *hb* in *Tribolium* results in larvae missing the head, thorax, and anterior abdomen, as is seen in flies mutant for *bcd*.

However, perhaps the most striking difference between *T. castaneum* and *D. melanogaster* are their respective methods of germband elongation and segmentation (Chapman, 1998; Davis & Patel, 2002; Liu & Kaufman 2005; Brown *et al.*, 2009). *D. melanogaster* possesses a derived form of long-germ extension, where all segments are determined almost all at once during the blastoderm stage, prior to cellularization (Akam, 1987; Nasiadka *et al.*, 2002). Therefore, diffusion of regulatory elements such as transcription factors (TFs) and ligands is largely responsible for early patterning mechanisms (Sulston & Anderson, 1996; Liu & Kaufman, 2005; Peel *et al.*, 2005). Positional information from gap and pair rule factors establish the boundaries of each segment by regulating segment polarity gene expression upon cellularization (Patel, 1994; Liu & Kaufman, 2005). (Figure 1.4.1).

This is in contrast to the short-germ extension of *T. castaneum*, whereby segments are sequentially added from a posterior growth zone (reviewed by Peel *et al.*, 2005; Schröder *et al.*, 2008) (Figure 1.4.2), and patterned in the reverse sequential manner from anterior to posterior (older to younger) (Patel, 1994; Choe *et al.*, 2006; Schröder *et al.*, 2008; Clark & Peel, 2017). This process involves pair-rule genes being expressed in periodic oscillations in the posterior growth zone of the embryo (Sarrazin *et al.*, 2012; Brena & Akam, 2013). Loss of function experiments in *T. castaneum* embryos of the pair-rule genes *even-skipped*, *odd-skipped*, and *runt* result in acutely truncated embryos, demonstrating their necessity for growth zone maintenance.

This is purported to be controlled by a segmentation clock analogous to that of vertebrates (Sarrazin *et al.*, 2012). Oscillating patterns of gene expression shown in the growth zone of vertebrates play an integral role in vertebrate segmentation (*e.g.* see Palmeirim *et al.*, 1997; Masamizu *et al.*, 2006; Oates *et al.*, 2012). In an elegant experiment, Sarrazin *et al.* (2012)

showed that *Tc-odd* expression oscillates with two-segment periodicity, providing compelling evidence of a segmentation clock for the first time in *Tribolium* (such a mechanism has been demonstrated in other arthropods, such as the myriapod *Strigamia maritima* (Brena & Akam, 2013) and the cockroach (Pueyo *et al.*, 2008)). The researchers used live cell tracking techniques to demonstrate that different cell populations expressed *Tc-odd* transcripts at different time points. They demonstrated that this was not simply a case of intraspecific variation between embryos; by bisecting live embryos along the anterior-posterior axis, they were able to show that *Tc-odd* expression oscillates within an individual embryo, with the two halves showing differential expression patterns at different stages of development (Sarrazin *et al.*, 2012).

The segmentation process has also been shown to mostly be a consequence of changing cell behaviours as opposed to primarily by cell proliferation. Throughout elongation, the increase in total germband area is relatively modest, and cell tracing experiments reveal that large cellular proliferation is unnecessary for posterior germband elongation (Nakamoto *et al.*, 2014), with anterior cellular migration from the growth zone being chiefly responsible (Figure 1.4.3). Moreover, the addition of segments in *Tribolium* is not uniform in its regularity: segment addition slows significantly in the early stages of germband extension, before rapidly increasing midway through the elongation process, coinciding with thoracic and abdominal identity transitions (Nakamoto *et al.*, 2014). Together, these studies suggest that cellular rearrangement is primarily responsible for germband elongation during abdominal segmentation, and that the ‘growth zone’ of the embryo, while exhibiting modest levels of mitosis, is largely regulating cellular organisation.

However, recent work suggests that the segmentation process is more conserved than previously thought (Clark & Peel, 2017; Clark, 2017). For example, work by Clark and Peel (2017) has investigated the role of *Caudal*, *Dichaete*, and *Odd-paired* in both *Drosophila* and *Tribolium* segmentation. They found that these three genes all have temporally distinct functions on the *Drosophila* pair-rule network: primary pair-rule genes are expressed in the context of *Caudal* and *Dichaete* expression, and secondary pair-rule genes are activated as *Caudal* is deactivated, with *Odd-paired* expression activating frequency doubling and segment polarity gene activity. In *Tribolium*, these genes are expressed in a concomitant manner, with the temporal activity of each gene correlating with analogous phases of segmentation in

Drosophila, implying a conserved function in coordinating the segmentation process (Clark & Peel, 2017).

Clark (2017) has therefore put forward a new model in an attempt to reconcile the divergent modes of long and short germband segmentation. Clark proposes that these mechanisms are not dichotomous modes, but rather represent differences in the regulation and deployment of a highly conserved pair-rule network. In long germ insects such as *Drosophila*, Clark posits that this pair-rule network is patterned by gap gene inputs, whereas in short germ insects such as *Tribolium*, the pair-rule network is under the control of oscillating clock enhancers (Clark, 2017). Therefore, the evolution of the derived long germ formation is not such a significant 'jump', which might explain why many paraphyletic orders within Insecta have seemingly evolved long germband development independently of one another (Liu & Kaufman, 2005; Lynch *et al.*, 2012).

Similarly, CNS development in *T. castaneum* has been shown to be largely conserved with *D. melanogaster* (Wheeler *et al.*, 2003; Wheeler *et al.*, 2005; Kux *et al.*, 2013; Biffar & Stollewerk, 2015). For example, *Tribolium* neuroblast formation is conserved with *Drosophila*; proneural cell clusters are under the control of an Achaete-Scute Complex (AS-C) comprising a single proneural gene, *Tc-achaete-scute homologue* (*Tc-ASH*), and a homologue of the neural precursor gene *asense* (*Tc-ase*) (Wheeler *et al.*, 2003). While this is discordant with *Drosophila* (which possess 3 proneural genes of the AS-C), *Tc-ASH* performs their collective function as a single gene (Wheeler *et al.*, 2003). *Tc-ASH* is expressed in all proneural clusters and neuroblasts and is necessary for neuroblast formation, and *Tc-ase* is expressed only in neuroblasts, both acting homologously to the *Drosophila* genes. Moreover, the early patterning of neuroblast positioning is conserved between *Drosophila* and *Tribolium*, with homologues of 3 of the columnar genes of *Drosophila*, *vnd*, *ind*, and *msh*, expressed in 3 longitudinal columns in the developing *Tribolium* neuroectoderm (Wheeler *et al.*, 2005; Biffar & Stollewerk, 2015) (Figure 1.4.4). The general arrangement of neuroblasts is also highly conserved between not just *Tribolium* and *Drosophila*, but other insect species also (Biffar & Stollewerk, 2014, Figure 1.4.5).

Therefore, comparative studies utilizing beetles such as *T. castaneum* have the power to elucidate the deep evolutionary innovations within insects, and address pressing questions on how those innovations might have arisen. *T. castaneum* is closely enough related to *D. melanogaster* (they are both holometabolous insects) to draw meaningful comparisons and identify orthologous features, whilst being evolutionarily distant enough to study how those orthologous features have changed and adapted over time.

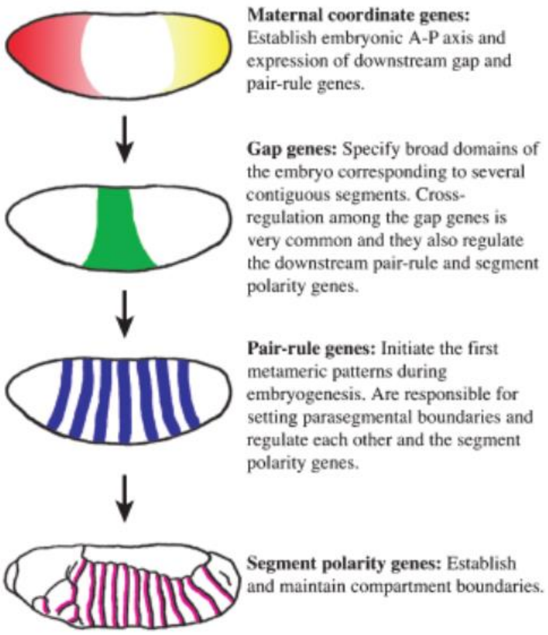

Maternal coordinate genes: Establish embryonic A-P axis and expression of downstream gap and pair-rule genes.
Gap genes: Specify broad domains of the embryo corresponding to several contiguous segments. Cross-regulation among the gap genes is very common and they also regulate the downstream pair-rule and segment polarity genes.
Pair-rule genes: Initiate the first metameric patterns during embryogenesis. Are responsible for setting parasegmental boundaries and regulate each other and the segment polarity genes.
Segment polarity genes: Establish and maintain compartment boundaries.

Figure 1.4.1. Schematic diagram of long germ development of *Drosophila melanogaster*. The maternal genes establish the embryonic axes, and initiate the expression of gap genes. Gap genes in turn activate and repress the pair-rule genes, which are expressed in a two-segmental periodicity. Finally, pair-rule genes regulate the expression of segment polarity genes which establish cell fate in each of the segments. (Figure modified from Liu & Kaufman, 2005).

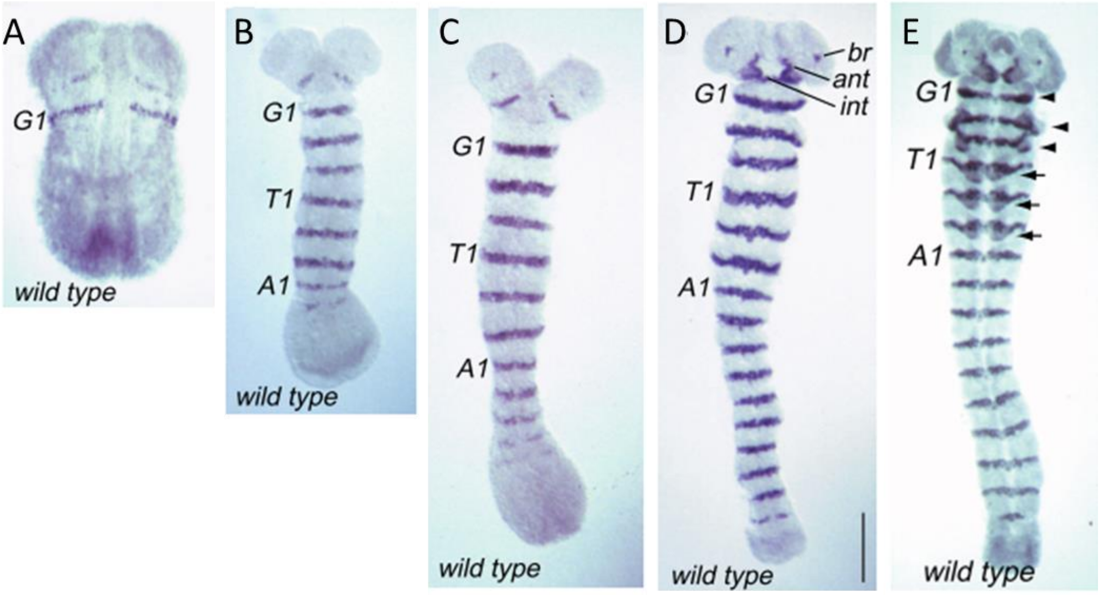


Figure 1.4.2. Wildtype *Tribolium castaneum* embryos stained for *engrailed* expression during germband elongation. A bulbous posterior growth zone is visible in early to mid-stage embryos, from which segments are sequentially added, thereby lengthening the embryo. G1 = gnathal segment 1; T1 = thoracic segment 1; A1 = abdominal segment 1. Anterior up. Figure reproduced from Sulston & Anderson, 1996.

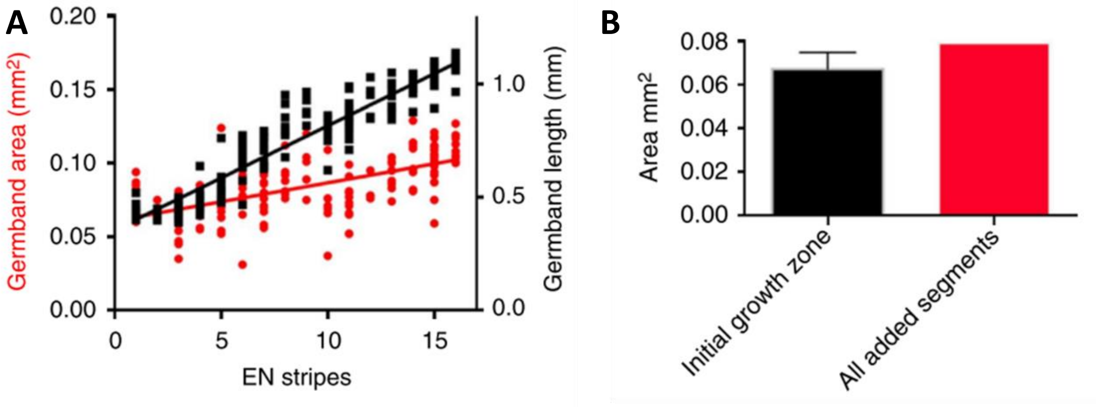


Figure 1.4.3. The elongation of *T. castaneum* germbands arises primarily through cellular migration as opposed to cellular proliferation. (A) The increase in overall germband area is comparatively lower than the overall increase in germband length. (B) The area of the posterior growth zone in early embryos is similar to the area of all segments of the elongated germband. Figure reproduced from Nakamoto *et al.*, 2015.

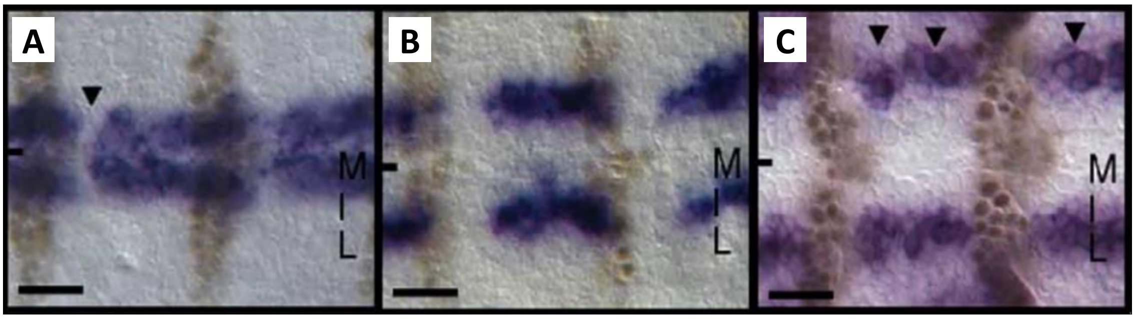


Figure 1.4.4. The longitudinal columns of the neuroectoderm in *T. castaneum*. The 3 longitudinal columns medial (M), intermediate (I), and lateral (L) observed in *D. melanogaster* are also observed in *T. castaneum*. Embryos are stained for *Tc-vnd* (A), *Tc-ind* (B), and *Tc-msh* (C) expression (blue/purple), and Engrailed expression (brown). Scale bars = 25 µm. Figure modified from Wheeler *et al.*, 2005.

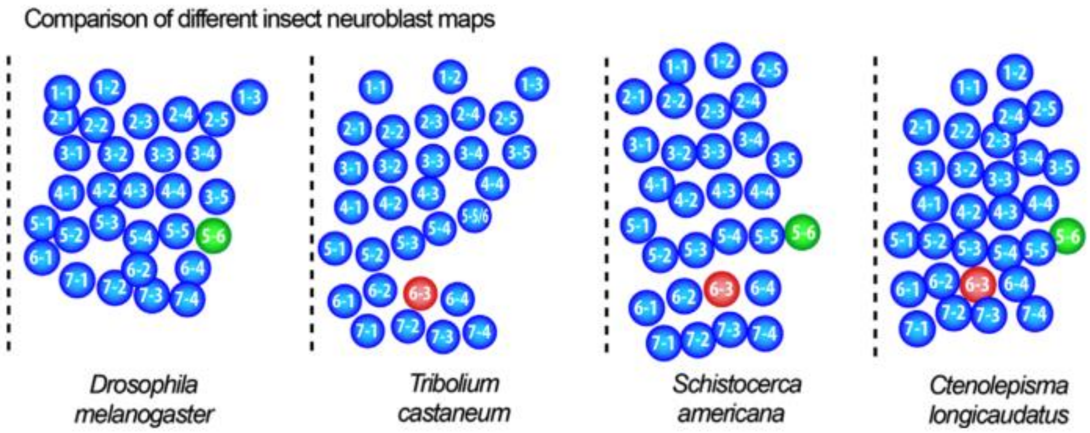


Figure 1.4.5. The neuroblast arrangements of 4 different insect species along a hemineuromere. NB numbers and arrangement are highly conserved across hemimetabolous and holometabolous insects, except for the presence of an additional NB in row 5 (green) that is missing in *T. castaneum*, and an additional NB in row 6 (red) that is missing in *D. melanogaster*. Anterior up, and the ventral midline is represented by the dashed line to the left. Figure reproduced from Biffar & Stollewerk, 2014.

1.5 Research aims

Sox are a fascinating and diverse family of genes that have provided the basis for much engaging research. Group B Sox, in particular, are widely studied across the animal kingdom and have been identified in all metazoan taxa examined to date. The majority of invertebrate SoxB research has been conducted in the model organism *Drosophila melanogaster*, illuminating the indispensable role that two SoxB genes, *Dichaete* and *SoxNeuro*, play in early CNS development. However, *Drosophila* is unrepresentative of insects for the reasons discussed in this chapter. It is therefore important to broaden the scope of research to better understand Sox evolution, and consequently, efforts to identify and characterise Sox across species are ongoing.

There are conflicting models explaining SoxB expansion within the protostomes that are yet to be resolved. Moreover, many of the early patterning genes in the CNS have been characterised in species other than *Drosophila*, yet research into *Dichaete* and *SoxNeuro* in the wider arthropods has been neglected. Finally, much of the research into the evolution of TF binding have been between relatively closely related species, separated by up to 90 million years (*e.g.* Odom *et al.*, 2007; He *et al.*, 2011; Carl & Russell, 2015), however, there are relatively few investigations of TF evolution across deep evolutionary time. In light of these points, the purpose of this research project was to address the following questions:

- **How did the early expansion of SoxB genes in protostomes and arthropods transpire? Which of the two conflicting models of SoxB evolution, if either, is valid?** To test each model, I identified and annotated the SoxB genes of 20 invertebrate species, and examined signature residues and clustering behaviours of the respective HMG domains against the assumptions of each model.

- **How do the expression patterns of *Dichaete* and *SoxNeuro* compare between insects with long germ development and those with short germ development?** To address this question, I selected the short germ insect *Tribolium castaneum* as a model in which to study both *Dichaete* and *SoxNeuro* expression, using *in situ* hybridisation.

- **How do the genome-wide binding profiles of two TFs, Dichaete and SoxNeuro, compare between species separated across deep evolutionary time?** To answer this question, I also selected *T. castaneum* as a model, as its genome is published and relatively well-annotated. However, this involved endeavouring to establish the first genome-wide study of TF binding in *Tribolium* embryos, and given the scarcity of antibodies available for this species, I elected to attempt DamID to achieve this.

This research is therefore highly exploratory in nature. The chief aim of this project is to broaden our understanding of Group B Sox genes within arthropods. The principal focus is establishing DamID as a technique in *T. castaneum*, to investigate the conservation/divergence of genome-wide binding activity of two integral SoxB proteins. This research will not only help address long-standing questions on Sox evolution within the arthropods, but also help establish this species as a model organism for wider genomics research beyond the *Drosophila* paradigm.

Chapter 2

Materials & Methods

2.1 Beetle husbandry and stock-keeping

Wild type (*WT*) Georgia GA2 and *vermillion*^{white} (*V*^{w}) (Berghammer *et al*., 1999) *Tribolium castaneum* strains were used for all experiments, reared on medium containing 1kg organic grain flour + 50g yeast powder, at 35^{o}C in a lightly humidified (40-60%) tower incubator. Food medium was pre-sieved with a 700 µm sieve, enabling separation of beetles from flour using an 800 µm sieve. *WT* embryo collections were conducted over a 24 hour period at 32^{o}C, where adults were removed from grain flour and transferred to organic white flour for the overnight lay.

Embryo injections to generate transgenic lines were conducted by Dr Julia Ulrich and myself at the University of Göttingen, and by Johannes Schinko at the Tribolium Genome Editing Service (TriGenES), part of the Institut de Génomique Fonctionnelle de Lyon (IGFL). *V*^{w} adults were transferred to white flour to lay embryos for 1 hour at 25^{o}C, and embryos were collected and left to develop for a further hour at 25^{o}C before being prepared for injection. Collected embryos were washed in deionized H_2O, and carefully dechorionated by washing two times in 1% bleach in a 200 µm mesh basket. Embryos were delicately washed once more in deionized H_2O, and, using a paintbrush, were lined up with the posterior tip facing outwards along a glass slide. Embryos were injected in the posterior third along the transverse plain using a glass needle loaded on a Leitz micromanipulator, with the injection mix consisting of a *piggyBac* helper plasmid at 0.4 µg / µl, and a *piggyBac* plasmid containing the construct of interest at 0.6 µg/µl. Slides were then placed onto apple juice agar plates (for humidity), and plates were sealed in a plastic box and left to develop at 32^{o}C. When the first larvae hatched, embryos were transferred to a dry box, and hatched larvae transferred to grain flour using a fine brush. Larvae were then backcrossed with *V*^{w} individuals, and F1 progeny were scored for eye-specific *GFP* expression. *GFP*-positive F1 progeny from a single injected adult were crossed with each other to establish an inbred transgenic population. Due to the lack of balancer chromosomes for *T. castaneum*, populations were monitored for *GFP* expression continuously until allele fixation in each population.

2.2 Phylogenetic analysis

Reference Genomes

Reference genomes were selected to represent major taxa across Insecta and Arthropoda, where genomes were available. Genomes were downloaded either from EnsemblMetazoa (http://metazoa.ensembl.org) or Ensembl (http://www.ensembl.org/) for 21 invertebrate and 3 vertebrate species. The invertebrate reference genomes used were: *Drosophila melanogaster* July 2014 (FlyBase, Release BDGP6)*, Drosophila pseudoobscura* 2012 (FlyBase, Release Dpse_3.0)*, Anopheles gambiae* February 2006 (VectorBase, Release AgamP4)*, Bombyx mori* February 2013 (SilkDB, Release ASM15162v1)*, Heliconius melpomene* February 2012 (Heliconius Genome Consortium, Release Hmel1)*, Tribolium castaneum* February 2010 (BeetleBase, Release Tcas3)*, Dendroctonus ponderosae* April 2013 (TRIA-Project, Release DendPond_male_1.0)*, Apis mellifera* February 2011 (BeeBase, Release Amel_4.5)*, Atta cepolates* July 2012 (Ant Genomes Portal, Release Attacep1.0)*, Nasonia vitripennis* November 2012 (NasoniaBase, Release Nvit_2.1)*, Pediculus humanus* November 2008 (VectorBase, Release PhumU2) *Acyrthosiphon pisum* June 2010 (AphidBase, Release Acyr_2.0)*, Rhodnius prolixus* December 2010 (Vector Base, Release RproC1)*, Zootermopsis nevadensis* June 2014 (Zootermopsis nevadensis Genome Project, Release ZooNev1.0)*, Daphnia pulex* February 2011 (JGI, Release V1.0)*, Strigamia maritima* February 2013 (EnsemblGenomes, Release Smar1)*, Ixodes scapularis* August 2007 (VectorBase, Release IscaW1)*, Tetranychus urticae* November 2011 (ORCAE, Release ASM23943v1)*, Parastaetoda tepidariorum* September 2013 (Baylor College of Medicine, i5k Initiative: Common House Spider Genome Project, Release Ptep_1.0)*, Hypsibius dujardini* August 2016 (Nematode and Neglected Genomics, IEB, Release Release LRSR01.1)*,* and *Caenorhabditis elegans* December 2012 (WormBase, Release WBcel235). The vertebrate reference genomes used were: *Gallus gallus* December 2013 (Gallus_gallus-5.0, INSDC Assembly), *Mus musculus* January 2012 (Genome Reference Consortium Mouse Reference 38; GRCm38.p5 INSDC Assembly), and *Homo sapiens* December 2013 (Genome Reference Consortium Human Build 38; GRCh38.p10 INSDC Assembly). Divergence times for 16 of the 19 arthropods studied were estimated using TimeTree (Hedges *et al*., 2015).

Identifying SoxB Homologs

It was not possible to acquire the full protein sequence for all genes across all species due to varying quality of genome assemblies. Instead, the HMG domain of *Drosophila melanogaster's* Dichaete protein (**QEGHIKRPMNAFMVWSRLQRRQIAKDNPKMHNSEISKRLGAE WKLLAESEKRPFIDEAKRLRALHMKEHPDYKYRPRRKPKNPLT**) was aligned against each target genome using the BLAST-like alignment tool (BLAT) (Kent, 2002). To preserve potentially conserved gene neighbourhoods, an R script was then used to extend coordinates by 200kb both upstream and downstream of hits generated by the BLAT report (or, if present in a <400kb sequence, i.e. a small shotgun sequence, the whole contig was selected), and the relevant region extracted from the target genome in DNA fasta format. Sequences were analysed using the Artemis genome browser (Rutherford *et al*., 2000). Highly conserved regions of the *Drosophila* HMG domain, which spans introns in the respective SoxB genes, were used to query the fasta files in Artemis, and HMG domains were annotated and saved in a separate fasta file. 20 amino acids upstream and downstream of the HMG were included for each sequence.

Sequence alignment, domain identification, amino acid distributions, and phylogenetics

Sequences were sorted as Dichaete-like or SoxNeuro-like according to sequence homology, and Dichaete-like sequences were subsequently categorised into candidates for Sox21a, Sox21b, or SoxB5 homologs according to a combination of their sequence homology, chromosomal positioning, intron structure, and closest hits according to the Basic Local Alignment Search Tool (Johnson *et al*., 2008). Amino acid sequence alignment was performed using the MAFFT multiple alignment software (Katoh & Standley, 2013), using the sorted fasta option and L-INS-I strategy, enabling the alignment of a set of flanking sequences around one alignable domain (in this case, the HMG-box domain). Sequences were sorted according to the alignment output. Sox1, Sox2, Sox3, Sox14, and Sox21 HMG domains from *Homo sapiens*, *Mus musculus*, and *Gallus gallus*, and SRY HMG domains from *H. sapiens* and *M. musculus*, were also included in the alignment.

As it was not possible to acquire/identify the full protein sequences of all species, a subset of 12 species was selected based on the quality of their genome assemblies. The 12 species selected were: *Drosophila melanogaster, Bombyx mori*, *Tribolium castaneum, Apis mellifera, Strigamia maritma, Ixodes scapularis, Tetranychus urticae, Caenorhabditis elegans, Gallus gallus, Mus musculus,* and *Homo sapiens*. Whole protein sequences were identified for each

Sox gene, with the exception of the *Dichaete-2* gene of *Strigamia maritima* due to an incomplete shotgun sequence. Unaligned sequences were manually sorted according to orthology and species, and queried for conserved protein domains using the NCBI BatchConservedDomain tool (Marchler-Bauer *et al*., 2017).

Consensus sequences for HMG domains were generated using WebLogo (Crooks *et al*., 2004), and an R script was used to count the proportion of residues conforming to the consensus sequence at each position. An R script was also used to categorize the R-group of each amino acid; heatmaps were generated to visualise these data, and the proportion of residues conforming to the consensus R-group at each position was counted, as above.

Finally, the PhyML package (Guindon *et al*., 2010) was used to generate Maximum Likelihood trees with 100 bootstraps using the WAG substitution model (Whelan & Goldman, 2001).

2.3 *In situ* hybridisation & Immunohistochemistry

Probe Synthesis

Primers were selected to amplify the *Dichaete* and *SoxNeuro* loci from the *Tribolium castaneum* genome, including 5' and 3' untranslated regions (Table 2.3.1). PCR products were then cloned into the TOPO vector (Invitrogen) using TA cloning (Holton & Graham, 1991). Once cloning was achieved, plasmids were linearized using *Not*I (NEB) and *Bam*HI (NEB) restriction enzymes for sense and anti-sense transcription, respectively.

In vitro transcription utilized the DIG RNA Labelling Mix (Roche) and the Fluorescein RNA Labelling Mix (Roche). T7 RNA polymerase (Thermo) was used to synthesise sense RNA, and Sp6 RNA polymerase (NEB) for anti-sense RNA. 0.5 µg of linearized plasmid DNA was added as template, with 2 µl DNA labelling mix, 2 µl transcription buffer, 0.4 µl RiboLock^{TM} RNase Inhibitor (Thermo), and 1 µl or 2 µl of the respective polymerase, with DEPC-treated water bringing the total volume up to 20 µl. In total, ~20 µg of RNA was synthesised from the 0.5 µg starting DNA template for each probe. The only probe to successfully generate signal was *Tc-SoxNeuro_2* (Table 2.3.1).

Table 2.3.1. Primers used to generate the template for DIG-labelled riboprobe synthesis. Tm = annealing temperature.

	5’ Primer	3’ Primer	Tm (^oC)
Tc-Dichaete_1	CAAGATGCACAACTCGGAGA	TGCATTTGCACTATTGATGGA	59
Tc-Dichaete_2	CTGCCCACGGCGCTCAAG	CATAACTGGGACCGGCCTGC	62
Tc-Dichaete_3	AAGACGGGGGTGGGTTTC	CCTGCGGATGTCCAGCTCT	62
Tc-SoxNeuro_1	GTCCAGCTTGATCCCGACTA	GGCGACGCACTGTACTGCT	60
Tc-SoxNeuro_2	AGTACCGGCCTAGGAGGAAG	AATAAATGGCGACGGATTCA	60

Embryo Fixation

24 hour embryos were collected following an overnight lay as described above. Embryos were dechorionated in 50% bleach using a 200 µm mesh basket, rinsed well with deionized H_2O, and fixed by shaking for 25 minutes in 3ml *Tribolium* fixation buffer (13 ml 1x PBS, 13.4 ml 0.5 M EGTA, pH 8.0, 73.6 ml H_2O), 6 ml heptane, and 450 µl formaldehyde. The aqueous phase was removed and 8 ml of methanol was added. Embryos were vigorously shaken for 30 seconds, and then left to settle. This osmotic shock liberated embryos from their vitelline membrane: embryos that sank to the bottom of the vial were collected and transferred into a 1.5 ml Eppendorf tube with a cut 1 ml pipette tip. Manual devitellinization is required for subsequent steps: an 0.8 µm canula was fitted to a 10 ml syringe, and embryos remaining at the interphase were sucked up and expelled with moderate force back into the vial. Any embryos that sank at this stage were collected and transferred to a 1.5 ml tube, as described above. This procedure was repeated three more times, each time increasing the force applied to the syringe. Fixed devitellinized embryos were washed three times with methanol, and stored at -20^{o}C for later use.

Embryo Staining

For colourimetric *in situ* hybridisation, embryos were re-hydrated via successive washes in 1:1 PBT/MeOH, and then 1x PBT (PBT = Phosphate Buffered Saline + 0.4% Triton X100). Embryos were post-fixed for 15 minutes in 1 ml PBT containing 140 µl 37% formaldehyde, and washed with PBT; followed by 6 minutes incubation in 1 ml PBT containing 8 µg of Proteinase K, and then an additional post-fix step for 15 minutes in 1 ml PBT containing 140 µl 37% formaldehyde. Embryos were incubated at 65^{o}C for 1 hour in Hyb solution (10 ml deionized formamide, 5 ml 20x SSC pH 5.5, 5 ml deionized H_2O, 400 µl of 10 mg/ml boiled sonicated salmon testis DNA, 100 µl of 20 mg/ml tRNA, 20 µl of 50 mg/ml heparin). Probes were meanwhile diluted 1:1000 in Hyb solution, incubated at 95°C for 2 minutes, and immediately transferred to ice, before being added to the embryos and left overnight at 65°C to hybridise. Successive washes in PBT and BBT blocking solution (0.1% bovine serum albumin in 1x PBT) were performed, and embryos were then incubated for 1 hour in 1:2000 anti-DIG or anti-fluorescein antibody in BBT at room temperature. Successive washes in PBT were performed again, and the subsequent alkaline phosphatase staining reaction took place using 20 µl of NBT/BCIP (Roche) or INT/BCIP (Roche) stock solution in 1ml PBT. Colour was left to develop in the dark at room temperature for up to three hours until a strong signal could be detected under the dissecting microscope, before stopping the reaction with successive washes in PBT. For double colourimetric *in situ* hybridisation, an additional antibody incubation was performed at this stage with a 1:2000 dilution of anti-DIG or anti-fluorescein in BBT for 1 hour at room temperature, and a second alkaline phosphatase reaction using NBT/BCIP (Roche) or INT/BCIP (Roche). A final post-fix was performed on embryos by adding 1 ml PBT containing 140 µl of 37% formaldehyde.

For immunohistochemistry, antibody staining was performed with the Engrailed/Invected antibody (4D9; Santa Cruz Biotechnology) on embryos which had already been stained via colourimetric *in situ* hybridisation. Embryos were incubated overnight at 4°C with the primary antibody at a 1:5 dilution in PBT. Successive washes with PBT were performed, and embryos were incubated for 2 hours at room temperature with the secondary antibody diluted 1:200 in PBT. Primary antibodies were detected with biotin-conjugated secondary antibodies (goat anti-mouse IgG) using the ABC Elite Kit (Vectastain), with horseradish peroxidase colourimetric detection.

Stained embryos were mounted in glycerol and transferred to Single Frost Micro Slides (Corning) for imaging, using Openlab v.4.0.2 software on a Zeiss Axioplan microscope with 10x and 20x objectives.

2.4 DamID

Cloning

Three constructs were generated for the purposes of DamID: pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-Dichaete-Myc-Dam-SV40], pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-SoxN-Myc-Dam-SV40], and pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-Dam-Myc-SV40].

Dichaete and *SoxN* gene regions were amplified from *T. castaneum* genomic DNA, and constructs containing C-Myc-Dam and N-Dam-Myc sequences were provided by the van Steensel lab (van Steensel and Henikoff, 2000; Greil *et al*., 2006). The *D. melanogaster* HSP70 region was initially used as a promoter, however this proved lethal for the embryos. Instead, the endogenous *Tc-HSP68* basal promoter was selected to facilitate 'leaky' expression, having demonstrated its ability act as a more reliable expression driver than *Dm-HSP70* (Schinko *et al*., 2010). Fragments were assembled in the *piggyBac* vector (Horn & Wimmer, 2000) using the NEBuilder High-Fidelity DNA Assembly Cloning Kit (NEB); the pBac vector was linearized using the *Asc*I and *Fse*I restriction enzymes (NEB), and fragments generated using Phusion High-Fidelity DNA Polymerase (NEB). Primers were designed to contain overlapping sequences with intended neighbouring regions, and the NEBuilder High-Fidelity DNA Assembly Cloning Kit utilises the principle of Gibson Assembly cloning (Gibson *et al.*, 2009), whereby a 5' exonuclease, DNA polymerase, and DNA ligase are introduced in a single reaction, assembling multiple DNA fragments into circular DNA in as little as 60 minutes. The primers used for cloning were generated by NEBuilder software (NEB) can be found in Tables 2.4.1-3.

Table 2.4.1. Primers used to generate pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-Dichaete-Myc-Dam-SV40]. Primers were generated using the NEBuilder software to overlap joining fragments; the overlapping sequences are shown in lower case lettering.

Anneals to	Primer	Overlaps with	Tm (°C)
5xUAS-Tc'HSP	tgtatcttaagcttatcgatacgcgtacggcgcgccATCGATATCTGCAGGTCG	*pBac*	55
5xUAS-Tc'HSP	catggtggcgaattcCGGTACCACTTTGAATTC	*Dichaete*	55
Dichaete	ttcaaagtggtaccgGAATTCGCCACCATGTCTAATTTATA	5xUAS-Tc'HSP	64
Dichaete	tctgttcgcggccgcACATAACTGGGACCGGCC	*Dam*	64
Dam	cggtcccagttatgtGCGGCCGCGAACAGAAAC	*Dichaete*	68
Dam	gacgtcccatggccattcgaattcggccggccAGGCCTTCTAGACTTGAGAATTATTTTTTCG	*pBac*	68

Table 2.4.2. Primers used to generate pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-SoxNeuro-Myc-Dam-SV40]. Primers were generated using the NEBuilder software to overlap joining fragments; the overlapping sequences are shown in lower case lettering.

Anneals to	Primer	Overlaps with	Tm (°C)
5xUAS-Tc'HSP	tgtatcttaagcttatcgatacgcgtacggcgcgccATCGATATCTGCAGGTCG	*pBac*	55
5xUAS-Tc'HSP	catcgtcaacatggtCGGTACCACTTTGAATTC	*SoxNeuro*	55
SoxNeuro	ttcaaagtggtaccgACCATGTTGACGATGGAAACGGACCTCAAAG	5xUAS-Tc'HSP	72
SoxNeuro	tctgttcgcggccgcTGTGCGCGAGGGGCGCCA	*Dam*	72
Dam	cgcccctcgcgcacaGCGGCCGCGAACAGAAAC	*SoxNeuro*	68
Dam	gacgtcccatggccattcgaattcggccggccAGGCCTTCTAGACTTGAGAATTATTTTTTCG	*pBac*	68

Table 2.4.3. Primers used to generate pBac[3xP3-EGFP-SV40;-UAS-Tc'Hsp68-Dam-Myc-SV40]. Primers were generated using the NEBuilder software to overlap joining fragments; the overlapping sequences are shown in lower case lettering.

Anneals to	Primer	Overlaps with	Tm (^{o}C)
5xUAS-Tc'HSP	tgtatcttaagcttatcgatacgcgtacggcgcgccATCGATATCTGCAGGTCG	*pBac*	55
5xUAS-Tc'HSP	ggtggcgttgaattcCGGTACCACTTTGAATTC	*Dam*	55
Dam	ttcaaagtggtaccgGAATTCAACGCCACCATGAAGAAAAATC	5xUAS-Tc'HSP	71
Dam	gacgtcccatggccattcgaattcggccggcCGACCGGCGCTCAGCTGG	*pBac*	71

Isolation of genomic DNA and qRT-PCR of samples

For the pilot study and first two attempts at DamID, adults from each transgenic line were left to lay eggs for 24 hours on organic white flour. The adults were then separated from the flour and returned to their respective vials, and eggs were separated from flour and collected in a petri dish. Residual flour was removed first using a fine paintbrush, and then by rinsing embryos with deionized H_2O embryos three times in a 200 µm mesh basket for 60-90 seconds. Washed embryos were transferred using a paintbrush to a 1.5 ml Eppendorf, and frozen at -20^{o}C. Multiple egg collections had to be performed to generate sufficient biological material for subsequent steps: 3 biological replicates were collected for each transgenic line, with each replicate consisting of ~100 µl settled volume of moist embryos.

For the third attempt at DamID, adults from each transgenic line were collected and euthanized at -20^{o}C. Adult heads were then dissected using a scalpel while the beetles were still frozen, and residual flour was manually removed using a fine paintbrush. The heads were then rinsed three times in deionized H_2O for 60 seconds each, and then subsequently three times in 100% ethanol for 60 seconds, in a 200 µm mesh basket. Heads were transferred to 1.5 ml Eppendorf tubes and frozen at -20^{o}C.

To extract genomic DNA, embryos were suspended in 180 µl homogenization buffer (140 µl 1x phosphate buffer saline, 40 µl 500 mM EDTA). Using blue polypropylene pellet pestles (Sigma-Aldrich) and a pellet pestle motor (Kimble Chase), 30 seconds of motorized homogenization was applied to each sample. DNA was extracted from each sample using the Qiagen DNeasy Blood & Tissue Kit: 20 µl 12.5 µg / µl RNase was added to each sample and pipette mixed; 20 µl of Proteinase K (Qiagen DNeasy kit) was added, pipette mixed, and left to stand for 1 minute at room temperature; 200 µl of Buffer AL (Qiagen DNeasy Kit) was added and pipette mixed 50 times, and left to incubate at 56^{o}C for 10 minutes on a heat block. The samples were then cooled to room temperature, 200 µl of 100% ethanol was added and pipette mixed, and samples were transferred to a spin column. The columns were spun at 6000 x *g* for 1 minute, and the flow-through was discarded. 500 µl AW1 solution (Qiagen DNeasy Kit) was added to the column and spun for 6000 x *g* for 1 minute, the flow-through discarded, and the column was transferred to a new collection tube. 500 µl AW2 solution (Qiagen DNeasy Kit) was added to the column and spun for 6000 x *g* for 1 minute, and the flow through was discarded. The column was additionally spun at 20,000 x *g* for 3 minutes to dry the column. The column was transferred to a new 1.5 ml Eppendorf tube, and 200 µl MilliQ H_2O was added to the centre of the column and left to incubate at room temperature for 30 minutes. Finally, the column was spun at 6000 x *g* for 1 minute: the eluate was stored and the column discarded. The quantity and purity of each sample was measured by loading on Nanodrop (Thermo Scientific), and 5 µl of each sample was run on a 1% agarose gel with Tris Acetate EDTA buffer to determine the quality of the DNA before proceeding.

Following the second attempt at DamID, a quantitative real-time PCR step was used to determine the relative content of wheat and beetle DNA in different experimental samples. Primers were selected to amplify 134bp and 139bp targets in the wheat and beetle genomes, respectively (Table 2.4.4). The SYBR Green Master Mix (Thermo Fischer) real-time PCR system was used according to the manufacturers' instructions. 3 replicates from each sample were included, along with a 5-fold concentration gradient of respective amplicons in order to determine the absolute standard curve.

Table 2.4.4. Primers used in quantitative real-time PCR analysis to identify beetle and wheat DNA. B = beetle, W = wheat, Tm = annealing temperature.

	5' Primer	3' Primer	Tm (^{o}C)
Tc'D_Amplicon_B	CACCCCAACTCGCACGGA	GCAATGGCACACAGACCCCT	60
Tv'R_Amplicon_W	CGTCGTGGACGGAAGTTGA	ACGTGGTTTTGCCCAGTTTT	60

Enrichment of methylated DNA, sonication & library preparation of DamID samples for sequencing

Molecular biology was performed as described by Vogel *et al.* (2007) for the first attempt, with 17 cycles of amplification during the PCR. Molecular biology for the second and third attempts was performed essentially as described by Marshall *et al.* (2016), with some minor modifications. 2.5 µg of DNA for each sample was transferred to a new 1.5 ml Eppendorf tube, and pelleted using a Speed-Vac at 55^{o}C for 60 minutes. The pellets were re-suspended in 43.5 µl MilliQ H_2O, and 5 µl CutSmart Buffer (NEB) and 1.5 µl *Dpn*I enzyme (NEB) was added to each sample and pipette mixed. The samples were left to incubate at 37^{o}C overnight, washed using the Qiagen PCR Purification Kit and eluted in 32 µl MilliQ H_2O. DNA concentration and purity was measured using Nanodrop (Thermo Scientific), and ~400 ng DNA from each sample was pelleted using Speed-Vac at 55^{o}C for 60 minutes. Pellets were re-suspended in 15 µl MilliQ H_2O, and adapter ligation and *Dpn*II digestion were performed as described (Marshall *et al.,* 2016). 15 cycles of PCR amplification was used with the MyTaq Polymerase (Bioline) and samples were purified using Quiagen PCR Purification Kit.

Sonication & Library preparation of DamID samples for sequencing

Following purification in the first attempt at DamID, 1 µl of each sample was run on a 1% agarose gel in TAE buffer to check the quality of DNA, and quantified using Qubit (Thermo Scientific). 1 µg of each sample was transferred to a fresh 1.5 ml Eppendorf tube, and diluted in 100 µl MilliQ H_2O. A 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) was used for subsequent clean-up to remove high and low molecular weight DNA, thereby removing any residual genomic DNA, primers or adapters. The purified DNA was then measured on a 2100 BioAnalyzer using a High Sensitivity DNA chip (Agilent) to determine average fragment size. Libraries were prepared using the ThruPLEX DNA-seq Kit (Rubicon Genomics) according to the manufacturer's protocol, with 10 cycles used during the PCR amplification stage. Libraries were purified once more using the Agencourt AMPure XP beads, and measured on a BioAnalyzer using a High Sensitivity DNA chip (Agilent), whereupon sharp peaks were visible in the 150-250bp range, indicating significant concatemer formation. Libraries therefore underwent a size selection step using the Agencourt AMPure XP beads, whereby a 0.6:1 ratio of libraries:beads was used in order to eliminate/significantly reduce DNA fragments <200bp. Libraries were then multiplexed and submitted to the CRUK Cambridge Institute Genomics Core for 50bp single-end-reads on the HiSeq 4000 platform.

For the second and third attempts at DamID, 1 µl of each sample was run on a 1% agarose gel in TAE buffer to check the quality of DNA, and quantified using Qubit (Thermo Scientific). 2 µg of each sample was transferred to a fresh 1.5 ml Eppendorf tube and diluted in 90 µl MilliQ H_2O, and 10 µl of CutSmart Buffer (NEB). Samples were then sonicated at 4^oC using a Diagenode Bioruptor Plus for 6 cycles of 30 seconds on, 30 seconds off, on high power. Fragment sizes were then measured on the 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent) to ensure successful sonication, and 1 µl *Alw*I enzyme (NEB) was added to the samples. Samples were incubated overnight at 37^oC.

200 ng of each sample was transferred to a fresh 1.5 ml Eppendorf tube, and diluted in 100 µl MilliQ H_2O. A 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) was used for subsequent clean-up to remove high and low molecular weight DNA, thereby removing any residual genomic DNA, primers or adapters. DNA was eluted in 25 µl MilliQ H_2O, which was then subsequently pelletized using Speed-Vac, and re-suspended in 15 µl MilliQ H_2O. Libraries were prepared using the ThruPLEX DNA-seq Kit (Rubicon Genomics) according to the manufacturer's protocol, with the exception that 5 cycles were used during the PCR amplification stage. Samples were cleaned up using a 1:1 ratio of Agencourt AMPure XP beads (Beckman Coulter) and eluted in 40 µl H_2O. The average size of each library was determined on the 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent), and the concentration of each sample was determined using Qubit (Thermo Scientific). The ng / µl concentration was used to calculate the molarity per L using the following equation:

$$DNA\ molarity\ (nmol/L) = \left(\frac{1500}{average\ bp\ of\ library}\right) \times Concentration\ in\ ng/\mu l$$

Libraries were then multiplexed: the concentration of the multiplex was determined using Qubit and the average library size was once more determined on the 2100 Bioanalyzer using a High Sensitivity DNA chip (Agilent). 15 µl of the multiplex was submitted to the CRUK Cambridge Institute Genomics Core for 50bp single-end-reads on a HiSeq 4000.

Sequencing Data Analysis

All high-throughput sequencing data supplied by the CRUK Cambridge Institute Genomics Core were received in FastQ format. A multiple genome alignment was performed against 30 reference genomes by the Institute using bowtie software, however the reference genome of *Tribolium castaneum* was not included in this preliminary analysis. Once data was downloaded from the Institute's servers, the following pipeline, modified from Bardet *et al.* (2011), and used by Dr Sarah Carl to analyse the data published by Carl & Russell in 2015, was used to analyse my data. For the first attempt at DamID only, each library had the adapter sequences trimmed using the cutadapt tool (Martin, 2011) *in silico*. A bowtie (Langmead *et al*., 2009) index was generated for the 2016 *Tribolium castaneum* 5.2 genome assembly and for the genome of *Triticum aestivum*, the Chinese spring wheat variety (see Clavijo *et al*., 2017). All libraries were aligned to the reference genomes using bowtie v0.12.8. Mapped sam files were converted to bam files, sorted and indexed using samtools (Heng *et al*., 2009). Reads were converted to bed files and extended using BEDtools (Quinlan & Hall, 2010) according to average fragment length prior to the library preparation stage. Reads were then visualized by converting to wig and then bigwig file formats, and viewed using the Integrated Genome Browser (IGB) (Freese *et al*., 2016). A FastQC analysis (Andrews, 2015) was also performed for each library.

Chapter 3

SoxB Evolution and Divergence in Arthropods

3.1 Motivations for research

All animals possess Sox genes (Phochanukul & Russell, 2010), from the earliest metazoans such as sponges and cnidarians, through to more complex animals including vertebrates and insects (Prior & Walter, 1996; Wegner, 1999; Bowles *et al.,* 2000; Guth & Wegner, 2008). Given the ancient ubiquity of Sox genes in metazoan development, and their absence in the closest relatives of metazoans, the choanoflagellates, they are speculated to have played a critical role in the emergence of metazoan multicellularity (Larroux *et al.*, 2006; Phochanukul & Russell, 2010). Understanding the phylogenetic origins of the Sox family is therefore of considerable importance.

Sox genes are categorised in groups A through to J across metazoans, and all possess a signature High Mobility Group (HMG)-box encoding domain, which is involved in DNA binding to the minor groove (Ferrari *et al.*, 1992), DNA bending, protein interactions, and nuclear transport (Lefebvre *et al.*, 2007). The HMG domain contains ~79 amino acid residues (Gubbay *et al.,* 1990; Bowles *et al.*, 2000), and are typically found in the first half of the protein sequence.

Multiple paralogues exist within groups A-J, and Group B Sox are a particularly well-studied group. Within the vertebrates, SoxB genes can be clearly classified into two distinct subgroups in terms of function and orthology: Groups B1 and B2 (Bowles *et al.*, 2000; Lefebvre *et al.*, 2007; Guth & Wegner, 2008). Within the chicken, Group B1 genes, *Sox1*, *Sox2*, and *Sox3*, act as transcriptional activators, whereas B2 genes, *Sox14* and *Sox21*, act as transcriptional suppressors (Uchikawa *et al.*, 1999; Kamachi *et al*., 2000; Pevny & Paczek, 2005). However, there has been some recent research confounding these functional classifications (Popovic *et al.*, 2014).

Within insects, there are typically fewer Sox paralogues found within each group (Phochanukul & Russell, 2010). However, Group B Sox genes are more numerous: for example, the model organism *Drosophila melanogaster* possesses four SoxB genes: *SoxNeuro*, *Dichaete*, *Sox21a*, and *Sox21b*. These four SoxB genes are present in 10 species of *Drosophila* studied thus far (Wei *et al*., 2011), with an 11th species, *Drosophila persimilis*, lacking *Sox21b*. These four SoxB genes appear to be typical across Diptera, as they are also found in *Anopheles gambiae* (McKimmie *et al*., 2005). Within the hymenoptera, the Sox genes of *Apis mellifera* and *Nasonia vitripennis* have been identified, which also possess homologues of the four SoxB genes identified above (McKimmie *et al*., 2005; Wilson & Dearden, 2008; Wei *et al*., 2011). However, within the Coleopteran *Tribolium castaneum* and Lepidopteran *Bombyx mori*, a fifth SoxB gene has also been identified, which I call *SoxB5* throughout this chapter. (Wilson & Dearden, 2008; Wei *et al*., 2011). As the Hymenoptera, a more basal branch, do not possess this gene, it has been assumed to not be ancestral to the insects.

There has been much debate as to the phylogenetic origins of Group B Sox, and competing models have been proposed to explain the emergence of SoxB within both the deuterostomes and protostomes. Phylogenetic origins can be difficult to resolve in Sox genes due to the highly conserved nature of the HMG domain and the poorly conserved flanking regions across the rest of the protein. This can impede phylogenetic inferences, as regions that are too highly conserved or too poorly conserved are difficult to infer relationships from (Goldman, 1998; Yang, 1998); both appear to be the case for Group B proteins. Nonetheless, phylogenetic trees generated from amino acid sequence alignments along the HMG domain have proven useful in grouping the proteins into discrete classifications, enabling identification of evolutionary relationships (Bowles *et al.*, 2000; McKimmie *et al.*, 2005; Wilson & Dearden, 2008; Zhong *et al.,* 2011). The two main competing explanations of SoxB evolution are proposed by Bowles *et al.* (2000) (with a model later developed by Zhong *et al.* (2011)), and McKimmie *et al.* (2005) (Figure 3.1.1), and have yet to be fully resolved, especially in arthropod lineages. Both of these analyses focus exclusively on the HMG domain of the proteins.

The McKimmie model proposes that *Dichaete* and *SoxNeuro* are the ancestral SoxB genes and that *Dichaete* duplicated to produce *Sox21a*. Early SoxB expansion is proposed to have arisen through a genome duplication followed by a tandem duplication, with the 3 SoxB genes (*SoxNeuro*, *Dichaete*, and *Sox21a*) present at the deuterostome-protostome split, all being paralogues of vertebrate *Sox3*. Following this, within the insects *Sox21a* and *SoxNeuro* retained their ancestral-like state, orthologous to vertebrate *Sox3*; whereas *Dichaete* diversified and subsequently duplicated to produce *Sox21b*, going on to form an insect-unique class of proteins. Within the vertebrates, *Sox21* and *Sox14* arise and form the B2 class, and *Sox1* and *Sox2* derive from *Sox3* and form the B1 class (summarised Figure 3.1.1A).

In contrast, the Zhong model proposes that an ancient tandem duplication of the ancestral SoxB gene laid the foundations for the SoxB1 and SoxB2 groups, predating the deuterostome/protostome split. SoxB1 is hypothesised to be orthologous to *SoxNeuro* within the insects, and SoxB2 orthologous to *Dichaete*, which in turn underwent two rounds of tandem duplication to produce *Sox21a* and *Sox21b*. Within the vertebrates, the Zhong model proposes both genome duplications and tandem duplications being responsible for the vertebrate expansion of B1 and B2 genes (Figure 3.1.1B).

There is supporting evidence for each model (Figure 3.1.2), with Zhong *et al.* (2011) demonstrating that arthropod SoxB form distinct subgroups, with *SoxNeuro* genes clustering with human SoxB1, and *Dichaete*, *Sox21a*, and *Sox21b* genes clustering with human SoxB2 (Figure 3.1.2A). Contrary to this, Wilson & Dearden's (2008) investigation into insect SoxB genes clusters *SoxNeuro* most closely with *Sox21a*, and *Dichaete* with *Sox21b* (Figure 3.1.2B). More recent analysis of chelicerates recapitulates the phylogenetic grouping exhibited by Wilson & Dearden (2008), with *Dichaete* and *Sox21b,* and *SoxNeuro* and *Sox21a*, forming distinct clusters (S. Russell, unpublished data).

Much of this analysis also appears to rest on the respective authors focusing on different amino acid residues within the HMG domain (Figure 3.1.3). For the McKimmie model, the respective residues are at positions 16 and 21. At position 16, the Dichaete/Sox21b class tends to possess a Leucine/Isoleucine (L/I), whereas the SoxNeuro/Sox21a class a Glycine (G). At position 21, the Dichaete/Sox21b class possesses an Isoleucine (I), and the SoxNeuro/Sox21a class a Methionine (M).

In contrast, Zhong *et al.* (2011) and Bowles *et al.* (2000) focus on the residues at positions 2 and 78: at position 2, SoxB1 possess an Arginine (R) residue where SoxB2 proteins possess a Histidine (H); and at position 78, SoxB1 a Threonine (T) and B2 a Proline (P).

Over the past five years, many metazoan genomes have been sequenced and made available. Sox genes continue to generate considerable interest in the literature, with the ongoing characterisation of SoxB in vertebrates (Kamachi & Kondoh, 2013; Sarkar & Hochedlinger, 2013; Popovic *et al*., 2014; Heenan *et al*., 2016) and several recent studies from this research group on their genome-wide binding profiles in drosophilids (Aleksic *et al.*, 2013; Ferrero *et al.*, 2014; Carl & Russell, 2015). The increased availability of metazoan genomes enables further investigation into metazoan Sox gene sequence and arrangement. The principle motivation for the research discussed here is to test the two models above against emerging genomic evidence, particularly within pan-arthropod taxa. This investigation utilises a combination of phylogenetic tree construction approaches and conserved domain analyses to investigate the supporting evidence for each model, in an attempt to resolve the phylogenetic origins of Group B Sox genes in the Bilateria.

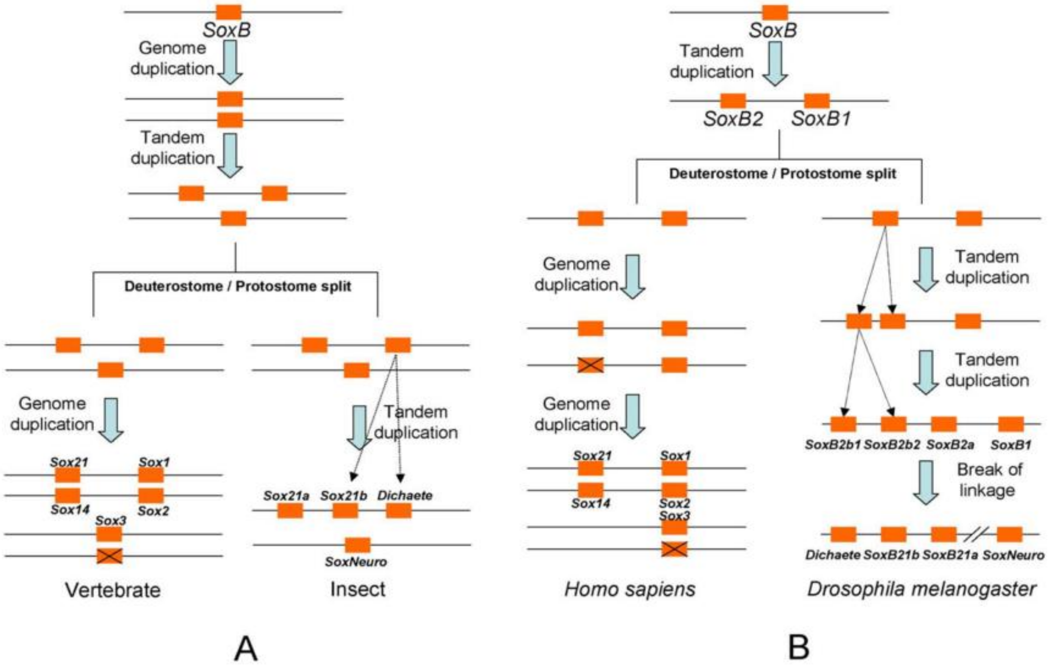


Figure 3.1.1. The McKimmie model vs Zhong model of SoxB evolution in the Bilateria. (A) The McKimmie model; (B) the Zhong model (reproduced from Zhong *et al.,* 2011). The McKimmie model proposes three SoxB genes present at the Deuterostome-Protostome split, which are *Dichaete*, *Sox21a*, and *SoxNeuro*. These are not orthologous to the B1 and B2 subgroupings in vertebrates; instead insect and vertebrate SoxB followed separate evolutionary trajectories in terms of phylogeny and function. The Zhong model proposes just two: SoxB1 and SoxB2 which are orthologous to the B1 and B2 subgroupings found within vertebrates.

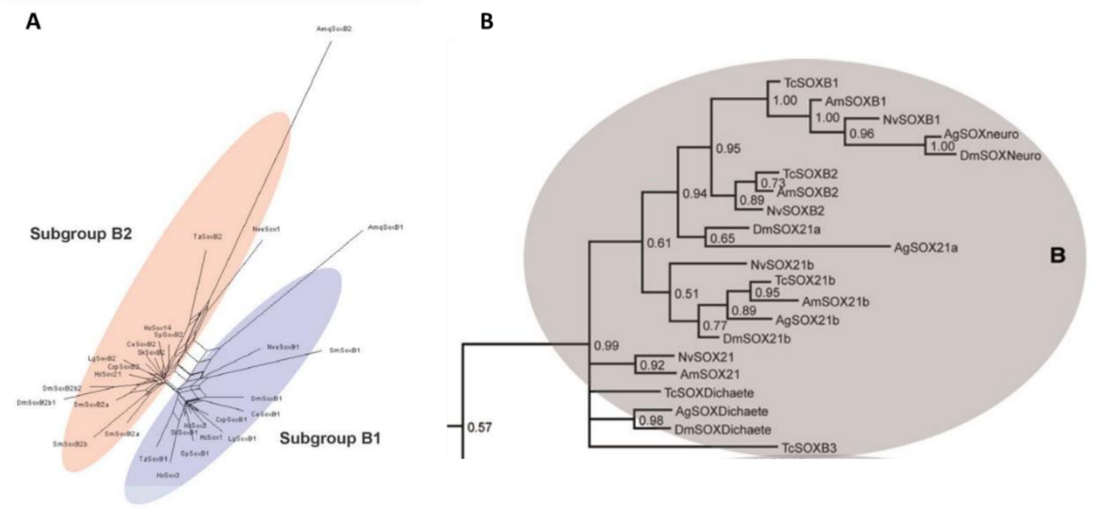


Figure 3.1.2. Supporting evidence for the McKimmie and Zhong models. (A) Supporting evidence for the Bowles model (reproduced from Zhong *et al*., 2011); here, there are clear and distinct subgroupings of arthropod SoxB genes into the vertebrate B1 and B2 groups. (B) Supporting evidence for the McKimmie model (reproduced from Wilson & Dearden, 2008); the cluster of insect *SoxNeuro* with *Sox21a* supports the McKimmie model of SoxB evolution.

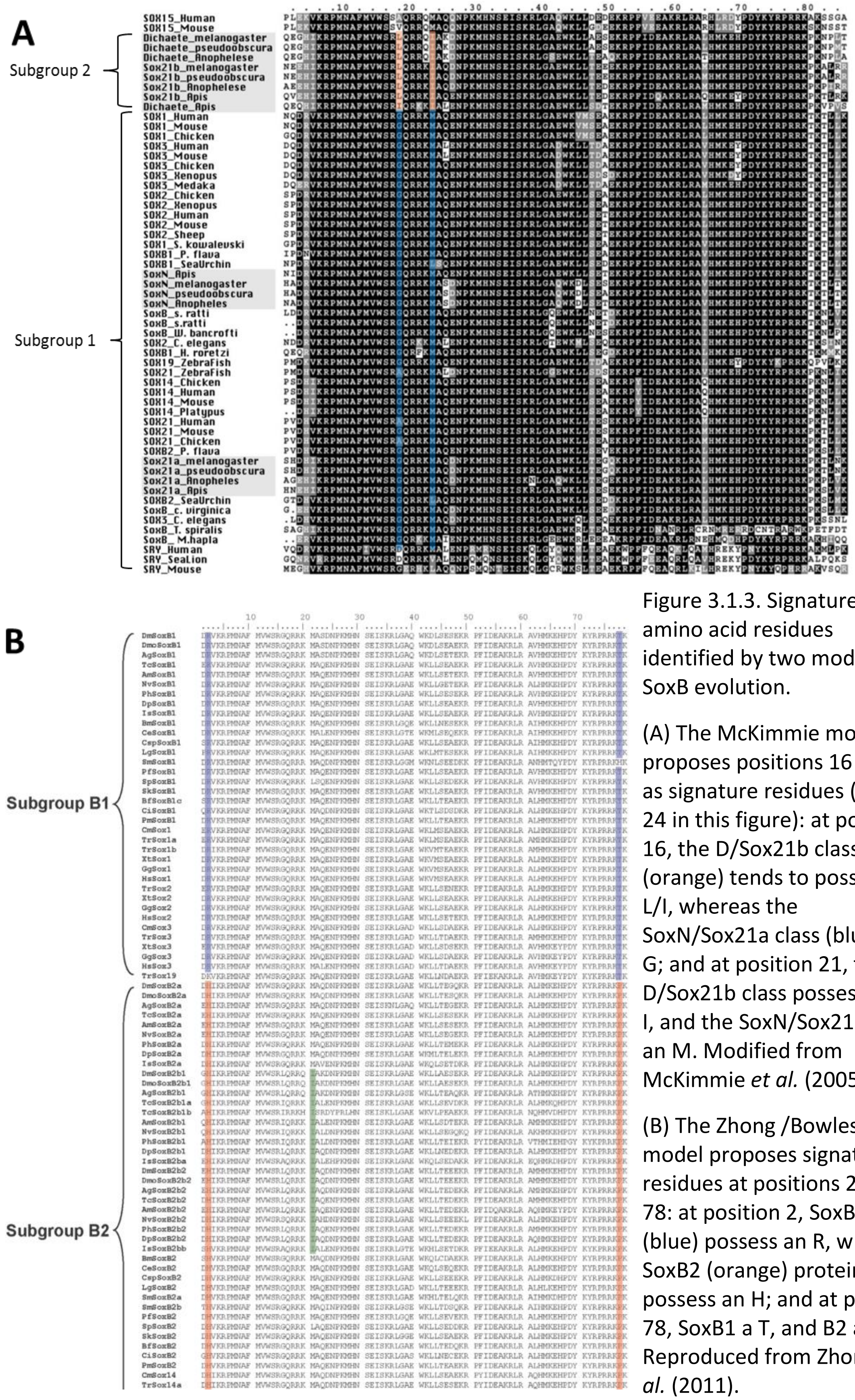


Figure 3.1.3. Signature amino acid residues identified by two models of SoxB evolution.

(A) The McKimmie model proposes positions 16 and 21 as signature residues (19 and 24 in this figure): at position 16, the D/Sox21b class (orange) tends to possess an L/I, whereas the SoxN/Sox21a class (blue) a G; and at position 21, the D/Sox21b class possesses an I, and the SoxN/Sox21a class an M. Modified from McKimmie *et al.* (2005).

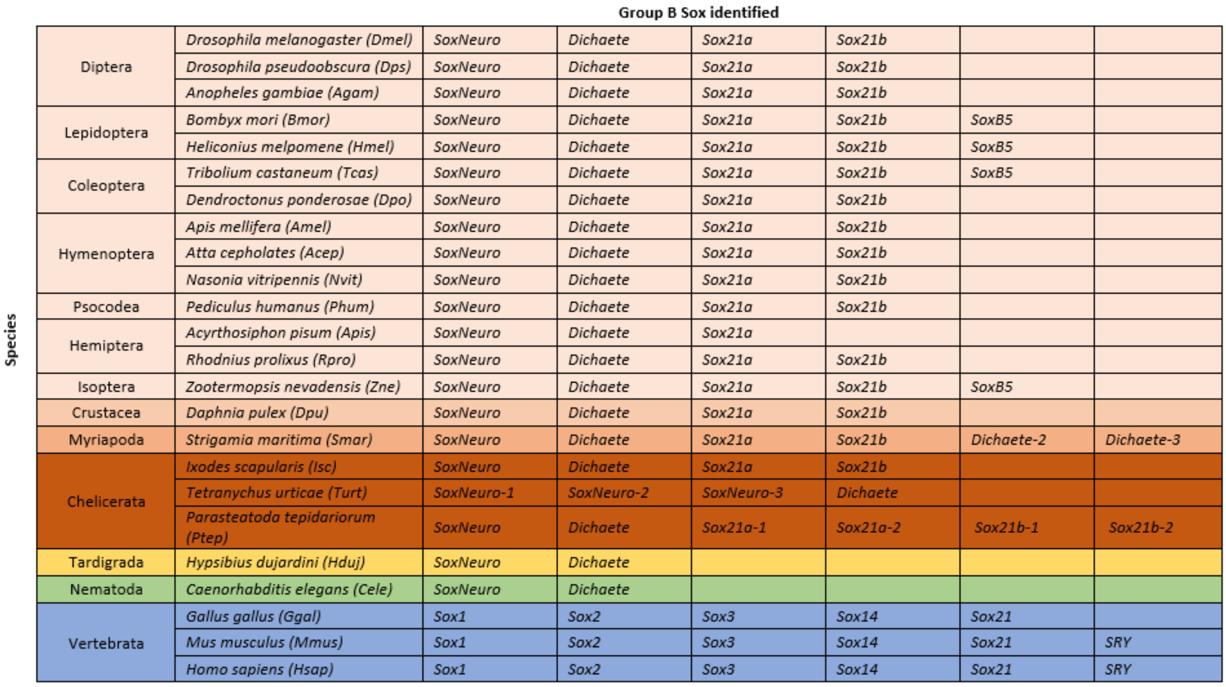
(B) The Zhong /Bowles model proposes signature residues at positions 2 and 78: at position 2, SoxB1 (blue) possess an R, where SoxB2 (orange) proteins possess an H; and at position 78, SoxB1 a T, and B2 a P. Reproduced from Zhong *et al.* (2011).

3.2 Metazoans analysed and their phylogenetic relationships

In total, 21 invertebrate species and three vertebrate species were selected for phylogenetic analysis of protein evolution (summarised in Table 3.2.1), 13 of which did not have prior SoxB gene annotations. The majority of the species analysed were insects (14) due to the availability of genome sequences. Phylogenetic relationships between insects are based on phylogenomic analysis by Misof *et al.* (2014), summarised in Figure 3.2.1C. The phylogeny of the Ecdysozoa is based on phylogenomic analysis of nuclear protein-coding sequences by Regier *et al.* (2010) and is summarised in Figure 3.2.1B. The divergence times representing the 11 orders of the 19 arthropod species included here were estimated using TimeTree software (Hedges *et al.,* 2015), and are shown in Figure 3.2.1A.

Divergence times within the insects, arthropods, invertebrates, and Bilateria are estimated as 358my, 601my, 743my, and 797my, respectively, also using TimeTree software, which takes the mean divergence time estimations across relevant phylogenetic and phylogenomic studies (Hedges *et al.,* 2015).

Table 3.2.1. Group B Sox proteins identified in 24 metazoan species. Listed in descending order according to their relatedness to *Drosophila melanogaster* (top). The abbreviations used in figures for each species throughout this chapter are shown in brackets following the Latin binomials.



Species		Group B Sox identified					
Diptera	*Drosophila melanogaster (Dmel)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Drosophila pseudoobscura (Dps)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Anopheles gambiae (Agam)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Lepidoptera	*Bombyx mori (Bmor)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
	Heliconius melpomene (Hmel)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
Coleoptera	*Tribolium castaneum (Tcas)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
	Dendroctonus ponderosae (Dpo)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Hymenoptera	*Apis mellifera (Amel)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Atta cephalotes (Acep)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Nasonia vitripennis (Nvit)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Psocodea	*Pediculus humanus (Phum)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Hemiptera	*Acyrthosiphon pisum (Apis)*	*SoxNeuro*	*Dichaete*	*Sox21a*			
	Rhodnius prolixus (Rpro)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Isoptera	*Zootermopsis nevadensis (Zne)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
Crustacea	*Daphnia pulex (Dpu)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Myriapoda	*Strigamia maritima (Smar)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*Dichaete-2*	*Dichaete-3*
Chelicerata	*Ixodes scapularis (Isc)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Tetranychus urticae (Turt)	*SoxNeuro-1*	*SoxNeuro-2*	*SoxNeuro-3*	*Dichaete*		
	Parasteatoda tepidariorum (Ptep)	*SoxNeuro*	*Dichaete*	*Sox21a-1*	*Sox21a-2*	*Sox21b-1*	*Sox21b-2*
Tardigrada	*Hypsibius dujardini (Hduj)*	*SoxNeuro*	*Dichaete*				
Nematoda	*Caenorhabditis elegans (Cele)*	*SoxNeuro*	*Dichaete*				
Vertebrata	*Gallus gallus (Ggal)*	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	
	Mus musculus (Mmus)	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	*SRY*
	Homo sapiens (Hsap)	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	*SRY*

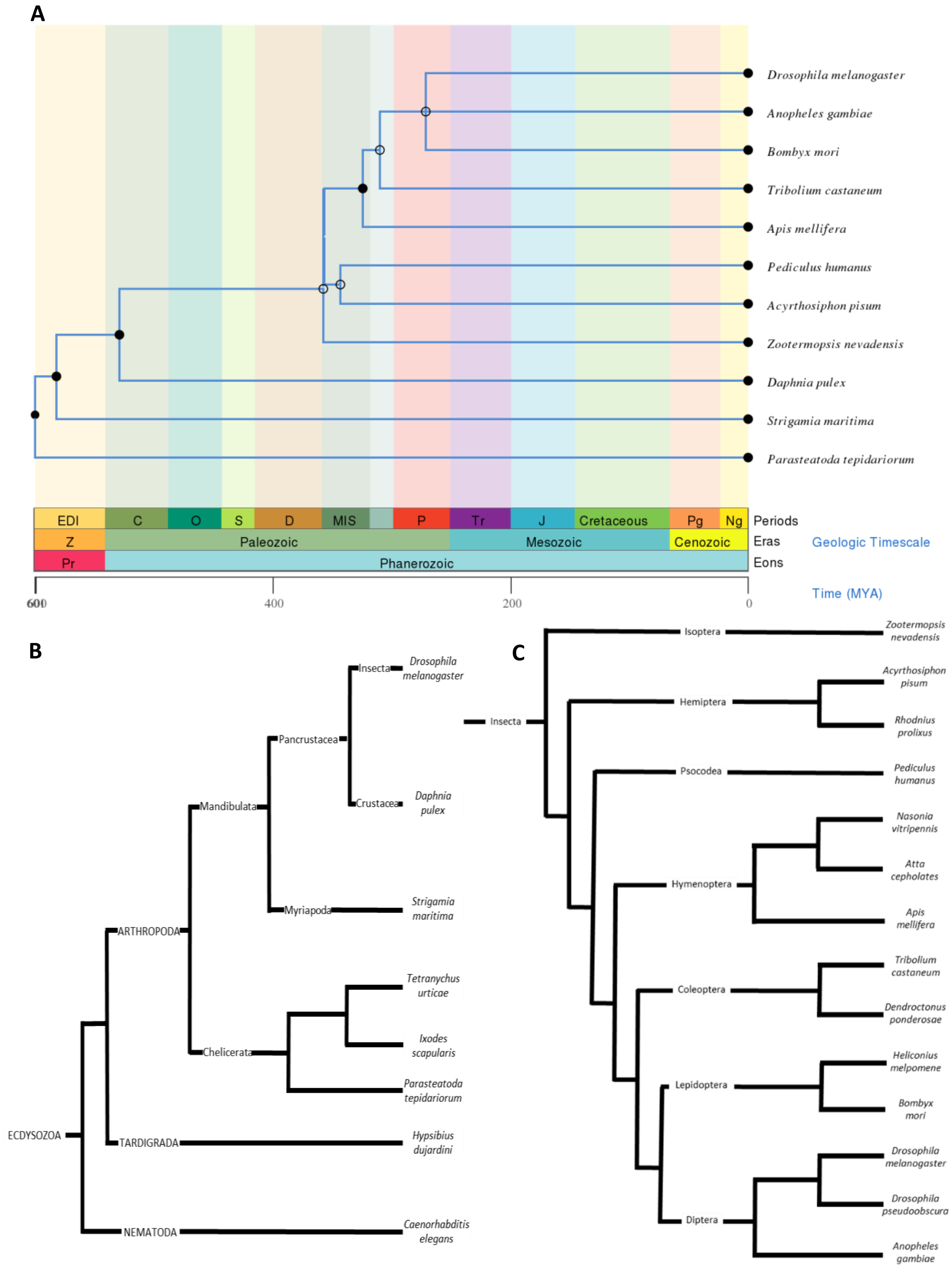


Figure 3.2.1. The cladistic relationships of invertebrate species analysed for SoxB evolution. (A) TimeTree (Hedges *et al.,* 2015) generated phylogenetic tree of taxonomy representing 11 arthropod clades analysed and the estimated geological timescale for divergence. (B & C) Cladograms of the 21 invertebrate species analysed.

3.3 Conserved features across deep evolutionary time

Here I feel it is pertinent to first discuss naming conventions before proceeding with gene identification. There have been multiple attempts in the literature to rename insect SoxB genes to better reflect vertebrate Sox (Bowles *et al*., 2000; Schepers *et al*., 2002; Wilson & Dearden, 2008; Zhong *et al*., 2011), however many of the names given to *Drosophila* genes predate the discovery of their vertebrate orthologues, including the vertebrate *Sry* gene (Sinclair *et al*., 1990). One example is the *Dichaete* gene which was first discovered in 1915 by Calvin Bridges and described by Bridges & Morgan in their 1923 catalogue of *Drosophila* mutations (Bridges & Morgan, 1923). Therefore, renaming invertebrate Sox with unapproved nomenclature is unhelpful (Phochanukul & Russell, 2010). In light of this, I have named genes according to inferred homology with the *Drosophila melanogaster* SoxB repertoire across the protostomes I have analysed. Moreover, I have referred to the fifth SoxB gene found in *T. castaneum* and *B. mori* as *SoxB5* throughout this study. I feel that *SoxB3*, the name used in several papers, is misleading as it implies that there is a third subgroup, along with B1 and B2 genes, which this fifth gene belongs to. Given the ambiguity of whether subgroups B1 and B2 even apply to insects, I do not feel it helpful to potentially implicate a third subgroup.

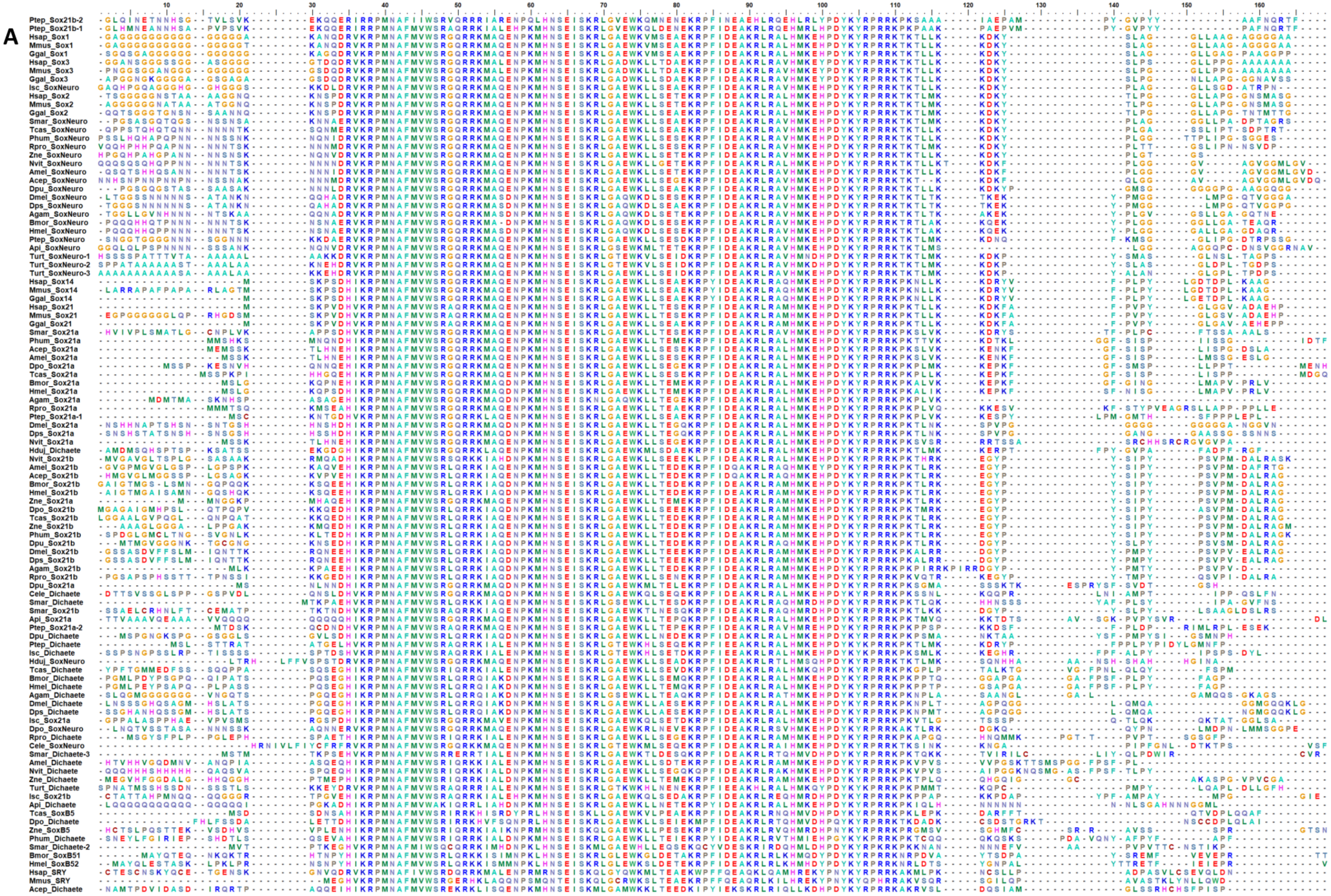
Within the arthropods, HMG domains of almost all Dichaete-homologue proteins (Dichaete, Sox21a, Sox21b, SoxB5) begin with the amino acid sequence **E/DHIKRP**, whereas all the SoxNeuro-like HMG domains start with **E/DRVKRP**. This enabled the classification of each gene according to sequence homology: genes were accordingly sorted as being either a '*Dichaete-homologue*' or '*SoxNeuro-homologue*' (or just '*Dichaete*' or '*SoxNeuro*' within species that contained only two SoxB genes, such as *H. dujardini* and *C. elegans*, as these genes are assumed to be ancestral). Within the arthropods, '*Dichaete-homologue*' genes were subsequently categorised into orthologues/paralogues of *Dichaete*, *Sox21a*, *Sox21b*, and *SoxB5*. This was achieved primarily via identifying intron structures within the HMG domain: the HMG-encoding region of almost all arthropod *Sox21a* and *Sox21b* genes (and all those of the insects) share conserved introns in the same locations, and almost all *Dichaete* and *SoxNeuro* genes of arthropods are intronless. Where there was ambiguity, clustering behaviours and BLAST reports were used to identify the gene. The *Dichaete* gene of *D. pulex* identified here, for example, possesses the same intron as the *Sox21b* gene of insects, yet clusters best with the *Dichaete* genes of arthropods. In contrast, the *Sox21b* gene of *D. pulex* clusters best with the *Sox21b* genes of arthropods and BLAST reports identify it most closely with the *Sox21b* gene of *D. melanogaster*.

Interestingly, within the more basal arthropods *Strigamia maritima*, *Parasteatoda tepidariorum*, and *Ixodes scapularis,* and the nematode *Caenorhabditis elegans*, the HMG domains of Dichaete-homologue proteins begin with the sequence **HVKRP**, similar to that of Sox21 and SRY in the vertebrates. Consensus analysis reveals that position 47 appears to be the most plastic across all proteins (Figure 3.3.1). However, all other positions in the HMG domain exhibit some degree of conservation.

All sequences were aligned with MAFFT alignment software (Katoh & Standley, 2013) using the L-INS-I option, which enables the alignment of flanking sequences around one common alignable domain (in this case, the HMG domain: positions 34-112 in Figure 3.3.1). The signature residues proposed in each of the two models by Zhong and McKimmie (discussed in Section 3.1 and shown in Figure 3.1.3) are highlighted in Figure 3.3.2. Here, sequences are sorted according to the respective signature residues at positions 2 & 78 for the Zhong model and 16 & 21 for the McKimmie model. From this analysis, there is more support for the signature residues identified in the Zhong model, which appears to be the most representative for the proposed subgroups; in contrast, there are many more exceptions to the signature residues proposed in the McKimmie model.

It is also important to query the entirety of the HMG domain to investigate which model is most representative, as opposed to merely examining a handful of residues. One would assume, for example, that with *bona fide* functional and evolutionary subgroupings, the respective consensus sequence of the HMG domains for each subgroup should be the most representative. The chief difference between the two subgroup models within arthropods is the placement of Sox21a proteins: *i.e.* whether they are grouped with SoxNeuro proteins, or with Dichaete and Sox21b proteins. Selecting only arthropod sequences, proteins were separated into the two respective subgroups proposed by Zhong and McKimmie, and consensus sequences (Crooks *et al.*, 2004) were generated for each subgroup (Figures 3.3.3A and 3.3.4A). Support for each consensus motif is shown in the graphs underneath (Figures 3.3.3B and 3.3.4B), where the proportional frequency for each consensus amino acid residue is shown.

Here, it is difficult to conclude one way or the other which subgroup consensus sequence is the most representative, and therefore I elected to perform a more high-level examination of peptide conservation. Amino acids can be classified by their physiochemical properties, *e.g.* by R-group (side-chain) status, to examine functional conservation at the protein level. Each of the 20 amino acids found in eukaryotes can be sorted according to their alpha carbon-attached side chains (R-groups), which are: nonpolar, aliphatic; polar, uncharged; aromatic; positively charged; and negatively charged. Heat maps were thus generated for the respective subgroupings by categorising each amino acid according to its respective R group. Data is shown in Figures 3.3.5 and 3.3.6 for each proposed subgroup, respectively. Here, the level of conservation across the HMG domain is more clearly demonstrated than at the amino acid level. However, in this analysis, it is not any clearer whether the Zhong subgroupings are more representative than those of McKimmie across the entire HMG domain, with similar proportional support for the consensus R groups across both subgroups. The two proposed subgroupings consequently appear to have a comparable explanatory power regarding motif conservation.


A
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
Ptep_Sox21b-2
Ptep_Sox21b-1
Hsap_Sox1
Mmus_Sox1
Ggal_Sox1
Hsap_Sox3
Mmus_Sox3
Ggal_Sox3
Isc_SoxNeuro
Hsap_Sox2
Mmus_Sox2
Ggal_Sox2
Smar_SoxNeuro
Tcas_SoxNeuro
Phum_SoxNeuro
Rpro_SoxNeuro
Zne_SoxNeuro
Nvit_SoxNeuro
Amel_SoxNeuro
Acep_SoxNeuro
Dpu_SoxNeuro
Dmel_SoxNeuro
Dps_SoxNeuro
Agam_SoxNeuro
Bmor_SoxNeuro
Hmel_SoxNeuro
Ptep_SoxNeuro
Api_SoxNeuro
Turt_SoxNeuro-1
Turt_SoxNeuro-2
Turt_SoxNeuro-3
Hsap_Sox14
Mmus_Sox14
Ggal_Sox14
Hsap_Sox21
Mmus_Sox21
Ggal_Sox21
Smar_Sox21a
Phum_Sox21a
Acep_Sox21a
Amel_Sox21a
Dpo_Sox21a
Tcas_Sox21a
Bmor_Sox21a
Hmel_Sox21a
Agam_Sox21a
Rpro_Sox21a
Ptep_Sox21a-1
Dmel_Sox21a
Dps_Sox21a
Nvit_Sox21a
Hduj_Dichaete
Nvit_Sox21b
Amel_Sox21b
Acep_Sox21b
Bmor_Sox21b
Hmel_Sox21b
Zne_Sox21a
Dpo_Sox21b
Tcas_Sox21b
Zne_Sox21b
Phum_Sox21b
Dpu_Sox21b
Dmel_Sox21b
Dps_Sox21b
Agam_Sox21b
Rpro_Sox21b
Dpu_Sox21a
Cele_Dichaete
Smar_Dichaete
Smar_Sox21b
Api_Sox21a
Ptep_Sox21a-2
Dpu_Dichaete
Ptep_Dichaete
Isc_Dichaete
Hduj_SoxNeuro
Tcas_Dichaete
Bmor_Dichaete
Hmel_Dichaete
Agam_Dichaete
Dmel_Dichaete
Dps_Dichaete
Isc_Sox21a
Dpo_SoxNeuro
Rpro_Dichaete
Cele_SoxNeuro
Smar_Dichaete-3
Amel_Dichaete
Nvit_Dichaete
Zne_Dichaete
Turt_Dichaete
Isc_Sox21b
Api_Dichaete
Tcas_SoxB5
Dpo_Dichaete
Zne_SoxB5
Phum_Dichaete
Smar_Dichaete-2
Bmor_SoxB51
Hmel_SoxB52
Hsap_SRY
Mmus_SRY
Acep_Dichaete

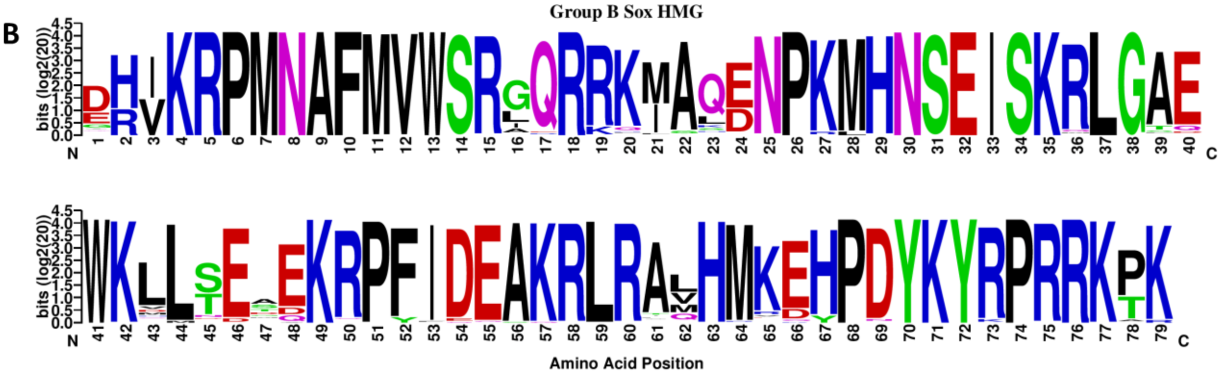


Figure 3.3.1. (A) MAFFT alignment (L-INS-i) of 104 amino acid sequences identified in 24 species of metazoans. The HMG domain is shown in positions 34-112. (B) The consensus sequence of the HMG domain from all 104 seuences. The Consensus logo was generated using WebLogo, developed by Crooks *et al.* (2004).

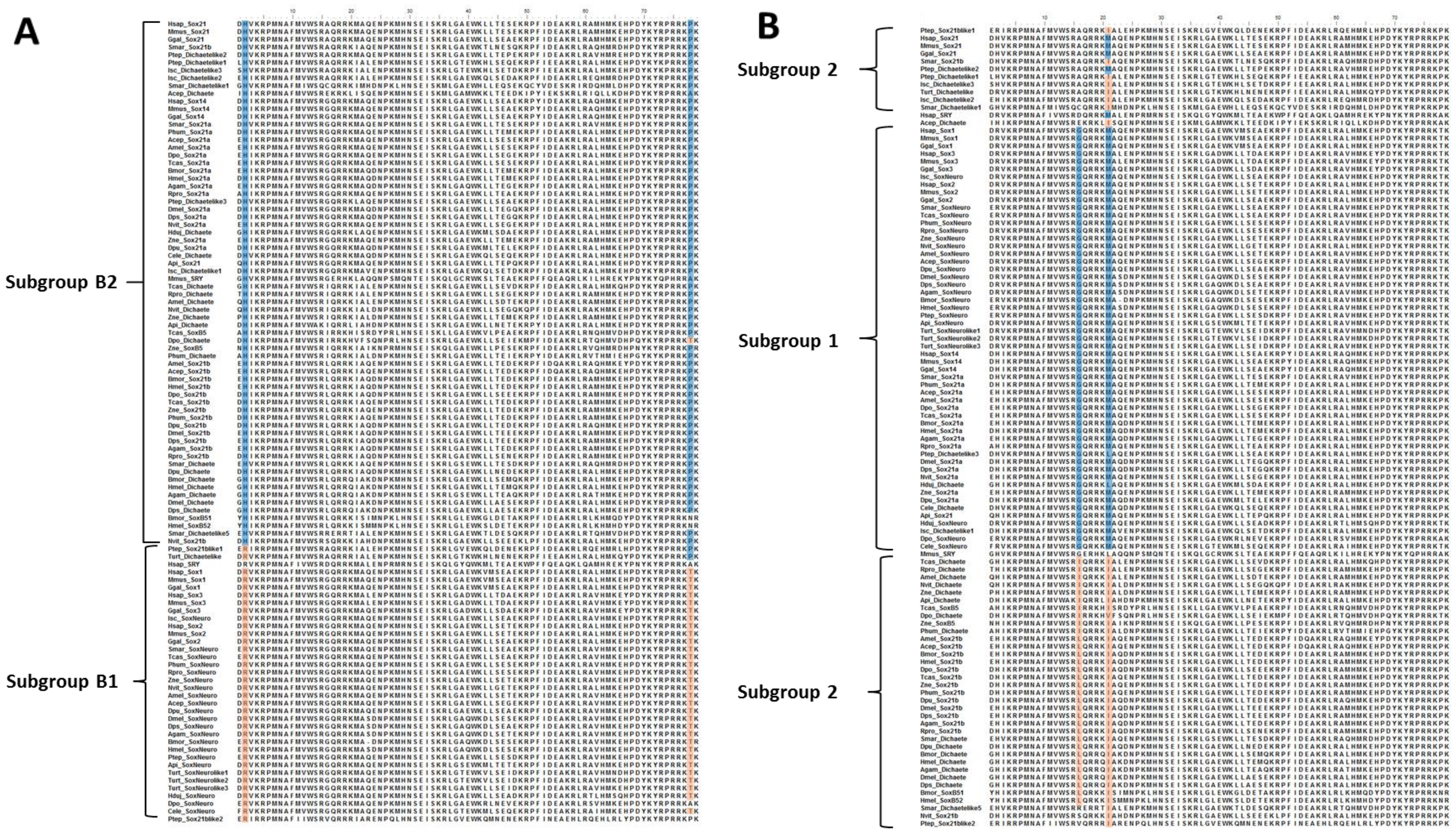


Figure 3.3.2. The signature residues proposed by the Zhong (A) and McKimmie (B) models mapped onto the HMG alignment of arthropod sequences. The McKimmie model proposes positions 16 and 21 as signature residues: at position 16, the D/Sox21b class (orange) tends to possess an L/I, whereas the SoxN/Sox21a class (blue) a G; and at position 21, the D/Sox21b class possessing an I, and the SoxN/Sox21a class an M. The Zhong /Bowles model proposes signature residues at positions 2 and 78: at position 2, SoxB1 (orange) possess an R, where SoxB2 (blue) proteins possess a H; and at position 78, SoxB1 a T, and B2 a P.

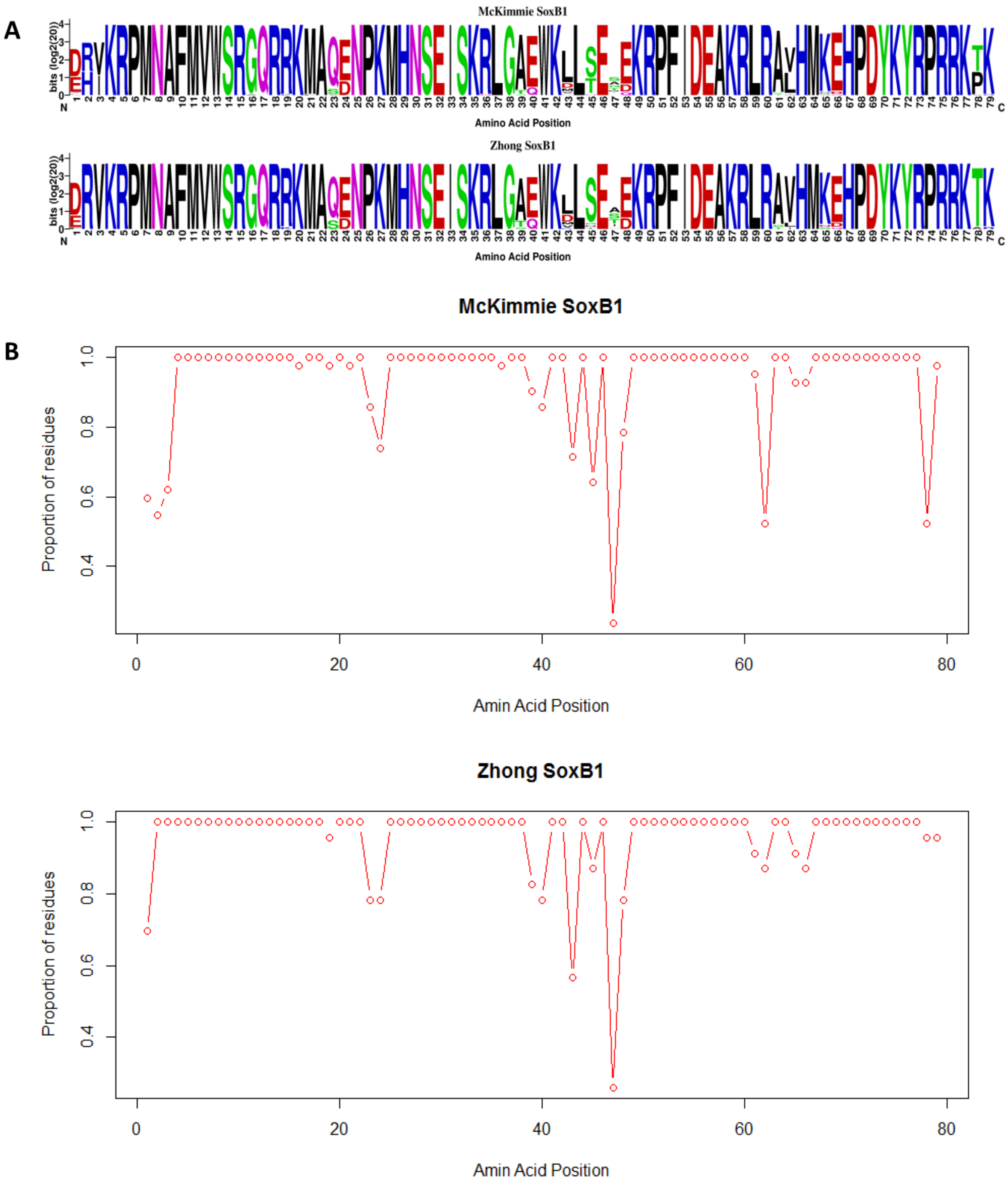


Figure 3.3.3. Consensus sequences and their proportional support for the subgroup 1 proposed by the McKimmie and Zhong models. (A) Consensus HMG sequences generated according to the McKimmie subgroup 1 of proteins, and the Zhong subgroup B1. (B) Proportional support at each amino acid residue for the consensus sequences shown in (A). The Consensus logos were generated using WebLogo, developed by Crooks *et al.* (2004).

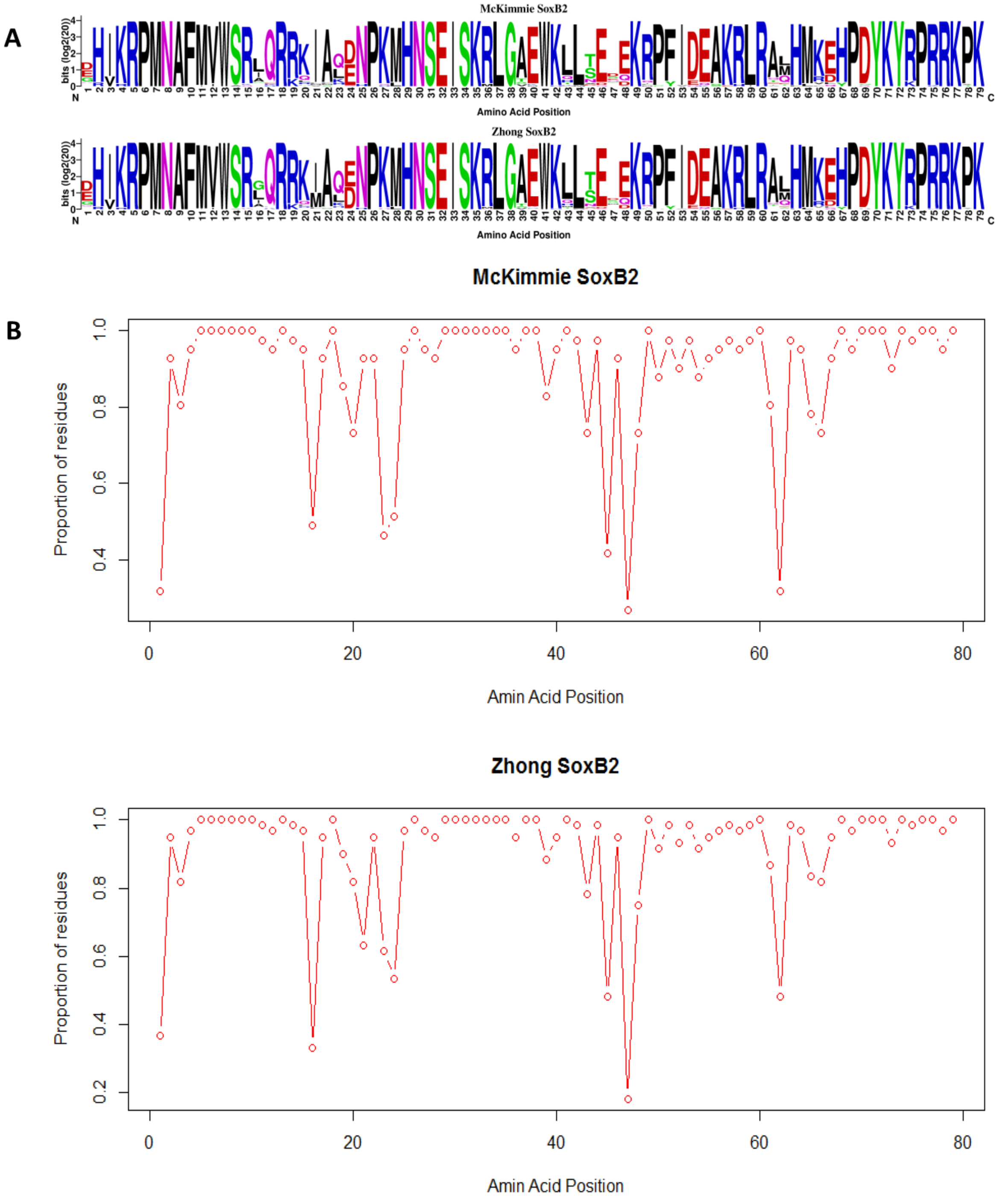


Figure 3.3.4. Consensus sequences and their proportional support for the subgroup 2 proposed by the McKimmie and Zhong models. (A) Consensus HMG sequences generated according to the McKimmie subgroup 1 of proteins, and the Zhong subgroup B2 (B) Proportional support at each amino acid residue for the consensus sequences shown in (A). The Consensus logos were generated using WebLogo, developed by Crooks *et al.* (2004).

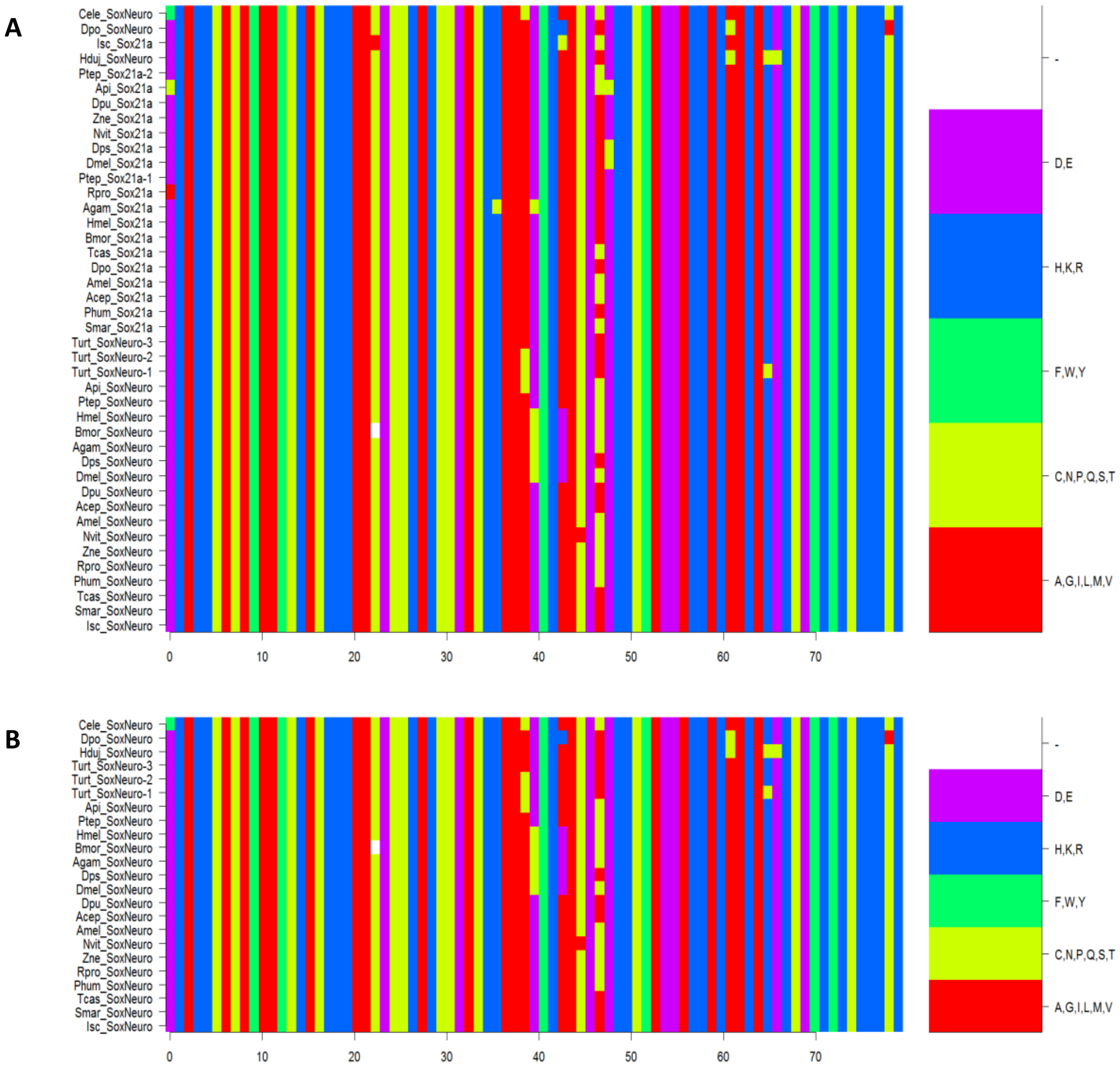


Figure 3.3.5. R-group representations for McKimmie's subgroup 1 (A) and Zhong subgroup B1 (B). *Red* = Nonpolar, aliphatic R group (A, G, I, L, M, V); *Yellow* = Polar, uncharged R group (C, N, P, Q, S, T); *Green* = Aromatic R group (F, W, Y); *Blue* = Positively charged R group (H, K, R), *Purple* = Negatively charged R group (D, E), and *White* = gap (-).

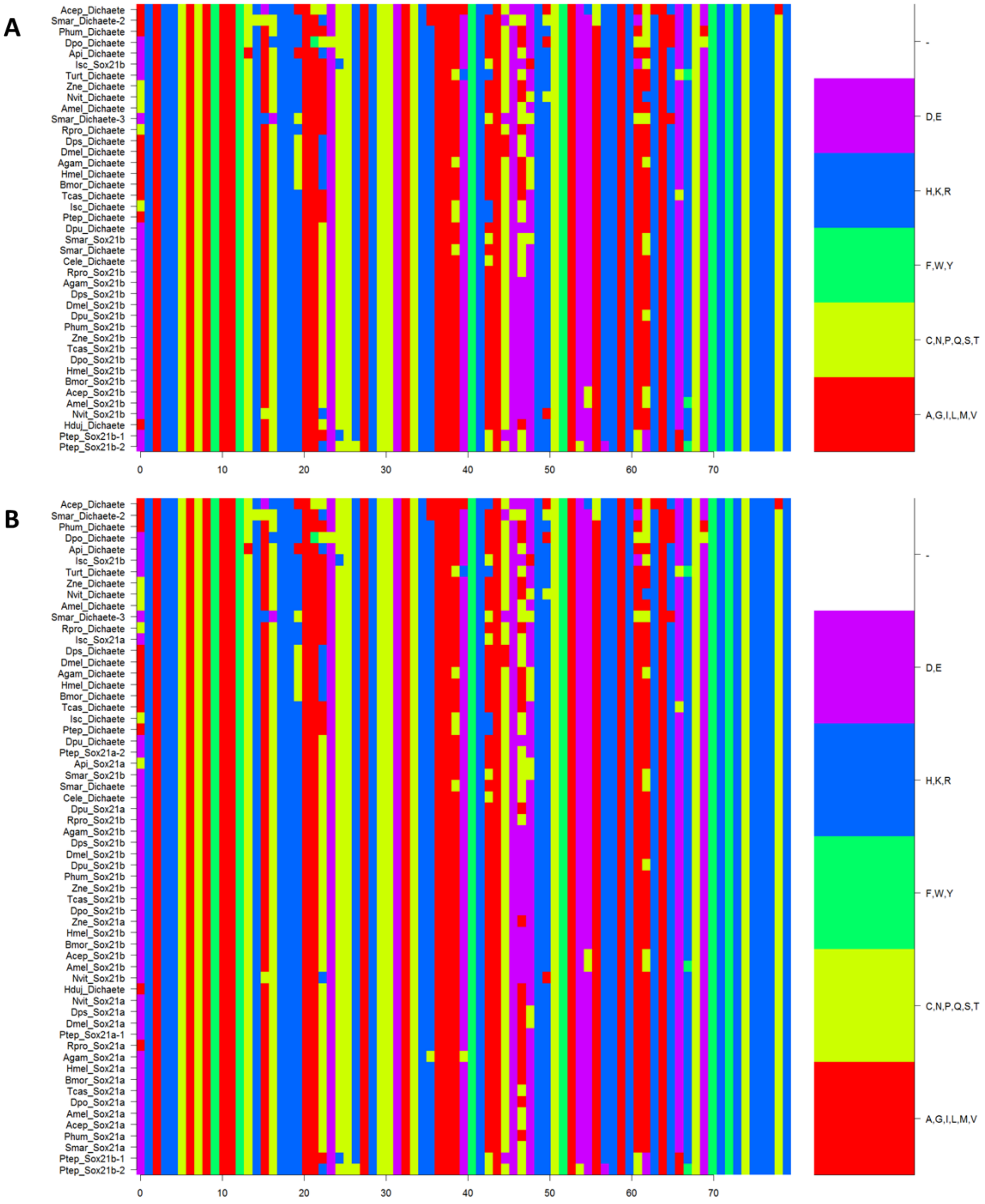


Figure 3.3.6. R-group representations for McKimmie's subgroup 2 (A) and Zhong subgroup B2 (B). *Red* = Nonpolar, aliphatic R group (A, G, I, L, M, V); *Yellow* = Polar, uncharged R group (C, N, P, Q, S, T); *Green* = Aromatic R group (F, W, Y); *Blue* = Positively charged R group (H, K, R), *Purple* = Negatively charged R group (D, E), and *White* = gap (-).

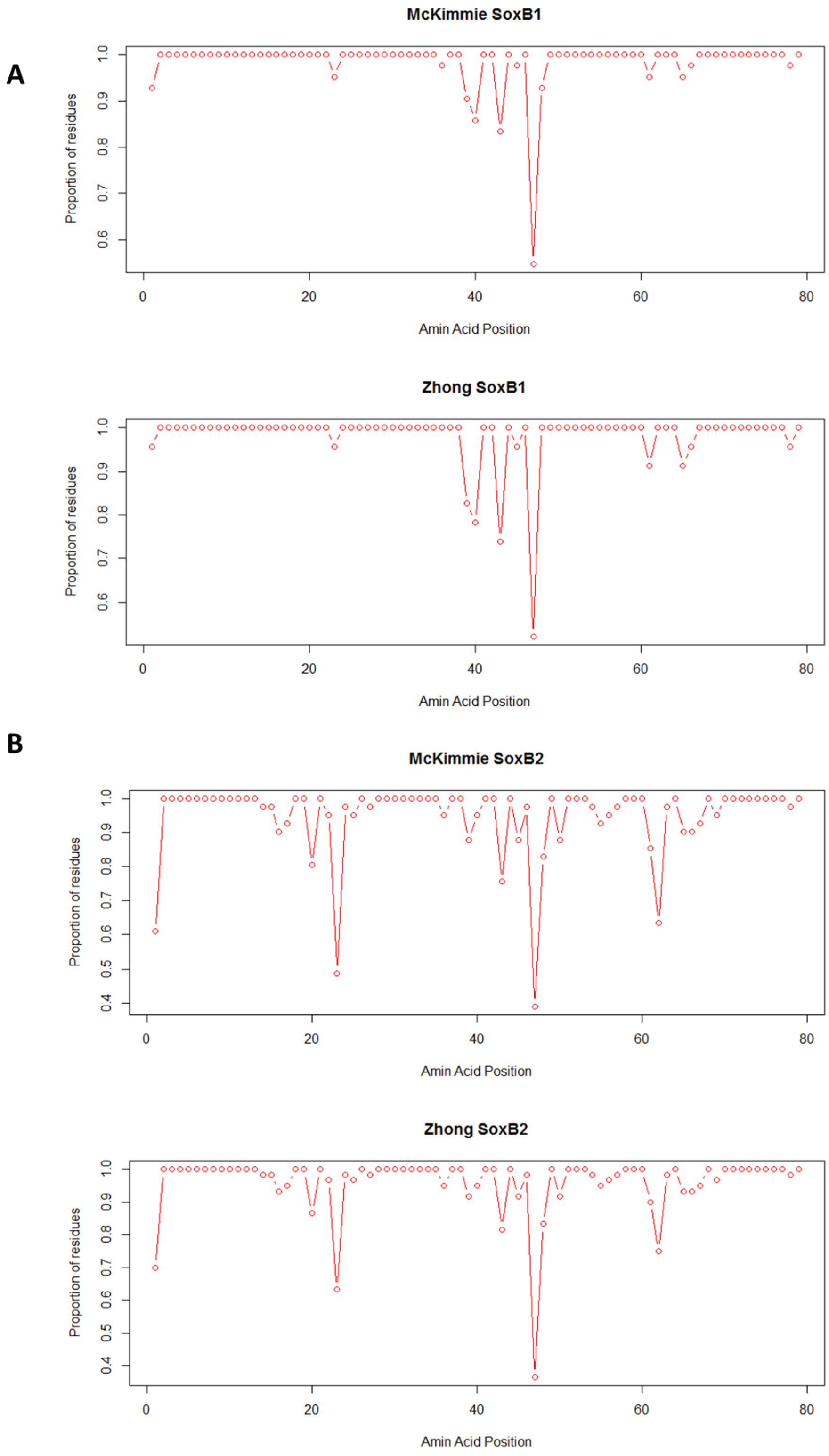


Figure 3.3.7. Proportional support at each amino acid residue for the consensus R-group representations for the McKimmie and Zhong subgroups. (A) Proportional support for subgroups 1 proposed by McKimmie and Zhong; (B) proportional support for subgroups 2 proposed by McKimmie and Zhong.

3.4 Extra-HMG-box domains of SoxB

To date, the majority of work on SoxB evolution has focused on the HMG domain of the proteins (Bowles *et al.*, 2000; McKimmie *et al.*, 2005; Wilson & Dearden, 2008; Zhong *et al.,* 2011), as this is the most highly conserved region, with the rest of the protein assumed to be poorly alignable. In this analysis, where possible, I also extracted 20 amino acids both upstream and downstream of the HMG domain, in order to search for additional conserved domains that might have been previously missed.

One such domain that, to the best of my knowledge, has been so far undocumented, is an extra-HMG domain, C-terminal to the HMG domain, which is found exclusively within insect Sox21b proteins. The *Sox21b* genes encode for the consensus peptide sequence **EGYPYSIPYPSVPMDALRAG** (positions 122-163) (Figure 3.4.1), with little variation, suggesting a conserved function of this region which is presently unknown. This domain does not appear within the Sox21b proteins of the non-insect arthropods analysed here, nor any of the vertebrate proteins, suggesting that it is a synapomorphy unique to the last common ancestor of insects. Somewhat peculiarly, both the Sox21a and Sox21b proteins of *Zootermopsis nevadensis* contain this conserved region. This implies that *Z. nevadensis*'s *Sox21a* gene might have undergone recent gene conversion, whereby a portion of the sequence of *Sox21b* has replaced a homologous portion of *Sox21a*.

To explore this further, representative species with more robust genome annotations were selected for whole protein queries to identify additional conserved domains. 12 species were chosen: five insect species, three non-insect arthropods, one non-arthropod invertebrate, and three vertebrates. In total 53 protein sequences were queried for conserved domains, using the NCBI BatchConservedDomain (BatchCD) tool (Marchler-Bauer *et al.*, 2017); the results are shown in Figure 3.4.2. As expected, the HMG-box domain is present in all the protein sequences. For the vertebrate proteins, there is remarkable preservation of the position and organisation of the HMG domain; and for Sox21a and Sox21b, within the invertebrates there appears to be a high degree of conservation also. However, for the Dichaete and SoxNeuro proteins, the HMG domains appear to be notably more variable in position. It also seems that for the majority of SoxB proteins, the HMG domain appears in the first third of the peptide sequence. The absence of any aforementioned Sox21b domain within the insects represented (discussed above) confirms that this domain has been thus far uncharacterised.

Perhaps more interesting, however, is the presence of another extra-HMG domain immediately to the HMG domain's C-terminus, 'SOXp'. This region putatively possesses two conserved consensus motifs: **KKDKY** and **LPG** (Gao *et al.*, 2013; Gao *et al.*, 2015a; Gao *et al.*, 2015b). Vertebrate Sox14, Sox21, Sox2 and Sox3 possess the putative domain, but Sox1 does not. For the invertebrates, the SOXp domain is present in SoxNeuro in all species except for *Drosophila melanogaster* and *Apis mellifera*, and the Sox21a protein of *Strigamia maritima*.

Remarkably little appears to have been published on the SOXp domain, with the protein family (id: pfam12336) listing just one citation; a study demonstrating the role of mouse Sox2 and POU proteins in upregulating the neural promoter *Nestin* via enhancer interaction (Tanaka *et al.,* 2004). However, this study does not mention or discuss any extra-HMG domain, so whether or not this is a truly separate domain may perhaps be questionable. Its absence in all Dichaete-homologue proteins and vertebrate Sox1 suggests that the peptide sequence is certainly conserved and exclusive, yet the function of this domain remains elusive.

However, what is interesting is the presence of the SOXp region within the Sox21a protein of *Strigamia maritima*. Inspection of the relevant portion of *S. maritima*'s Sox21a protein reveals the sequence **KKDRY** and **LPC** – a highly orthologous match. One might be tempted to conclude that this is an example of convergent evolution, however, upon inspection of all pan-arthropod Sox21a proteins, most possess an orthologous sequence exhibiting an insertion for the first portion of the SOXp domain, **KKE-KF**, which is strikingly similar to the SOXp domain of arthropod SoxNeuros, **KKDKY** (Figure 3.4.3). The presence of this domain in arthropod SoxNeuro and Sox21a proteins strongly supports the McKimmie model of these two genes being more closely related than *Dichaete* and *Sox21b*.

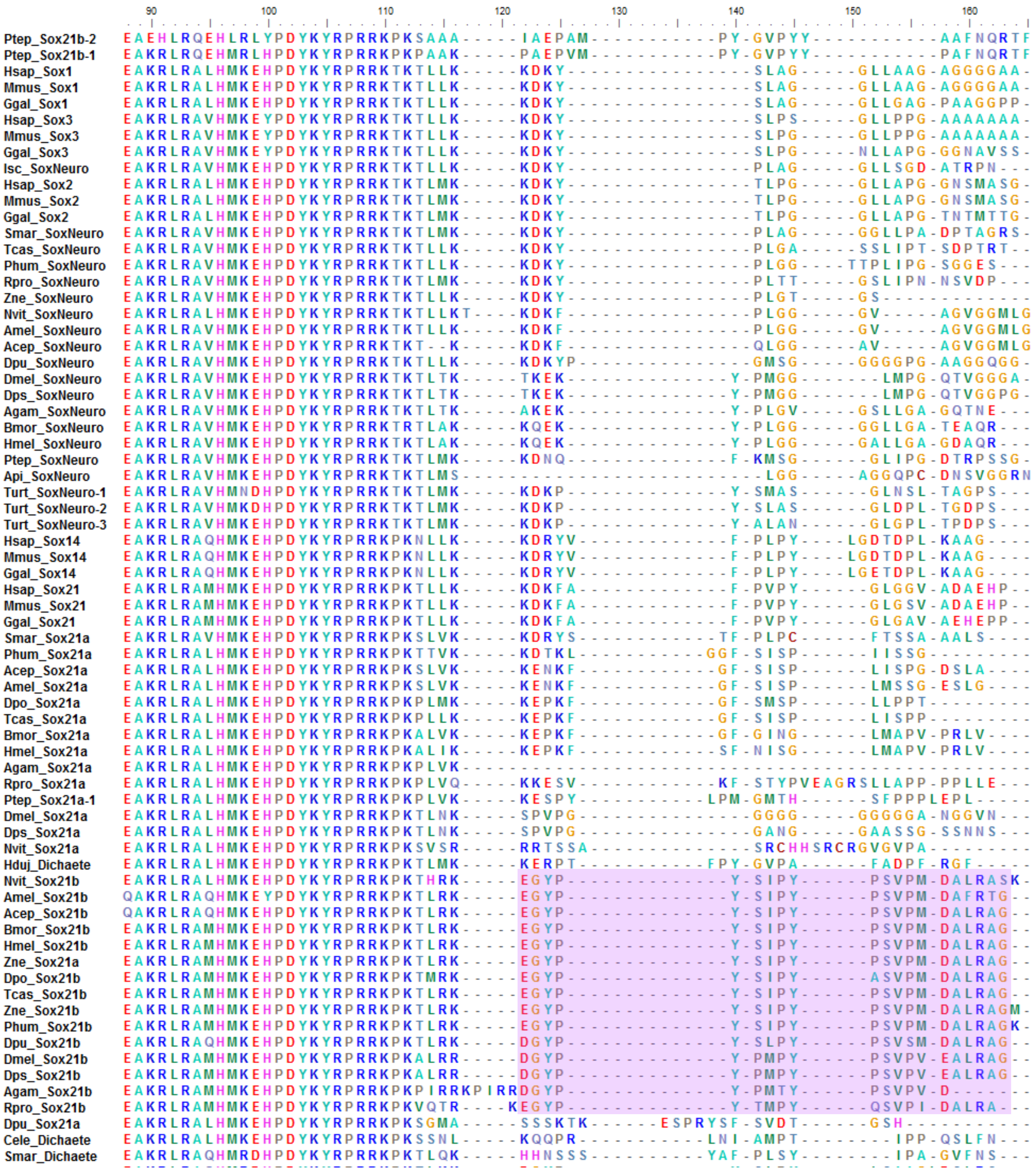


Figure 3.4.1. Zoomed in portion of the MAFFT alignment shown in Figure 3.3.1, from positions 88-165. Highlighted within the purple box is an extra-HMG domain which is exclusive to insect Sox21b proteins.

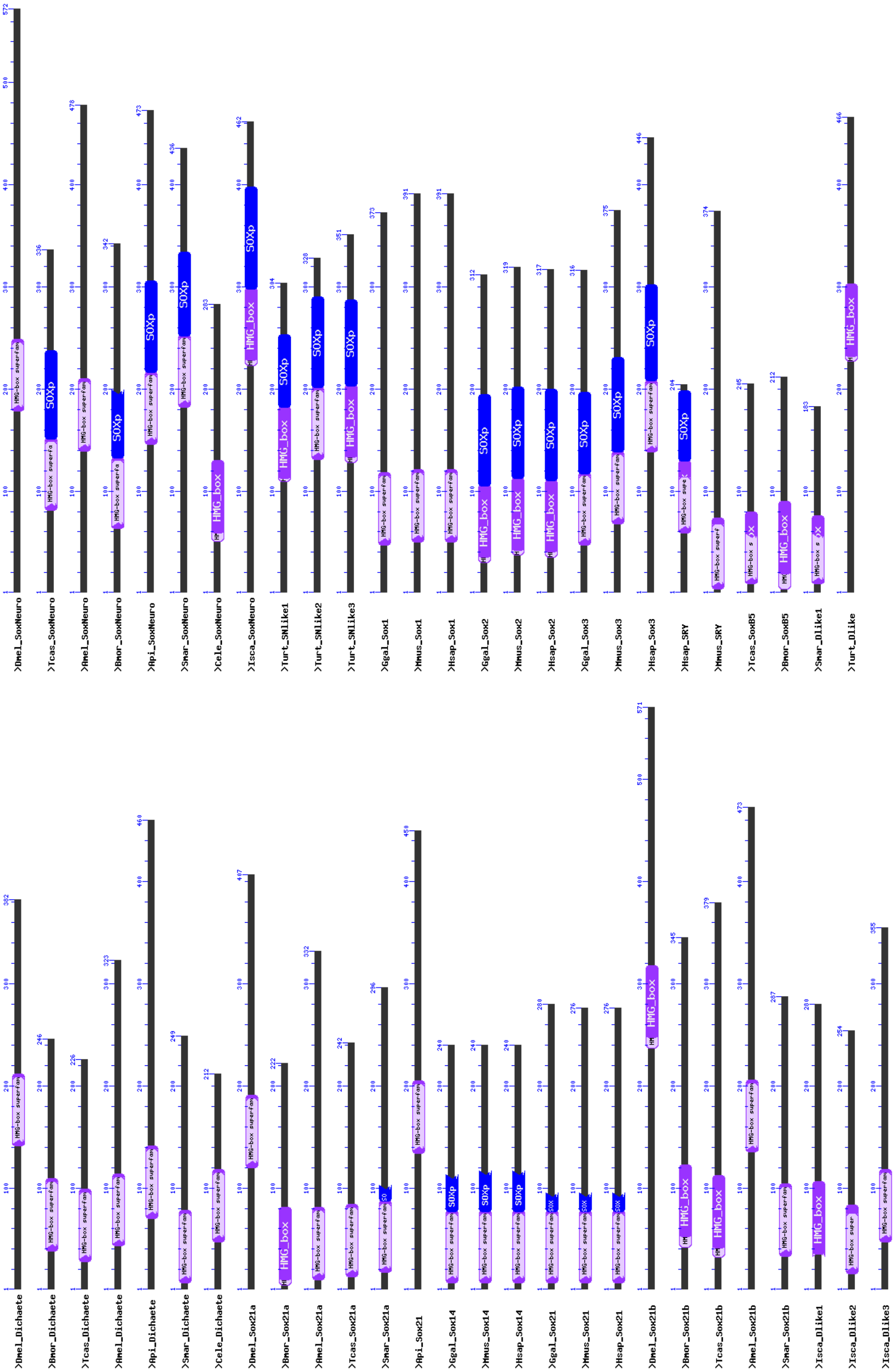


Figure 3.4.2. Conserved domains of 53 group B proteins from 12 species. In purple/pink, the HMG-box is identified, and in blue, an additional "SOXp" domain is identified.

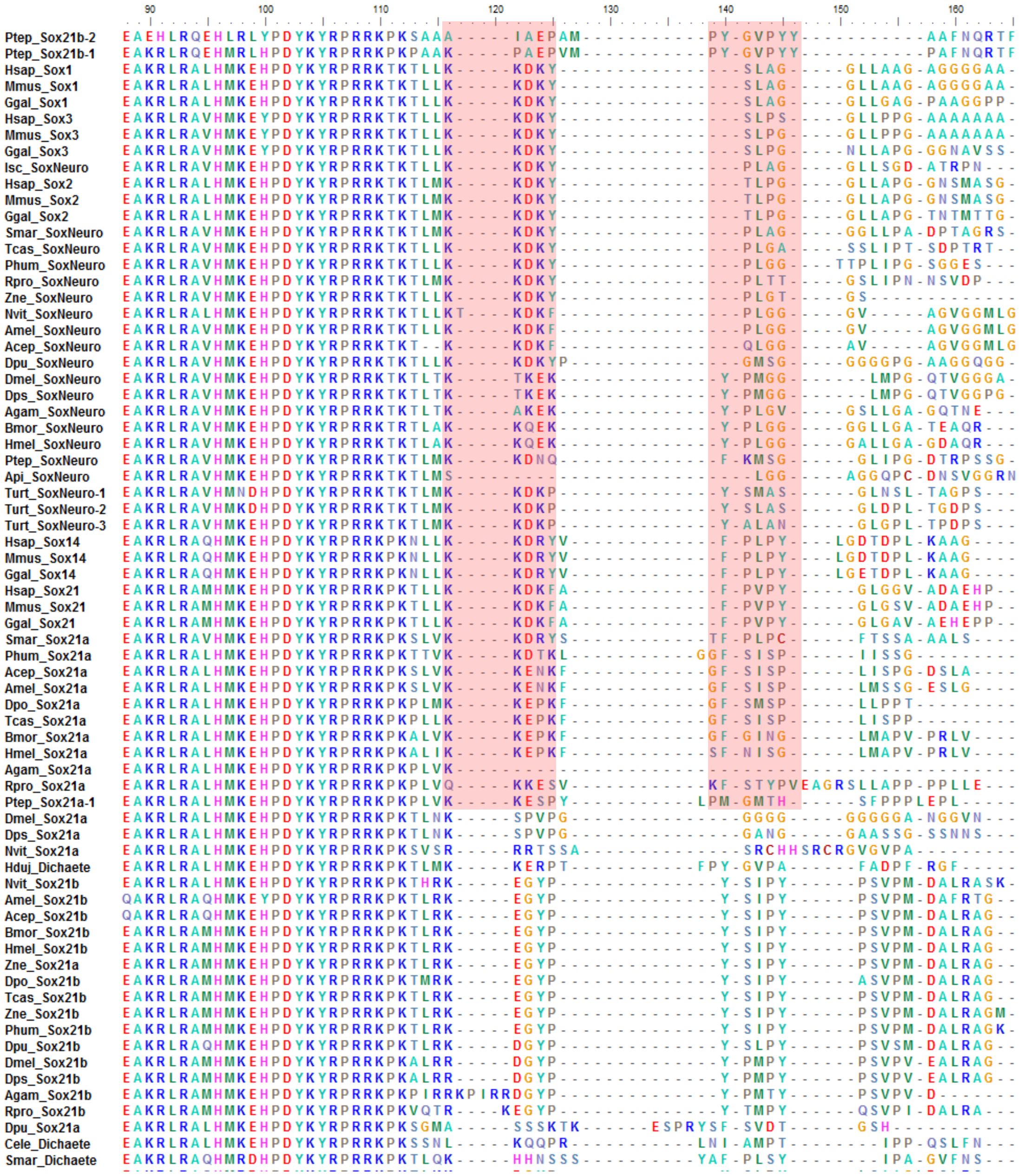


Figure 3.4.3. Zoomed in portion of the MAFFT alignment shown in Figure 3.3.1, from positions 88-165. Highlighted within the red boxes is the extra-HMG domain 'SOXp' which is found in vertebrate Sox14, Sox21, Sox2 and Sox3 and invertebrate SoxNeuro and Sox21a.

3.5 Phylogenetic Relationships of SoxB

SoxB genes were plotted against an established phylogenetic tree for the Ecdysozoa and Insecta (Figure 3.5.1). Here, it is difficult to resolve the expansion of invertebrate SoxB as all invertebrates analysed have either a single copy of a *Dichaete-homologue*, or at least 3 copies (with the exception of *Tetranychus urticae*). All species analysed possess a single copy of *SoxNeuro-homologue*, except *T. urticae* which has 3. The *Dichaete* conserved gene neighbourhood (CGN) that exists in *D. melanogaster* (comprising *Dichaete, Sox21a,* and *Sox21b*, see Figure 3.5.2), exists at least as far back as *Daphnia pulex* and *Strigamia maritima* with the same chromosomal arrangement. It may also be present in the other arthropods analysed here, however the genome assemblies are yet to be sufficiently assembled to allow examination of chromosomal clustering behaviours of these genes.

However, the evidence from these data does contradict McKimmie's model, which suggests that there were three SoxB genes present at the deuterostome/protostome split (*Dichaete*, *Sox21a*, and *SoxNeuro*): *C. elegans* and *H. dujardini* only possess two SoxB each. Instead, these data support the Zhong model which proposes two ancestral SoxB genes, SoxB1 and SoxB2. This may, of course, be an example of gene loss, and further examination of non-arthropod protostomes may support the McKimmie model; yet the most parsimonious conclusion from this data supports just two SoxB existing at the deuterostome/ protostome split.

Next, Maximum Likelihood (ML) trees were generated using the PhyML package (Guindon *et al.*, 2010) with 100 bootstraps under the WAG substitution model (Whelan & Goldman, 2001). Trees were generated with amino acid sequences for just insect taxa (Figure 3.5.3) and all taxa (Figure 3.5.4).

These phylogenetic trees appear to provide conflicting support for each model. The ML tree generated using just insect sequences shows a distinct subgrouping of SoxNeuro and Sox21a (with bootstrap support of 91%) (Figure 3.5.3) and the clustering of Sox21b and Dichaete. This is evocative of the Wilson & Dearden tree in 3.1.2B and supports of the McKimmie model of arthropod SoxB expansion.

In contrast, the ML tree generated for all taxa (Figure 3.5.4) supports the Zhong model: the Dichaete-homologues of arthropods cluster with the vertebrate B2 proteins Sox14 and Sox21 (with 87% bootstrap support for the monophyletic clade), and arthropod SoxNeuro proteins cluster closely with vertebrate B1 proteins, Sox1, Sox2, and Sox3. The phylogenetic tree inclusive of all species is likely the most accurate, simply by possessing an increased sample size. Consequently, these data imply phylogenetic support for the Zhong model of divergence, where vertebrate subgroups B1 and B2 also apply to insects (and, as demonstrated here, the wider arthropods).

Within the insect tree, the branches appear to suggest SoxB5 to be the ancestral group giving rise to Sox21b, which in turn gives rise to Dichaete and Sox21a, and Sox21a giving rise to SoxNeuro. However, within the tree with all Bilateria analysed, the arthropod phylogeny appears to suggest Sox21a as the parent group of the SoxB2 giving rise to a paraphyletic Dichaete, Sox21b, and SoxB5 subclade.

Within the insect ML tree, the Sox21b proteins appear to be monophyletic with the exception of the Sox21a protein of *Zootermopsis nevadensis*, which is nested within the Sox21b cluster, albeit with only moderate bootstrap support (63%). This is hardly surprising, as the *Sox21a* gene of *Zootermopsis nevadensis* appears to have undergone some degree of gene conversion, as discussed in Section 3.4; this appears to be an example of concerted evolution (Liao, 1999).

The SoxB5 proteins, while not conforming to monophyletic principles, do appear to cluster together within the Bilaterian tree with moderate support (63%), but this support increases to 72% within the holometabolous insects represented here (Coleoptera and Lepidoptera). Both *Dendroctonus ponderosae* and the more basally branching *Pediculus humanus* appear to have once possessed a fourth *Dichaete-homologue* in their genome, now a pseudogene (represented in their lineages in Figure 3.5.1). The most parsimonious explanation is, therefore, that *SoxB5* is an ancestral gene to the insects, at least as far back as the Isoptera branch, which has been independently lost multiple times. What is peculiar, however, is the SoxB5 gene of *Tribolium castaneum* and the *Dichaete* gene of *Dendroctonus ponderosae* clustering together with high bootstrap support (92%). The *Dichaete* gene of *D. ponderosae* possesses an intron, which is most unusual, being the only example of an insect *Dichaete* gene to do so. The *SoxB5* gene of *T. castaneum* also possesses an intron, although not at the same site. With *D. ponderosae* showing evidence of once having a fourth *Dichaete-homologue* pseudogene, one might speculate that perhaps during the early evolution and sub-functionalisation of *SoxB5* within the insects, the *SoxB5* gene of *D. ponderosae* continued to act redundantly with *Dichaete*, and consequently in this lineage it was *Dichaete* that decayed and ultimately lost, not *SoxB5*.

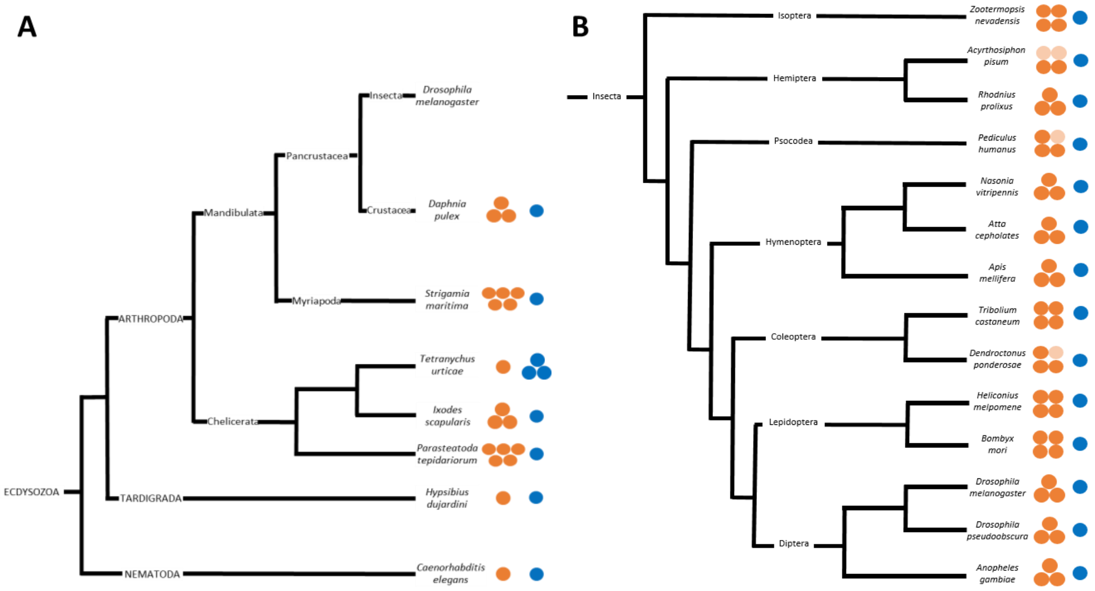


Figure 3.5.1. SoxB genes identified plotted against TimeTree (Hedges et al., 2015) established cladograms for (A) Ecdysozoa and (B) Insecta. Dichaete-homologous genes are shown in orange, and SoxNeuro-homologous genes in blue. Gene loss events are shown in light orange.

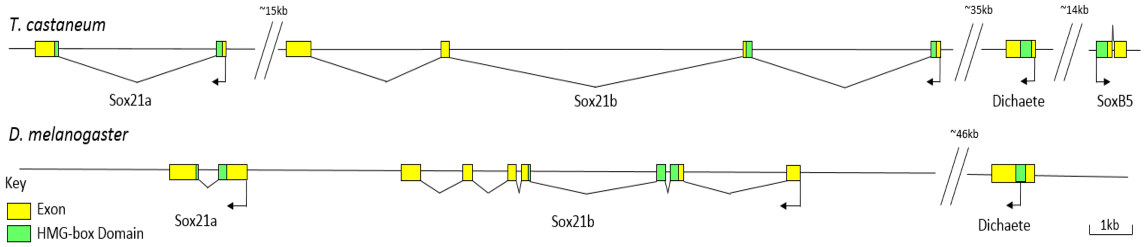


Figure 3.5.2. Gene models for the '*Dichaete* conserved gene neighbourhood (CGN)' in *Tribolium castaneum* and *Drosophila melanogaster*. For both species, *Sox21b* and *Sox21a* contain one intron running through the HMG domain, whereas *Dichaete* is intronless. *T. castaneum* contains a fourth gene in this cluster, *SoxB5*, located on the positive strand.

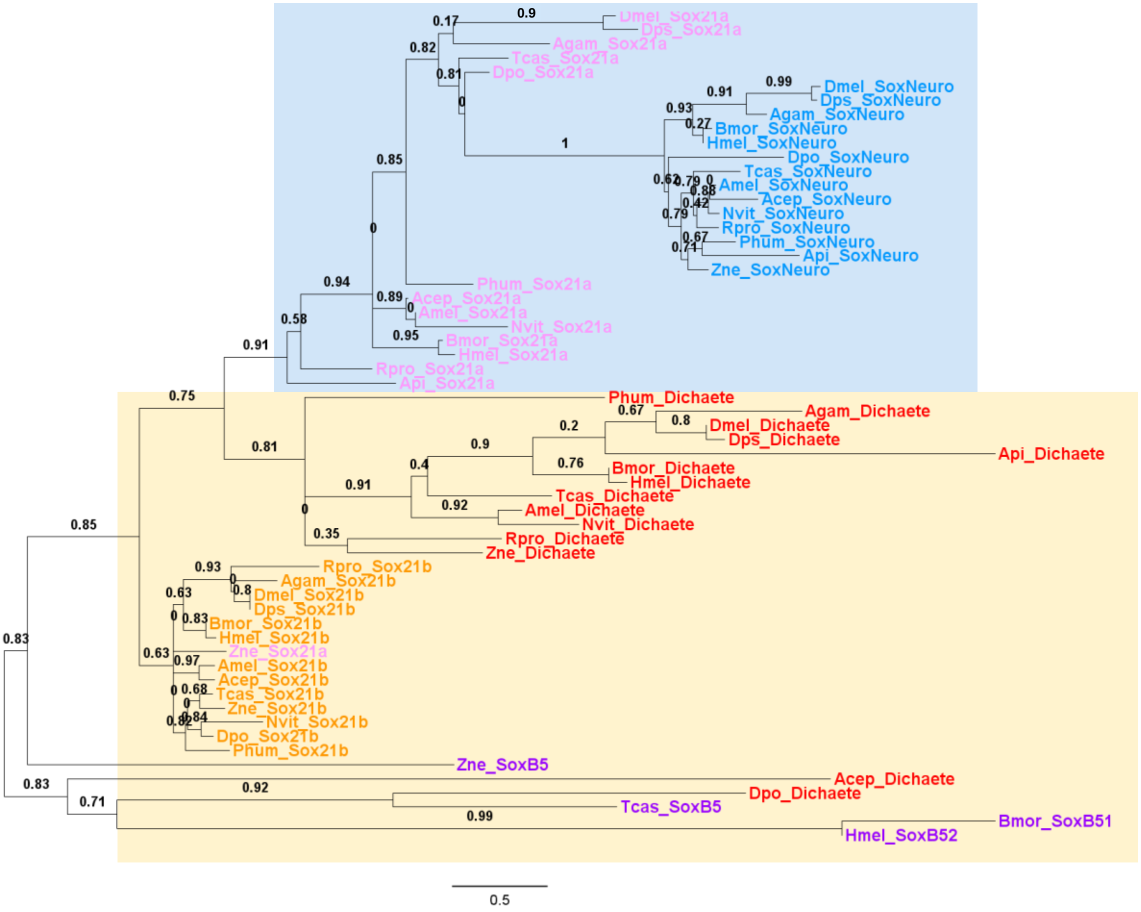


Figure 3.5.3. Unrooted Maximum Likelihood Tree of 59 SoxB sequences from 14 insect taxa. SoxNeuro orthologues are coloured blue, Dichaete orthologues red; Sox21b orthologues orange, Sox21a orthologues pink, and SoxB5 orthologues purple. Bootstrap support values are displayed above their respective branches, and the scale bar corresponds to branch length. The subgroupings proposed by the McKimmie model overlaid: orange represents the Dichaete and Sox21b Subgroup, and blue the SoxNeuro and Sox21a subgroup.

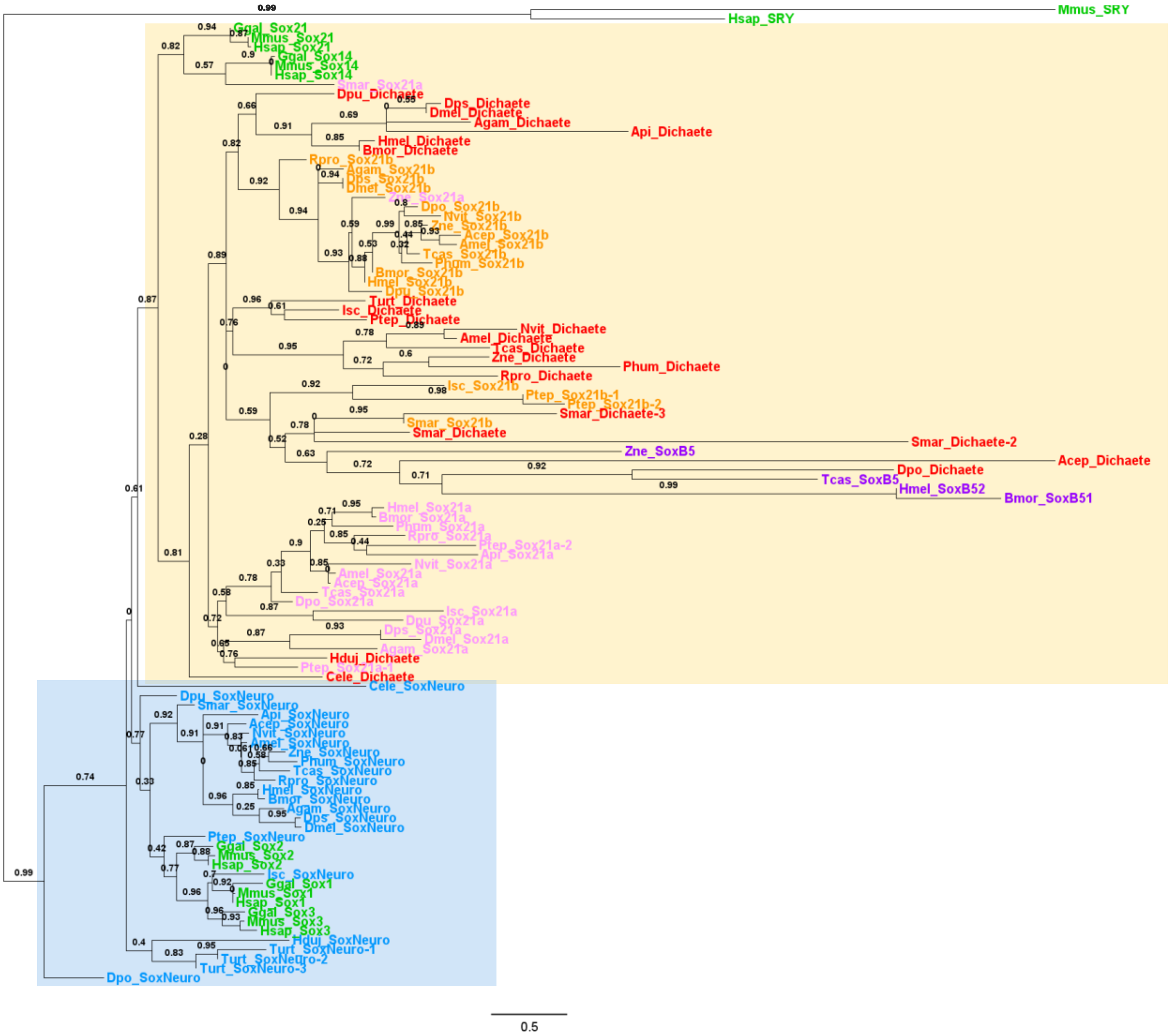


Figure 3.5.4. Unrooted Maximum Likelihood Tree of 104 SoxB sequences from 24 bilaterian taxa. SoxNeuro orthologues are coloured blue, Dichaete orthologues red; Sox21b orthologues orange; Sox21a orthologues pink; SoxB5 orthologues purple; and vertebrate Sox are in green. Bootstrap support values are displayed above their respective branches, and the scale bar corresponds to branch length. The subgroupings proposed by the Zhong model overlaid: orange represents the B2 subgroup, blue the B1 subgroup, and purple for SRY as an out-group.

3.6 Discussion of Results

In this investigation, the Group B Sox genes from several new metazoan genomes have been analysed in terms of their protein alignments and phylogenetic relationships. To date, phylogenetic research has primarily focused on just the HMG domains of proteins (Bowles *et al.*, 2000; McKimmie *et al.*, 2005; Wilson & Dearden, 2008; Zhong *et al.*, 2011), within a limited selection of animal taxa (Phochanukul & Russell, 2010; Wei *et al*., 2011). The data presented here imply that the two competing models proposed by McKimmie *et al.* (2005) and Zhong *et al.* (2011) are each insufficient to explain the evolutionary subgroupings of SoxB genes fully; phylogenetic clustering of insect SoxB most closely groups *Sox21a* with *SoxNeuro*, and *Dichaete* with *Sox21b*, supporting the McKimmie model, whereas phylogenetic clustering of all species supports the Zhong model, clustering arthropod *Dichaete*, *Sox21a*, and *Sox21b* with the vertebrate B2 subgroup, and arthropod *SoxNeuro* with the vertebrate B1 subgroup.

The evolutionary emergence of SoxB genes can also be plotted against a cladogram of the species analysed: with single distinct B1 and B2 genes present for both *C. elegans* and *H. dujardini*, these data support the Zhong model of SoxB phylogeny, parsimoniously suggesting these to be ancestral to the Ecdysozoa (Figure 3.6.1A). Expansion of the Dichaete-like homologues, *Dichaete*, *Sox21a*, and *Sox21b*, must have occurred very early within the arthropods, with all arthropods analysed possessing at least 3 Dichaete-like homologues, except for *T. urticae*, which is highly atypical in its SoxB distribution and therefore likely to be an anomaly.

Analysis of amino acid conservation at signature residues of the HMG domain, and across the entire HMG domain, implies further ambiguity in the groupings. However, extra-HMG domain residues have been identified for the first time within arthropods in this study, and strongly support the McKimmie model of evolutionary divergence of arthropod SoxB, with SoxNeuro and Sox21a both possessing a putative SOXp domain downstream of the HMG-domain, implying that they are likely to be most closely related. This is unlikely to be explained by convergent evolution as so many taxa evolving this domain independently is not parsimonious.

In light of these data, both models have proven unsatisfactory. Instead, I propose a new model for the phylogenetic emergence and divergence of SoxB genes within the Bilateria that attempts to resolve the issues discussed in this chapter. This new model is presented in Figure 3.6.2 and combines elements from both the McKimmie model and the Zhong model. In this model, Zhong's proposal for the existence of subgroups B1 and B2 prior to the deuterostome/ protostome split is retained, as the phylogenetic support for this scenario is strong in the data described above. However, this new model posits that the SOXp domain is ancestral to the bilaterian SoxB as it is present in both vertebrates and arthropods. Within the vertebrates, expansion occurs essentially as described in the Zhong model, through a combination of both whole genome duplications and tandem duplications. However, within the arthropods, the model proposes that instead of *Dichaete* being the ancestral gene within the *Dichaete* conserved gene neighbourhood, it is in fact *Sox21a*. This model resolves the paraphyly encountered in the two previous models and explains the tendency for *Sox21a* to be grouped with *SoxNeuro* in the analyses described here and elsewhere. It is also supported by the phylogeny shown in Figure 3.5.3, with the *Sox21a* genes being more basal than *Dichaete*, *Sox21b*, and *SoxB5*. Also apparent is the '*Dichaete*' gene of *H. dujardini* and *C. elegans* clustering most closely with the *Sox21a* branches, implying that these are in fact *Sox21a* genes and not *Dichaete*. Further evidence to support this is the presence of an intron in the HMG encoding region of these respective genes just 15bp downstream of the intron found in all *Sox21a* genes of the insects.

Arthropod *Sox21a* is likely to be most closely related to vertebrate *Sox21*, and not *Sox14*, given the position of a Valine (V) in the signature residue at position 2 in vertebrate Sox21 and in the Sox21a protein of non-insect arthropods *S. maritima* and *P. tepidariorum*. Indeed, McKimmie *et al*. (2005) first proposed the name *Sox21a* for this gene in *Drosophila* because BLAST reports indicated its orthology with vertebrate *Sox21*. After *Sox21a* duplicated to generate *Dichaete* within arthropods, *Dichaete* lost the SOXp domain, and then through a tandem duplication produced *Sox21b*, explaining the tendency for *Dichaete* and *Sox21b* to form a subgroup within arthropods. Within the insects, *Dichaete* duplicates one more time to generate *SoxB5*, which is retained in several insect taxa but lost in many others (see Figure 3.6.1B).

Together, these data have shed new light on the evolution of Group B Sox genes within the Bilateria, uncovering a novel model for SoxB expansion. This new model resolves the two conflicting preceding models by reconciling aspects of each, and was primarily achieved by expanding the search beyond the HMG domain of Sox and examining their intron structures. As increasing numbers of metazoan genomes become available in the public domain, Sox genes can be further characterised and categorised, and the model proposed here can be further examined. However, this model best explains the available data for SoxB evolution and expansion, representing the most complete explanation for arthropod SoxB phylogeny to date.

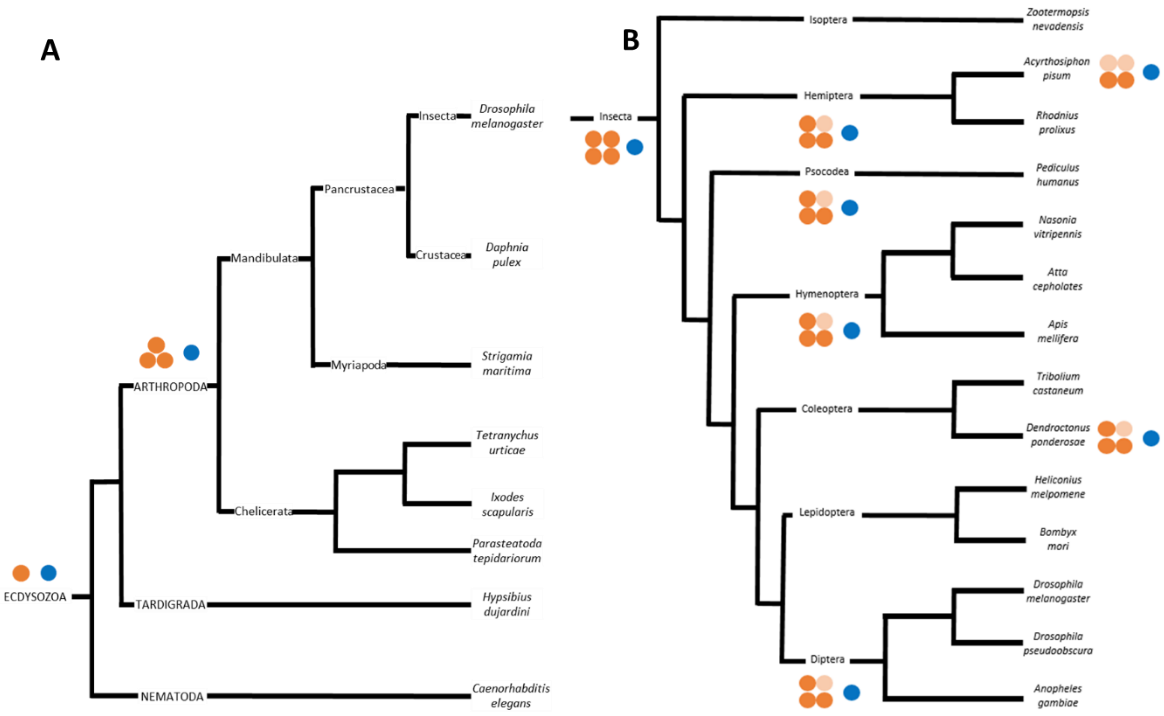


Figure 3.6.1. The most parsimonious SoxB emergence events, plotted on Ecdysozoa and Insecta cladograms. (A) Two SoxB genes, B1 (orange) and B2 (blue), are ancestral to the Ecdysozoa. Within the arthropods, two additional B2 genes appear early in arthropod evolution. (B) Within the insects, another B2 gene appears, meaning there are four B2 genes and one B1 gene ancestral to Insecta. Within several insect lineages, B2 genes are lost (light orange).

Table 3.6.1. Updated Group B Sox proteins identified in 24 metazoan species, listed in descending order according to their relatedness to *Drosophila melanogaster* (top). The abbreviations used in figures for each species are shown in brackets following the Latin binomials.

Species		Group B Sox identified					
Diptera	*Drosophila melanogaster (Dmel)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Drosophila pseudoobscura (Dps)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Anopheles gambiae (Agam)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Lepidoptera	*Bombyx mori (Bmor)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
	Heliconius melpomene (Hmel)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
Coleoptera	*Tribolium castaneum (Tcas)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
	Dendroctonus ponderosae (Dpo)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Hymenoptera	*Apis mellifera (Amel)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Atta cepholates (Acep)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Nasonia vitripennis (Nvit)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Psocodea	*Pediculus humanus (Phum)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Hemiptera	*Acyrthosiphon pisum (Apis)*	*SoxNeuro*	*Dichaete*	*Sox21a*			
	Rhodnius prolixus (Rpro)	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Isoptera	*Zootermopsis nevadensis (Zne)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*SoxB5*	
Crustacea	*Daphnia pulex (Dpu)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
Myriapoda	*Strigamia maritima (Smar)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*	*Dichaete-2*	*Dichaete-3*
Chelicerata	*Ixodes scapularis (Isc)*	*SoxNeuro*	*Dichaete*	*Sox21a*	*Sox21b*		
	Tetranychus urticae (Turt)	*SoxNeuro-1*	*SoxNeuro-2*	*SoxNeuro-3*	*Dichaete*		
	Parasteatoda tepidariorum (Ptep)	*SoxNeuro*	*Dichaete*	*Sox21a-1*	*Sox21a-2*	*Sox21b-1*	*Sox21b-2*
Tardigrada	*Hypsibius dujardini (Hduj)*	*SoxNeuro*	*Sox21a*				
Nematoda	*Caenorhabditis elegans (Cele)*	*SoxNeuro*	*Sox21a*				
Vertebrata	*Gallus gallus (Ggal)*	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	
	Mus musculus (Mmus)	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	*SRY*
	Homo sapiens (Hsap)	*Sox1*	*Sox2*	*Sox3*	*Sox14*	*Sox21*	*SRY*

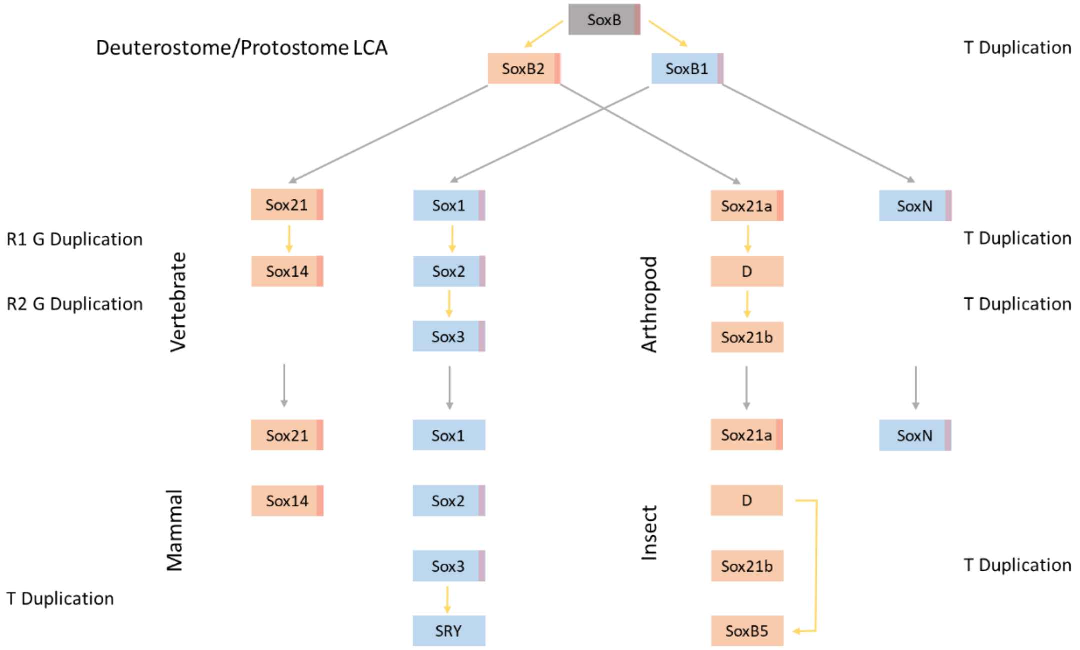


Figure 3.6.2. Proposed model for SoxB evolution within Bilateria. A single SoxB gene, ancestral to the Bilateria, undergoes a tandem duplication to produce SoxB1 and SoxB2. This SoxB gene contains the signature HMG-box encoding domain, as well as the extra-HMG SOXp domain (shown in red). SoxB1 within the protostomes remains mostly static, however the B2 gene, *Sox21a*, undergoes 3 rounds of tandem duplication, giving rise to the expansion of SoxB we observe today. Firstly, *Sox21a* duplicates to produce *Dichaete*, and then *Dichaete* duplicates to produce *Sox21b*, within the arthropods. Then within insects, *Dichaete* undergoes a further tandem duplication to produce *SoxB5*. Within the vertebrates, several rounds of whole genome duplications and tandem duplications gives rise to the multiple paralogues of B1 and B2 genes we can observe today, as well as the sex-determining gene, *Sry*. (SoxB1 genes are shown in blue, SoxB2 in orange; the SOXp-encoding domain is shown as a red stripe. Grey arrows link a new subclade, and yellow arrows a duplication event. G = Genome, T = Tandem).

Chapter 4

Expression Patterns of *Dichaete* and *SoxNeuro* within *Tribolium castaneum*

4.1 Motivations for research

Understanding how cells and tissues acquire distinct identities is a fundamental question driving research in developmental biology. From totipotent to pluripotent, and to eventual specialised cell types, the mechanisms driving these cellular decisions are of great importance and integral to animal development (Wolpert *et al*., 2015). Insect models present a powerful opportunity to elucidate developmental mechanisms due to their comparatively short life cycles and high fecundity, and the vast majority of work using insects in developmental biology to date has focused on the *Drosophila melanogaster* model (Bate & Martinez-Arias, 1993; Wolpert *et al*., 2015). The genetic pathways governing cell fate have been studied extensively in *Drosophila*, elucidating how cell-fate specification acts in a position-dependent manner under the control of a host of regulatory gene networks (Bate & Martinez-Arias, 1993).

One particular area of interest is the central nervous system (CNS) of insects, where it appears that many of the regulatory genes controlling CNS development are highly conserved with vertebrates (Bhat, 1999; Skeath, 1999; Wolpert *et al*., 2015). In *Drosophila*, the CNS develops from neural stem cells, called neuroblasts (NBs), which arise from the neuroectoderm, a ventrolateral ectodermal layer that forms during gastrulation (Skeath, 1999). Gene regulatory networks organise neural stem cells in a precise and tightly controlled manner within symmetrical hemisegments aligned along the ventral midline, with NBs delaminating in five successive waves from proneural cell clusters in the neuroectoderm. Segment polarity genes pattern NBs along the anterior-posterior (AP) axis into four transverse rows, and columnar genes pattern and specify neural precursors along the dorsal-ventral axis (DV) into three longitudinal columns: medial; intermediate; and lateral columns (see reviews by Bhat, 1999; Skeath, 1999). Examples of the *Drosophila* segment polarity genes are *wingless* (*wg*), *hedgehog* (*hh*), *patched* (*ptc*), *gooseberry* (*gsb*), *engrailed* (*en*), and *invected* (*inv*) (reviewed by Bhat, 1999). Columnar genes pattern the DV axis of the *Drosophila* CNS, and the genes identified thus far are *Epidermal growth factor receptor* (*Egfr*), *ventral nerve cord defective* (*vnd*), *intermediate nerve cord defective* (*ind*), and *muscle segment homeodomain* (*msh*) (reviewed by Skeath, 1999: see Figure 4.1.1), along with *Dichaete* (*D*) and *SoxNeuro* (*SoxN*) (Zhao & Skeath, 2002; Zhao *et al*., 2007).

vnd, *ind*, and *msh* were the first genes to be identified that pattern the CNS along the DV axis (Zhao *et al*., 2007). The activity of *Egfr* is required within the medial and intermediate columns prior to the first wave of NB formation and establishes *vnd* in the medial column and *ind* in the intermediate column. *vnd* expression characterises the medial column, and inhibits *ind* expression, promoting medial column fates (Skeath *et al*., 1994, Chu *et al*., 1998). *ind* expression characterises the intermediate column during the first two waves of NB formation, determining their NB fates, and inhibits *msh* (Weiss *et al*., 1998). *msh* expression identifies the lateral column during the first two waves of NB formation, yet appears to have no effect on lateral column gene expression (Buescher & Chia, 1997; Skeath, 1999). It is worth noting that the vertebrate genes *Msx*, *Gsh1*, and *Nkx2.1* and *Nkx2.2* are orthologous to *msh*, *ind*, and *vnd*, respectively. These genes pattern the vertebrate neural plate dorsoventrally, along three longitudinal columns either side of the midline, in an orthologous manner to the neuroectodermal patterning of *Drosophila* embryos (Wolpert *et al*., 2015).

The *Dichaete* and *SoxNeuro* genes act in parallel to *vnd* and *ind* during DV CNS patterning (Buescher *et al*., 2002; Overton *et al*., 2002; Zhao & Skeath, 2002; Zhao *et al*., 2007). *Dichaete* is initially expressed in a broad domain encompassing the anlage of the entire trunk region, before resolving transiently into seven transverse pair-rule stripes in the blastoderm embryo (Nambu & Nambu, 1996; Russell *et al*., 1996). Its expression later becomes confined to the midline glia, and the medial and intermediate columns of the ventral neuroectoderm throughout all waves of NB formation. *Dichaete* mutants exhibit defects in the differentiation of glial lineages within the midline (Sánchez-Soriano & Russell, 1998), yet neural phenotypes are relatively weak in the medial and intermediate columns (Nambu & Nambu, 1996; Overton *et al*., 2002). *Dichaete* has been shown to interact with *vnd* and *ind*, contributing towards the specification of cell fates (Zhao & Skeath, 2002).

Dichaete is also active during the early segmentation of the *Drosophila* embryo, with primary pair-rule genes *even-skipped*, *hairy*, and *runt* dependent on the Dichaete TF for correct expression. *Dichaete* has also been shown to be necessary for correct brain development, most notably within the neural cells of the tritocerebrum. However, strong *Dichaete* expression is found throughout the protocerebrum, deuterocerebrum, and tritocerebrum (Sánchez-Soriano & Russell, 2000). Together, these demonstrate the activity of Dichaete as an integral modulator of insect development: most notably embryonic segmentation and DV patterning in the neuroectoderm; the latter role acting in parallel with *vnd* and *ind*, and upstream and in parallel to proneural gene activity.

SoxNeuro is expressed in a pan-neuroectodermal manner throughout neurogenesis (Cremazy *et al*., 2000), across all three DV columns (Buescher *et al*., 2002; Overton *et al*., 2002).

Mutations result in severe hypoplasia in the lateral regions of the developing CNS, yet the medial column forms almost normally and intermediate neural phenotypes are less severe (Buescher *et al*., 2002; Overton *et al*., 2002). There are also severe defects in head formation. Similarly to *Dichaete*, *SoxNeuro* acts in parallel to *vnd* and *ind*, and upstream and in parallel to the proneural genes of the *ac/sc* complex (Buescher *et al*., 2002; Overton *et al*., 2002; Zhao *et al*., 2007).

Moreover, *Dichaete* and *SoxNeuro* double mutants show more severe defects than either single mutant. While *SoxNeuro* mutants show a loss of NB formation in the lateral column, within the medial and intermediate columns (where *SoxNeuro* expression overlaps with *Dichaete*) the phenotype is less severe. Double mutants, however, exhibit strong neural hypoplasia throughout the CNS, with longitudinal axons almost entirely absent. These results strongly suggest that the two genes act in a partially redundant manner (Buescher *et al*., 2002; Overton *et al*., 2002). Further genomic studies imply strong redundancy between these two TFs through common genome binding intervals in *Drosophila* species (Ferrero *et al*., 2014; Carl & Russell, 2015).

Within arthropods, columnar genes have only been characterised in a handful of species thus far, the majority of which are drosophilids. In the honeybee, *Apis mellifera*, *Am-SoxNeuro* is expressed along ventral gastrulation folds and the procephalic neurogenic region in gastrulating embryos. Post-gastrulation, *Am-SoxNeuro* expression continues in NBs arising from the neuroectoderm along the ventral midline, and strong expression is observed in the neurons of the cephalic lobes in the embryonic brain (Wilson & Dearden, 2008). No mRNA expression is detected for *Am-Dichaete* in embryos, ovaries, or adults, by *in situ* hybridisation or RT-PCR, and is consequently suggested to be a pseudogene (Wilson & Dearden, 2008). Nevertheless, in the absence of selection pressures, genes rapidly accumulate mutations and decay (Qian *et al*., 2010); thus whether *Am-Dichaete* is truly a pseudogene needs to be examined further as its open reading frame is still intact. The honeybee embryo performs germband elongation and retraction in a similar long-germ manner to *Drosophila melanogaster* (Walldorf *et al.*, 2000; Wilson *et al*., 2010). Whether this is an example of convergent or homologous evolution is unresolved, as long germband extension is a derived characteristic for Diptera (Liu & Kaufman, 2005), yet its existence in wider Hymenopteran species (Lynch *et al*., 2012) would imply paraphyly if orthologous. RT-PCR experiments for SoxB genes have also been carried out in the Lepidopteran *Bombyx mori* (Wei *et al*., 2011), however further work is required improve the spatiotemporal resolution within embryos.

The brains of different insect orders exhibit significant heterochrony (reviewed in Boyan & Williams, 2011; Dieter *et al*., 2016; Koniszewski *et al*., 2016); for example, the central complex of the brain is fully formed in the embryos of orthopteran species (Boyan & Williams, 1997), partially formed in coleopteran embryos (Wegerhoff *et al*., 1992; Wegerhoff *et al*., 1996), and does not appear in *Drosophila* until late larval stages (Renn *et al*., 1999). However, anlagen of the optic and antennal lobes are more common in insect embryos. Therefore, the expression patterns of *Dichaete* and *SoxNeuro* in the brain of the *Drosophila* embryo may be a derived feature, due to its incomplete brain development.

Within the Coleopteran *Tribolium castaneum*, neural development has been shown to be largely conserved with *Drosophila*, albeit with some minor variance (Wheeler *et al*., 2003; Wheeler *et al*., 2005; Kux *et al*., 2013; Biffar & Stollewerk, 2015). Two homologues of the *Drosophila achaete-scute* complex genes have been identified and characterised in *Tribolium*: *achaete-scute homolog* (*Tc-ASH*) and *asense* (*Tc-ase*). *Tc-ASH* expression is observed in all neuroblasts and proneural clusters, becoming restricted to presumptive neural precursor cells in developmentally older segments, and RNAi experiments demonstrate it is necessary for neuroblast formation (Wheeler *et al*., 2003). *Tc-ase* expression is limited to neural precursors, and is therefore expressed downstream of *Tc-ASH*, suggesting functional conservation with their homologues in the fruit fly. Similarly, two homologues of the *Enhancer of split E(spl*) complex also show conserved functions in *T. castaneum*, whereby expression is observed in the neuroectoderm during germband extension in response to *Tc-Notch* and *Tc-ASH*, maintaining lateral inhibition of proneural cluster cells that do not acquire NB status (Kux *et al*., 2013).

Five of the columnar genes have also been characterised in *Tribolium castaneum*. Expression of the columnar genes *Egfr*, *vnd*, *ind*, and *msh* have been described by Wheeler *et al*. (2005), and are found in the medial, intermediate, and lateral columns of the developing *Tribolium* CNS in similar (but not identical) patterns to *Drosophila*. The initiation of *Tc-vnd* has been shown to be conserved, whereas *Tc-ind* has diverged. As in *Drosophila*, *Tc-vnd* and *Tc-ind* modulate neural precursor formation in the medial and intermediate columns, respectively, and *Tc-vnd* inhibits *Tc-ind* within the medial column, establishing the first columnar borders (Wheeler *et al*., 2005). *Tc-Egfr* and *Tc-vnd* are both active in thin longitudinal stripes either side of the pre-gastrula embryo. Expression is also visible in the growth zone of the early germband embryo, with *vnd* expressed in longitudinal columns towards the posterior tip, while *Tc-Egfr* expression appears as a more solid single band in the growth zone of 15hr embryos. *Tc-ind* and *Tc-msh* expression is absent from the growth zone across all stages of germband extension (Figure 4.1.2; Wheeler *et al*., 2005).

The fifth columnar gene to be characterised in *T. castaneum* is *Tc-Dichaete*, with a preliminary characterisation of *Tc-Dichaete* within the context of Wnt/β-catenin signalling performed by Oberhofer *et al*. (2014), and a more thorough characterisation recently carried out by Clark and Peel (2017) within the context of insect segmentation. Oberhofer *et al*. (2014) investigated the *hedgehog* and Wnt pathways in the beetle, and found that *Tc-Dichaete*, amongst other genes, is down-regulated in the absence of the Wnt pathway. They performed whole-mount *in situ* hybridisation experiments on 27 of these Wnt regulated genes in early germband embryos, including *Tc-Dichaete* (Figure 4.1.3). Their experiments show that *Tc-Dichaete* is strongly expressed in the growth zone, except for the posterior-most region. Clark and Peel (2017) performed a more extensive analysis of *Tc-Dichaete* expression, covering the majority of stages during germband extension. The authors compared *Tc-Dichaete* expression to that observed in the fly, concluding that the segmentation process in insects is temporally regulated by the expression sequence of Caudal, Dichaete, and Odd-paired, after demonstrating their necessity for correct primary pair rule expression in *Drosophila*, and identifying similar expression patterns between *Drosophila* and *Tribolium*.

When I first began my experiments, neither *Tc-Dichaete* nor *Tc-SoxNeuro* expression had been fully characterised in short germband embryos, and thus marks the purpose of the investigation presented here. I cloned the orthologues of both genes from *T. castaneum* and synthesised several complementary DIG-labelled RNA probes for each. However, generating an effective probe for the *Tc-Dichaete* gene proved to be problematic with several attempts yielding non-specific signal. While troubleshooting this problem, work from Clark and Peel (2017) was published showing the expression pattern of *Tc-Dichaete* in the beetle. I thus elected to use this data instead for my analysis, as their investigations only considered *Tc-Dichaete* within the context of embryo segmentation. Here I present the *in situ* hybridisation data I generated for *Tc-SoxNeuro*, and draw comparisons with the published expression patterns of *Tc-Dichaete* generated by Clark and Peel (2017). Collectively, these investigations suggest that the *Dichaete* and *SoxNeuro* genes of *Tribolium* are perhaps operating in a similar, and therefore conserved, manner to their orthologues in *Drosophila*.

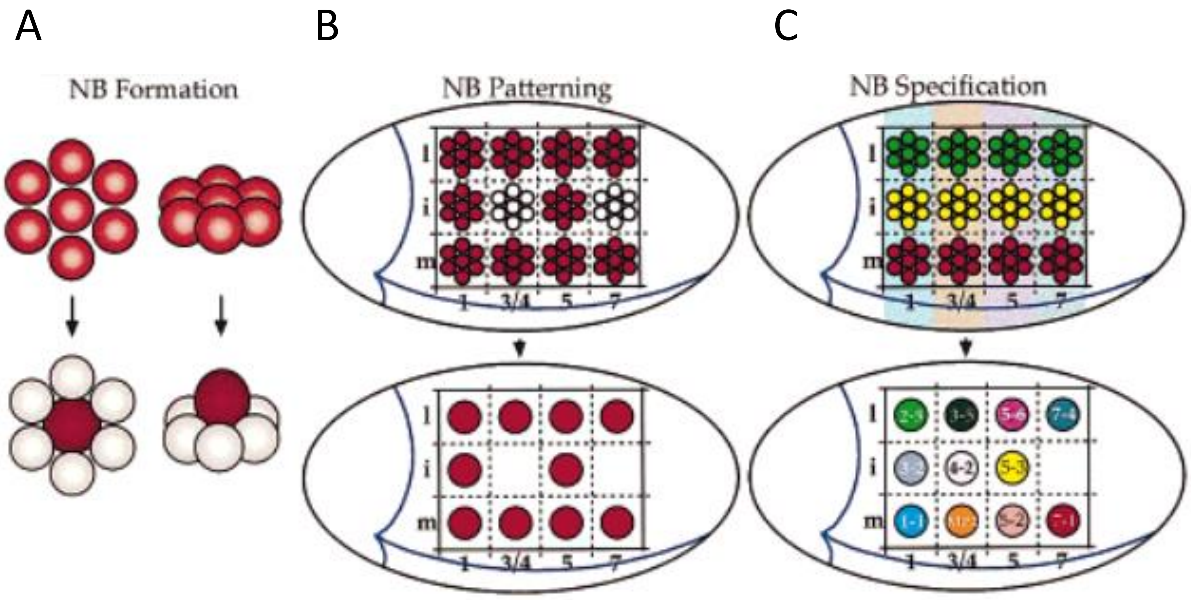


Figure 4.1.1. Neuroblast (NB) formation, patterning, and specification in *D. melanogaster* embryo hemisegments of the developing neuroectoderm. (A-B) A cluster of proneural cells (in light red) are initially equipotent; however a single cell from this cluster (dark red) is selected to become the presumptive NB. The cell with the highest level of *achaete-scute complex* gene expression is fated to become the NB cell. *Notch* signalling in the adjacent cells causes the lateral inhibition of proneural genes, and the presumptive NB enlarges and delaminates into the interior of the embryo. The rest of the proneural cell cluster go on to form the epidermis. (C) NB specification is determined by their respective positions along 3 longitudinal columns (medial (red), intermediate (yellow), and lateral (green)), governed by columnar genes: *vnd* in the medial column, *ind* in the intermediate column, and *msh* in the lateral column. NBs are also arranged along 4 transverse rows (1, 3/4, 5, and 7, named after the respective transverse rows found in the grasshopper), governed by the segment polarity genes. Reproduced from Skeat, 1999.

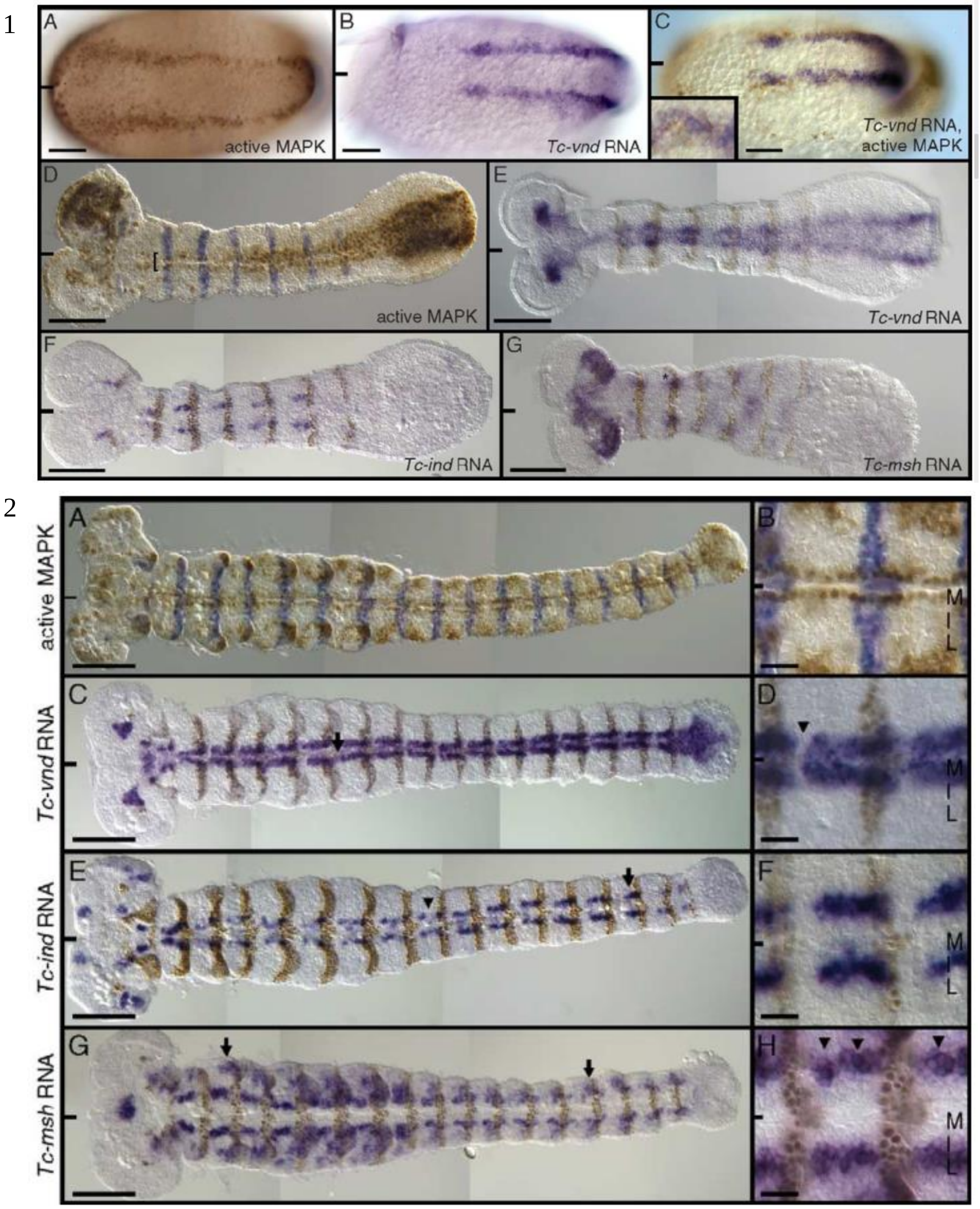


Figure 4.1.2. Ventral views of *Tc-vnd*, *Tc-ind*, *Tc-msh* expression and Tc-MAPK presence (marking *Tc-Egfr* expression) in the *T. castaneum* embryo. (1A-C): Pre-gastrula embryos (<~4 hr). (1D-G) Post-gastrula germ anlagen (~15 hr) (2A-H) Extended germbands (~22 hr). (1A-C) *Tc-vnd* and Tc-MAPK are expressed in overlapping pairs of longitudinal stripes in the pre-gastrula embryo, either side of the ventral midline. (2C-H) *Tc-vnd*, *Tc-ind*, and *Tc-msh* are expressed in the medial, intermediate, and lateral columns of the neuroectoderm, respectively. Only Tc-MAPK and *Tc-vnd* are expressed in the growth zone of the embryos (1D-E; 2A,C). Scale bars in 2B,D,F,H = 25 µm; 1A-C = 50 µm; and 11D-G = 100 µm. Figure reproduced from Wheeler *et al*., 2005.

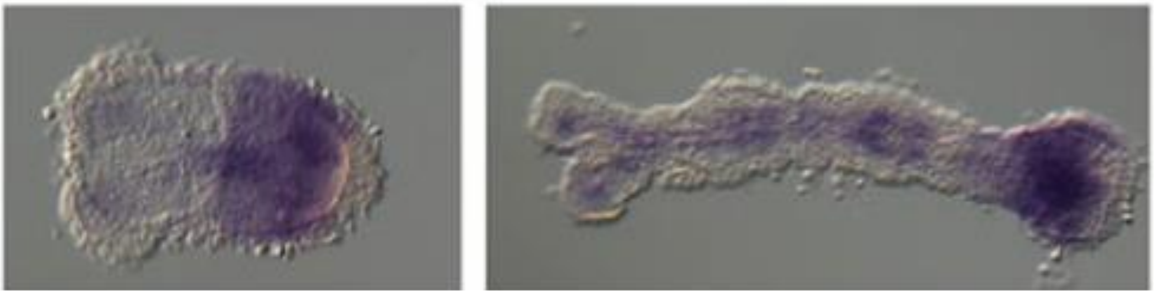


Figure 4.1.3. Ventral views of *Tc-Dichaete* expression in the early *T. castaneum* embryo. Expression is visible in the posterior growth zone at the onset of germband extension, and along two longitudinal stripes extending towards the procephalic region of the embryo. Anterior = left. Reproduced from Oberhofer *et al*., 2014.

4.2 Published data for *Dichaete* and *SoxNeuro* in *T. castaneum*

I first began by exploring quantified expression data for *Tc-Dichaete* and *Tc-SoxNeuro*, although RNA-seq data for *Tribolium castaneum* embryos is sparse. Just three time points at 32^{o}C are available (Accession: PRJNA275195: iBeetle RNA-seq, Schmitt-Engel *et al*., 2015), representing gastrulation up to early germband extension (Figure 4.2.1). Expression levels are apparently low for both *Dichaete* and *SoxNeuro* immediately post-gastrulation, with increased *Dichaete* expression at 9-11 hrs. In contrast, *SoxNeuro* expression remains comparatively lower across the short time series represented here.

There is also some very preliminary RNAi data for these genes conducted by the iBeetle project (Dönitz *et al*., 2015; Schmitt-Engel *et al*., 2015). For *Tc-SoxNeuro* RNAi, lethality occurred in 20% of individuals 11 days after pupal injection; for *Tc-Dichaete*, the figure is 30%. Knock-downs for each gene also exhibits irregular musculature patterns in the developing embryo and segmentation defects. In the first larval instar, *Tc-Dichaete* knock-downs lack a thorax, and some abdominal segments are also absent. For *Tc-SoxNeuro* knock-downs in the first larval instar, shape is irregular, larval appendages are mostly absent, and the larvae are partially everted (Dönitz *et al*., 2015; Schmitt-Engel *et al*., 2015).

4.3 Published *Dichaete* expression patterns within the *T. castaneum* embryo

From the work of Clark and Peel (2017). *Dichaete* mRNA transcripts are expressed extensively in the growth zone of the early gastrula except for the posterior-most region (Figure 4.3.1A-B), with expression resolving into the central-most region at the early onset of elongation (Figure 4.3.1C-D). As the germband extends, growth zone expression retracts anteriorly, where in late germbands expression is only present in the most anterior region of the growth zone before becoming altogether absent (Figure 4.3.1E-F). Within the trunk of the gastrula, *Dichaete* is observed in two transverse stripes in the early germband, which intersect with two longitudinal stripes along the ventral midline (Figure 4.3.1A). During early to mid-stage germband extension, expression resolves solely to the ventral neuroectoderm in the developmentally older (anterior) segments (Figure 4.3.1C). These longitudinal stripes extend to the anterior-most tip of the procephalic region of the gastrula, where they begin to diverge (Figure 4.3.1A-CB). Within the head, Dichaete expression is initially confined to the diverging longitudinal stripes, with each stripe broadening in mid-stage germbands (Figure 4.3.1C-D), before branching in the head of older embryos (Figure 4.3.1E). In fully extended germbands, transcripts become more diffuse across various cell populations in the head, and *Dichaete* is strongly expressed in the posterior regions of the lobular anlagen. (Figure 4.3.1F).

4.4 *SoxNeuro* expression patterns within the *T. castaneum* embryo

SoxNeuro expression is first detected in the early gastrula, with no expression observable in the syncytial blastoderm. Within the early gastrula, expression is observed in stripes along the ventral furrow towards the posterior of the embryo, and in a symmetrical ring-like pattern in the head (Figure 4.4.1A-C). Transcripts resolve into a thick transverse stripe within the posterior third of the embryo, and expression is notably absent from the posterior portion of the growth zone of the early germ anlage (Figure 4.4.1D-F). This stripe is then elongated during germband extension, eventually resolving longitudinally throughout the neuroectoderm (Figure 4.4.1F-H & Figure 4.5.1G). Within the procephalic region, the symmetrical ring-like expression of *SoxNeuro* observed in the early germband expands throughout the presumptive brain, before once again being isolated to specific regions; most likely the mushroom body and lobe anlagen in fully extended germbands (see Figure 4.4.1 & Figure 4.5.2A,C,E). Throughout germband elongation, expression is absent in the posterior-most portion in the growth zone and retracts anteriorly until becoming absent from the growth zone altogether, and appears ubiquitous in the youngest segments of the extending trunk (Figure 4.4.1F-H & Figure 4.5.1C). Expression continues in a pan-neuroectodermal manner throughout all stages of germband extension. No signal was detected using sense probes synthesised at the *SoxNeuro* locus (Figure 4.4.2).

4.5 Comparisons of *SoxNeuro* & *Dichaete* expression patterns

Dichaete expression is present within the beetle growth zone except for the posterior-most tip at gastrulation and throughout germband elongation (Figure 4.5.1D-E). In contrast, *SoxNeuro* expression is only present in the growth zone briefly during elongation; it is absent from the growth zone at gastrulation, and its expression is more anterior than that of *Dichaete* during the early stages of germband extension (Figure 4.5.1A-B). Expression of both genes in the growth zone retracts anteriorly during germband elongation, and is absent entirely from the growth zone in fully elongated germ bands (Figure 4.5.1C,F).

Within the trunk of the embryo, *SoxNeuro* is expressed posteriorly along what appear to be the gastrulation folds of the early germ anlage (halting at the boundary of the emerging growth zone) (Figure 4.5.1A), whereas *Dichaete* is expressed in longitudinal stripes along the ventral midline (Figure 4.5.1D). During early germband extension, *SoxNeuro* expression resolves into a thick band toward the posterior of the embryo which extends into the growth zone (Figure 4.5.3B), and as extension progresses, expression appears to be ubiquitous in developmentally younger segments (Figure 4.5.1C), resolving only to the neuroectoderm in developmentally older segments (Figure 4.5.1G). In contrast, *Dichaete* is expressed longitudinally along the neuroectoderm in the developmentally youngest segments, albeit absent at the segment boundaries (Figure 4.5.1E), and is uninterrupted in developmentally older segments (Figure 4.5.1F). The longitudinal expression of *SoxNeuro* also appears to extend more laterally than *Dichaete* in the developmentally older segments (Figure 4.5.1G-H).

Within the head, *SoxNeuro* is expressed in a ring-like pattern in the procephalic region of the gastrula (Figure 4.5.2A), whereas *Dichaete* expression appears in diverging longitudinal stripes extending towards the anterior tip (Figure 4.5.2B). During germband extension, *SoxNeuro* expression is more evenly distributed in the head of younger germbands (Figure 4.5.2C), resolving in later stages into what might be the developing anlagen of the antennal and optic lobes in the anterior protocerebrum and deuterocerebrum, and the mushroom body (Figure 4.5.2E). *Dichaete*, on the other hand, continues to be expressed as diverging longitudinal stripes in the head of the younger germbands, which begin to broaden and branch, with strong expression in the posterior extremities of the head (Figure 4.5.2F).

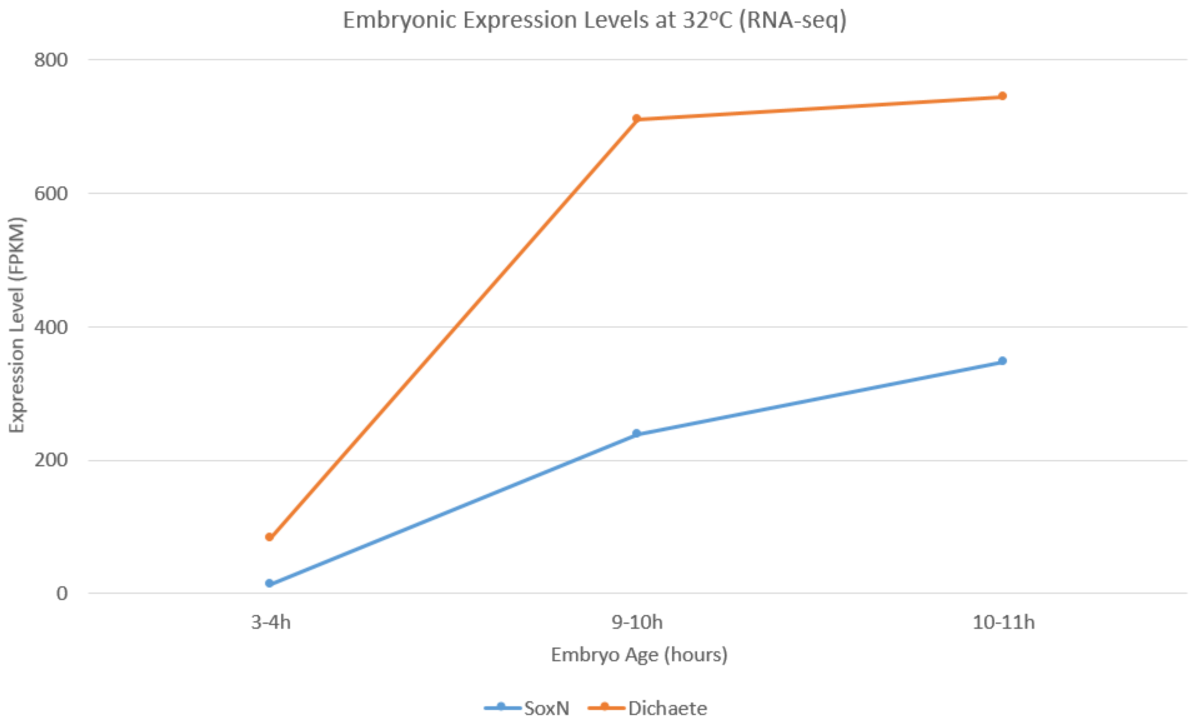


Figure 4.2.1 RNA-seq data for *T. castaneum* embryos at 32°C, from gastrulation (3-4 hr) up to early germband extension (11 hrs). *Tc-Dichaete* (orange) shows a sharp rise in expression after gastrulation, whereas *Tc-SoxNeuro* expression (blue) is comparatively lower. (Accession: PRJNA275195, iBeetle RNA-seq, Schmitt-Engel *et al*., 2015).

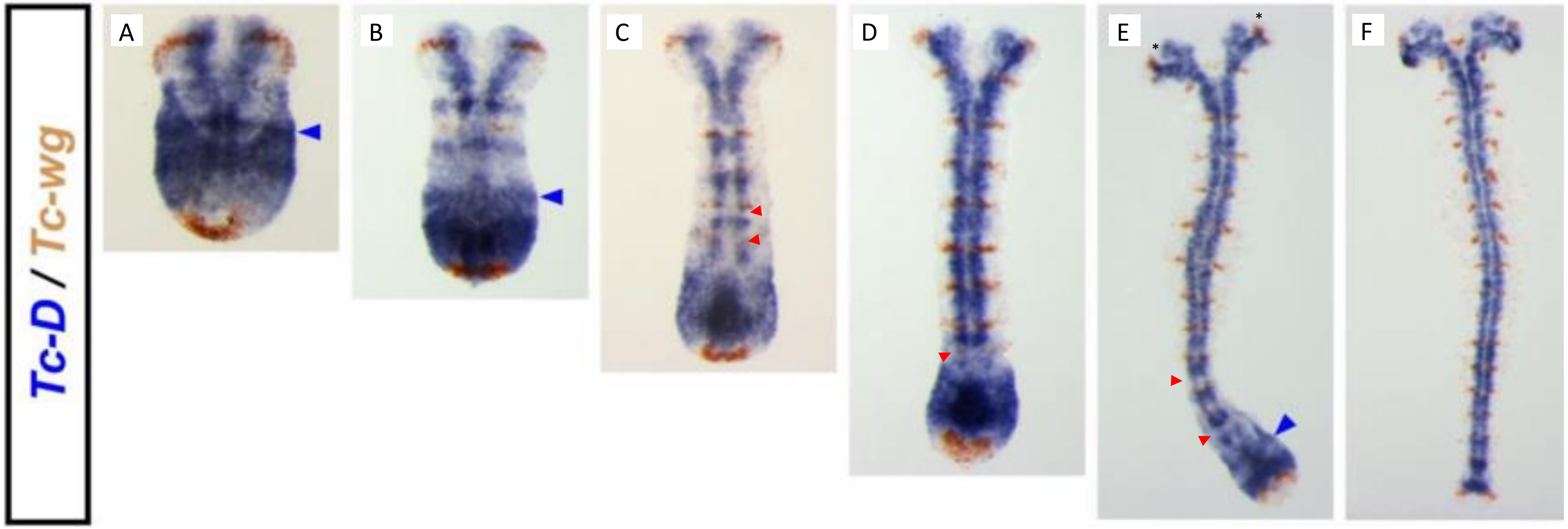


Figure 4.3.1. (A-F) *Dichaete* (purple) and *Wingless* (brown) mRNA expression in *T. castaneum* embryos. *Dichaete* expression is strong in the central-most region of the posterior growth zone during germband elongation (B-D), however retracts in an anterior-fashion in the fully extended germband (F). Anterior to the growth zone, transverse stripes are detected at early stages of development (A-B, blue arrow head). Expression is strong in the neuroectoderm of the developing embryo in all segments, however in developmentally younger segments neuroectodermal expression is absent at the posterior boundaries of each segment (red arrow heads). In the procephalic region, the parallel longitudinal stripes either side of the ventral midline expand and diverge towards the anterior tip (A-C), with branches beginning to form in later stage embryos (asterisk, E). Ventral views, anterior = top; Modified from Clark & Peel 2017.

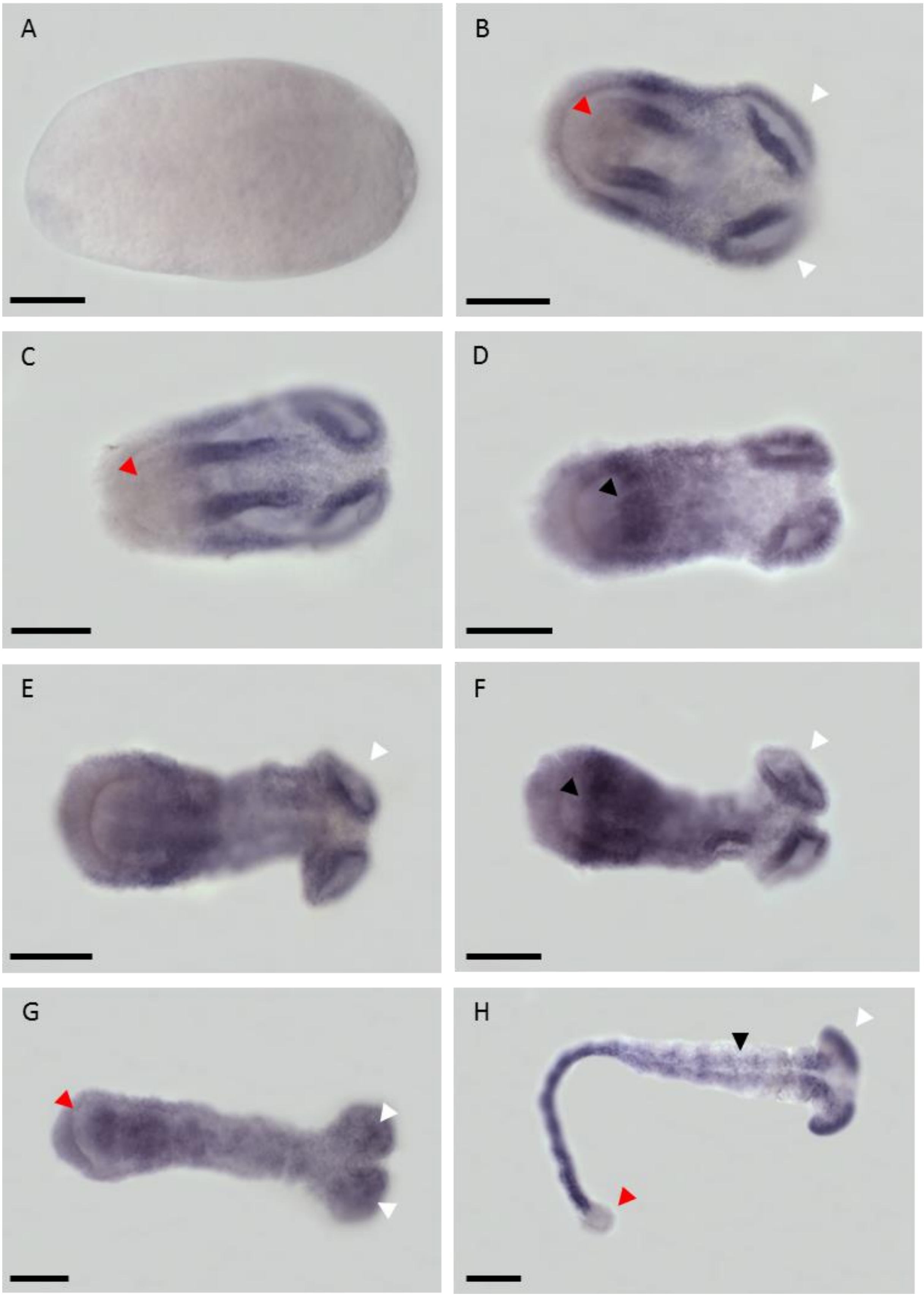

A
B
C
D
E
F
G
H

Figure 4.4.1. *SoxNeuro* mRNA expression (purple) in 0-24hr *T. castaneum* embryos. (A) No expression is observable in the syncytial embryo. (B) In the early gastrula, *SoxNeuro* is expressed in a symmetrical ring-like fashion in the developing head (white arrow heads), and along a pair of longitudinal stripes towards the posterior of the embryo (halting at the anterior region of the growth zone), along what might be the gastrulation folds, following ventral furrow formation. (C) The expression pattern observed in (B) is continued at the onset of elongation, and is absent from the majority of the growth zone (red arrow heads). (D) Expression is maintained in its ring-like pattern in the developing head, however a strong and ubiquitous expression domain in the posterior third is established, again halting at the growth zone, replacing the pair of stripes. Expression appears to be confined outside the growth zone (black arrow), continuing in an anterior direction along an expression gradient (becoming gradually weaker towards the head). (E-F) The expression observed in (D) is maintained as the elongation of the blastoderm continues with the boundary in the anterior region of the growth zone (black arrow) and ring-like expression in the brain (white arrow), and becoming stronger in the mid-section of the trunk. (G) The expression domain of the developing germband expands during elongation; expression appears to be confined to cell clusters within the rudimentary segments beginning to form, and within distinct cell populations of the head (white arrow heads), and continues to be absent from the posterior portion of the growth zone (red arrow head). (H) In the fully elongated germband, expression is ubiquitous throughout the developing neuroectoderm as NBs begin to form along each side of the ventral midline (black arrow head). However, within the developmentally younger segments (towards the posterior of the embryo), expression is ubiquitous. In the head, signal is strong within the developing lobular regions and what might be the developing mushroom bodies. No expression is observed in the growth zone in extended germbands (red arrow head). Ventral views, anterior to the right, scale bars = 100 µm.

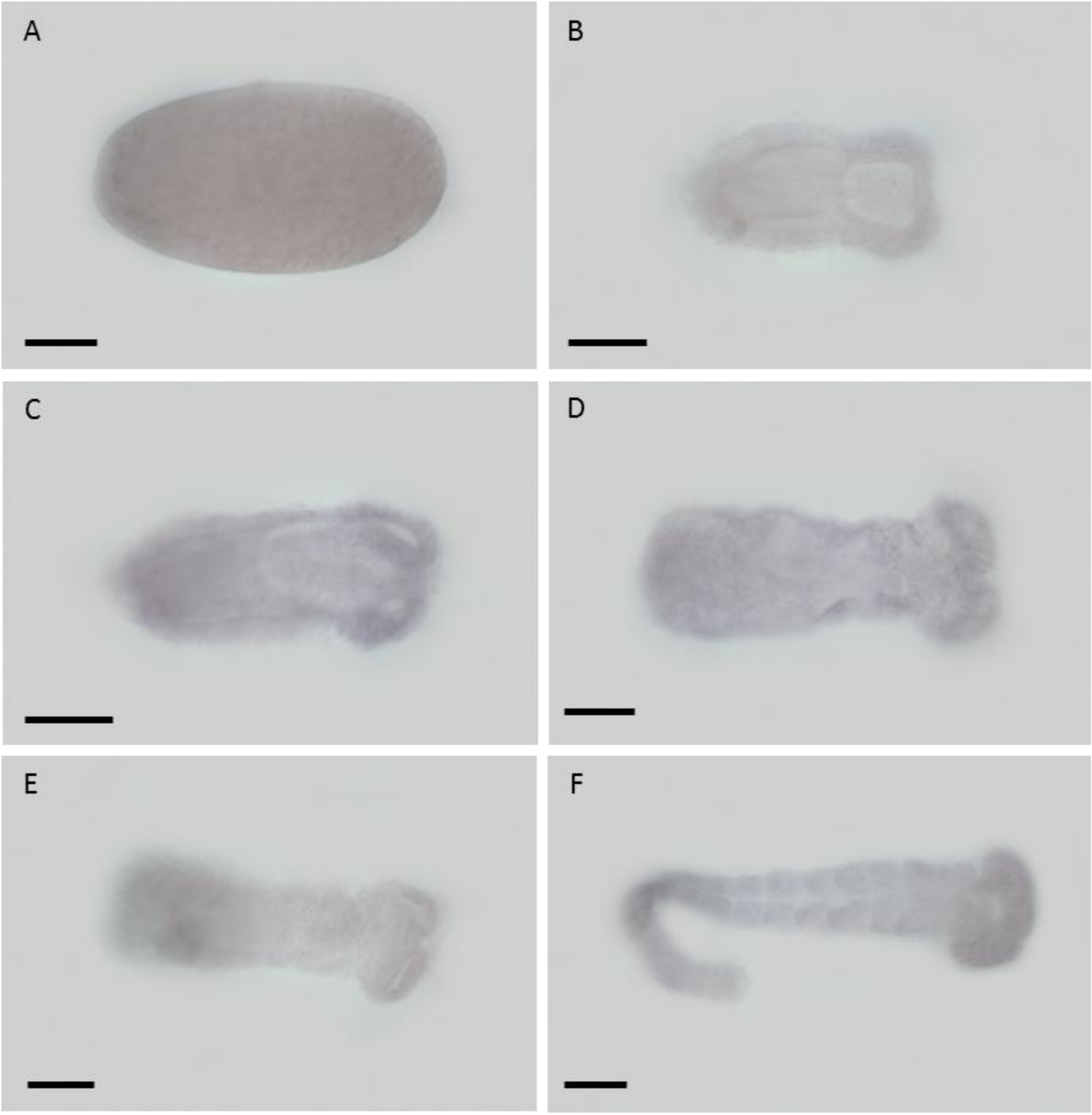


Figure 4.4.2. Sense probes (negative controls) synthesised from the *SoxNeuro* locus in 0-24hr *T. castaneum* embryos. No signal is detectable across all stages of development. Anterior to the right, scale bars = 100 µm.

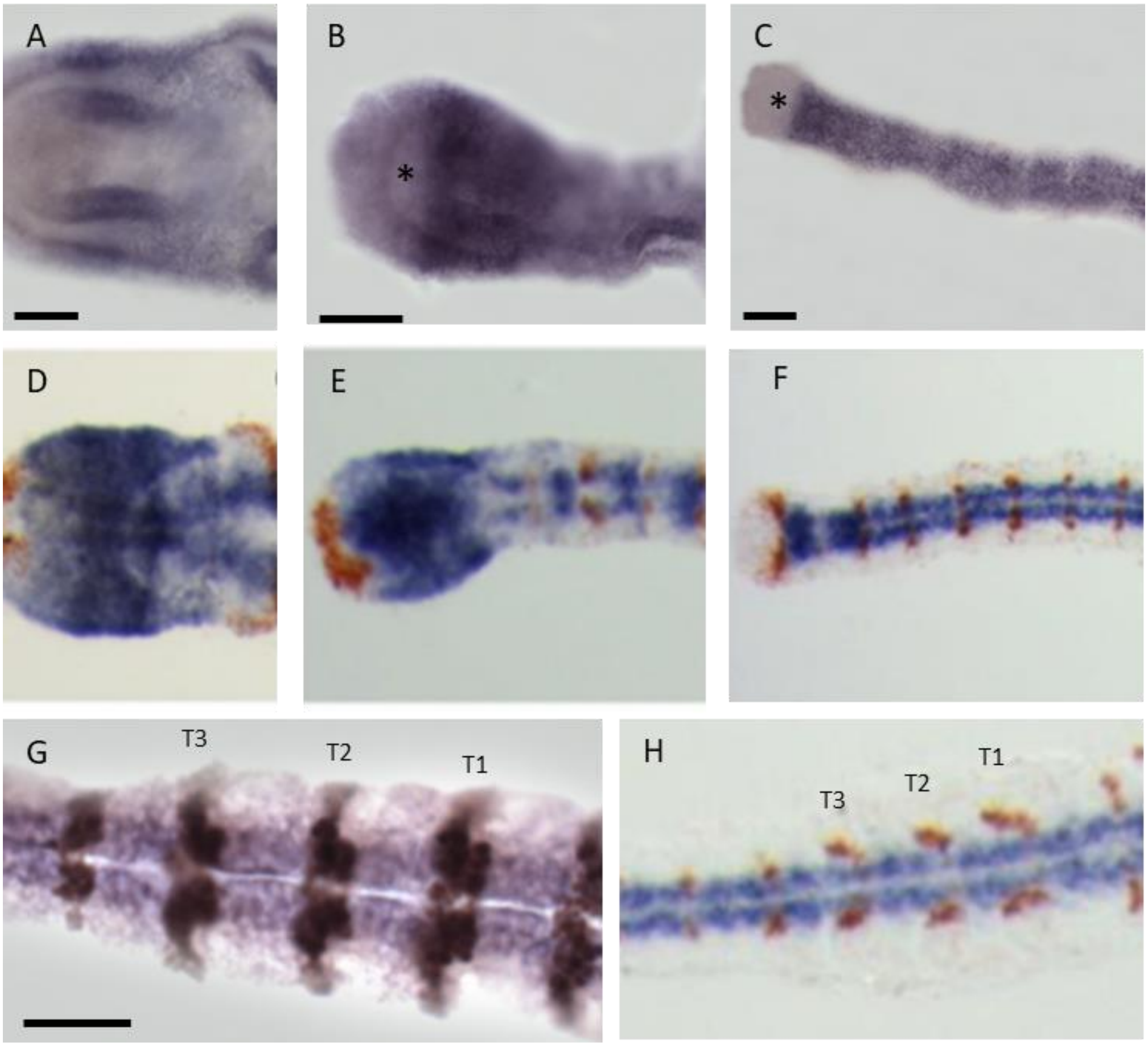


Figure 4.5.1. The growth zone and extending trunk of the *T. castaneum* embryo stained for *SoxNeuro* (purple: A-C,G), *Dichaete* (purple: D-F,H), and *Wingless* (brown: D-F, H) mRNA and Engrailed protein (brown: G), across similarly staged embryos. (A) *SoxNeuro* expression appears along the gastrulation folds either side of the ventral furrow formed during gastrulation, extending in a posterior fashion. (B) *SoxNeuro* expression is ubiquitous in the developmentally younger segments at a boundary in the growth zone (asterisks), retracting anteriorly in elongated germbands (C), where it continues to be expressed ubiquitously in developmentally younger segments. (G) Expression resolves into the neuroectoderm in developmentally older segments (thoracic segments shown here). (D) *Dichaete* expression is strong in the posterior of the germ anlage, and two longitudinal and transverse stripes are observed, with the longitudinal stripes extending into the head. (E-F) *Dichaete* expression is strongest in the central region of the growth zone, and absent on the posterior-most region. Expression is comparably less intense beyond the growth zone, resolving into thin longitudinal stripes along the ventral midline. (H) Expression continues as thin longitudinal stripes along the ventral midline in developmentally older segments, and is confined to the neuroectoderm. Ventral views, anterior to the right, scale bars = 50 µm. (D-F, and H were generated by Clark & Peel, 2017).

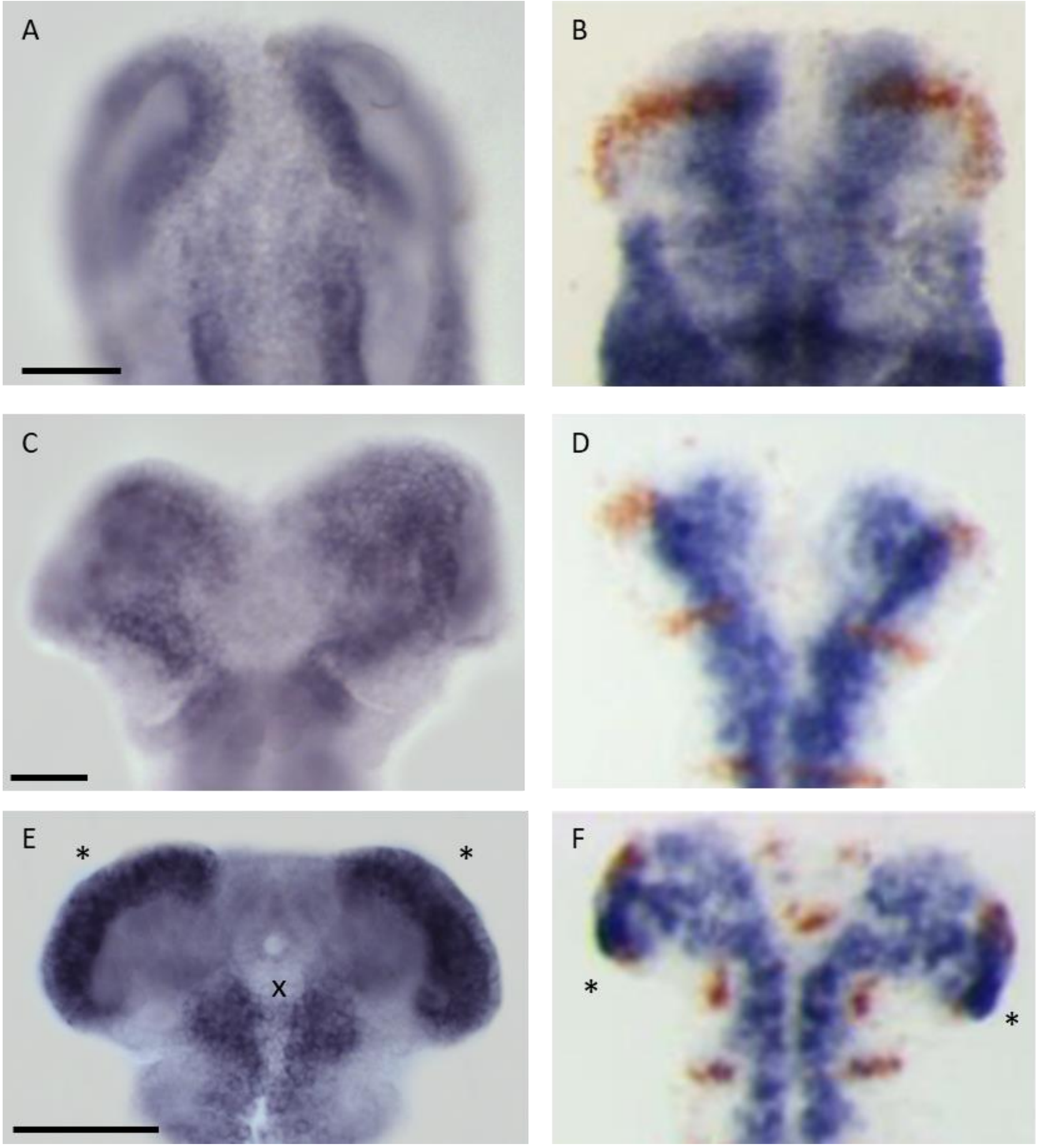


Figure 4.5.2. The head of the extending germband labelled for *SoxNeuro* (purple: A,C,E), *Dichaete* (purple: B,D,F), and *Wingless* (brown: B,D,F) mRNA, across similarly-staged embryos. (A-B) *SoxNeuro* expression appears ring-like in the head anlagen, with the longitudinal stripes observed in the trunk absent in the anterior region of the embryo. In contrast, *Dichaete* expression in the head anlagen appears as thick longitudinal stripes. (C-D) *Dichaete* and *SoxNeuro* expression appear to be similar during mid-stage germband elongation. (E) *SoxNeuro* expression resolves into specific domains in the lobular regions (asterisks), and what might be the mushroom bodies (cross). (F) *Dichaete* also exhibits strong expression in the lobular domains however these do not extend as anteriorly as they do for *SoxNeuro* (asterisks), and expression remains more diffuse throughout the brain. Ventral views, anterior top, scale bars = 50 µm. (B,D and F were generated by Clark & Peel, 2017.)

4.6 Discussion of results

In this investigation, I have characterised the spatiotemporal expression patterns of the DV patterning gene *SoxNeuro* in *Tribolium castaneum* across germband elongation, and have re-appraised the expression pattern of *Dichaete* within the context of central nervous system development. These expression patterns appear to be highly conserved with those of the long germ insect *D. melanogaster*, despite their different modes of germband elongation. *Tc-SoxNeuro* expression is similar to its orthologue in *Drosophila*, with expression observed in a pan-neuroectodermal manner throughout development. *Tc-Dichaete* exhibits some plasticity, however, with an absence of expression in the developing midline; yet its expression in the growth zone and neuroectoderm implies conservation where insect segmentation and VNC development is concerned. In the developing head, there is overlapping expression of *Tc-SoxNeuro* and *Tc-Dichaete* in the anlagen of the antennal and optic lobes and mushroom bodies, although *Tc-Dichaete* continues to be expressed more expansively in late-stage embryos, whereas *Tc-SoxNeuro* expression appears to resolve more to the lobular extremities of the head.

The sparse RNA-seq data available is broadly supported by these *in situ* hybridisation data, yet the RNA-seq data shows *Tc-Dichaete* expression being significantly higher than that of *Tc-SoxNeuro* in the early embryo. This may be reflected by each gene's respective activity in the growth zone, which contains substantially more cells than the rest of the embryo during earlier stages of development (Nakamoto *et al*., 2015), and as such transcript copy numbers may be more abundant for *Tc-Dichaete* than *Tc-SoxNeuro*.

There is also preliminary data available for these genes as part of a mass RNAi screen by the iBeetle project (Dönitz *et al*., 2015; Schmitt-Engel *et al*., 2015). *Tc-Dichaete* RNAi experiments generate segmentation defects in embryos and first instar larvae, suggesting that there is conserved function for this gene in insect segmentation. However, these data are a 'first pass screen', whereby experiments are performed just once with few replicates, and off-target controls are not included (Dönitz *et al*., 2015; Schmitt-Engel *et al*., 2015). These data therefore need to be replicated and validated, and double knock-downs may need to be performed to uncover phenotypes given the redundancy exhibited by *Dichaete* and *SoxNeuro* in *D. melanogaster*.

Tc-Dichaete and *Tc-SoxNeuro* expression does appear to overlap in the neuroectoderm of developmentally older segments, and *Tc-SoxNeuro* expression seems to extend more laterally than *Tc-Dichaete*, although this is difficult to be sure of this from single-stainings, which lack the resolution required for columnar identification. However, if it is the case that *Tc-SoxNeuro* extends more laterally than *Tc-Dichaete*, this would be conserved with the activity of these two genes in *Drosophila*, where *SoxNeuro* expression is found within the medial, intermediate, and lateral columns, and *Dichaete* expression in just the medial and intermediate. Fluorescent *in situ* hybridisation would be more optimal to study this in future, as fluorescent channels may be overlaid with each other to define overlapping expression fully, and confocal microscopy yields higher resolution images better enabling the detection of expression within individual cells.

In *Tribolium*, *Tc-Egfr* and *Tc-vnd* are both active in thin longitudinal stripes either side of the pre-gastrula embryo. These may be the precursors to the gastrulation folds, where *Tc-SoxNeuro* expression is observable post-gastrulation. Data is absent for *Tc-Dichaete* in the syncytial blastoderm. *Tc-Dichaete*, similar to *Tc-Egfr* and *Tc-vnd*, is expressed in the growth zone throughout all stages of germband extension. *Tc-Dichaete* expression is much more diffuse throughout the growth zone, whereas *Tc-Egfr* and *Tc-vnd* is limited to longitudinal stripes either side the ventral midline. Nonetheless, *Tc-Dichaete*, *Tc-Egfr*, and *Tc-vnd* expression appear to be at least partially overlapping in the post-gastrula embryo throughout the neuroectoderm. The more diffuse expression of *Tc-Dichaete* in the growth zone may reflect the earlier role that *Dm-Dichaete* has in embryo segmentation (Sánchez-Soriano & Russell, 1998), as commented by Clark and Peel (2017). For example, within *Tribolium*, expression of the segmentation-regulating gap gene *Tc-hunchback* (*Tc-hb*) shares similarities with the expression of *Tc-Dichaete* during germband extension (see Wolff *et al*., 1995). *Tc-hb* is initially expressed in transverse stripes in the early embryo; subsequent expression along a u-shaped rim in the growth zone is observed at the onset of elongation, before expanding through the entire growth zone until the germband is fully extended. *Tc-hb* is also expressed in NBs in later stages of development, most notably in the head and older segments of the germ anlage (Wolff *et al*., 1995). The similar expression patterns of *Tc-Dichaete* therefore strengthens the hypothesis of a conserved role in embryo segmentation (Clark & Peel, 2017).

Tc-SoxNeuro expression is absent in the posterior portion of the growth zone, similarly to *Tc-ind* and *Tc-msh*, whereas *Tc-vnd* is not. *Tc-vnd* expression is also observed in parallel with *Egfr* activity in two longitudinal pre-gastrulation stripes, suggesting that *Tc-vnd* is acting more upstream in *Tribolium* than its orthologue in *Drosophila* in the early establishment of neuroectoderm; *Tc-SoxNeuro*, *Tc-ind*, and *Tc-msh* activity, on the other hand, appears to be limited to later patterning roles in the CNS. *Tc-SoxNeuro* is also expressed more homogeneously in developmentally younger segments at the onset of and during germband extension. This might imply an additional function in cell populations neighbouring the ventral ectoderm – the neuroectodermal columns are already established by this stage, as evidenced by the longitudinal columnar patterning of *Tc-vnd*, *Tc-Egfr*, and *Tc-Dichaete* in these segments in similarly-staged embryos. *Tc-Dichaete* expression is absent in the ventral midline of *Tribolium* embryos, suggesting a departure from its role in midline glial formation in *Drosophila*. As in *Drosophila*, both *Tc-SoxNeuro* and *Tc-Dichaete* are expressed upstream and in parallel to the proneural genes of the *Tc-achaete/scute complex* and the lateral inhibition genes of the *Tc-E(spl)* complex (Zhao & Skeath, 2002; Zhao *et al*., 2007; Kux *et al*., 2013; Hartenstein & Stollewerk, 2015).

The genes involved in CNS patterning and development described here have yet to be fully characterised within the developing *Drosophila* brain. However, *Dm-Dichaete* is necessary for correct brain development, most notably within the tritocerebrum neural cells, and strong expression is present throughout the protocerebrum, deuterocerebrum, and tritocerebrum (Sánchez-Soriano & Russell, 2000). Both *Tc-SoxNeuro* and *Tc-Dichaete* show expression in the embryonic brain during development, overlapping in what appear to be the protocerebral optic lobes and the deuterocerebral antennal lobes at the extremities of the developing head, and the mushroom bodies in the central posterior region. *Tc-Dichaete* expression appears more intense in the posterior lobular regions however, while *Tc-SoxNeuro* is strong throughout the lobes.

The expression patterns of *SoxNeuro* appear to be conserved beyond *Drosophila* and *Tribolium*: in the honeybee, *Am-SoxNeuro* expression, for example, is found in the neuroectoderm and cephalic lobes of the developing embryo. *Am-SoxNeuro* expression is also seen in the ventral gastrulation folds of the early gastrula, similarly to the expression of *Tc-SoxNeuro* observed in the early *Tribolium* gastrula. Moreover, in adults, *Am-SoxNeuro* is expressed in the mushroom bodies of the male worker bees (Wilson & Dearden, 2008). *Tc-SoxNeuro* expression thus appears to be highly conserved across the insects, within the Hymenoptera, Coleoptera, and Diptera. Preliminary RT-PCR data also exists in the Lepidopteran *Bombyx mori*, which shows that *Bm-SoxNeuro* is expressed throughout embryonic development when examined across 24hr intervals during embryogenesis (Wei *et* *al*., 2011). However, spatial expression patterns have yet to be characterised, and the resolution of temporal expression at distinct stages is lacking.

Dichaete expression patterns across the insects are more ambiguous and yet to be fully resolved. Wilson and Dearden (2008) failed to detect any *Am-Dichaete* expression in the honeybee via both *in situ* hybridisation and RT-PCR, and conclude that it may be a pseudogene. However, this is unlikely as the ORF of the *Am-Dichaete* locus is still intact, which suggests that it is still genetically functional as pseudogenes accumulate mutations and decay swiftly in the absence of selection pressures (Qian *et al*., 2010). In *Bombyx mori*, *Bm-Dichaete* expression is detected via RT-PCR: it is absent at the onset of embryogenesis, present after 24hrs, and then absent again at 48hrs; all subsequent time points up to hatching show evidence of *Bm-Dichaete* expression. Similarly to *Bm-SoxNeuro*, *Bm-Dichaete* expression has yet to be characterised at finer temporal resolution and any spatial resolution in embryos.

Interestingly, *Sox21b* expression is markedly different in *A. mellifera* when compared to *D. melanogaster*. In the fruit fly, *Sox21b* expression is limited to the cells of the intestinal anlagen and ventral epidermis of the embryo (Fisher *et al*., 2012: FlyBase report), whereas in the honeybee *Sox21b* expression is observed in symmetrical ganglia cells across hemisegments of the developing CNS in later-stage embryos (Wilson & Dearden, 2008). During the later stages of embryogenesis, *Am-SoxNeuro* and *Am-Sox21b* expression does not overlap in the developing VNC, however there is overlapping expression in the mushroom bodies of the brain anlage (Wilson & Dearden, 2008). If *Am-Dichaete* is a pseudogene, perhaps *Am-Sox21b* has convergently evolved to replace its function? Further work characterising *Sox21b* expression in *T. castaneum* will help elucidate whether the expression observed in *A. mellifera* is ancestral, or whether it is derived.

Also of interest is the *SoxB5* gene of *T. castaneum* which has been annotated in the previous chapter, yet remains to be characterised fully in any species. Its orthologue in *B. mori* has been shown to be expressed only during late embryogenesis (Wei *et al*., 2011), although its expression is not associated in particular with any organs or tissues during larval stages. During metamorphosis expression of *Bm-SoxB5* is observed in late-stage pupae and adults (Wei *et al*., 2011). As this gene appears to be the most recent SoxB2 paralogue within the insects, characterising its expression and function may help illustrate how sub-functionalization of SoxB genes might occur.

In conclusion, *SoxNeuro* and *Dichaete* expression is highly conserved across the species it has been characterised in thus far, although *Dichaete* exhibits some degree of plasticity. *SoxNeuro* expression is conserved throughout the neuroectoderm of the insects studied, despite being separated by more than 350 million years of evolution across three different insect orders. *Dichaete* expression in the beetle and fly appear to be highly similar, with expression patterns associated with both CNS development and segmentation; however, further work is required to characterise *Dichaete* in other orders such as Lepidoptera and Hymenoptera. Future investigations should make use of fluorescent *in situ* hybridisation techniques to better visualise the respective expression patterns in the longitudinal columns of the developing neuroectoderm. Genetic techniques, such as RNAi or CRISPR, are also required to study the function of these two genes further, establishing the conservation/divergence of CNS patterning between short- and long-germband insect development. It appears that despite the conspicuously different mechanisms behind segmentation and germband elongation in *Tribolium castaneum* and *Drosophila melanogaster*, the toolkit governing cellular specification and patterning in the embryonic neuroectoderm is likely to be ancient and highly conserved. Moreover, although this investigation has only considered *Dichaete* and *SoxNeuro*, the three other SoxB genes present in insects, *Sox21a*, *Sox21b*, and *SoxB5* are also of interest: especially given the divergence of *Sox21b* expression in the honeybee. Therefore, to gain a more complete understanding of insect SoxB evolution and function, expression and functional studies ought to be carried out for these genes across insect taxa.

Chapter 5

DamID in *Tribolium castaneum*

5.1 Motivations for research

Investigations into mechanisms governing genomic regulation is a particularly exciting field of research in genetics, and a vast complexity in these regulatory networks has been exposed in recent years (Mardis, 2008; Celniker *et al*., 2009; Conaway, 2012; Dunham *et al*., 2012). Research has typically focused on the activity of proteins, called transcription factors (TFs), which regulate the transcription of other genes via interactions with *cis* regulatory modules (CRMs) in the genome (Maris, 2008; Dunham *et al*., 2012). Identifying functional elements in the genome beyond individual genes has consequently become a major focus for many molecular biologists. For example, the modENCODE project aims to elucidate the functional and regulatory elements in the model species *Drosophila melanogaster* and *Caenorhabditis elegans*, and provide a publicly-accessible and comprehensive encyclopaedia for this data (Celniker *et al*., 2009) and the ENCODE project shares the same goal for the human genome (Dunham *et al*., 2012). The vast majority of these studies utilise chromatin immunoprecipitation (ChIP), which involves cross-linking the TF of interest to DNA *in vivo*, fragmenting chromatin via sonication, and enriching bound fragments using a highly specific antibody, thereby enabling the targeted retrieval of DNA bound by the protein (Aparicio *et al*., 2005). DNA is then typically either hybridized to a microarray (ChIP-chip) or sequenced (ChIP-seq). At the time of writing, modENCODE has 343 entries of ChIP studies mapping chromatin binding sites of various TFs within the *Drosophila melanogaster* genome, vastly augmenting our understanding of the regulatory networks contained within the fly genome. High-throughput sequencing technologies have precipitated a revolution in biological research (Schuster, 2008), enabling the sequencing of entire genomes in a matter of days (Schendure & Ji, 2008; Graveley, 2008). These technologies have consequently made approaches to understanding regulatory features available to study in any organism with a sequenced genome, and are often much faster and cheaper than array-based approaches.

Another independent method of studying TF binding *in vivo* is Dam identification (DamID). DamID aims to achieve a similar result to ChIP, and leaves a historical ‘footprint’ of TF binding in the genome by methylating nearby adenine regions in the context of GATC motifs (van Steensel & Henikoff, 2000; Vogel *et al.*, 2007; Marshall *et al.*, 2016). It does not rely on the use of an antibody; instead, methylated adenine regions are recovered using methylation-sensitive restriction enzymes, and enriched DNA either hybridised to a tiling array or sequenced via an NGS platform (van Steensel & Henikoff, 2000; Vogel *et al*., 2007; Southall *et al*., 2013; Aughey

& Southall, 2015; Marshall *et al.*, 2016)). DamID has been performed successfully in *Drosophila melanogaster* (van Steensel & Henikoff, 2000; Ferrero *et al*., 2014; Carl & Russell, 2015; Marshall *et al*., 2016), mammalian cells (Vogel *et al*., 2007), *Caenorhabditis elegans* (Schuster *et al*., 2010), and *Arabidopsis thaliana* (Germann & Gaudin, 2011). (A more thorough comparison of these two techniques is discussed in Chapter 1.3.)

These techniques mapping genome-wide binding patterns of TFs help elucidate regulatory regions of the target species' DNA, however, interspecific comparative studies enable researchers to explore evolutionary changes at the genomic level. For example, ChIP investigations in closely related fungal species reveal significant divergence of TF binding within the genomes of 3 species of the *Saccharomyces* genus (Borneman *et al*., 2007); this effect is yet more pronounced in more evolutionarily distant species of fungi (Tuch *et al*., 2008). Comparative studies have also been conducted between humans and mice to explore vertebrate TF divergence: *e.g.*, despite the highly conserved function of four shared TFs, *in vivo* mapping of binding activity in hepatocyte cells reveals extensive variation in binding site turnover between humans and mice for each TF investigated (Odom *et al*., 2007). These comparative binding studies on hepatocyte cells have been extended to more distantly related vertebrate species also, including *Canus familiaris*, *Monodelphis domesticus*, and *Gallus gallus*, with approximate evolutionary distances of up to ~300 million years, with findings suggesting that binding divergence between species is largely driven by changes to the target motifs of TFs (Schmidt *et al*., 2010). As discussed above, mapping regulatory elements in the genome is one of the principle aims of genomic studies (Mardis, 2008; Conaway, 2012; Dunham *et al*., 2012), and investigations across 20 different mammalian studies have identified that enhancer evolution is rapid in the mammals, whereas promoter evolution is slower (Villar *et al*., 2015).

Within invertebrates, much work has been carried out in *Drosophila* species (MacArthur *et al*., 2009; Bradley *et al*., 2010; He *et al*., 2011; Paris *et al*., 2013; Villar *et al*., 2014; Ferrero *et al*., 2015; Carl & Russell, 2015). MacArthur *et al*. (2009) investigated the genome-wide interactions of 31 TFs involved in early embryonic patterning in *Drosophila* whilst simultaneously examining chromatin accessibility data. MacArthur *et al*. argue that chromatin accessibility, as opposed to TF specificity, is chiefly responsible for TF regulatory activity, at least within the *Drosophila* genome (MacArthur *et al*., 2009). Interspecific comparative studies within the *Drosophila* genus have also been conducted, with Bradley *et al*. (2010) mapping the genome-wide binding sites of 6 TFs: Bicoid, Hunchback, Krüppel, Giant, Knirps, and Caudal, in *Drosophila* *melanogaster* and *Drosophila yakuba*, using ChIP-seq. The researchers find evidence of binding conservation, gain and loss of binding, changes in binding location, and changes in binding intensity, between the two species studied (Figure 5.1.1). Moreover, He *et al*. (2011) investigated the binding of the TF Twist in 6 *Drosophila* species, showing that binding conservation recapitulates evolutionary distances, with the most closely related species exhibiting greater conservation of binding intervals than the least closely related species. Recently, Prasad *et al*. (2016) have explored the binding of the Hox protein Ultrabithorax (Ubx) across different insect orders: the investigators found that there are substantial differences in the targets of Ubx between *Bombyx mori*, *Apis mellifera*, and *D. melanogaster*, however, a significant number of genes enriched for wing-patterning ontology are retained despite being separated by >300 million years of evolution (Prasad *et al*., 2016).

The genomic activity of the SoxB proteins in *Drosophila* has also been extensively studied in previous investigations by members of the Russell lab (Aleksic *et al*., 2013; Ferrero *et al*., 2014; Carl & Russell 2015). For example, Ferrero *et al*. (2014) identified areas of common binding for Dichaete and SoxNeuro in *Drosophila melanogaster* using DamID (Figure 5.1.2), and Carl & Russell (2015) find widespread examples of binding site turnover between 4 *Drosophila* species for Dichaete, which correlate with phylogenetic distances. The binding motifs identified across the 4 *Drosophila* species for Dichaete, and 2 *Drosophila* species for SoxNeuro, are also evolutionarily conserved (Figure 5.1.3). Moreover, sites commonly bound by Dichaete and SoxNeuro exhibit the strongest binding site conservation, implying that despite the redundancy of these genes, selection pressures have maintained the ability of these two proteins to bind at the same loci (Carl & Russell, 2015). The redundancy of these two proteins was demonstrated through an elegant experiment by Ferrero *et al*. (2014), whereby the genome-wide activities of each protein exhibit evidence of functional compensation, *de novo* binding, and loss-of-binding events in *Drosophila* embryos mutant for the orthologous gene. That is, in *SoxNeuro* mutants, Dichaete binding is shown to be stronger, novel, or absent at various loci when compared to wildtype. The same is true for SoxNeuro binding in *Dichaete* mutants (Figure 5.1.4). This extraordinary evidence for compensation, redundancy, and dependency builds upon previous research demonstrating their phenotypic functional redundancy during embryonic development (Buescher *et al*., 2002; Overton *et al*., 2002)

The purpose of my investigation was twofold. First, I wished to establish *Tribolium castaneum* as a model organism for genomics research. The genome of *T. castaneum* shares significant homology with *D. melanogaster* with a similar genome size and many orthologous regions (Richards *et al*., 2008) (Figure 5.1.5), however at present, there have been no genome-wide TF binding studies conducted in *Tribolium* embryos, despite proposals for such experiments first appearing 9 years ago (Roth & Hartenstein, 2008). ChIP-seq has been carried out in *Tribolium* larvae on a trans viral histone, CpBV-H4, which is endogenous to the parasitoid wasp *Cotesia plutellae* (Hepat *et al*., 2013). The researchers introduced CpBV-H4 to late-stage larvae of *T. castaneum* in order to investigate its involvement in epigenetic control of gene expression in eukaryotic organisms, exploring its effect on total transcript content via RNA-seq, and its incorporation sites in insect chromosomes. However, the ChIP-seq assay identified just 16 sites of interaction within the genome of *T. castaneum*, with no conserved target motif detectable (Hepat *et al*., 2013). ChIP-on-chip has been performed in *Tribolium* embryos, however this was only on a specific locus of 240kb, using a custom-made tiling array (Cande *et al*., 2009).

Because ChIP investigations rely on a robust and highly specific antibody (Gilmour & Lis, 1985; Orlando, 2000; Buck & Lieb, 2004), these experiments are less suitable for non-model organisms where the repertoire of specific polyclonal antibodies is less complete. DamID therefore represents an attractive method to study protein-DNA interactions in such genomes as it does not rely on antibodies. However, DamID presents challenges of its own: high expression of the Dam protein in eukaryotes is almost invariably toxic, and as such tolerance is poor (van Steensel & Henikoff, 2000; Southall *et al*., 2013). Consequently, low, 'leaky' levels of expression are only tolerated in the genomes of metazoans (Vogel *et al*., 2007; Southall *et al*., 2013), and a suitable promoter must be identified that will allow sufficient expression as to methylate the host genome in detectable quantities, whilst simultaneously avoiding toxicity and saturation of methylation (Vogel *et al*., 2007). DamID is a well-established technique in *Drosophila*, however to the best of my knowledge it has yet to be established in another arthropod species.

Cytosine methylation has been identified in both *Tribolium* (Feliciello *et al*., 2013; Song *et al*., 2017) and *Drosophila* (see Takayama *et al*., 2014): in each species, both symmetrical and non-CpG methylation is observable, with the methylome revealing novel and unique methylation patterns in the animal kingdom which function in contrast to the methylomes of vertebrates (Song *et al*., 2017). Until recently, it was assumed that adenine methylation did not occur in metazoans (Luo *et al*., 2015; Zhang *et al*., 2015; Wu *et al*., 2016). Recent investigations have confounded this, however, detecting N^6-methyladenine (6mA) presence in mouse embryonic stem cells (Wu *et al*., 2016), in *C. elegans* (Greer *et al*., 2015), and in *D. melanogaster* (Zhang *et al*., 2015). Nonetheless, 6mA methylation is present within specific motifs in *C. elegans*, none of which are GATC regions; this implies that DamID is still a suitable methodology to identify TF binding in metazoans as the protocol enriches GA^{m}TC fragments only.

The second aim of this investigation was to examine the genome-wide activity of the Tc-Dichaete and Tc-SoxNeuro proteins in *Tribolium castaneum*. Unlike *Drosophila melanogaster*, regulatory regions within *T. castaneum* are not well-represented in published genome annotations. However, the binding of Dichaete and SoxNeuro within the *Drosophila* genome map to gene loci and are most often associated with mapped regulatory elements (Ferrero *et al*., 2014; Carl & Russell, 2015), therefore analyses identifying Sox-bound genes ought to be possible in the *Tribolium* genome. Moreover, these experiments will identify potential regulatory elements within the *Tribolium* genome that can be explored further using enhancer trap lines (Trauner *et al*., 2009).

Within the beetle genome, I hoped to identify whether there was significant overlap between Tc-Dichaete and Tc-SoxNeuro binding, as is observed in *Drosophila*. The overlapping expression patterns of the transcripts of these two genes (discussed in Chapter 4) implies that there might be conserved function with *D. melanogaster*, in which Dichaete and SoxNeuro binding does overlap (Ferrero *et al*., 2014; Carl & Russell, 2015). This suggests that there might be areas of common binding between the two proteins in *T. castaneum* also. Moreover, the speculated additional function of *Tc-Dichaete* in insect segmentation (Chapter 4; Clark & Peel, 2017) would lead one to predict unique binding with regions associated with *Tribolium* primary pair-rule genes (Clark & Peel, 2017). Evidence implicating functional properties for each protein would be obtained by performing gene ontology enrichment on the genes associated with Tc-Dichaete and Tc-SoxNeuro binding, elucidating whether enriched regions are associated most strongly with CNS development for both proteins as they are in *Drosophila*. Furthermore, I wished to identify *de novo* target motifs for both Tc-Dichaete and Tc-SoxNeuro within the beetle genome, and draw comparisons with the conserved motifs discovered in *Drosophila* species (Figure 5.1.3). The DNA-binding HMG domains of Tc-Dichaete and Tc-SoxNeuro are highly conserved with those of Dm-Dichaete and Dm-SoxNeuro (89.9% and 92.4% sequence identities, respectively), which suggests high conservation in the DNA binding mechanism: one might, therefore, predict very similar, if not identical, target motifs would be identified. Finally, by taking a selection of the gene loci associated with the strongest binding intervals for each protein, a comparison with the binding intervals observed in *Drosophila* may be drawn to quantify the level of conservation/divergence across 350 million years of evolution.

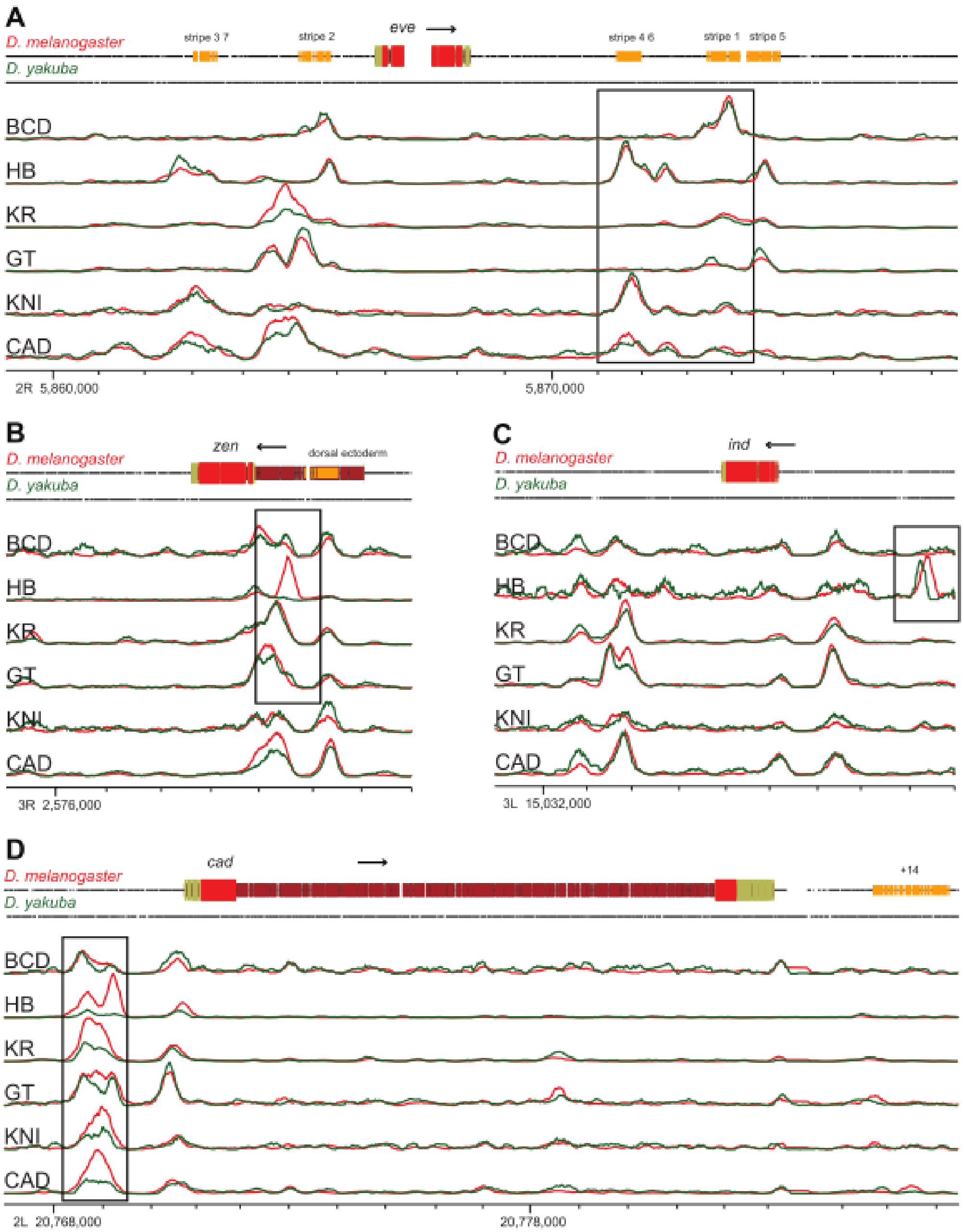


Figure 5.1.1. Binding intervals of 6 common transcription factors in two species of *Drosophila*: *D. melanogaster* and *D. yakuba*. (A) Common binding events, (B) Unique binding events, (C) Shift in binding peak, and (D) Change in binding intensity. BCD = Bicoid; HB = Hunchback; KR = Krüppel; GT = Giant; KNI = Knirps; CAD = Caudal. (Reproduced from Bradley *et al*., 2010.)

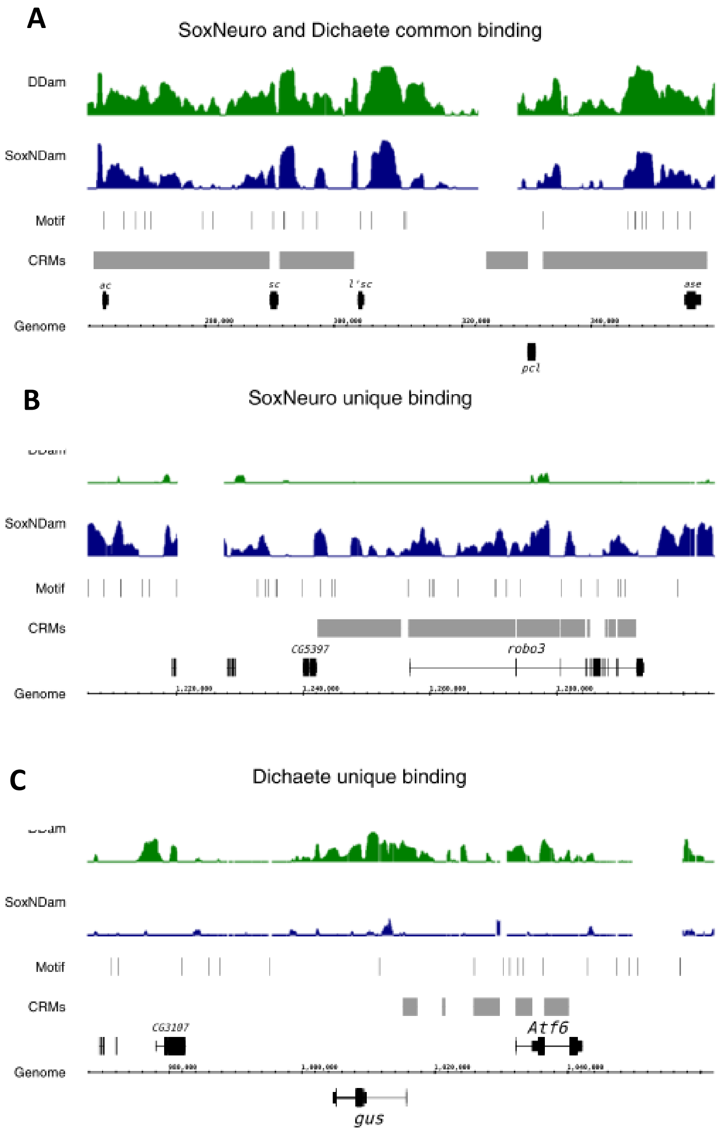


Figure 5.1.2. SoxNeuro and Dichaete binding profiles in *D. melanogaster* embryos (dark blue and dark green, respectively). Matches to the SoxN binding motif are displayed as thin bars, FlyLight and REDfly enhancers are displayed in light grey. (A) SoxN and Dichaete common binding at the *achaete-scute complex* (AS-C) locus. (B) Unique binding of SoxN across *robo3*. (C) Unique binding of Dichaete in the *gus* and *Atf6* region. (Figure reproduced from Ferrero *et al*., 2014.)

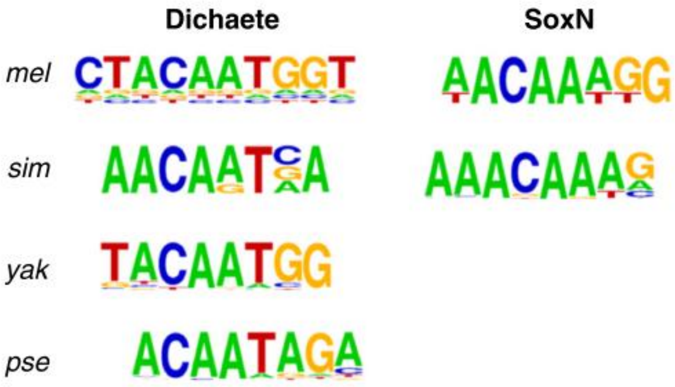


Figure 5.1.3. Target motifs identified for Dichaete and SoxNeuro proteins in different *Drosophila* species. *mel* = *melanogaster*; *sim* = *simulans*; *yak* = *yakuba*; *pse* = *pseudoobscura*. (Figure reproduced from Carl & Russell, 2015.)

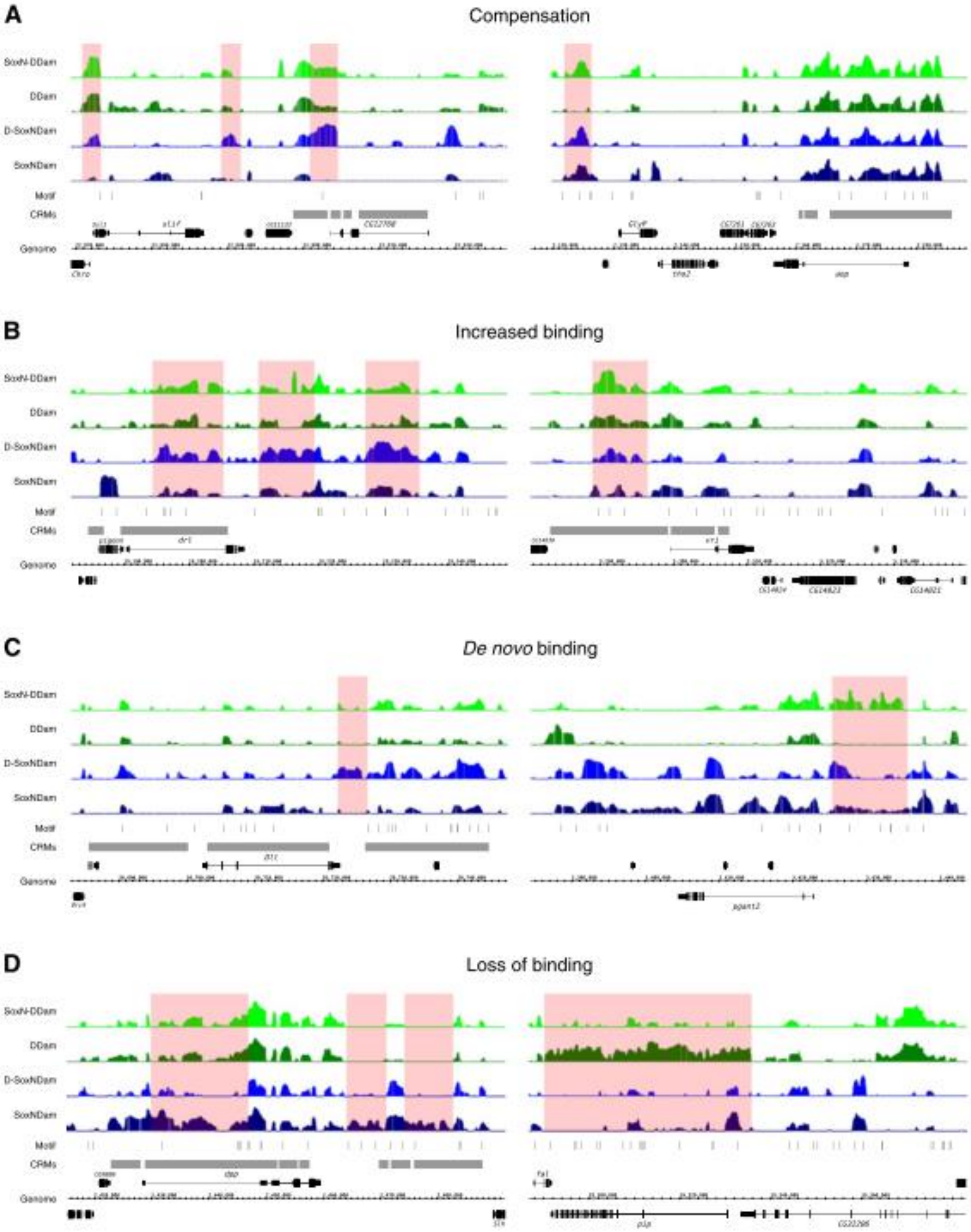


Figure 5.1.4. The redundancy of Dichaete and SoxNeuro is exhibited in *D. melanogaster* mutants. Green = Dichaete binding, light green = Dichaete binding in *SoxNeuro* mutants. Blue = SoxNeuro binding, light blue = SoxNeuro binding in *Dichaete* mutants. Instances of (A) Compensation, (B) increased binding, (C) *de novo* binding, and (D) loss of binding are highlighted in the red boxes. (Figure reproduced from Ferrero *et al*., 2014.)

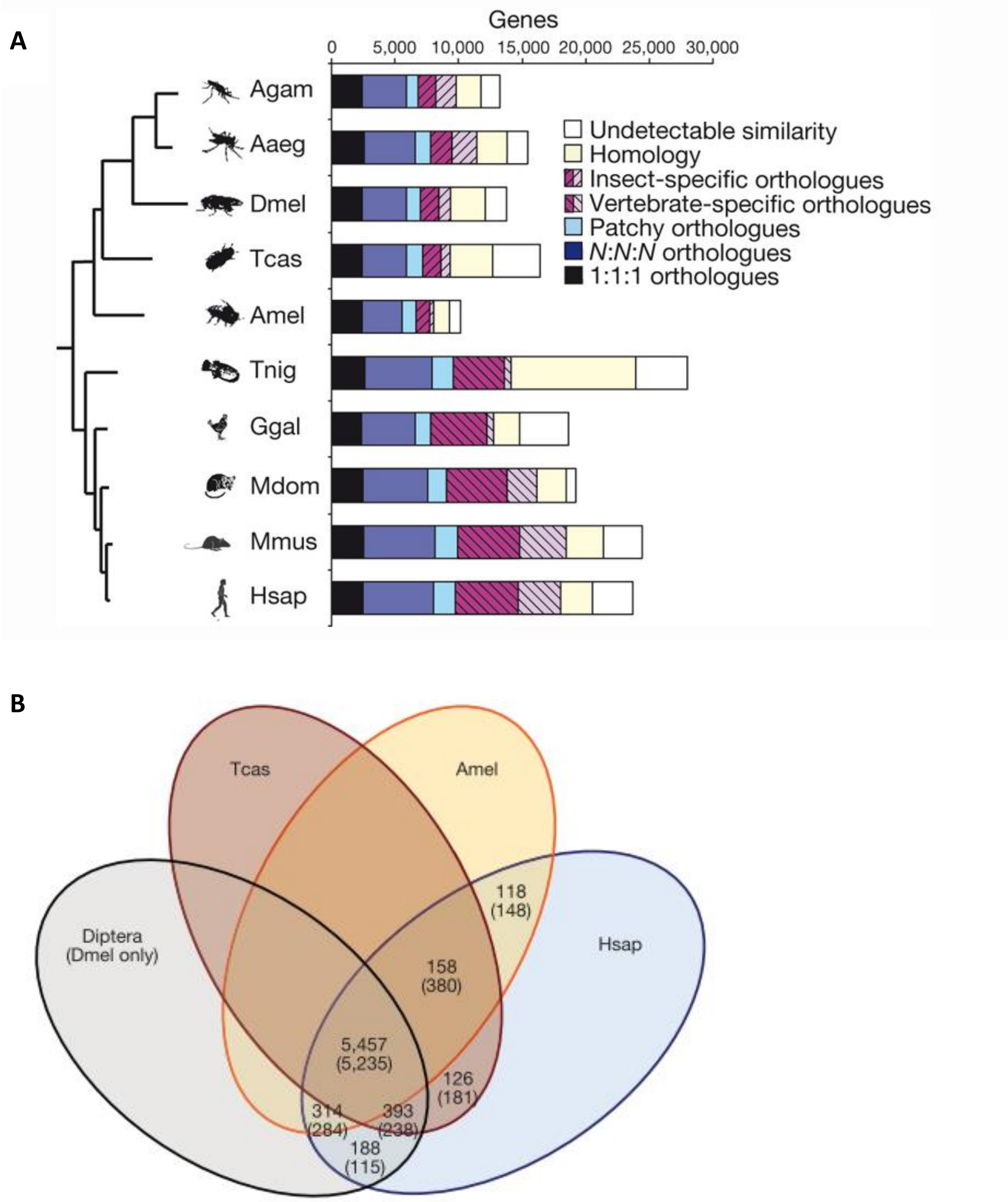


Figure 5.1.5. Gene orthology across metazoan genomes. (A) The conservation of genes by their similarities and numbers are characterised across different metazoan species. (B) Venn diagram showing the orthologous genes shared between 3 insect species and humans. *Drosophila melanogaster* and *Tribolium castaneum* share 5,473 orthologous genes. Agam = *Anopheles gambiae*, Aaeg = *Aedes aegypti*, Dmel = *Drosophila melanogaster*, Tcas = *Tribolium castaneum*, Amel = *Apis mellifera*, Tnig = *Tetraodon nigroviridis*, Ggal = *Gallus gallus*, Mdom = *Monodelphis domestica*, Mmus = *Mus musculus*, Hsap = *Homo sapiens*. The Diptera in (B) is represented here by *Anopheles gambiae*, *Aedes aegypti* and *Drosophila melanogaster* (with numbers considering only *D. melanogaster* shown in parentheses). (Figure reproduced from Richards *et al*., 2008.)

5.2 Feasibility assay

I first wished to assess the feasibility of performing DamID in *Tribolium castaneum*. The tethered Dam protein used in DamID will only methylate adenine in the context of GATC, and is thus dependent on the abundance of GATC sites in the target genome (van Steensel & Henikoff, 2000; Vogel *et al*., 2007). In the genome of *D. melanogaster*, GATC density provides sufficient resolution to map to nearby genomic features, and GATC sites have been reported to occur on average every ~200-300 base pairs (bp) (van Steensel & Henikoff, 2000). I therefore wished to calculate GATC distribution in *T. castaneum* and the average distances between GATC sites. Using an R script, I calculated the mean and median, and minimum and maximum, distances in bp between each GATC site, as well as the total number of occurrences. This R script was applied to the genomes of both *D. melanogaster* and *T. castaneum*, as well as 10 other arthropod species, in order to ascertain typical arthropod GATC occurrences (Figure 5.2.1, Table 5.2.1). For all arthropods, the mean occurrence of GATC sites is 480bp, whereas the median is 272bp. For *D. melanogaster*, the mean occurrence of GATC sites is 355bp, and the median 195bp; for *T. castaneum*, the mean is 567bp and the median 330bp. This suggests that GATC occurrence in *Tribolium* is less frequent on average than in *Drosophila* and most other arthropods. However, the methylation activity of tethered Dam proteins has been shown to act significantly up to ~2.5kb upstream or downstream from the TF binding site (van Steensel & Henikoff, 2000), and 98.3% of GATC sites in the *Tribolium* genome occur within 2.5kb of one another (99.35% for *Drosophila*, and 98.61% on average across arthropods – see Figure 5.2.1A). This suggests that GATC occurrence is more than sufficient in the genome of *T. castaneum* for DamID experiments.

There is also a significant negative correlation (R^2 = -0.699; *p* = 0.0115) between average GC content of the genome and the mean distances between GATC sites (Figure 5.2.1B). Whether this fully explains the differences in GATC distributions between arthropods remains to be determined: there are likely other factors influencing the relative GATC distributions such as repetitive elements, and selection pressures on the structural organisation of the genome, and the quality of the genome builds.

Out of curiosity, I also used the same R script to investigate the occurrence of the Dm-Dichaete and Dm-SoxNeuro target motifs (ACAATG and ACAAAG, respectively) in the *Tribolium* genome. I found that the Dm-Dichaete motif occurs 108,280 times, with a mean distance of 1505bp separating each motif, and the Dm-SoxNeuro motif occurs 128,445 times, with a mean distance of 1268bp separating each occurrence. Their occurrence in the *Drosophila melanogaster* genome is thus: the Dm-Dichaete motif occurs 96,273 times, with a mean distance of 1461b, and the Dm-SoxNeuro motif occurs 110,944 times, with a mean distance of 1268bp. The occurrences between each species are therefore very similar.

Next, I wished to test the feasibility of using *Dm-HSP70*, a basal promoter endogenous to *D. melanogaster*, to allow low level, 'leaky' expression in *T. castaneum in vivo*. This promoter has been used successfully in DamID experiments in different drosophilid species representing ~25 million years of divergence (Carl & Russell, 2015); although beetles and flies are separated by ~350 million years. Use of this promoter would have greatly streamlined the cloning required to generate *Tc-Dichaete-Dam* and *Tc-SoxNeuro-Dam* constructs. Berghammer *et al*. (2009) demonstrated that the *Drosophila HSP70* promoter was sufficient to provide basal promoter function in the GAL4/UAS system (Brand *et al*., 1993; Brand, 1994) in *Tribolium*, however, its efficacy can be inconsistent (Schinko *et al*., 2010). Using *piggyBac* constructs (Horn & Wimmer, 2000) previously generated for DamID in *D. melanogaster* by Dr Sarah Carl (Carl & Russell, 2015), I sought to generate transgenic lines with *Dm-SoxN-Dam* in *T. castaneum* (and a Dam-only negative control), which were under the control of the *HSP70* promoter and UAS sequences, in the absence of GAL4. This not only served as a useful pilot experiment, but I was also interested if the binding data of *Drosophila* SoxB was similar to that of *Tribolium* SoxB *in vivo*. I sought the assistance of Professor Gregor Bucher's expertise in *T. castaneum*, and Dr Julia Ulrich from Professor Bucher's lab assisted me in the microinjections of *T. castaneum* embryos, with the above *piggyBac* constructs. These microinjections proved unsuccessful, however; the results from this pilot study are shown in Table 5.2.2. Zero transgenic lines were obtained, and the survival rate to adulthood was extremely poor (<2% for each construct).

Professor Bucher and Dr Ulrich expressed surprise at the low survival rates (especially as Dr Ulrich had performed ~50% of the injections herself), and suggested that the *Dm-HSP70* promoter might be unsuitable for DamID experiments in *T. castaneum*. I was therefore directed towards using another promoter, *HSP68*, this time endogenous to *Tribolium*. This is a basal promoter that has been shown to efficiently and consistently perform with the GAL4/UAS system in *Tribolium* (Schinko *et al*., 2010). Using the principles of Gibson Assembly (Gibson *et al*., 2009), I assembled the following constructs with *Tribolium Dichaete* and *SoxNeuro*, and the basal *Tribolium* promoter *HSP68* and UAS: pBac[3xP3-EGFP;SV40;5xUAS-Tc-

Hsp68-Tc-Dichaete-Myc-Dam;SV40], pBac[3xP3-EGFP;SV40;5xUAS-Tc-Hsp68-Tc-SoxNeuro-Myc-Dam;SV40], and pBac[3xP3-EGFP;SV40;5xUAS-Tc-Hsp68-Dam-Myc;SV40] (see Figure 5.2.2).

Transgenesis is less well-established in *Tribolium* than it is for *Drosophila* (Berghammer *et al*., 2009), and balancer chromosomes are available for just ~30% of the *Tribolium* genome (Brown *et al*., 2009). However, using transposable elements as vectors along with helper plasmids achieves transgenesis at a comparable efficiency to *Drosophila* (Berghammer *et al*., 2009), and the *piggyBac* transposable element used by Dr Carl in her experiments is also effective in the beetle (Lorenzen *et al*., 2003). The dominant marker Enhanced Green Fluorescent Protein (EGFP) has been reported as a universal marker for insect transgenesis, and under the control of the *P3* promoter is expressed in the eyes (Berghammer *et al*., 1999). Together, *piggyBac*-mediated mutagenesis and the use of EGFP as a dominant marker eliminate the need for balancer chromosomes, as recombination and artificial selection can lead to allele fixation fairly rapidly. I therefore designed the Tc-Sox-Dam constructs in the *piggyBac* vector by cloning the respective SoxB loci, omitting their stop codons, and using the Myc tag (Terpe, 2002) to fuse the Sox and Dam proteins. *EGFP* was used as the reporter gene, and upstream of the *Sox-Dam* sequence was a 5xUAS sequence and the basal *Tribolium* promoter *HSP68*. These constructs were modelled on constructs successfully used for DamID in drosophilids by Dr Carl (Carl & Russell, 2015).

The constructs were then submitted to the Tribolium Genome Editing Service (TriGenES: http://trigenes.com) who performed the embryo microinjections and screenings. The TriGenES service experienced greater success with this new promoter, however survival rates were nonetheless well below those observed with the positive control (see Table 5.2.3). I was fortunate in that transgenic lines were obtained for all three constructs: for the *Dichaete-Dam* construct, 2 transgenic lines were obtained; for *SoxNeuro-Dam*, 3 lines, and for the *Dam*-only negative control, just 1 transgenic line. This was significantly lower than the positive control (just 0.1% of embryos injected with the *Dam*-only construct produced a transgenic line, in contrast to 2.8% with the positive control), suggesting that adenomethylation may indeed be poorly tolerated in *T. castaneum*.

Table 5.2.1. GATC occurrence across 12 arthropod species, ordered according to relatedness to *D. melanogaster*. *D. melanogaster* is highlighted in blue, and *T. castaneum* in orange.

Genome	Genome Size (Mb)	GC Content	Mean (bp)	Median (bp)	Min. (bp)	Max. (bp)	% <500bp	% <2500bp
D.mel	143.7	42.14%	355	195	1	54279	77.58	99.35
B.mor	481.8	37.80%	449	247	1	137280	70.94	98.91
H.mel	273.8	32.60%	617	399	1	24019	57.25	97.71
T.cas	165.9	35.19%	567	330	1	1201383	63.54	98.30
D.pon	252.8	38.45%	511	271	1	374131	69.48	97.38
A.cep	317.7	34.40%	381	205	1	21221	76.69	98.85
N.vit	295.8	43.21%	409	205	1	137182	78.08	99.30
A.pis	541.7	31.20%	630	369	1	86956	59.46	97.09
R.pro	702.6	37.10%	561	288	1	232941	68.13	99.07
Z.nev	485.0	38.60%	459	294	1	23158	67.99	99.21
D.pul	197.2	42.40%	408	210	1	143138	77.53	98.94
S.mar	176.2	35.80%	415	246	1	8202	71.60	99.27
Mean	336.18	37.41%	480	272	1	203657.50	69.86	98.61

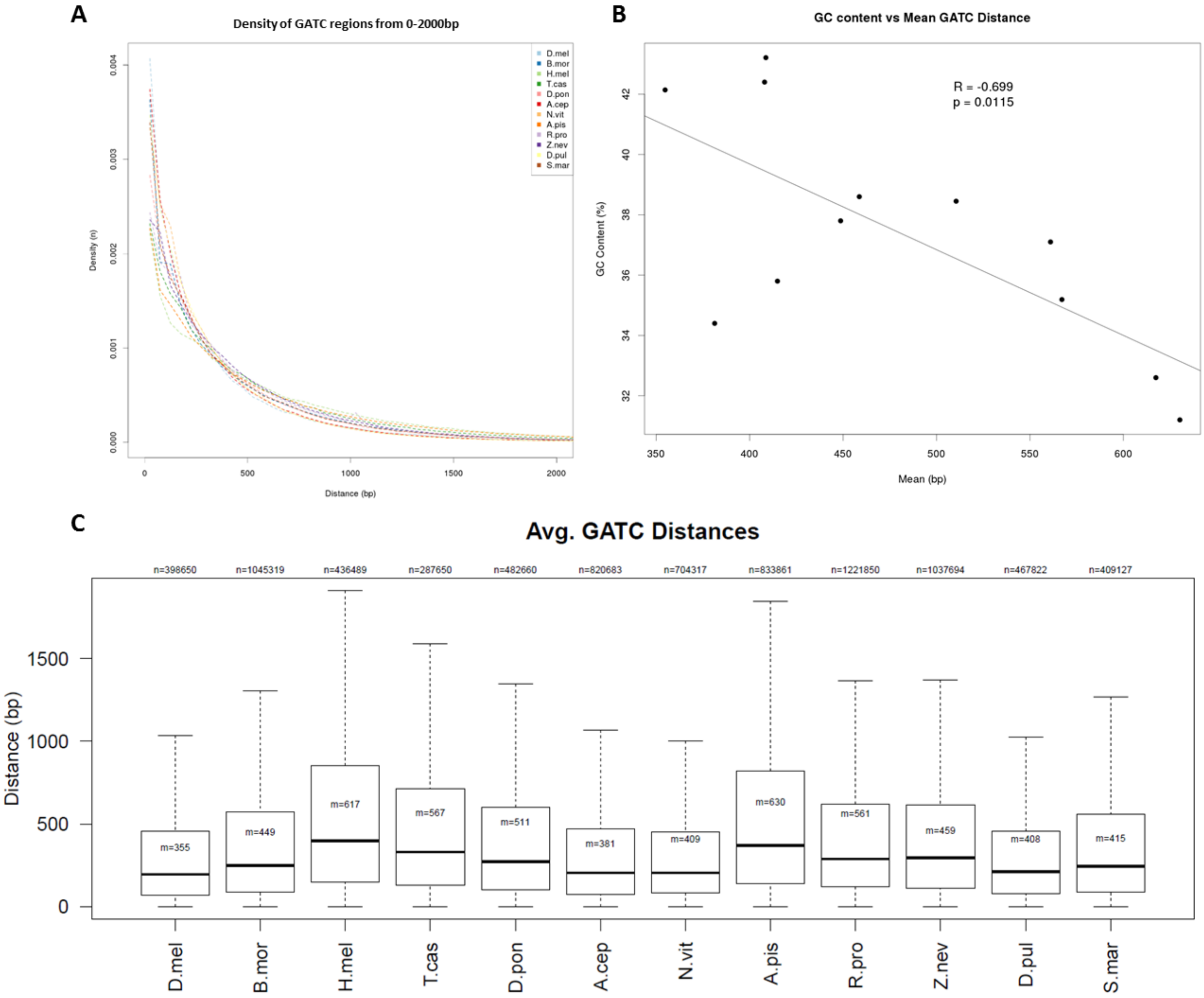


Figure 5.2.1. GATC occurrence and distance across 12 arthropod species. (A) Graph showing the density of GATC regions from 0-2000bp of the different arthropod species. (B) Scatterplot of GC content vs mean GATC distance. (C) Boxplot showing the average distances between GATC occurrences across 12 genomes. The median is represented as a solid black line, and the mean (m) displayed above it. The total number of occurrences are also given (n).

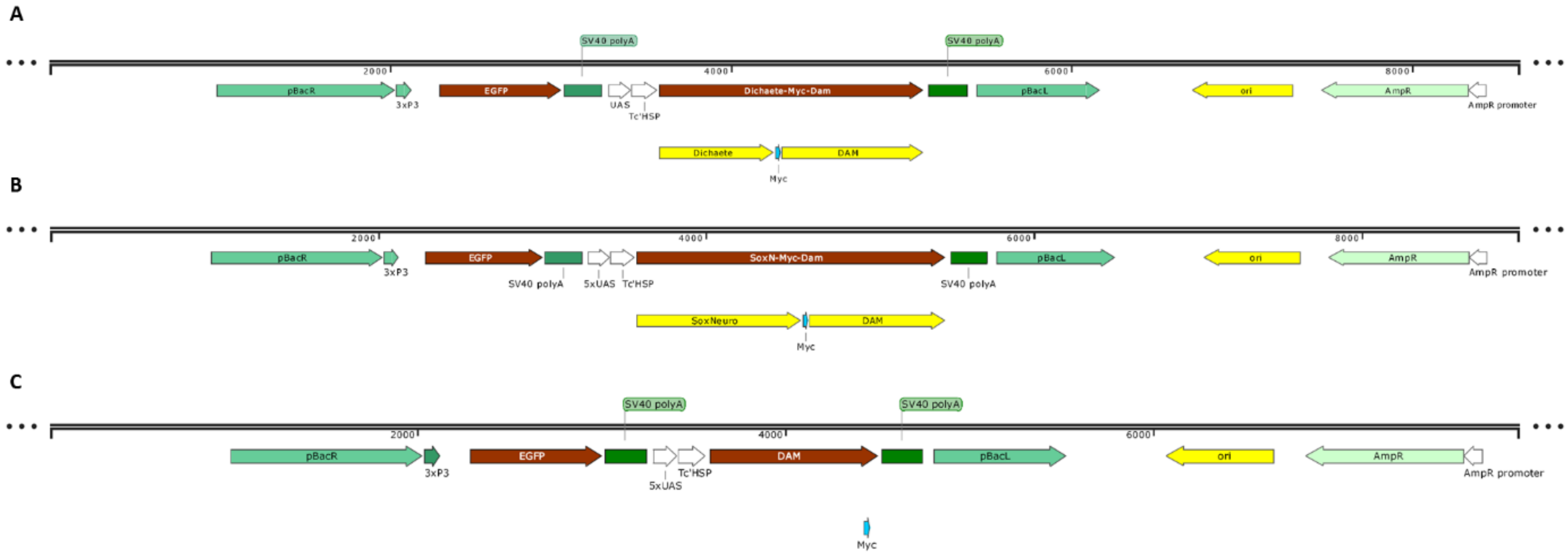


Figure 5.2.2 *piggyBac* constructs containing Tc-Sox-Dam fusions that were used for *T. castaneum* transgenesis. (A) The *Dichaete-Dam* construct pBac[3xP3-EGFP;SV40;5xUAS-Tc-Hsp68-Tc-Dichaete-Myc-Dam;SV40]; (B) The *SoxNeuro-Dam* construct pBac[3xP3-EGFP;SV40;5xUAS-Tc-Hsp68-Tc-SoxNeuro-Myc-Dam;SV40]; and (C) The *Dam*-only construct pBac[3xP3-EGFP;SV40;5xUAS-Tc-Hsp68-Dam-Myc;SV40].

Table 5.2.2. *Tribolium castaneum* microinjection table, using the *Drosophila melanogaster pigygyBac* constructs with the *Dm-HSP70* promoter.

	# Injected	# Hatched	% Hatched	# Adulthood	% Adulthood	# Transgenic	# Transgenic offspring	% Transgenic of injected	% Transgenic of hatched	% Transgenic of adults
Dm-SoxN-Dam	1700	71	4.18	27	1.59	0	0	0	0	0
Dm-Dam	2300	82	3.57	39	1.70	0	0	0	0	0

Table 5.2.3. *Tribolium castaneum* microinjection table, using the *Tribolium castaneum pigygyBac* constructs with the *Tc-HSP68* promoter.

	# Injected	# Hatched	% Hatched	# Adulthood	% Adulthood	# Transgenic	# Transgenic offspring	% Transgenic of injected	% Transgenic of hatched	% Transgenic of adults
Dichaete-Dam	1010	285	28.22	208	20.59	5	2	0.20	0.70	0.96
SoxN-Dam	250	133	53.20	103	41.20	4	3	1.20	2.26	2.91
Dam	1010	316	31.29	116	11.49	1	1	0.10	0.32	0.86
Positive Control	250	142	56.80	118	47.20	7	7	2.80	4.93	5.93

5.3 DamID: Attempt 1

Transgenic populations were given time to grow and expand in optimum growth conditions at 32°C, with fresh media being regularly administered. However, adults from these populations appeared in poor health, and their lifespan was not comparable to wildtype beetles; population growth was consequently substantially slower than would be expected with wildtype populations. Given the absence of balancer chromosomes for *T. castaneum*, populations had to be continuously monitored for GFP expression to prevent allele loss; eventually, when populations were of a satisfactory size, GFP-expressing adults were positively selected until allele fixation. From these allele-fixed populations, three smaller populations were selected for the purpose of creating distinct replicates, and the parent population kept at 25°C for reserve stock purposes.

A pilot experiment was conducted on each population, taking 100 µl of embryos laid by each parent population, and one wildtype negative control (see Chapter 2.4 for embryo collection methodology), to determine methylation presence/absence. The protocol used is essentially as described in Vogel *et al*., (2007) with 17 cycles used for amplification (see Chapter 2.4.2-3 for genomic isolation and methylation enrichment methodology), where 'No *Dpn*I' and 'No T4 DNA ligase' samples were included as double negative controls, in order to establish the presence/absence of genomic adenomethylation (see Vogel *et al*., 2007). The expected result if adenine methylation is successful and amplification is optimal is a visible smear of DNA on agarose gel from 200bp-2kb in the experimental samples, and no DNA product in the negative controls. If amplification is not optimal, and over-amplification occurs, then larger background artefacts may also be amplified (Greil *et al*., 2006; Vogel *et al*., 2007).

This pilot study yielded mixed results (Figure 5.3.1); the *Dichaete-Dam* and *Dam* samples showed evidence of methylation, as evidenced by the DNA smears ranging from 250bp-1.5kb. However the *SoxNeuro-Dam* sample appeared the same as the wildtype sample and negative controls, especially upon further amplification, with smears characteristic of amplification of background artefacts (Figure 5.3.1B). Fortunately, there were 3 independent transgenic lines generated for *SoxNeuro-Dam*, and so another population was selected. This population exhibited evidence of methylation in a similar pattern to *Dichaete-Dam* and *Dam*, and thus was used for all subsequent experiments. This pilot study also helped optimise the number of amplification cycles required during the PCR step; over-amplification is to be avoided to limit background amplification effects (van Steensel & Henikoff, 2000; Vogel *et al*., 2007; Marshall *et al*., 2016).

I also performed a PCR amplification for each of the three transgenic lines to determine the presence of each transgene, and amplified product was submitted for Sanger sequencing to establish integrity of the inserts. For all three inserts, the sequences were identical to those of the constructs, suggesting that no mutations had occurred. (For the *Dichaete-Dam* and *SoxNeuro-Dam* constructs, synonymous mutations were initially detected in the cloning vectors, however these were likely due to population-level variation with the published gene sequences.)

Embryos were collected from the 3 populations for each transgenic line and methylation enrichment performed as described above, with 16 cycles of amplification (results in Figure 5.3.2). In this experiment, DNA smears of 250bp-1.5kb were present in all the experimental samples, and no product was detected in the negative controls or the wildtype sample. These results were promising and were characteristic of the successful results described by Vogel *et al*. (2007). Libraries were prepared using the ThruPLEX® DNA-seq Kit, with 10 cycles of PCR amplification. BioAnalyzer analysis on these samples revealed sharp peaks in fragments <250bp (Figure 5.3.2A), indicating that there might be concatemer formation and contamination present. Libraries were pooled into a multiplex, and a size-selection step was successfully carried out using Agencourt AMPure XP beads to remove these smaller fragments (Figure 5.3.3B). Samples were submitted CRUK Cambridge Institute Genomics Core for 50bp single-end-reads on an Illumina HiSeq 4000, with a 50% PhiX control included to smooth the low complexity arising from the DamID adapters present at the start of each sequence.

Sequencing yielded 393,401,639 total sequences. A bowtie (Langmead *et al*., 2009) index was generated for the 2016 *T. castaneum* 5.2 genome assembly, and each library had the adapter sequences trimmed using the cutadapt tool (Martin, 2011). The libraries were then aligned to the reference genome using bowtie v0.12.8 (Table 5.3.1). The results of this mapping proved extremely disappointing; less than 0.3% of reads from each library mapped to the reference genome. Mapped sam files were converted to bam files, sorted and indexed using samtools (Heng *et al*., 2009). Reads were converted to bed files and extended using BEDtools (Quinlan & Hall, 2010) according to average fragment length prior to the library preparation stage. Reads were then visualized by converting to wig and then bigwig file formats, and viewed using the

Integrated Genome Browser (IGB) (Freese *et al*., 2016). (Note: this pipeline is modified from Bardet *et al*. (2011)). The visualization revealed repeated sequences scattered throughout portions of the genome in no discernible pattern (Figure 5.3.4). These data proved unusable: typically, the binding data from the Dam-only control is 'subtracted' from the Dam-fusion data, and the differential binding is recognized as authentic binding of the TF (van Steensel & Henikoff, 2000; Vogel *et al*., 2007). Since there is virtually no overlap between the Dichaete-Dam fusion and Dam-only control, I was unable to assess this differential binding. Moreover, the reads that are present are merely narrow 'stacks'; a sequence being mapped repeatedly to the same locus.

A sample of 100 unmapped reads each from a replicate of *Dichaete-Dam*, *SoxN-Dam*, and *Dam* was queried using the Basic Local Alignment Search Tool (BLAST), and although some fungal and bacterial DNA was reported, no single species was significantly represented in the unmapped sequences, and BLAST alignment scores were nonetheless poor for those that were present. Finally, a FastQC analysis (http://bioinformatics.babraham.ac.uk/projects/fastqc/) was performed for each library, and this revealed significant contamination with overrepresented sequences – likely concatemers – which accounted for at least 22%-41% of each library (Table 5.3.2), thereby explaining the 'stacked' mapping observed in Figure 5.3.4.

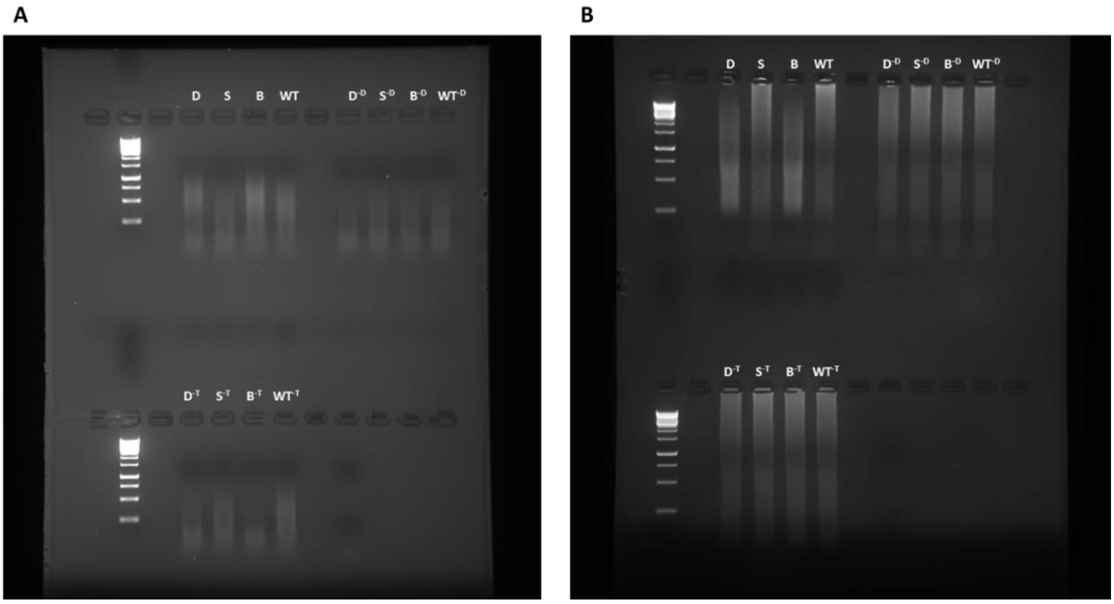


Figure 5.3.1. Gel electrophoresis of the pilot DamID enrichment of methylated DNA. (A) The DNA products generated after 17 cycles of amplification. (B) The products from (A) were amplified for a further 10 cycles to fully characterise background. D = *Dichaete-Dam*, S = *SoxNeuro-Dam*, B = *Dam*, WT = wild type; -D = no *Dpn*I negative control; -T = no T4 ligase negative control.

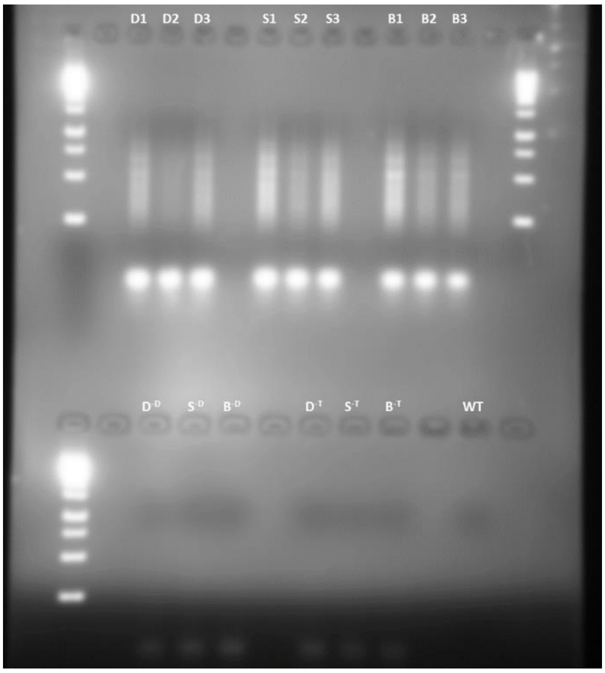


Figure 5.3.2. Gel electrophoresis of the DamID products used for library construction. D = *Dichaete-Dam*, S = *SoxNeuro-Dam*, B = *Dam*, WT = wild type; -D = no *Dpn*I negative control; -T = no T4 ligase negative control.

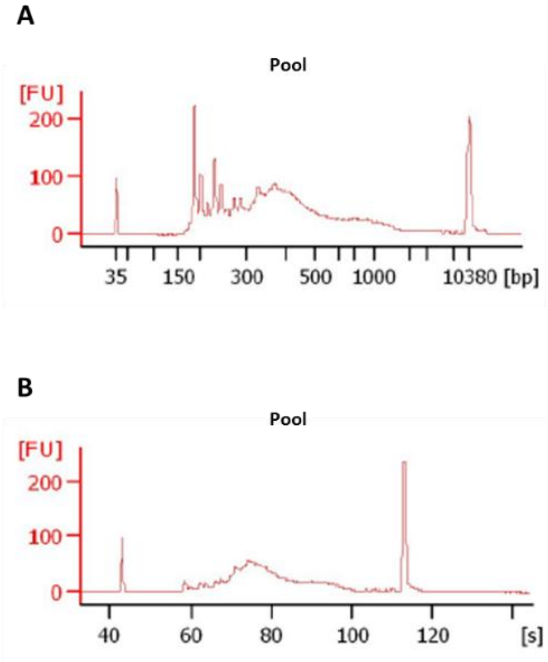


Figure 5.3.3. BioAnalyzer traces of pooled libraries. (A) Initially, libraries showed a sharp peak in fragments below 250bp. (B) A size selection step was performed to remove fragments below 250bp.

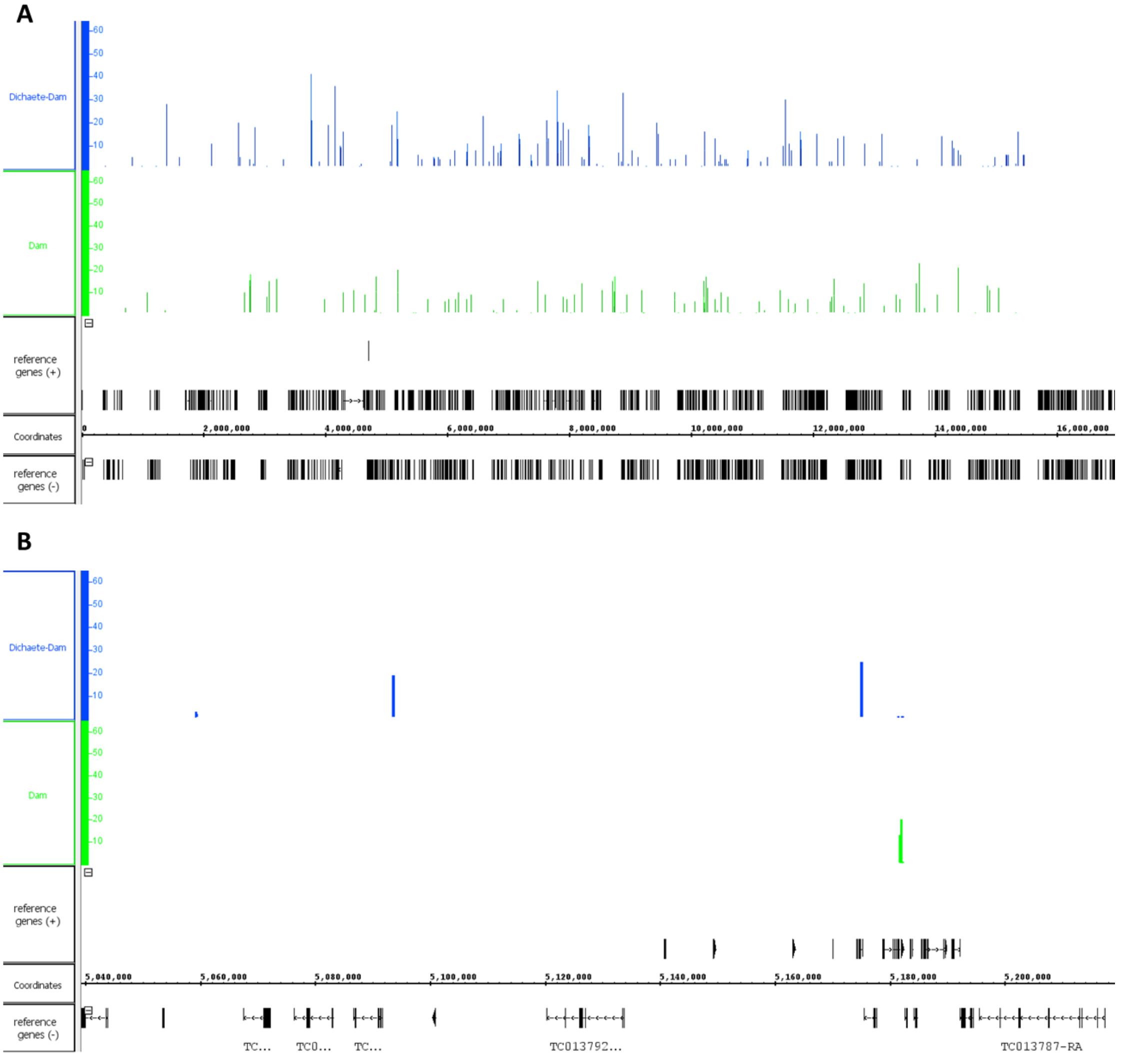


Figure 5.3.4. Screenshot of reads mapped to the *T. castaneum* genome, visualized using the Integrated Genome Browser (IGB; Freese *et al*., 2016). (A) Linkage Group 5 (LG5) of the *T. castaneum* genome. (B) Zoomed in screenshot of a locus on LG5 showing the dispersal of mapped reads as narrow stacks of the same sequence. Y axis = sequence read count.

Table 5.3.1. Bowtie mapping of libraries from DamID Attempt 1 to the *Tribolium castaneum* genome.

	Reads Processed	Reads with at least 1 reported alignment	Reads that failed to align
Dichaete_1	20530628	26292 (0.13%)	20504336 (99.87%)
Dichaete_2	22454454	47647 (0.21%)	22406807 (99.79%)
Dichaete_3	21505392	59296 (0.28%)	21446096 (99.72%)
SoxN_1	22095068	14877 (0.07%)	22080191 (99.93%)
SoxN_2	30282886	39494 (0.13%)	30243392 (99.87%)
SoxN_3	18487888	12673 (0.07%)	18475215 (99.93%)
Dam_1	19892114	16687 (0.08%)	19875427 (99.92%)
Dam_2	20669641	17747 (0.09%)	20651894 (99.91%)
Dam_3	15719624	10213 (0.06%)	15709411 (99.94%)

Table 5.3.2. FastQC report of overrepresented sequences (concatemers) present in each library from DamID Attempt 1.

	Reads Processed	Over-represented sequences (#)	Over-represented sequences (%)
Dichaete_1	20530628	8513669	41.47%
Dichaete_2	22454454	7943037	35.37%
Dichaete_3	21505392	6168404	28.68%
SoxN_1	22095068	6289200	28.46%
SoxN_2	30282886	9469977	31.27%
SoxN_3	18487888	4490494	24.29%
Dam_1	19892114	4372817	21.98%
Dam_2	20669641	5927960	28.68%
Dam_3	15719624	4291302	27.30%
Total	191637695	57466860	29.99%

5.4 DamID: Attempt 2

As concatemer contamination was deemed to be the issue chiefly responsible for unmapped reads, a new protocol was sought to eliminate this problem. Discussions with Dr Tony Southall revealed that the polymerase I had been using (Advantage cDNA Polymerase, Clontech) had also caused them issues with concatemer formation, and he instead suggested that I use the MyTaq polymerase (Bioline) system which yielded better results in their experiments (T. Southall, personal communication). Dr Southall further suggested I follow their protocol on targeted DamID (TaDa) for my isolation and amplification steps which was published after I had finished my first experiments (Marshall *et al*., 2016). This protocol introduces additional steps where cut DNA is passed through a size-selecting column, and the restriction enzyme *Alw*I is used to cleave the DamID amplification adapters prior to library preparation (and sequencing); previously I had removed the adapters *in silico*. I also elected to use fewer cycles of amplification – 15 during DamID amplification, and 5 during library preparation, to further attempt to mitigate any over-amplification effects – and include a wildtype library as a positive control for bowtie mapping.

The results from this experiment were more illuminating than the first attempt. FastQC analysis on each library proved more promising – concatemer contamination had been notably reduced, with over-represented sequences accounting for just 1.73% of all reads (Table 5.4.1). However, reads still failed to align to the *Tribolium* genome (Table 5.4.2), with the majority of the Dam libraries exhibiting <1% mapped reads. More concerning was the wildtype control, where only 16.68% of reads mapped to the *Tribolium* genome. The Dichaete-Dam samples appeared to exhibit a higher percentage of mapped reads than the other DamID samples, and thus I performed a Student's T test to determine if the mean percentage of mapped reads for Dichaete-Dam and SoxNeuro-Dam differed significantly; however they do not (p = 0.081343).

Upon examination of the unmapped sequences for the DamID samples, querying them with BLAST obtained similar results to those in the previous attempt – no species was notably more represented than others, and alignment scores were poor. However, querying the unmapped wildtype reads yielded a different story; the vast majority of the reads were reported as belonging to *Triticum aestivum*, the domesticated wheat species. This implicated a source of contamination: the medium used to rear the beetles is organic flour produced from wheat grain.

Despite attempts to mitigate flour contamination during embryo collections, which involved manually removing flour debris with a paintbrush and rinsing 3x with ddH_2O, wheat DNA appeared to be significantly overrepresented in DNA samples. The genome of the Chinese spring wheat variety has just been released earlier this year (see Clavijo *et al*., 2017). The wheat genome is 13,427,354,022bp in length, which is 81x larger than the 165,944,000bp long genome of *T. castaneum*. Moreover, wheat is hexaploid, whereas *T. castaneum* is diploid, so wheat has 3x as many copies of each locus as the beetle. This means that for a single beetle cell and a single wheat cell, there is 243x as much DNA present in the wheat cell. This represents a significant challenge for DNA isolation experiments in beetles (particularly in embryos which possess relatively few cells), and, prior to the recent publication of the wheat genome, appears to have remained unnoticed in *Tribolium* research until now.

A wheat index was generated using bowtie, and reads from each library aligned to it. The results are shown in Table 5.4.3. 77.25% of reads from the wildtype library map to the wheat genome. Together with the beetle alignment, 93.93% of reads map to the beetle or wheat genome, verifying that sterile conditions were achieved throughout the experiments. However, very few reads from the DamID samples aligned to the wheat genome (less than 1% in all cases). This is likely because non-GA^{m}TC DNA is removed during the DamID protocol, and there is no evidence suggesting adenomethylation occurs in wheat *in natura*.

I therefore hypothesized that if the vast majority of DNA in each sample was in fact wheat DNA, the input of methylated beetle DNA into the DamID protocol was insufficient, and indiscernible from any other background DNA present during PCR amplification. This hypothesis accounts for the fact that the majority of the unmapped reads from Attempts 1 & 2 do not exhibit notable association with any particular organism, and instead are likely low-level background from various DNA fragments contaminating the samples. If true, this low-level background likely had a comparable presence to any methylated beetle DNA, and was therefore comparably amplified.

Table 5.4.1. FastQC report of overrepresented sequences (concatemers) present in each library from DamID Attempt 2.

	Reads Processed	Over-represented sequences (#)	Over-represented sequences (%)
Dichaete_1	47243317	1505499	3.19%
Dichaete_2	37525180	744511	1.98%
Dichaete_3	32381347	1722633	5.32%
SoxN_1	35008284	163193	0.47%
SoxN_2	37666808	257643	0.68%
SoxN_3	33758970	567257	1.68%
Dam_1	24728239	85046	0.34%
Dam_2	33903946	271617	0.80%
Dam_3	30598114	735085	2.40%
Wildtype	36468766	0	0.00%
Total	349282971	6052484	1.73%

Table 5.4.2. Bowtie mapping of libraries from DamID Attempt 2 to the *Tribolium castaneum* genome.

	Reads Processed	Reads with at least 1 reported alignment	Reads that failed to align
Dichaete_1	47243317	922287 (1.95%)	46321030 (98.05%)
Dichaete_2	37525180	235483 (0.63%)	37289697 (99.37%)
Dichaete_3	32381347	441286 (1.36%)	31940061 (98.64%)
SoxN_1	35008284	19416 (0.06%)	34988868 (99.94%)
SoxN_2	37666808	22150 (0.06%)	37644658 (99.94%)
SoxN_3	33758970	38861 (0.12%)	33720109 (99.88%)
Dam_1	24728239	9989 (0.04%)	24718250 (99.96%)
Dam_2	33903946	14795 (0.04%)	33889151 (99.96%)
Dam_3	30598114	26203 (0.09%)	30571911 (99.91%)
Wildtype	36468766	6082372 (16.68%)	30386394 (83.32%)

Table 5.4.3. Bowtie mapping of libraries from DamID Attempt 2 to the *Triticum aestivum* genome.

	Reads Processed	Reads with at least 1 reported alignment	Reads that failed to align
Dichaete_1	47243317	340276 (0.72%)	46903041 (99.28%)
Dichaete_2	37525180	293194 (0.78%)	37231986 (99.22%)
Dichaete_3	32381347	238358 (0.74%)	32142989 (99.26%)
SoxN_1	35008284	120557 (0.34%)	34887727 (99.66%)
SoxN_2	37666808	143245 (0.38%)	37523563 (99.62%)
SoxN_3	33758970	183749 (0.54%)	33575221 (99.46%)
Dam_1	24728239	54791 (0.22%)	24673448 (99.78%)
Dam_2	33903946	86580 (0.26%)	33817366 (99.74%)
Dam_3	30598114	125985 (0.41%)	30472129 (99.59%)
Wildtype	36468766	28171313 (77.25%)	8297453 (22.75%)

5.5 DamID: Attempt 3

I sought to perform one final attempt at DamID in *Tribolium castaneum*. The egg collections that were performed for the previous 2 attempts took a significant amount of time due to the poor health of the transgenic populations and comparatively low fecundity. As I was nearing the completion of my funding at this point, and did not feel I had sufficient time to collect enough biological material for another experiment with embryos, I instead elected to perform the experiments with adult heads. The purpose of this investigation was to illuminate the binding properties of Tc-Dichaete and Tc-SoxN in the central nervous system, and to establish DamID as a resource for genomic studies in the beetle. Since DamID identifies historical binding events in the genome, genomic DNA from adults should still possess the methylation signatures of these binding events during embryonic development (and subsequent stages of the beetle life cycle). Moreover, beetle heads were chosen as they contain substantial amounts of CNS tissue, a known site of *SoxNeuro* and *Dichaete* expression, and very little of the digestive tract, which likely contains a substantial amount of wheat flour. Finally, adult beetles possess significantly more cells than embryos, maximising potential DNA yields.

I therefore dissected the heads of ~200 adults from each transgenic line (and a wildtype population as a control); however, there was insufficient material for replicates. To mitigate wheat flour contamination, I removed residual flour with a paintbrush, and thoroughly rinsed the heads 3x in ddH$_2$O, and then added additional washes 3x in 100% ethanol, in an attempt to wash clear any flour adhering to beetle mouthparts and the surface of the exoskeleton. DNA was then extracted from these samples, and processed according to the same DamID protocol used in Section 5.4, which had successfully diminished concatemer contamination effects. 100% of the DNA from each of the DamID samples was used at each step of the DamID protocol in an attempt to maximize DNA input.

Prior to submitting the samples for a final, and expensive round of sequencing, I sought to test the extent of wheat as a contaminating factor in the DamID investigations by performing a quantitative real-time PCR. I selected the *TaRca2-α* locus of the wheat genome (Saeed *et al.*, 2016), common to several wheat strains, in order to maximize the likelihood that the (unknown) wheat strain used in my flour medium contained the target amplicon. Primers were designed to generate a 134bp amplicon from the wheat genome, and the *Dichaete* locus of the

T. castaneum genome was selected to target a 139bp amplicon (see primers used in Chapter 2.4.1).

The samples I tested were as follows: 1 replicate for the genomic DNA from embryos of Dichaete-Dam, SoxN-Dam, and Dam-only, and wildtype embryonic gDNA; each of these samples were used in the second experimental attempt described in Section 5.4. I also included the genomic DNA isolated from wildtype adult heads, which were prepared in parallel to the genomic DNA from the adult DamID samples. Serial dilutions of the target amplicons for both beetle and wheat DNA were performed in order to generate a standard curve.

The results of this experiment are shown in Figure 5.5.1. The qPCR revealed that wheat DNA is indeed present in detectable quantities in each sample. The Dichaete-Dam replicate shows approximately 2.5-3.7x as much target DNA present for the beetle amplicon (Figure 5.5.1A). This may be reflected by the fact that the amplicon used is from the *Dichaete* locus, and within the Dichaete-Dam transgenic line, there would be 2 such loci present in the genome. However, more promising was the fact that the heads from wildtype adults exhibited substantial presence of the beetle amplicon, and very low presence of the wheat amplicon (Figure 5.5.1A-B).

However, as these two amplicons are of comparable sizes (134bp and 139bp); this does not reflect the true quantities of wheat and beetle of DNA present, as the wheat genome is 81x as large as the beetle genome (discussed in Section 5.4). The size of the respective amplicon relative to the size of the entire genome can therefore be used to calculate total DNA presence. Once the total DNA presence is calculated for each species, the relative percentages of wheat and beetle DNA content for each sample can be calculated.

The following equation was used to determine total quantity of genomic DNA (***Q****t*) in each sample, for wheat and beetle DNA respectively: $\boldsymbol{Q_t = Q_a * \frac{G}{A}}$ where ***Q****a* = the quantity of amplicon present (pg), ***G*** = total genome size (bp), and ***A*** = the amplicon size (bp). The relative percentages of each can then be calculated by dividing each ***Q****t* value with the starting concentration of the sample (***C***) and multiplying by 100: $\boldsymbol{\frac{Q_t}{C} * 100}$.

The results are summarised in Figure 5.5.2. These results demonstrate that, with the exception of the Dichaete-Dam sample, wheat DNA represents >90% of the total DNA for each of the embryo samples examined. This supports the hypothesis proposed at the end of Section 5.4 that wheat was a significant contaminating factor; beetle DNA is highly underrepresented in these samples, and consequently methylated DNA likely even more so.

Moreover, the finding that wheat DNA represents just 7.5% of total DNA in the heads of wildtype adults was very promising, as this DNA was isolated in in parallel with the DamID samples. I therefore decided to generate sequencing libraries for the 4 DNA samples isolated from adult heads, and submit them for one final round of Illumina sequencing.

However, the sequencing results disprove the hypothesis of wheat contamination being the chief confounding variable: once again, very few reads (<1%) mapped to the beetle genome from the DamID samples (Table 5.5.1). The percentage of mapped reads for the wildtype sample (85%) demonstrates that I successfully mitigated wheat as a contaminating factor, however, indicating that the more stringent washing conditions with 100% ethanol and the greater cell numbers present in the heads make a significant difference to DNA yields. With <1% of all reads mapping to the wheat genome (Table 5.5.2), and concatemers representing just 1.26% of the total reads across libraries, the troubleshooting experiments detailed here are shown to have been successful in mitigating contaminating factors.

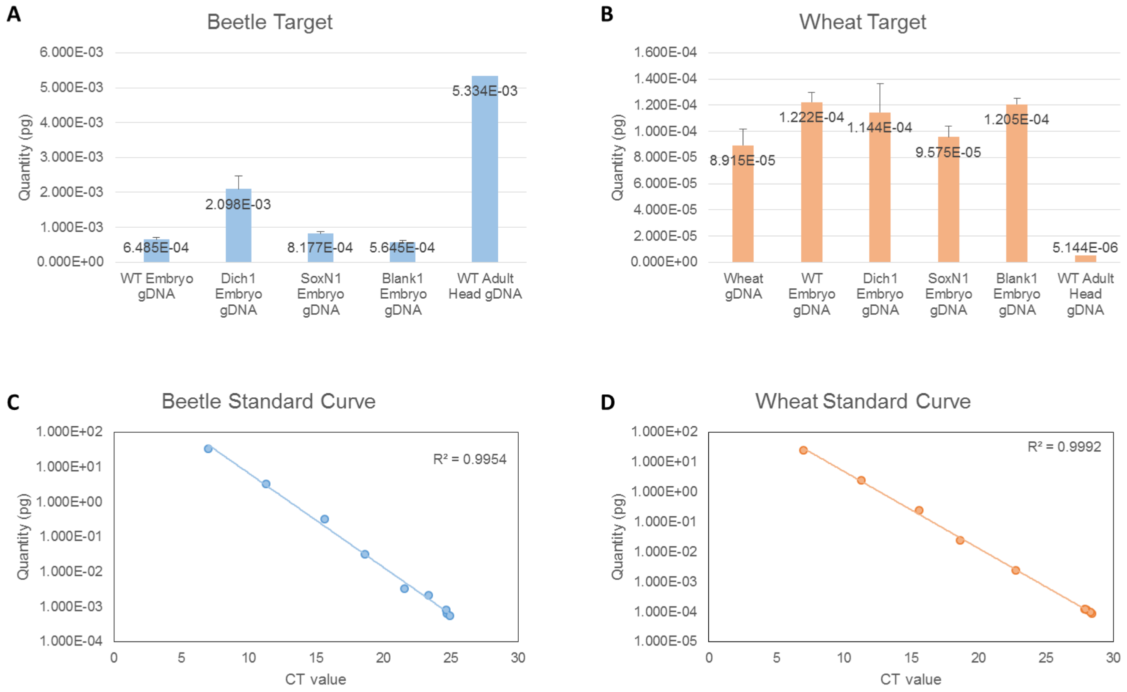


Figure 5.5.1. qRT-PCR results in DamID samples using a beetle amplicon (A,C) and a wheat amplicon (B,D). The standard curve for these experiments was extremely significant, with R^2 values of 0.9954 and 0.9992, respectively, indicating high efficiency in PCR amplification.

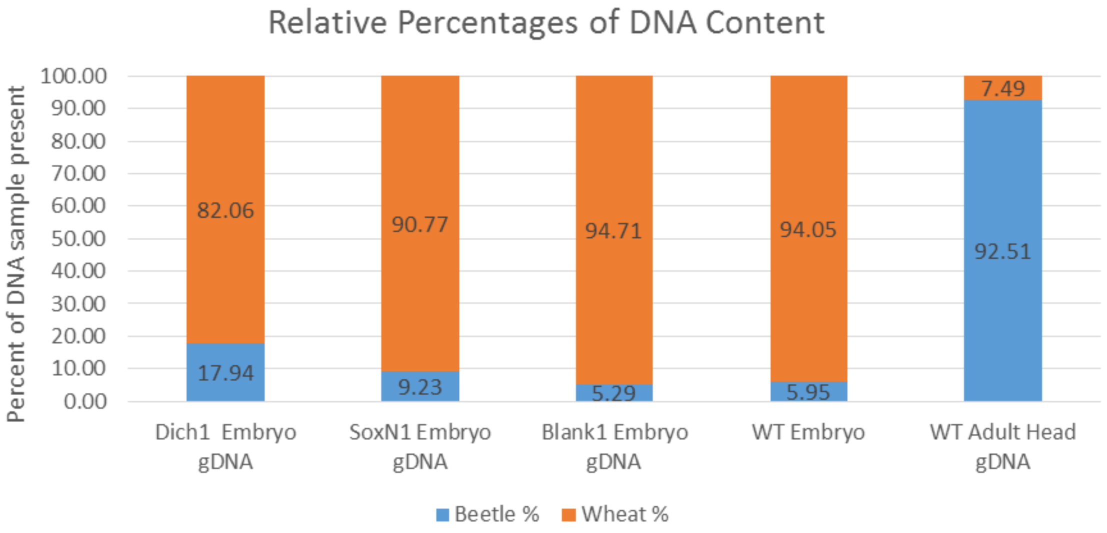


Figure 5.5.2. Relative percentages of total DNA content from wheat DNA (orange) and beetle DNA (blue) in DamID samples.

Table 5.5.1. Bowtie mapping of Adult Head (AH) libraries from DamID Attempt 3 to the *Tribolium castaneum* genome.

	Reads Processed	Reads with at least 1 reported alignment	Reads that failed to align
Dichaete_AH	75676064	424505 (0.56%)	75251559 (99.44%)
SoxN_AH	77394517	578618 (0.75%)	76815899 (99.25%)
Dam_AH	92126641	490723 (0.53%)	91635918 (99.47%)
WT_AH	97887504	83074213 (84.87%)	14813291 (15.13%)

Table 5.5.2. Bowtie mapping of Adult Head (AH) libraries from DamID Attempt 3 to the *Triticum aestivum* genome.

	Reads Processed	Reads with at least 1 reported alignment	Reads that failed to align
Dichaete_AH	75676064	186366 (0.25%)	75489698 (99.75%)
SoxN_AH	77394517	42632 (0.06%)	77351885 (99.94%)
Dam_AH	92126641	27453 (0.03%)	92099188 (99.97%)
WT_AH	97887504	941281 (0.96%)	96946223 (99.04%)

Table 5.5.3. FastQC report of overrepresented sequences in the Adult Head (AH) libraries from DamID Attempt 3.

	Reads Processed	Over-represented sequences (#)	Over-represented sequences (%)
Dichaete_AH	75676064	1651906	2.18%
SoxN_AH	77394517	810701	1.05%
Dam_AH	92126641	1853120	2.01%
WT_AH	97887504	0	0.00%
Total	343084726	4315727	1.26%

5.6 Discussion of results

In this investigation, I have attempted to establish the first genome-wide study of TFs in *Tribolium castaneum* embryos, and attempted the first use of DamID within arthropods beyond drosophilids. However, despite significant troubleshooting and several experimental attempts, I have been unsuccessful in achieving these aims. Nevertheless, DamID is a complex and sensitive technique, and the negative results generated here provide a significant contribution to establishing *Tribolium castaneum* as a future resource for genomic studies.

I have tested the feasibility of DamID in *T. castaneum* by demonstrating that GATC occurrence is sufficiently abundant for DamID experiments, and I have tested two different core promoters, one endogenous to *D. melanogaster* and the other to *T. castaneum*, for their suitability with *Sox-Dam* fusion transgenes. I have found that the *Drosophila* promoter *HSP70* may be unsuitable for DamID experiments, as injected embryos exhibited extremely poor survival rates and zero transgenic individuals. In contrast the *Tribolium* basal promoter *HSP68* appeared to be more compatible with the *Sox-Dam* fusions in yielding transgenic lines; however, these were still much more difficult to generate than would normally be expected from non-toxic constructs. This suggests that *Tribolium* are perhaps less tolerant of adenomethylation than *Drosophila*, in which DamID transgenesis is mostly routine with normal survival rates (S. Chan & S. Carl, personal communication).

Preliminary observations with these lines appeared to confirm that adenomethylation was present at some level, as the isolation and enrichment of DNA via methylation-sensitive enzymes produced distinct gel distributions when compared with the negative controls and wildtype DNA (Figure 5.3.2). In the protocol devised by Marshall *et al*. (2016), following digestion with the methylation-sensitive enzyme *Dpn*I, DNA is passed through a size-selection column meaning that only the methylated DNA cleaved by *Dpn*I should pass through, whereas uncut genomic DNA is left behind. Double selection of methylated DNA occurs with *Dpn*II, which in turn cleaves only unmethylated DNA, meaning it cannot be enriched by PCR amplification. Over-amplification can lead to background artefacts being amplified, as was observed in the initial pilot study (Figure 5.3.1), however this over-amplification has a distinct pattern on the gel that varies with the negative controls. These results lead me to conclude that some degree of adenomethylation was present in these beetle populations.

Concatemer formation also appeared to pose a significant challenge in Attempt 1 of the experiments, with FastQC reports indicating that concatemers represented anywhere between 22%-41% of these reads. These challenges were mitigated by the adoption of a different polymerase for the enrichment of methylated fragments, and also from the addition of a digestion step with the restriction enzyme *Alw*I, which recognises and cleaves the adapter sequence used for PCR amplification (Marshall *et al*., 2016). Fewer cycles of amplification also likely helped reduce concatemer formation. Together, these steps reduced concatemer contamination from ~30% of the total reads to just 1.73%.

Another important novel finding, however, comes from the second experimental attempt: wheat is a significant contaminating factor in DNA preparations from *T. castaneum*. Initially I had elected not to include a wildtype library in Attempt 1 as I wanted to guarantee sufficient read depth across the samples. However, by including the wildtype as a positive control in Attempt 2, I was able to determine that the vast majority of this DNA library was actually wheat DNA (77.25%). The wheat genome is 81x larger than the beetle genome, and is hexaploid as opposed to diploid, meaning that for each wheat cell, there is 243x as much DNA present compared to each beetle cell. Research in embryos is likely to be significantly more challenging as a consequence given their relatively low cell numbers. Future research should therefore attempt to eliminate flour medium from samples as much as possible via rigorous washing protocols, perhaps in ethanol or bleach. Maximising the starting quantity of DNA is also likely to aid in this.

These results may explain why the only genomic study to date in *T. castaneum* has been conducted in beetle larvae, which have many more cells than embryos; however the fact that the researchers identified just 16 binding sites in the genome (Hepat *et al*., 2013) leads one to speculate whether they too encountered similar problems in obtaining sufficient enrichment of genuine binding events in their investigations. In contrast, many RNA-seq studies have been performed in *T. castaneum* (*e.g.* see Schmitt-Engel *et al*., 2015), where contamination from wheat transcripts does not appear to have proven an issue. This is likely to due to the fact that RNA is far less stable than DNA; the preparation of flour from wheat grain involves both rigorous heat and mechanical conditions which may lead to the loss of RNA; and the sterilisation techniques used to prepare the flour as a medium for the beetles (-80^{o}C freezing or 80^{o}C heating) is likely to lead to a further loss of wheat transcripts.

That wheat was a confounding factor in the investigations discussed here is unquestionable; the vast majority of genomic DNA extracted from embryos from each population (~90%) proved to be wheat when examined via qRT-PCR. The quantity of input DNA to the DamID enrichment protocol is critical given you are positively selecting for methylated sequences only; lower input DNA means that the ratio of methylated DNA to inevitably-occurring background DNA is far lower, and thus they are amplified in similar proportions. Alternatively, when there is an abundance of adapter sequence present relative to genomic DNA, adapters are more likely to ligate to one another and be amplified at the expense of everything else, which explains the severe concatemer presence in each of the DamID samples in Attempt 1.

However, both of these confounding factors – concatemer formation and wheat contamination – were controlled for in the final DamID experiments I performed and yet they still proved unsuccessful. This suggests that there is likely to be a problem with the activity of the Sox-Dam fusion protein. The transgenesis itself was successful, as 100% of individuals exhibited GFP expression and the transgene was PCR amplified and sequenced for each population, showing that mutation had not occurred and that the transgene was intact. Positional effects of the insertion site may provide an explanation in terms of silencing the transgene (*e.g.* if the insertion occurred in regions of open or closed chromatin), although this is unlikely given the successful expression of EGFP. Moreover, the experiments for enrichment of methylated DNA indicate that methylation is indeed present. Determining the extent of this methylation is far more difficult, and it could be that the promoter, while not driving expression so much as to be toxic, may be too tight and only allows minute levels of expression, which are insufficient for detection via the protocols I used. In future, determining the expression levels of Sox-Dam fusions via an RT-PCR in both *Drosophila* and *Tribolium*, and comparing the two, might yield insights into whether expression is sufficient under the control of the *Tribolium* promoter.

Other speculative explanations for the lack of detection of methylated regions exist; *i.e.* the beetle *T. castaneum* might be biologically incompatible with DamID as a technique. Methylation is very poorly understand in invertebrates; it has only recently been shown to occur in insects (Feliciello *et al*., 2013; Takayama *et al*., 2014; Zhang *et al*., 2015), and insect CpG methylation is dissimilar to the better-understood mechanisms in vertebrates (Song *et al*., 2017). In *Drosophila*, the demethylating enzyme DMAD is responsible for maintaining low levels of adenomethylation in the genome (Zhang *et al*., 2015). It is plausible for example that

Tribolium may have evolved a genomic mechanism orthologous or independent to DMAD that demethylates the beetle genome to a greater extent than that which occurs in *Drosophila*. Perhaps *T. castaneum* is simply less tolerant to adenomethylation than *D. melanogaster*, even at minute levels, and that for the Sox-Dam fusion to sufficiently methylate the genome in an identifiable manner, such levels of methylation are toxic.

An alternative method might be to potentially detect DNA methylation directly. If methylation levels are insufficient to be enriched via traditional techniques, then directly sequencing genomic DNA to detect methylation might prove a more sensitive assay. Bisulfite sequencing is routinely used to detect 5-methylcytosine, however this technique relies on converting all unmethylated cytosine residues to uracil (reviewed by Fraga & Estella, 2002), and thus is inappropriate for detecting N6-methyladenine (Flusberg *et al*., 2010). An alternative method is to use single-molecule, real-time (SMRT) sequencing, which utilises the incorporation of fluorescent nucleotides. Fluorescent 'pulses' are measured at incorporation, which enables direct detection of modified nucleotides, and discrimination between N6-methyladenine, 5-methylcytosine, and 5-hydroxymethylcytosine modifications is possible (Flusberg *et al*., 2010).

In *T. castaneum*, cytosine methylation occurs in a mosaic pattern in the genome; hypermethylated regions are interspersed amongst larger unmethylated regions (Song *et al*., 2017), and overall cytosine methylation is very low (in contrast to vertebrates). Research by Song *et al*. (2017) used deep sequencing to detect methylation. However, the authors report that their data of *T. castaneum* methylation are likely incomplete, and that it may require sequencing at extreme depths to fully uncover the beetle methylome. Moreover, Zhang *et al*. (2015) report low-level adenomethylation in the *Drosophila* genome, although these have yet to be fully quantified. Collectively, these reports of low methylation levels suggest that SMRT sequencing is likely to unsuitable because, despite its single-molecule sensitivity (Flusberg *et al*., 2010), scaling it up to the extreme depth likely required may prove very challenging.

The detection of N^6-methyladenine (6mA) in *C. elegans* and *D. melanogaster* described in Section 5.1 was achieved using an antibody specifically against the 6mA base in DNA (Greer *et al*., 2014; Zhang *et al*., 2015). Zhang *et al*. (2015) and Greer *et al*. (2014) each used dot blot analyses with different 6mA antibodies to first detect whether adenomethylation was present, and subsequently used an extremely sensitive mass spectrometry assay on genomic samples to detect single base modifications via changes in mass (Yin *et al*., 2013). Future DamID experiments in non-model organisms could perhaps make use of a dot blot assay to independently confirm the presence of adenomethylation. The high sensitivity mass spectrometry assays would be unsuitable for providing sequence resolution of 6mA, however.

Instead, 6mA methylated DNA IP (MeDIP-seq) can be used to identify methylated adenine sites using the 6mA antibody (Greer *et al.*, 2015); although these would not discriminate between GATC adenomethylation and non-GATC adenomethylation. This might be possible to achieve *in silico*, however, as only sites containing GATC motifs could be selected. MeDIP-seq may therefore provide an alternative method for identifying TF-Dam binding sites. However, the DamID protocol requires a negative, Dam-only control, because of the high affinity of the Dam protein for DNA, and this will still need to be included in the MeDIP-seq experiments. Therefore, direct detection of methylated adenine regions is a potential alternative to the methods currently used in DamID protocols, and might also prove a useful method for independently confirming the presence of adenomethylation.

DamID is a sophisticated method for detecting historical binding events of TFs *in vivo*, however this investigation reveals that significant optimisation must be performed for it to be established in a non-model organism. The two promoters tested here may have each been ‘too hot’ and ‘too cold’ respectively, and as such the search should be widened to discover a ‘Goldilocks’ promoter, allowing just the right amount of expression. However, whether such a Goldilocks promoter exists is difficult to determine, as adenomethylation in *T. castaneum* may not be sufficiently tolerated to allow the broad methylation across the genome necessary for its identification. A more viable option for future genomics research may well be to develop new antibodies, and test the specificity and robustness of existing ones, so that ChIP experiments may be performed, or use existing antibodies to directly identify adenomethylation in the genome. Nonetheless, the findings from this study implicating wheat DNA as a significant contaminant originating from the flour medium will be indispensable in such genomic studies, and especially so where embryonic development is concerned. Therefore, the contributions from this research will hopefully aid future investigations that seek to establish *Tribolium castaneum* as a model organism for genomics research.

Chapter 6

Discussion & Future Directions

6.1 SoxB Evolution within the Bilateria

There has been considerable debate in the literature concerning the phylogenetic origins of SoxB paralogues within Bilateria (Bowles *et al*., 2000; McKimmie *et al*., 2005; Wilson & Dearden, 2008; Zhong *et al*., 2011). Determining the phylogeny of Sox genes is perhaps less straightforward than for other gene families, due to both highly and poorly conserved regions within each protein (Bowles *et al*., 2000), both of which can frustrate phylogenetic investigations (Goldman, 1998; Yang, 1998). The majority of work to date with Sox genes has focused on the amino acid sequence of their High Mobility Group-box (HMG) domain, classifying them according to orthology (Wegner, 1999; Bowles *et al*., 2000), and there is ongoing debate as to whether the subgroupings B1 and B2 that are found in vertebrate species can also be applied to insect SoxB. The two leading hypotheses for SoxB evolution are proposed by Bowles *et al*. (2000) and McKimmie *et al*. (2005), and subsequent work in Bilateria has provided support for each model (Wilson & Dearden, 2008; Zhong *et al*., 2011). The Bowles model (later reappraised by Zhong *et al*. (2011)) proposes that the two subgroups of vertebrate SoxB genes, B1 and B2, are ancestral to the emergence of the Bilateria: within the insects, *SoxNeuro* belongs to the B1 subgroup, and *Dichaete*, *Sox21a*, and *Sox21b* belong to the B2 subgroup, with *Dichaete* being the ancestral B2 gene. In contrast, the McKimmie model contends that the B1 and B2 subgroups are only applicable to vertebrates, and that *Dichaete* and *Sox21b* represent a unique subgroup within insect lineages, with four Group B genes present in insects. Each of these hypotheses rests upon two different signature residues within the HMG domains of SoxB genes for support, and recent analyses have yet to elucidate which model, if either, is valid. Zhong *et al*. (2011), for example, investigated a range of Bilaterian species and found support for the model proposed by Bowles; in contrast, recent investigations into various arthropod species seem to support the McKimmie model (Wilson & Dearden, 2008; S. Russell, unpublished data).

In light of this conflict, I have analysed the SoxB genes of 24 different metazoan species, and in 20 of the species I annotated the SoxB HMG domains myself. I tested these data against each of the above models and found that the signature residues identified by Bowles and Zhong are most representative across the Sox sequences included in my analysis. However, when examining the entirety of the HMG domain, neither model was more representative than the other, in terms of either amino acid conservation or R-group conservation. Phylogenetic tree construction was no more illuminating: when examining just the insect sequences, clustering behaviours support the McKimmie model, whereas when examining all Bilaterian sequences, clustering behaviours support the Zhong model.

Most phylogenetic research to date has focused on just the HMG domains of Sox proteins (Bowles *et al.*, 2000; McKimmie *et al.*, 2005; Wilson & Dearden, 2008; Zhong *et al.*, 2011); I therefore elected to include 20 amino acids upstream and downstream of the HMG domain to determine if there are other conserved regions that may have eluded prior research efforts. Through this, extra-HMG domain residues have been described for the first time within arthropods: a characteristic domain downstream of the HMG domain of Sox21b proteins has been identified. Moreover, a putative SOXp domain is reported for vertebrate Sox14, Sox21, Sox2, and Sox3 proteins (Gao *et al.*, 2013; Gao *et al.*, 2015a; Gao *et al.*, 2015b). This domain is also found in the SoxNeuro proteins of the arthropods examined here, and a strikingly similar domain also appears in the arthropod Sox21a proteins. This appears to support the McKimmie model of *Sox21a* and *SoxNeuro* forming a distinct subgroup within arthropod SoxB genes.

There are further assumptions that each model proposes that have been tested in this investigation. For example, the Zhong model suggests that just two genes, representing the B1 and B2 subgroups, were present at the deuterostome/protostome split; the McKimmie model proposes that there were three. Analysis in two basally branching protostomes, *Caenorhabditis elegans* and *Hypsibius dujardini*, reveals just two SoxB genes, which each cluster with the vertebrate B1 and B2 subgroupings, thereby supporting the Zhong model. The absence of any other SoxB genes in these species may, however, be due to gene loss events, but there is presently no evidence for this having occurred.

That there is contradictory evidence available for each of the models suggests that they might be insufficient to fully explain SoxB evolution. I have therefore proposed a new model for the evolution of SoxB genes within arthropods which reconciles aspects of each previous model. In this model, the B1 and B2 subgroups are indeed present at the deuterostome/protostome split, but I suggest a different candidate for the B2 ancestral gene: *Sox21a*. This is in light of the existence of the SOXp domain in both the SoxNeuro and Sox21a proteins of arthropods, and explains why SoxNeuro and Sox21a tend to cluster together in insect phylogenetic trees. Moreover, the SoxB proteins of *C. elegans* and *H. dujardini* that cluster within the B2 branch also most closely cluster with arthropod Sox21a, and share an intron at a very similar location to arthropod *Sox21a* genes, suggesting orthology. Finally, the expansion of SoxB genes appears to have occurred very early within the arthropod lineage, prior to their diversification, as all arthropods possess at least four SoxB genes.

Whether the B1 and B2 subgroups apply to arthropods in terms of function remains to be addressed. In vertebrates, the B1 and B2 subgroupings also perform homologous functions as well as being more closely related (Bowles *et al*., 2000; Lefebvre *et al*., 2007; Guth & Wegner, 2008), yet it is unclear whether such functional subgroupings exist in the insect *D. melanogaster* (McKimmie *et al*., 2005). Since the paralogues of protostome SoxB appear to have emerged in the arthropods, this implies that they do not share orthology with vertebrate SoxB, but instead, as the McKimmie model proposes, do indeed represent a unique lineage.

Nonetheless, the function of these genes in other arthropods needs to be further examined to be confident of this. Some preliminary work has been carried out in chelicerates (S. Russell, unpublished data) and myriapods (M. Akam, unpublished data) examining the expression patterns of SoxB genes. Functional studies in these species will hopefully illuminate whether *D. melanogaster* is representative of arthropods regarding SoxB function. One may also wish to investigate the function of Group B genes in more basally branching deuterostomes. For example, studying the function of SoxB TFs in non-vertebrate chordates in terms of activator/repressor activity would elucidate whether vertebrate SoxB functional subgroupings are representative of the chordate phylum.

Further efforts to annotate the SoxB genes of other Bilateria need to be made in order to test the model proposed in this study. For example, is the SOXp domain present in other arthropod species? For the conserved domain analysis I performed, I used sequences from a limited sample of species due to the relative quality of assemblies; for several species in my analysis, it was difficult to identify whole protein sequences because of incomplete sequences or shotgun fragments. As the quality of genome assemblies improves for these species, it should be possible to better annotate SoxB ORFs and investigate orthologous features. Moreover, an increasing number of genomes are being publicly released, which will only enrich our understanding of Sox evolution. For example, as mentioned above, the protostomes *C. elegans* and *H. dujardini* may have lost a SoxB gene, thereby confounding the model put forward here. Analysing the genomes of more basally branching protostomes might help elucidate this, especially in nematode worms. There are presently 11 nematode genomes available on EnsemblMetazoa (Kersey *et al*., 2016), as well as the genomes of two annelid worms, three molluscs, one brachiopod, and one platyhelminth. Characterising the SoxB genes across these

Bilateria would greatly augment our understanding of SoxB evolution and expansion, and new models of SoxB emergence are likely to be proposed incorporating each of these lineages. Expanding this to the Radiata could complete the puzzle; at present, there are six genomes available for Radiata species on EnsemblMetazoa.

Finally, in this investigation, I have focused solely on SoxB evolution within the arthropods, in an attempt to resolve the conflicting models for this group's evolutionary emergence. However, there are other core groups of Sox genes: B, C and E are present across metazoans (van de Wetering *et al*., 1993; Wright *et al*., 1993; Meyer *et al*., 1996), and Bilateria contain groups B through to F (Bowles *et al*., 2000). Understanding the early expansion of these groups is an exciting area for future research, as this family of genes appear to have played an indispensable role in the evolution of multicellularity in metazoans (Phochanukul & Russell, 2010).

6.2 *Dichaete* and *SoxNeuro* in *Tribolium castaneum*

Understanding Sox genes is not only important at the higher evolutionary level discussed above, but it is also important at the species level. Two of the insect SoxB genes, *Dichaete* and *SoxNeuro*, have been extensively studied in *D. melanogaster*, revealing a critical role in the patterning the early development of fly embryos (Nambu & Nambu, 1996; Russell *et al*., 1996; Sánchez-Soriano & Russell, 1998; Cremazy *et al*., 2000; Buescher *et al*., 2002; Overton *et al*., 2002) and extensive binding profiles throughout the fly genome (Aleksic *et al*., 2013; Ferrero *et al*., 2014; Carl & Russell, 2015). Both of these genes are expressed in the developing nervous system, within the ventral nerve cord and brain anlagen (Buescher *et al*., 2002; Overton *et al*., 2002; Overton, 2003). *Dichaete* is expressed in midline glia and the medial and intermediate columns of the ventral neuroectoderm, and *SoxNeuro* is expressed in all three longitudinal columns: medial, intermediate, and lateral (Nambu & Nambu, 1996; Russell *et al*., 1996; Sánchez-Soriano & Russell, 1998; Cremazy *et al*., 2000; Buescher *et al*., 2002; Overton *et al*., 2002). These genes have also been shown to be partially redundant in function, with milder phenotypes in single mutants than in double mutants (Buescher *et al*., 2002; Overton *et al*. 2002). However, *Dichaete* and *SoxNeuro* have scarcely been studied in other arthropods beyond *Drosophila*, whereas the other columnar genes, *Egfr*, *vnd*, *ind*, and *msh*, have (Wheeler *et al*., 2005; Biffar & Stollewerk, 2014; Biffar & Stollewerk, 2015).

The second purpose of this project was to address the scarcity of research this by studying *Dichaete* and *SoxNeuro* in the Coleopteran species *Tribolium castaneum*. I began by attempting to characterise the expression patterns of these two genes in beetle embryos using whole-mount *in situ* hybridisation. I managed to generate a probe for *SoxNeuro* that yields good signal, however, for *Dichaete* I tried several probes which were unsuccessful. While optimising these probes, Clark & Peel (2017) published a study detailing *Dichaete* expression in *Tribolium* embryos within the context of insect segmentation, and thus the novelty of my efforts disappeared. I therefore elected to use their data in my analysis and reappraise the data in the context of central nervous system development.

In this investigation, I found that there is considerable conservation of both *Dichaete* and *SoxNeuro* expression between *Tribolium* and *Drosophila* embryos. Both genes are expressed in what appear to be overlapping regions of the ventral neuroectoderm and brain anlagen, and *Dichaete* appears to be implicated in the early segmentation process of *Tribolium* embryos via its expression in the posterior growth zone (Clark & Peel, 2017). *Tc-Dichaete* expression appears to have diverged partially with *Dm-Dichaete*, however, in that expression is observed in the ventral midline of *Drosophila* embryos; no such expression is observed in *Tribolium* embryos. Whether this represents the ancestral state or not cannot be determined without examining the expression of *Dichaete* in other species.

This research would be strengthened by examining *Tc-Dichaete* and *Tc-SoxNeuro* expression in the longitudinal columns of the neuroectoderm; while preliminary data may show that *Tc-SoxNeuro* expression extends more laterally than *Tc-Dichaete*, it is difficult to be confident of this without using expression markers to precisely identify the longitudinal columns. Double-staining embryos for *Tc-Dichaete* and *Tc-SoxNeuro*, perhaps using fluorescently labelled probes, would definitively determine the overlapping expression patterns of these genes. Double-staining with *Tc-msh*, which identifies the NBs in lateral columns of the neuroectoderm, would also aid in the identification of the expression domains of *Tc-Dichaete* and *Tc-SoxNeuro*.

Future studies should investigate the functional roles for each of these genes in *Tribolium* to fully determine whether there is conservation with their orthologues in *Drosophila*. This can be achieved using RNAi (Posnien *et al*., 2009; Schmitt-Engel *et al*., 2015), or CRISPR, which has recently been established in *Tribolium* (Gilles *et al*., 2015). There exists some preliminary data for these genes in a mass RNAi screen by the iBeetle project (Dönitz *et al*., 2015; Schmitt-Engel *et al*., 2015), however these need to be validated and explored further. Moreover, both single and double knock-downs/knock-outs ought to be performed to mitigate phenotype masking, given the redundancy of these two genes in *Drosophila* (Buescher *et al*., 2002; Overton *et al*. 2002). Such experiments could determine whether the expression of pair-rule genes in *Tribolium* are dependent on *Tc-Dichaete* (Clark & Peel, 2017) and whether double mutants have stronger phenotypes in the CNS than either single mutant. One would predict, for example, that the phenotype observed in midline glia in *Drosophila Dichaete* mutants would be absent in *Tribolium*.

Future work should also examine the roles of the other SoxB genes in *Tribolium*: *Sox21a*, *Sox21b*, and *SoxB5*. SoxB genes show conserved expression patterns in drosophilids (McKimmie *et al*., 2005), but whether this expression is conserved in more distant taxa is yet to be fully explored. *Sox21a* is expressed in the intestinal cells of *D. melanogaster* (Cremazy *et al*., 2001; McKimmie *et al*., 2005; Meng & Bitaeu, 2015), but in the honeybee *Apis mellifera*, is putatively expressed in the Malpighian tubule anlagen (Wilson & Dearden, 2008). Moreover, although further work is required to assess whether *Am-Dichaete* is a true pseudogene in *Apis mellifera*, it appears that the *Am-Sox21b* gene is expressed in the developing CNS (Wilson & Dearden, 2008), unlike in *Drosophila* where its orthologue is expressed in the embryonic intestinal cells and ventral epidermis (Fisher *et al*., 2012: FlyBase report). The above investigation into SoxB evolution has demonstrated that *Dichaete* and *Sox21b* are closely related paralogues; perhaps within the honeybee there existed redundancy between these genes and if *Dichaete* expression has indeed been lost, then has *Sox21b* expression and function evolved to compensate for this? By examining the role of *Sox21b* in *Tribolium*, we can determine whether its expression in the honeybee CNS is likely to be ancestral or derived.

Furthermore, the existence of a fifth SoxB gene in *Tribolium*, *SoxB5*, is proposed in the above model to be a paralogue of *Dichaete* and ancestral to the insects, at least as far back as the divergence from Isoptera. Examining *SoxB5* expression and function is therefore of great interest, as it appears to be the most recent paralogue of *Dichaete* within the arthropods. For example, is *SoxB5* more redundant with *Dichaete* than *SoxNeuro*? In *Drosophila*, the neo-functionalization observed for *Sox21a*, and the novel expression patterns observed for *Sox21b*, would imply that redundancy is not merely the facet of a recent duplication event. The evolutionary model proposed above suggests that the *Dichaete* gene is a more recent SoxB paralogue than *Sox21a*, which implies that the redundancy observed between *Dichaete* and

SoxNeuro in *Drosophila* might actually be an example of convergent evolution as opposed to sub-functionalization following gene duplication (Lynch & Force, 2000; Qian *et al*., 2010). Investigation into *Tribolium SoxB5* function may therefore elucidate the process of sub- or neo-functionalization within SoxB genes by providing an example of how relatively recent paralogues are maintained in the genome. The fact that *SoxB5* appears to have been independently lost in several insect lineages strengthens the hypothesis that sub- or neo-functionalization of paralogues is necessary to maintain their open reading frame, or otherwise they decay (Lynch & Force, 2000; Qian *et al*., 2010).

The above investigation into arthropod SoxB evolution has also revealed the ancientness of the so-called '*Dichaete* cluster'; a conserved gene neighbourhood (CGN) comprising *Dichaete*, *Sox21a*, and *Sox21b* in neighbouring regions within a chromosome. This investigation has identified the *Dichaete* cluster within the genomes of the crustacean *Daphnia pulex* and the myriapod *Strigamia maritima*. (It may also be present in the genomes of other arthropod species analysed here, however, the genome assemblies comprised small shotgun fragments and as such CGNs are unlikely to be preserved in these contigs.) Broadening the focus of SoxB function to other arthropods will therefore help address questions surrounding the evolution of redundancy and sub- and neo-functionalization for SoxB genes, and establish whether the expression patterns and functions observed in *A. mellifera* and *D. melanogaster* are ancestral or derived.

6.3 *Tribolium castaneum* as a model for genomics research

I have also attempted to examine the genomic activity of Dichaete and SoxNeuro in the *T. castaneum* genome in order to draw an evolutionary comparison with their activity in *Drosophila*. This experiment, if successful, would have illustrated either divergence or conservation of these TFs across deep evolutionary time, broadening our understanding of how 'master regulators' might evolve (Prior & Walter, 1996; Chan & Kyba, 2013). It would also have augmented our understanding of SoxB activity and evolution beyond the *Drosophila* paradigm.

However, realising this aim meant that I first had to establish *T. castaneum* as a model for genomics research. As there are no antibodies available for Dichaete and SoxNeuro in *Tribolium*, standard cross-linked ChIP experiments were unfeasible. I therefore elected to attempt DamID followed by high-throughput sequencing (DamID-seq) to map the genome-wide binding profiles of these two TFs. DamID relies upon abundant GATC motifs within the target genome (van Steensel & Henikoff, 2000), and I began by examining GATC occurrence in *T. castaneum*. I found that the median distance between GATC sites is 330bp in the *Tribolium* genome, which is slightly less frequent than the median distance in the *Drosophila* genome of 195bp. However, the Dam-fusion protein has been shown to methylate regions up to ~2.5kb from the TF binding site (van Steensel & Henikoff, 2000), and 98.3% of GATC sites within the *Tribolium* genome are shown to exist <2.5kb from one another. These *prima facie* observations imply that DamID as a technique would provide sufficient resolution in the *T. castaneum* genome to examine TF binding. (The results also indicate that the GATC densities in the genomes of 10 other arthropod species provide sufficient resolution for DamID.)

DamID is reliant on low level, 'leaky' expression in the genome, to avoid a toxic saturation of adenine methylation. It is therefore imperative that ectopic expression is tightly regulated. van Steensel and Henikoff (2000) utilised the *HSP70* promoter from *D. melanogaster* with a 5x upstream activator sequence (UAS) to drive expression of the TF-Dam fusion, in the absence of the GAL4 transcriptional activator. This promoter was sufficient to drive expression resulting in detectable levels of methylation without becoming toxic to individuals. In my experiments, I wished to test the feasibility of using this promoter in *T. castaneum*. However, its use resulted in extremely poor survival rates of injected embryos and yielded no transgenic lines. This promoter was therefore deemed to be unsuitable for driving the expression of the Sox-Dam fusion genes, as it appeared to be too toxic. I subsequently elected to use the basal *T. castaneum* promoter *HSP68* to drive expression; injections proved more successful with this promoter (although the survival and transgenic rates were still very poor) and transgenic lines were obtained for *SoxNeuro-Dam* and *Dichaete-Dam* fusions, and a *Dam*-only negative control.

However, although methylation was detectable via gel electrophoresis, it appears that the methylation levels were insufficient for enrichment and detection via high-throughput sequencing. Troubleshooting these experiments led to the identification of the food medium of *T. castaneum* – wheat flour – as a source of significant contamination when attempting to isolate DNA from beetle embryos. This is likely because each wheat cell contains 243 times as much DNA as each beetle cell, and since embryos possess relatively few cells, samples are instead saturated with wheat tissue. This meant that the methylated DNA input was insufficient, which explains the concatemer formation in the first attempt: adapters used for amplification were ligating to one another in contrast to the methylated DNA. However, even once these confounding factors of wheat contamination and concatemer formation were controlled for, high-throughput sequencing still led to unsuccessful enrichment of methylated regions. This therefore implies that there might be a lack of sufficient Dam activity in the genome.

The two promoters tested in this study may thus have been 'too hot' and 'too cold' for use in DamID, respectively. Instead, a 'Goldilocks' promoter may be required to drive expression enough as not to be lethal, but to methylate the genome sufficiently to be enriched and detected via high-throughput sequencing. However, it remains to be demonstrated whether *T. castaneum* is biologically compatible with DamID as a technique, as such sufficient methylation may be inherently lethal to beetles. Moreover, one might speculate that presently unexplored mechanisms of demethylation of adenine regions in the insect genome (Zhang *et al*., 2015) might be variable across species, rendering DamID incompatible as a technique in certain models. DamID identifies binding events *post hoc*, meaning that methylation ought to be more abundant in adult DNA than embryo DNA. It is therefore odd that enrichment was unsuccessful even in adult tissue. Genome methylation is poorly understood in insects at present (Song *et al*., 2017), and the demethylating mechanisms observed in the *Drosophila* genome (Zhang *et al*., 2015) might be stronger and more active in *T. castaneum*.

In future, studies should investigate the use of other promoters; however it has been advised that the use of promoters endogenous to *T. castaneum* is preferable (Lorenzen *et al*., 2002; Brown *et al*., 2009). The P3 promoter has been successfully used to drive GFP expression in beetles in this experiment and by others (Berghammer *et al*., 1999; Brown *et al*., 2009; Berghammer *et al*., 2009; Schinko *et al*., 2010), yet its expression is eye-specific. Lorenzen *et al*. (2002) report the use of the endogenous *Polyubiquitin* promoter to drive *vermillion* expression, and another alternative might be the use of the endogenous *Tubulin* promoter (Siebert *et al*., 2008). Nonetheless, these are all reported to lead to high expression levels: such promoters may not be suitable for DamID, which is why the basal *HSP68* promoter in the absence of the GAL4 system was originally used in this study.

In hindsight, DamID might have been an ambitious choice in this experimental design. Generating antibodies against Tc-Dichaete and Tc-SoxNeuro for ChIP experiments is also an option; however, generating antibodies in-house can be exceedingly time-intensive. An alternative might have been to use CAST-ChIP, whereby one generates a fusion protein between the TF of interest and another protein, such as GFP, and drive expression ectopically. A standard IP is then performed using an antibody against GFP, enriching the TF-fusion and bound regions also (Schauer *et al*., 2013). However, this ectopic expression can influence cell fate and development, and result in abnormal protein-binding behaviour (Marshall *et al*., 2016). It also requires large quantities of starting material (Schauer *et al*., 2013), which may be difficult to isolate from embryo tissues, particularly in *T. castaneum*. While there are antibodies available for Dam protein, the expression would be far too low to detect via IP (van Steensel & Henikoff, 2000), so CAST-ChIP was not an available option in my experimental design.

DamID therefore remains an attractive protocol for studying TF binding interactions in non-model organisms, should a suitable promoter be discovered. MeDIP-seq experiments, which identify adenomethylation directly using IP, might also be used as an independent enrichment protocol for TF-Dam binding sites. However, for the purposes of studying Sox gene binding, a useful preliminary experiment might be to express Tc-Sox-Dam fusions in *D. melanogaster* ectopically: DamID is well-established in the fly, and thus studying the *in vivo* binding of *T. castaneum* TFs might be more feasible if carried out in *Drosophila* embryos. These binding events may not accurately reflect their endogenous binding within the *T. castaneum* genome, but would perhaps elucidate the binding motifs of Tc-Dichaete and Tc-SoxNeuro. Therefore, although it is important that the scope of research is widened beyond *Drosophila*, the fruit fly remains an attractive system in which to study basic biological questions.

Examining the target motifs of orthologous SoxB TFs across evolutionary time was a principal aim of my research efforts: future researchers may wish to clone the SoxB genes of multiple arthropods for use with DamID in *Drosophila*. Identifying the gradual changes of SoxB binding motifs in species increasingly distant to *Drosophila* would certainly make for an interesting investigation into SoxB evolution over time while guaranteeing the use of a reliable and tested system.

6.4 Conclusions

This study has provided a greater understanding of SoxB evolution in arthropod species and has elucidated the difficulties of establishing genomics techniques in a non-model organism. I have performed the most comprehensive phylogenetic analysis of arthropod SoxB to date, and examined the expression patterns of two well-studied SoxB genes, *Dichaete* and *SoxNeuro*, within the context of central nervous system development in the short germ insect *Tribolium castaneum*. The phylogenetic analysis tested two evolutionary models proposed for SoxB emergence and proposed a new model in light of this data. The principal component of this project, however, attempted to establish *T. castaneum* as a research model for genomics studies by using the DNA adenine methyltransferase identification (DamID) technique to map genome-wide binding profiles of the Dichaete and SoxNeuro transcription factors (TFs). In this investigation, I identified that GATC motifs occur with sufficient frequency in the beetle genome for DamID analysis; I identified that the food medium is a significant contaminating factor to be cautious of; and I have tested the viability of two basal promoters to drive the ectopic expression of Dam-fusion proteins and found them both to be inadequate. Collectively, these results will hopefully provide a solid foundation for future work aiming to develop DamID in *T. castaneum* and other non-model species. Further experiments are required to examine the utility of other promoters to use in conjunction with DamID, and future studies examining TF activity across deep evolutionary time may continue to make use of the *Drosophila* model.

Bibliography

Akam, M. (1987). The molecular basis for metameric pattern in the Drosophila embryo. *Development*, *101*(1), 1–22. https://doi.org/citeulike-article-id:10064978

Akiyama, H., Chaboissier, M. C., Martin, J. F., Schedl, A., & De Crombrugghe, B. (2002). The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes and Development*, *16*(21), 2813–2828. https://doi.org/10.1101/gad.1017802

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*(1), 861. https://doi.org/10.1186/1471-2164-14-861

Andrew, S. (2010). Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Retrieved July 10, 2017, from http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aparicio, O., Geisberg, J. V, & Struhl, K. (2004). Chromatin immunoprecipitation for determining the association of proteins with specific genomic sequences in vivo. *Current Protocols in Cell Biology*, *Unit 17.6*. https://doi.org/10.1002/0471142727.mb2103s69

Aughey, G. N., & Southall, T. D. (2015). Dam it's good! DamID profiling of protein-DNA interactions. *Wiley Interdisciplinary Reviews: Developmental Biology*, *5*(1), 25–37. https://doi.org/10.1002/wdev.205

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., … Zhang, J. (2013). Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Computational Biology*, *9*(11), e1003326. https://doi.org/10.1371/journal.pcbi.1003326

Bardet, A. F., He, Q., Zeitlinger, J., & Stark, A. (2012). A computational pipeline for comparative ChIP-seq analyses. *Nature Protocols*, *7*(1), 45–61. https://doi.org/10.1038/nprot.2011.420

Bate, C. M. (1976). Embryogenesis of an insect nervous system I . A map of the thoracic and abdominal neuroblasts in Locusta migratoria. *Journal of Embryology and Experimental Morphology*, *35*(1), 107–123.

Bate, M., & Martinez Arias, A. (1993). *The Development of Drosophila melanogaster*. Cold Spring Harbor Laboratory Press.

Berghammer, A. J., Klingler, M., & Wimmer, E. A. (1999). A universal marker for transgenic insects. *Nature*, *402*(6760), 370–371. https://doi.org/10.1038/46463

Berghammer, A. J., Weber, M., Trauner, J., & Klingler, M. (2009). Red flour beetle (Tribolium) germline transformation and insertional mutagenesis. *Cold Spring Harbor Protocols*, *4*(8), pdb.prot5259. https://doi.org/10.1101/pdb.prot5259

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

Bhat, K. M. (1999). Segment polarity genes in neuroblast formation and identity specification during Drosophila neurogenesis. *BioEssays*, *21*(6), 472–485. https://doi.org/10.1002/(SICI)1521-1878(199906)21:6<472::AID-BIES4>3.0.CO;2-W

Bhattaram, P., Penzo-Méndez, A., Sock, E., Colmenares, C., Kaneko, K. J., Vassilev, A., … Lefebvre, V. (2010). Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nature Communications*, *1*(9), 1–12. https://doi.org/10.1038/ncomms1008

Biffar, L. (2013). *Early neurogenesis in the flour beetle Tribolium castaneum*. Queen Mary, University of London.

Biffar, L., & Stollewerk, A. (2014). Conservation and evolutionary modifications of neuroblast expression patterns in insects. *Developmental Biology*, *388*(1), 103–116. https://doi.org/10.1016/j.ydbio.2014.01.028

Biffar, L., & Stollewerk, A. (2015). Evolutionary variations in the expression of dorso-ventral patterning genes and the conservation of pioneer neurons in Tribolium castaneum. *Developmental Biology*, *400*(1), 159–167. https://doi.org/10.1016/j.ydbio.2015.01.025

Borneman, A. R., Gianoulis, T. A., Zhang, Z. D., Yu, H., Rozowsky, J., Seringhaus, M. R., … Snyder, M. (2007). Divergence of transcription factor binding sites across related yeast species. *Science*, *317*(5839), 815–9. https://doi.org/10.1126/science.1140748

Borok, M. J., Tran, D. A., Ho, M. C. W., & Drewell, R. A. (2010). Dissecting the regulatory switches of development: lessons from enhancer evolution in Drosophila. *Development*, *137*(1), 5–13. https://doi.org/10.1242/dev.036160

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

Boyan, G. S., & Williams, J. L. D. (1997). Embryonic development of the pars intercerebralis/central complex of the grasshopper. *Development Genes and Evolution*, *207*(5), 317–329. https://doi.org/10.1007/s004270050119

Boyan, G., & Williams, L. (2011). Embryonic development of the insect central complex: Insights from lineages in the grasshopper and Drosophila. *Arthropod Structure and Development*, *40*(4), 334–348. https://doi.org/10.1016/j.asd.2011.02.005

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). https://doi.org/10.1371/journal.pbio.1000343

Brena, C., & Akam, M. (2013). An analysis of segmentation dynamics throughout embryogenesis in the centipede Strigamia maritima. *BMC Biology*, *11*(112), 112. https://doi.org/10.1186/1741-7007-11-112

Bridges, C. B.; Morgan, T. H. (1923). *Third-Chromosome Group Of Mutant Characters Of Drosophila Melanogaster*. Carnegie Institution Of Washington, Washington.

Brody, T., & Odenwald, W. F. (2000). Programmed Transformations in Neuroblast Gene Expression during Drosophila CNS Lineage Development. *Developmental Biology*, *226*(1), 34–44. https://doi.org/10.1006/dbio.2000.9829

Brody, T., & Odenwald, W. F. (2005). Regulation of temporal identities during Drosophila neuroblast lineage development. *Current Opinion in Cell Biology*, *17*(6), 672–675. https://doi.org/10.1016/j.ceb.2005.09.013

Broitman-Maduro, G., Maduro, M. F., & Rothman, J. H. (2005). The noncanonical binding site of the MED-1 GATA factor defines differentially regulated target genes in the C. elegans mesendoderm. *Developmental Cell*, *8*(3), 427–433. https://doi.org/10.1016/j.devcel.2005.01.014

Brooks, J. E., & Roberts, R. J. (1982). Modification profiles of bacterial genomes. *Nucleic Acids Research*, *10*(3), 913–934. https://doi.org/10.1093/nar/10.3.913

Brown, S. J., Patel, N. H., & Denell, R. E. (1994). Embryonic expression of the single Tribolium engrailed homolog. *Developmental Genetics*, *15*(1), 7–18. https://doi.org/10.1002/dvg.1020150103

Brown, S. J., Shippy, T. D., Miller, S., Bolognesi, R., Beeman, R. W., Lorenzen, M. D., … Klingler, M. (2009). The red flour beetle, Tribolium castaneum (Coleoptera): A model for studies of development and pest biology. *Cold Spring Harbor Protocols*, *4*(8), 1–10. https://doi.org/10.1101/pdb.emo126

Bucher, G., & Wimmer, E. A. (2005). Beetle a-head: Investigating embryonic head formation using a novel model organism. *B.I.F.Futura*, *20*, 164–169.

Buck, M. J., & Lieb, J. D. (2004). ChIP-chip: Considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, *83*(3), 349–360. https://doi.org/10.1016/j.ygeno.2003.11.004

Buescher, M., & Chia, W. (1997). Mutations in lottchen cause cell fate transformations in both neuroblast and glioblast lineages in the Drosophila embryonic central nervous system. *Development*, *124*(3), 673–81.

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), 4193–203.

Cande, J. D., Chopra, V. S., & Levine, M. (2009). Evolving enhancer-promoter interactions within the tinman complex of the flour beetle, Tribolium castaneum. *Development*, *136*(18), 3153–3160. https://doi.org/10.1242/dev.038034

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

Carl, S., & Russell, S. (2015). Comparative Genomics of Transcription Factor Binding in Drosophila. In *Short Views on Insect Genomics and Proteomics* (pp. 157–175). Springer International Publishing Switzerland. https://doi.org/10.1007/978-3-319-24235-4_7

Celniker, S. E., L Dillon, L. A., Gerstein, M. B., Gunsalus, K. C., Henikoff, S., Karpen, G. H., … Lieb, J. D. (2009). Unlocking the secrets of the genome. *Nature*, *459*(18), 927–930. https://doi.org/10.1038/459927a

Chakravarthy, H., & Rizzino, A. (2009). *Mouse Sox2*. *Transcription Factor Encyclopedia*. https://doi.org/http://cisreg.cmmt.ubc.ca/cgi-bin/tfe/articles.pl?tfid=531

Chipman, A. D., & Stollewerk, A. (2006). Specification of neural precursor identity in the geophilomorph centipede Strigamia maritima. *Developmental Biology*, *290*(2), 337–350. https://doi.org/10.1016/j.ydbio.2005.11.029

Choe, C. P., & Brown, S. J. (2009). Genetic regulation of engrailed and wingless in Tribolium segmentation and the evolution of pair-rule segmentation. *Developmental Biology*, *325*(2), 482–491. https://doi.org/10.1016/j.ydbio.2008.10.037

Choe, C. P., Miller, S. C., & Brown, S. J. (2006). A pair-rule gene circuit defines segments sequentially in the short-germ insect Tribolium castaneum. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(17), 6560–6564. https://doi.org/10.1073/pnas.0510440103

Chu, H., Parras, C., White, K., & Jiménez, F. (1998). Formation and specification of ventral neuroblasts is controlled by vnd in Drosophila neurogenesis. *Genes and Development*, *12*(22), 3613–3624. https://doi.org/10.1101/gad.12.22.3613

Clark, E. (2017). Dynamic patterning by the Drosophila pair-rule network reconciles long-germ and short-germ segmentation. *bioRxiv*, *January*, 1–94. https://doi.org/https://doi.org/10.1101/099671

Clark, E., & Peel, A. D. (2017). Evidence for the temporal regulation of insect segmentation by a conserved set of developmental transcription factors. *bioRxiv*, *June*, 1–58. https://doi.org/10.1101/145151

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., … Clark, M. D. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research*, *27*(5), 885–896. https://doi.org/10.1101/gr.217117.116

Collas, P. (2010, May 14). The current state of chromatin immunoprecipitation. *Molecular Biotechnology*. Humana Press Inc. https://doi.org/10.1007/s12033-009-9239-8

Conaway, J. W. (2012). Introduction to Theme "Chromatin, Epigenetics, and Transcription." *Annual Review of Biochemistry*, *81*(February), 19–22. https://doi.org/10.1146/annurev-biochem-090711-093103

Crémazy, F., Berta, P., & Girard, F. (2000). SoxNeuro, a new Drosophila Sox gene expressed in the developing central nervous system. *Mechanisms of Development*, *93*(1–2), 215–219. https://doi.org/10.1016/S0925-4773(00)00268-9

Crémazy, F., Berta, P., & Girard, F. (2001). Genome-wide analysis of Sox genes in Drosophila melanogaster. *Mechanisms of Development*, *109*(2), 371–375. https://doi.org/10.1016/S0925-4773(01)00529-9

Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, *14*, 1188–1190. https://doi.org/10.1101/gr.849004.1

Dichtel-Danjoy, M.-L., Caldeira, J., & Casares, F. (2009). SoxF is part of a novel negative-feedback loop in the wingless pathway that controls proliferation in the Drosophila wing disc. *Development*, *136*(5), 761–769. https://doi.org/10.1242/dev.032854

Doe, C. Q. (1992). Molecular markers for identified neuroblasts and ganglion mother cells in the Drosophila central nervous system. *Development*, *116*(4), 855–63. https://doi.org/10.1146/annurev.ge.24.120190.002131

Doeffinger, C., Hartenstein, V., & Stollewerk, A. (2010). Compartmentalization of the precheliceral neuroectoderm in the spider Cupiennius salei: Development of the Arcuate Body, Optic Ganglia, and Mushroom Body. *Journal of Comparative Neurology*, *518*(13), 2612–2632. https://doi.org/10.1002/cne.22355

Dong, Z., Shi, C., Zhang, H., Dou, H., Cheng, F., Chen, G., & Liu, D. (2014). The characteristics of sox gene in Dugesia japonica. *Gene*, *544*(2), 177–183. https://doi.org/10.1016/j.gene.2014.04.053

Dönitz, J., Schmitt-Engel, C., Grossmann, D., Gerischer, L., Tech, M., Schoppmeier, M., … Bucher, G. (2015). iBeetle-Base: A database for RNAi phenotypes in the red flour beetle Tribolium castaneum. *Nucleic Acids Research*, *43*(D1), D720–D725. https://doi.org/10.1093/nar/gku1054

Dove, H. L. (2003). *Neurogenesis in the millipede Glomeris marginata (Myriapoda:Diplopoda)*. University of Köln.

Dove, H., & Stollewerk, A. (2003). Comparative analysis of neurogenesis in the myriapod Glomeris marginata (Diplopoda) suggests more similarities to chelicerates than to insects. *Development*, *130*(10), 2161–2171. https://doi.org/10.1242/dev.00442

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

Duman-Scheel, M., & Patel, N. H. (1999). Analysis of molecular marker expression reveals neuronal homology in distantly related arthropods. *Development*, *126*(11), 2327–2334.

Englesberg, E., Irr, J., Power, J., & Lee, N. (1965). Positive control of enzyme synthesis by gene C in the L-arabinose system. *Journal of Bacteriology*, *90*(4), 946–957. https://doi.org/10.1128/JB.01571-08

Erwin, D. H., & Davidson, E. H. (2006). Gene regulatory networks and the evolution of animal body plans. *Science*, *311*(5762), 796–800. https://doi.org/10.1126/science.1113832

Fabritius-Vilpoux, K., Bisch-Knaden, S., & Harzsch, S. (2008). Engrailed-like immunoreactivity in the embryonic ventral nerve cord of the Marbled Crayfish (Marmorkrebs). *Invertebrate Neuroscience*, *8*(4), 177–197. https://doi.org/10.1007/s10158-008-0081-7

Farrell, B. D. (1998). “Inordinate Fondness” Explained: Why Are There So Many Beetles? *Science*, *281*(5376), 555–559. https://doi.org/10.1126/science.281.5376.555

Feliciello, I., Parazajder, J., Akrap, I., & Ugarković, Đ. (2013). First evidence of DNA methylation in insect Tribolium castaneum. *Epigenetics*, *8*(5), 534–541. https://doi.org/10.4161/epi.24507

Ferrari, S., Harley, V. R., Pontiggia, A., Goodfellow, P. N., Lovell-Badge, R., & Bianchi, M. E. (1992). SRY, like HMG1, recognizes sharp angles in DNA. *EMBO Journal*, *11*(12), 4497–4506.

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 Biology*, *15*(5), R74. https://doi.org/10.1186/gb-2014-15-5-r74

Fisher, B., Weiszmann, R., Frise, E., Hammonds, A., Tomancak, P., Beaton, A., … Celniker, S. (2012). Berkeley Drosophila Genome Project. Retrieved July 24, 2017, from http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl

Flusberg, B. A., Webster, D. R., Lee, J. H., Travers, K. J., Olivares, E. C., Clark, T. A., … Turner, S. W. (2010). Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nature Methods*, *7*(6), 461–5. https://doi.org/10.1038/nmeth.1459

Focareta, L., & Cole, A. G. (2016). Analyses of Sox-B and Sox-E family genes in the cephalopod Sepia officinalis: Revealing the conserved and the unusual. *PLoS ONE*, *11*(6), 1–21. https://doi.org/10.1371/journal.pone.0157821

Fortunato, S., Adamski, M., Bergum, B., Guder, C., Jordal, S., Leininger, S., … Adamska, M. (2012). Genome-wide analysis of the sox family in the calcareous sponge Sycon ciliatum: multiple genes with unique expression patterns. *EvoDevo*, *3*(14), 1–11. https://doi.org/10.1186/2041-9139-3-14

Fraga, M. F., & Esteller, M. (2002). DNA methylation: A profile of methods and applications. *BioTechniques*, *33*(3), 632–649. https://doi.org/10.3389/fgene.2011.00074

Freese, N. H., Norris, D. C., & Loraine, A. E. (2016). Integrated genome browser: Visual analytics platform for genomics. *Bioinformatics*, *32*(14), 2089–2095. https://doi.org/10.1093/bioinformatics/btw069

Gao, J., Li, P., Zhang, W., Wang, Z., Wang, X., & Zhang, Q. (2015a). Molecular cloning, promoter analysis and expression profiles of the sox3 gene in Japanese flounder, Paralichthys olivaceus. *International Journal of Molecular Sciences*, *16*(11), 27931–27944. https://doi.org/10.3390/ijms161126079

Gao, J., Wang, J., Jiang, J., Fan, L., Wang, W., Liu, J., … Wang, X. (2013). Identification and characterization of a nanog homolog in Japanese flounder (Paralichthys olivaceus). *Gene*, *531*(2), 411–421. https://doi.org/10.1016/j.gene.2013.08.030

Gao, J., Zhang, W., Li, P., Liu, J., Song, H., Wang, X., & Zhang, Q. (2015b). Identification, molecular characterization and gene expression analysis of sox1a and sox1b genes in Japanese flounder, Paralichthys olivaceus. *Gene*, *574*(2), 225–234. https://doi.org/10.1016/j.gene.2015.08.013

Gibson, D. G., Young, L., Chuang, R.-Y., Venter, J. C., Hutchison, C. A., & Smith, H. O. (2009). Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods*, *6*(5), 343–345. https://doi.org/10.1038/nmeth.1318

Gilles, A. F., & Schinko, J. B. (2017). Tribolium Genome Editing Service (TriGenES). Retrieved July 10, 2017, from http://www.trigenes.com/

Gilles, A. F., Schinko, J. B., & Averof, M. (2015). Efficient CRISPR-mediated gene targeting and transgene replacement in the beetle Tribolium castaneum. *Development*, *142*(16), 2832–2839. https://doi.org/10.1242/dev.125054

Gilmour, D. S., & Lis, J. T. (1984). Detecting protein-DNA interactions in vivo: Distribution of RNA polymerase on specific bacterial genes. *Biochemistry*, *81*(14), 4275–4279. https://doi.org/10.1073/pnas.81.14.4275

Gilmour, D. S., & Lis, J. T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. *Molecular and Cellular Biology*, *5*(8), 2009–18. https://doi.org/10.1128/MCB.5.8.2009

Goldman, N. (1998). Phylogenetic information and experimental design in molecular systematics. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1407), 1779–1786. https://doi.org/10.1098/rspb.1998.0502

Gonzalez, D. H. (2015). Plant Transcription Factors: Evolutionary, Structural and Functional Aspects. *Plant Transcription Factors: Evolutionary, Structural and Functional Aspects*, *754*, 1–422. https://doi.org/10.1016/C2013-0-19051-4

Granderath, S., & Klämbt, C. (1999). Glia development in the embryonic CNS of Drosophila. *Current Opinion in Neurobiology*, *9*(5), 531–536. https://doi.org/10.1016/S0959-4388(99)00008-2

Greer, E. L., Blanco, M. A., Gu, L., Sendinc, E., Liu, J., Aristizábal-Corrales, D., … Shi, Y. (2015). DNA methylation on N6-adenine in C. elegans. *Cell*, *161*(4), 868–878. https://doi.org/10.1016/j.cell.2015.04.005

Greil, F., Moorman, C., & van Steensel, B. (2006). DamID: Mapping of In Vivo Protein-Genome Interactions Using Tethered DNA Adenine Methyltransferase. *Methods in Enzymology*, *410*(6), 342–359. https://doi.org/10.1016/S0076-6879(06)10016-6

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. https://doi.org/10.1038/346245a0

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

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

H M Prior, M. A. W. (1996). SOX genes: architects of development. *Molecular Medicine*, *2*(4), 405.

Hanna-Rose, W., & Han, M. (1999). COG-2, a sox domain protein necessary for establishing a functional vulval-uterine connection in Caenorhabditis elegans. *Development*, *126*(1), 169–79.

Hartenstein, V., & Campos-Ortega, J. a. (1984). Early neurogenesis in wild-type Drosophila melanogaster. *Roux's Archives of Developmental Biology*, *193*(5), 308–325. https://doi.org/10.1007/BF00848159

Hartenstein, V., & Jan, Y. N. (1992). Studying Drosophila embryogenesis with P-lacZ enhancer trap lines. *Roux's Archives of Developmental Biology*, *201*(4), 194–220. https://doi.org/10.1007/BF00188752

Hartenstein, V., & Stollewerk, A. (2015). The evolution of early neurogenesis. *Developmental Cell*, *32*(4), 390–407. https://doi.org/10.1016/j.devcel.2015.02.004

Hartenstein, V., Technau, G. M., & Campos-Ortega, J. A. (1985). Fate-mapping in wild-type Drosophila melanogaster. *Roux's Archives of Developmental Biology*, *194*(4), 213–216. https://doi.org/10.1007/BF00848248

Harzsch, S. (2001). Neurogenesis in the crustacean ventral nerve cord: Homology of neuronal stem cells in malacostraca and branchiopoda? *Evolution and Development*, *3*(3), 154–169. https://doi.org/10.1046/j.1525-142X.2001.003003154.x

Harzsch, S. (2003). Ontogeny of the ventral nerve cord in malacostracan crustaceans: A common plan for neuronal development in Crustacea, Hexapoda and other Arthropoda? *Arthropod Structure and Development*, *32*(1), 17–37. https://doi.org/10.1016/S1467-8039(03)00008-2

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., … Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nature Genetics*, *43*(5), 414–420. https://doi.org/10.1038/ng.808

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., & Kumar, S. (2015). Tree of life reveals clock-like speciation and diversification. *Molecular Biology and Evolution*, *32*(4), 835–845. https://doi.org/10.1093/molbev/msv037

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

Heitzler, P., Bourouis, M., Ruel, L., Carteret, C., & Simpson, P. (1996). Genes of the Enhancer of split and achaete-scute complexes are required for a regulatory loop between Notch and Delta during lateral signalling in Drosophila. *Development*, *122*(1), 161–171.

Hepat, R., Song, J.-J., Lee, D., & Kim, Y. (2013). A viral histone h4 joins to eukaryotic nucleosomes and alters host gene expression. *Journal of Virology*, *87*(20), 11223–11230. https://doi.org/10.1128/JVI.01759-13

Holton, T. A., & Graham, M. W. (1991). A simple and efficient method for direct cloning of PCR products using ddT-tailed vectors. *Nucleic Acids Research*, *19*(5), 1156. https://doi.org/10.1093/nar/19.5.1156

Horn, C., & Wimmer, E. A. (2000). A versatile vector set for animal transgenesis. *Development Genes and Evolution*, *210*(12), 630–637. https://doi.org/10.1007/s004270000110

Hughes, C. L., & Kaufman, T. C. (2000). A diverse approach to arthropod development. *Evolution and Development*, *2*(1), 6–8. https://doi.org/10.1046/j.1525-142X.2000.00038.x

Hunt, T., Bergsten, J., Levkanicova, Z., Papadopoulou, A., John, O. St., Wild, R., … Vogler, A. P. (2007). A comprehensive phylogeny of beetles reveals the evolutionary origins of a superradiation. *Science*, *318*(5858), 1913–1916. https://doi.org/10.1126/science.1146954

Imai, K. S., Hikawa, H., Kobayashi, K., & Satou, Y. (2017). Tfap2 and Sox1/2/3 cooperatively specify ectodermal fates in ascidian embryos. *Development*, *144*(1), 33–37. https://doi.org/10.1242/dev.142109

Isshiki, T., Pearson, B., Holbrook, S., & Doe, C. Q. (2001). Drosophila neuroblasts sequentially express transcription factors which specify the temporal identity of their neuronal progeny. *Cell*, *106*(4), 511–521. https://doi.org/10.1016/S0092-8674(01)00465-2

Jackson, V. (1978). Studies on histone organization in the nucleosome using formaldehyde as a reversible cross-linking agent. *Cell*, *15*(3), 945–954. https://doi.org/10.1016/0092-8674(78)90278-7

Jackson, V., & Chalkley, R. (1981). A new method for the isolation of replicative chromatin: Selective deposition of histone on both new and old DNA. *Cell*, *23*(1), 121–134. https://doi.org/10.1016/0092-8674(81)90277-4

Jacob, F. F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*(3), 318–356. https://doi.org/10.1016/S0022-2836(61)80072-7

Jager, M., Quéinnec, E., Chiori, R., Le Guyader, H., & Manuel, M. (2008). Insights into the early evolution of SOX genes from expression analyses in a ctenophore. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *310*(8), 650–667. https://doi.org/10.1002/jez.b.21244

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. https://doi.org/10.1016/j.ympev.2005.12.005

Jenett, A., Rubin, G. M., Ngo, T. T. B., Shepherd, D., Murphy, C., Dionne, H., … Zugates, C. T. (2012). A GAL4-Driver Line Resource for Drosophila Neurobiology. *Cell Reports*, *2*(4), 991–1001. https://doi.org/10.1016/j.celrep.2012.09.011

Jiménez, F., & Campos-Ortega, J. A. (1990). Defective neuroblast commitment in mutants of the achaete-scute complex and adjacent genes of D. melanogaster. *Neuron*, *5*(1), 81–89. https://doi.org/10.1016/0896-6273(90)90036-F

Johnson, D. S., Mortazavi, A., Myers, R. M., & Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. *Science (New York, N.Y.)*, *316*(5830), 1497–502. https://doi.org/10.1126/science.1141319

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., & Madden, T. L. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Research*, *36*(Web Server issue), 5–9. https://doi.org/10.1093/nar/gkn201

Johnston, D. S., & Nüsslein-Volhard, C. (1992). The origin of pattern and polarity in the Drosophila embryo. *Cell*, *68*(2), 201–219. https://doi.org/10.1016/0092-8674(92)90466-P

Kamachi, Y., Iwafuchi, M., Okuda, Y., Takemoto, T., Uchikawa, M., & Kondoh, H. (2009). Evolution of non-coding regulatory sequences involved in the developmental process: reflection of differential employment of paralogous genes as highlighted by Sox2 and group B1 Sox genes. *Proceedings of the Japan Academy. Series B, Physical and Biological Sciences*, *85*(2), 55–68. https://doi.org/10.2183/pjab.85.55

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

Kamachi, Y., Uchikawa, M., & Kondoh, H. (2000). Pairing SOX off: With partners in the regulation of embryonic development. *Trends in Genetics*, *16*(4), 182–187. https://doi.org/10.1016/S0168-9525(99)01955-1

Karin, M. (1990). Too many transcription factors: positive and negative interactions. *The New Biologist*, *2*(2), 126–131.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Kent, W. J. (2002). BLAT — The BLAST -Like Alignment Tool. *Genome Research*, *12*, 656–664. https://doi.org/10.1101/gr.229202.

Kersey, P. J., Allen, J. E., Armean, I., Boddu, S., Bolt, B. J., Carvalho-Silva, D., … Staines, D. M. (2016). Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Research*, *44*(D1), D574–D580. https://doi.org/10.1093/nar/gkv1209

Khoury, G., & Gruss, P. (1983). Enhancer Elements. *Cell*, *33*(June), 313–314. https://doi.org/10.1016/0092-8674(83)90410-5

Kim, J., Lo, L., Dormand, E., & Anderson, D. J. (2003). SOX10 maintains multipotency and inhibits neuronal differentiation of neural crest stem cells. *Neuron*, *38*(1), 17–31. https://doi.org/10.1016/S0896-6273(03)00163-6

Kin Chan, S. S. (2013). What is a Master Regulator? *Journal of Stem Cell Research & Therapy*, *3*(2), 10–13. https://doi.org/10.4172/2157-7633.1000e114

King, B., & Denholm, B. (2014). Malpighian tubule development in the red flour beetle (Tribolium castaneum). *Arthropod Structure & Development*, *43*(6), 605–13. https://doi.org/10.1016/j.asd.2014.08.002

King, N., Westbrook, M. J., Young, S. L., Kuo, A., Abedin, M., Chapman, J., … Rokhsar, D. (2008). The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. *Nature*, *451*(7180), 783–8. https://doi.org/10.1038/nature06617

Klingler, M. (2004). Tribolium. *Current Biology*, *14*(17), 639–640. https://doi.org/10.1016/j.cub.2004.08.004

Koniszewski, N. D. B., Kollmann, M., Bigham, M., Farnworth, M., He, B., Büscher, M., … Bucher, G. (2016). The insect central complex as model for heterochronic brain development—background, concepts, and tools. *Development Genes and Evolution*, *226*(3), 209–219. https://doi.org/10.1007/s00427-016-0542-7

Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and DNA. *Science*, *184*(139), 868–871. https://doi.org/10.1126/science.184.4139.868

Kumar, S., & Hedges, S. B. (1998). A molecular timescale for vertebrate evolution. *Nature*, *392*(6679), 917–920. https://doi.org/10.1038/31927

Kux, K., Kiparaki, M., & Delidakis, C. (2013). The two Tribolium E(spl) genes show evolutionarily conserved expression and function during embryonic neurogenesis. *Mechanisms of Development*, *130*(4–5), 207–225. https://doi.org/10.1016/j.mod.2013.02.003

Landgraf, M., Bossing, T., Technau, G. M., & Bate, M. (1997). The origin, location, and projections of the embryonic abdominal motorneurons of Drosophila. *The Journal of Neuroscience*, *17*(24), 9642–9655.

Landgraf, M., & Thor, S. (2006). Development of Drosophila motoneurons: Specification and morphology. *Seminars in Cell and Developmental Biology*, *17*(1), 3–11. https://doi.org/10.1016/j.semcdb.2005.11.007

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Larroux, C., Fahey, B., Liubicich, D., Hinman, V. F., Gauthier, M., Gongora, M., … Degnan, B. M. (2006). Developmental expression of transcription factor genes in a demosponge: Insights into the origin of metazoan multicellularity. *Evolution and Development*, *8*(2), 150–173. https://doi.org/10.1111/j.1525-142X.2006.00086.x

Larroux, C., Luke, G. N., Koopman, P., Rokhsar, D. S., Shimeld, S. M., & Degnan, B. M. (2008). Genesis and expansion of metazoan transcription factor gene classes. *Molecular Biology and Evolution*, *25*(5), 980–996. https://doi.org/10.1093/molbev/msn047

Latchman, D. S. (1997). Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology*, *29*(12), 1305–1312. https://doi.org/10.1016/S1357-2725(97)00085-X

Laudet, V., Stehelin, D., & Clevers, H. (1993). Ancestry and diversity of the HMG box superfamily. *Nucleic Acids Research*, *21*(10), 2493–2501. https://doi.org/10.1093/nar/21.10.2493

Lee, N., D’Souza, C. A., & Kronstad, J. W. (2003). Signaling in Phytopathogenic Fungi. *Annual Review of Phytopathology*, *41*(1), 399–427. https://doi.org/10.1146/annurev.phyto.41.052002.095728

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. *International Journal of Biochemistry and Cell Biology*, *39*(12), 2195–2214. https://doi.org/10.1016/j.biocel.2007.05.019

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–1735. https://doi.org/10.1093/hmg/ddg180

Levine, M. (2010). Transcriptional enhancers in animal development and evolution. *Current Biology*, *20*(17), R754–R763. https://doi.org/10.1016/j.cub.2010.06.070

Li, A., Ahsen, O. O., Liu, J. J., Du, C., McKee, M. L., Yang, Y., … Tanzi, R. E. (2013). Silencing of the Drosophila ortholog of SOX5 in heart leads to cardiac dysfunction as detected by optical coherence tomography. *Human Molecular Genetics*, *22*(18), 3798–3806. https://doi.org/10.1093/hmg/ddt230

Li, H. H., Kroll, J. R., Lennox, S. M., Ogundeyi, O., Jeter, J., Depasquale, G., & Truman, J. W. (2014). A GAL4 driver resource for developmental and behavioral studies on the larval CNS of Drosophila. *Cell Reports*, *8*(3), 897–908. https://doi.org/10.1016/j.celrep.2014.06.065

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, X. Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., … Biggin, M. D. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biology*, *6*(2), 0365–0388. https://doi.org/10.1371/journal.pbio.0060027

Liao, D. (1999). Concerted evolution: molecular mechanism and biological implications. *The American Journal of Human Genetics*, *64*(1), 24–30. https://doi.org/10.1086/302221

Liu, G., Huan, P., & Liu, B. (2017). A SoxC gene related to larval shell development and co-expression analysis of different shell formation genes in early larvae of oyster. *Development Genes and Evolution*, *227*(3), 181–188. https://doi.org/10.1007/s00427-017-0579-2

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

Lorenzen, M. D., Berghammer, A. J., Brown, S. J., Denell, R. E., Klingler, M., & Beeman, R. W. (2003). piggyBac-mediated germline transformation in the beetle Tribolium castaneum. *Insect Molecular Biology*, *12*(5), 433–440. https://doi.org/10.1046/j.1365-2583.2003.00427.x

Lorenzen, M. D., Brown, S. J., Denell, R. E., & Beeman, R. W. (2002). Transgene expression from the Tribolium castaneum Polyubiquitin promoter. *Insect Molecular Biology*, *11*(5), 399–407. https://doi.org/10.1046/j.1365-2583.2002.00349.x

Lovell-Badge, R. (2010). The early history of the Sox genes. *International Journal of Biochemistry and Cell Biology*, *42*(3), 378–380. https://doi.org/10.1016/j.biocel.2009.12.003

Lowe, C. J., Wu, M., Salic, A., Evans, L., Lander, E., Stange-Thomann, N., … Kirschner, M. (2003). Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell*, *113*(7), 853–865. https://doi.org/10.1016/S0092-8674(03)00469-0

Lowe, T., Garwood, R. J., Simonsen, T. J., Bradley, R. S., & Withers, P. J. (2013). Metamorphosis revealed: time-lapse three-dimensional imaging inside a living chrysalis. *Journal of The Royal Society Interface*, *10*(84), 20130304–20130304. https://doi.org/10.1098/rsif.2013.0304

Luo, G.-Z., Blanco, M. A., Greer, E. L., He, C., & Shi, Y. (2015). DNA N(6)-methyladenine: a new epigenetic mark in eukaryotes? *Nature Reviews Molecular Cell Biology*, *16*(12), 705–10. https://doi.org/10.1038/nrm4076

Lynch, J. A., El-Sherif, E., & Brown, S. J. (2012). Comparisons of the embryonic development of Drosophila , Nasonia , and Tribolium. *Wiley Interdisciplinary Reviews: Developmental Biology*, *1*(1), 16–39. https://doi.org/10.1002/wdev.3

Lynch, M., & Force, A. (2000). The probability of duplicate gene preservation by subfunctionalization. *Genetics*, *154*(1), 459–473. https://doi.org/10.1371/journal.pgen.0040029

MacArthur, S., Li, X.-Y., Li, J., Brown, J. B., Chu, H. C., Zeng, L., … Eisen, M. B. (2009). Developmental roles of 21 Drosophila transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. *Genome Biology*, *10*(R80), 1–26. https://doi.org/10.1186/gb-2009-10-7-r80

Magie, C. R., Pang, K., & Martindale, M. Q. (2005). Genomic inventory and expression of Sox and Fox genes in the cnidarian Nematostella vectensis. *Development Genes and Evolution*, *215*(12), 618–630. https://doi.org/10.1007/s00427-005-0022-y

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., … Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222–D226. https://doi.org/10.1093/nar/gku1221

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, *24*(3), 133–141. https://doi.org/10.1016/j.tig.2007.12.007

Marshall, O. J., Southall, T. D., Cheetham, S. W., & Brand, A. H. (2016). Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing. *Nature Protocols*, *11*(9), 1586–98. https://doi.org/10.1038/nprot.2016.084

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. https://doi.org/10.14806/ej.17.1.200

Masamizu, Y., Ohtsuka, T., Takashima, Y., Nagahara, H., Takenaka, Y., Yoshikawa, K., … Kageyama, R. (2006). Real-time imaging of the somite segmentation clock: revelation of unstable oscillators in the individual presomitic mesoderm cells. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(5), 1313–8. https://doi.org/10.1073/pnas.0508658103

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, *7*(1), 29–59. https://doi.org/10.1146/annurev.genom.7.080505.115623

Matsui, T., Kanai-Azuma, M., Hara, K., Matoba, S., Hiramatsu, R., Kawakami, H., … Kanai, Y. (2006). Redundant roles of Sox17 and Sox18 in postnatal angiogenesis in mice. *Journal of Cell Science*, *119*(17), 3513–3526. https://doi.org/10.1242/jcs.03081

Maurange, C., Cheng, L., & Gould, A. P. (2008). Temporal Transcription Factors and Their Targets Schedule the End of Neural Proliferation in Drosophila. *Cell*, *133*(5), 891–902. https://doi.org/10.1016/j.cell.2008.03.034

McDonald, J. A., Holbrook, S., Isshiki, T., Weiss, J., Doe, C. Q., & Mellerick, D. M. (1998). Dorsoventral patterning in the Drosophila central nervous system: The vnd homeobox gene specifies ventral column identity. *Genes and Development*, *12*(22), 3603–3612. https://doi.org/10.1101/gad.12.22.3603

McKimmie, C., Woerfel, G., & Russell, S. (2005). Conserved genomic organisation of Group B Sox genes in insects. *BMC Genetics*, *6*(26), 1–15. https://doi.org/10.1186/1471-2156-6-26

Mendonça, A. G., Alves, R. J., & Pereira-Leal, J. B. (2011). Loss of genetic redundancy in reductive genome evolution. *PLoS Computational Biology*, *7*(2), e1001082. https://doi.org/10.1371/journal.pcbi.1001082

Meng, F. W., & 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. https://doi.org/10.1016/j.celrep.2015.09.061

Meulemans, D., & Bronner-Fraser, M. (2007). The Amphioxus SoxB Family: Implications for the Evolution of Vertebrate Placodes. *International Journal of Biological Science*, *3*(6), 356–364. https://doi.org/10.1016/j.dci.2009.07.002

Meyer, A., & Van De Peer, Y. (2005). From 2R to 3R: Evidence for a fish-specific genome duplication (FSGD). *BioEssays*, *27*(9), 937–945. https://doi.org/10.1002/bies.20293

Meyer, J., Wirth, J., Held, M., Schempp, W., & Scherer, G. (1996). SOX20, a new member of the SOX gene family, is located on chromosome 17p13. *Cytogenetic and Genome Research*, *72*(2–3), 246–249. https://doi.org/10.1159/000134200

Miller, S. W., Avidor-Reiss, T., Polyanovsky, A., & Posakony, J. W. (2009). Complex interplay of three transcription factors in controlling the tormogen differentiation program of Drosophila mechanoreceptors. *Developmental Biology*, *329*(2), 386–399. https://doi.org/10.1016/j.ydbio.2009.02.009

Minor, P. J., He, T.-F., Sohn, C. H., Asthagiri, A. R., & Sternberg, P. W. (2013). FGF signaling regulates Wnt ligand expression to control vulval cell lineage polarity in C. elegans. *Development*, *140*(18), 3882–91. https://doi.org/10.1242/dev.095687

Misof, B., Liu, S., Meusemann, K., Peters, R. S., & Al., E. (2014). Phylogenomics resolves the timing and pattern of insect evolution. *Science*, *346*(6210), 763–767. https://doi.org/10.1017/CBO9781107415324.004

Miya, T., & Nishida, H. (2003). Expression pattern and transcriptional control of SoxB1 in embryos of the ascidian Halocynthia roretzi. *Zoological Science*, *20*(1), 59–67. https://doi.org/10.2108/zsj.20.59

Morozova, O., & Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. *Genomics*, *92*(5), 255–264. https://doi.org/10.1016/j.ygeno.2008.07.001

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. https://doi.org/10.1038/nmeth.1226

Mukherjee, A., Melnattur, K. V., Zhang, M., & Nambu, J. R. (2006). Maternal expression and function of the Drosophila Sox gene Dichaete during oogenesis. *Developmental Dynamics*, *235*(10), 2828–2835. https://doi.org/10.1002/dvdy.20904

Nakamoto, A., Hester, S. D., Constantinou, S. J., Blaine, W. G., Tewksbury, A. B., Matei, M. T., … Williams, T. A. (2015). Changing cell behaviours during beetle embryogenesis correlates with slowing of segmentation. *Nature Communications*, *6*(11), 6635. https://doi.org/10.1038/ncomms7635

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.

Nanda, S., Defalco, T. J., Hui Yong Loh, S., Phochanukul, N., Camara, N., Van Doren, M., & Russell, S. (2009). Sox100B, a drosophila group e sox-domain gene, is required for somatic testis differentiation. *Sexual Development*, *3*(1), 26–37. https://doi.org/10.1159/000200079

Nasiadka, A., Dietrich, B. H., & Krause, H. M. (2002). Anterior-posterior patterning in the Drosophila embryo. *Advances in Developmental Biology and Biochemistry*, *12*, 155–204. https://doi.org/10.1016/S1569-1799(02)12027-2

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. *Journal of Biological Chemistry*, *275*(6), 3810–3818. https://doi.org/10.1074/jbc.275.6.3810

O'Kane, C. J., & Gehring, W. J. (1987). Detection in situ of genomic regulatory elements in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(24), 9123–9127. https://doi.org/10.1073/pnas.84.24.9123

O'Neill, L. P., & Turner, B. M. (1996). Immunoprecipitation of chromatin. *Methods in Enzymology*, *274*, 189–197. https://doi.org/10.1016/S0076-6879(96)74017-X

Oates, A. C., Morelli, L. G., & Ares, S. (2012). Patterning embryos with oscillations: structure, function and dynamics of the vertebrate segmentation clock. *Development*, *139*(4), 625–639. https://doi.org/10.1242/dev.063735

Oberhofer, G., Grossmann, D., Siemanowski, J. L., Beissbarth, T., & Bucher, G. (2014). Wnt/β-catenin signaling integrates patterning and metabolism of the insect growth zone. *Development*, *141*(24), 4740–4750. https://doi.org/10.1242/dev.112797

Odom, D. T., Dowell, R. D., Jacobsen, E. S., Gordon, W., Danford, T. W., MacIsaac, K. D., … Fraenkel, E. (2007). Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genetics*, *39*(6), 730–2. https://doi.org/10.1038/ng2047

Okuda, Y., Ogura, E., Kondoh, H., & Kamachi, Y. (2010). B1 SOX coordinate cell specification with patterning and morphogenesis in the early zebrafish embryo. *PLoS Genetics*, *6*(5), 36. https://doi.org/10.1371/journal.pgen.1000936

Oommen, K. S., & Newman, A. P. (2007). Co-regulation by Notch and Fos is required for cell fate specification of intermediate precursors during C. elegans uterine development. *Development*, *134*(22), 3999–4009. https://doi.org/10.1242/dev.002741

Orlando, V. (2000). Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends in Biochemical Sciences*, *25*(3), 99–104. https://doi.org/10.1016/S0968-0004(99)01535-2

Osaki, E., Nishina, Y., Inazawa, J., Copeland, N. G., Gilbert, D. J., Jenkins, N. A., … Semba, K. (1999). Identification of a novel Sry-related gene and its germ cell-specific expression. *Nucleic Acids Research*, *27*(12), 2503–2510. https://doi.org/10.1093/nar/27.12.2503

Overton, P. M. (2003). *The role of Sox genes in the development of Drosophila melanogaster. Department of Genetics*. University of Cambridge.

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.

Pai, A., Bennett, L., & Yan, G. Y. (2005). Female multiple mating for fertility assurance in red flour beetles (Tribolium castaneum). *Canadian Journal of Zoology*, *83*(7), 913–919. https://doi.org/10.1139/z05-073

Palmeirim, I., Henrique, D., Ish-Horowicz, D., & Pourquié, O. (1997). Avian hairy gene expression identifies a molecular clock linked to vertebrate segmentation and somitogenesis. *Cell*, *91*(5), 639–648. https://doi.org/10.1016/S0092-8674(00)80451-1

Pan, Y., Zhou, Y., Guo, C., Gong, H., Gong, Z., & Liu, L. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *Development*, *118*(5), 401–415. https://doi.org/10.1101/lm.1331809

Paris, M., Kaplan, T., Li, X. Y., Villalta, J. E., Lott, S. E., & Eisen, M. B. (2013). Extensive Divergence of Transcription Factor Binding in Drosophila Embryos with Highly Conserved Gene Expression. *PLoS Genetics*, *9*(9), e1003748. https://doi.org/10.1371/journal.pgen.1003748

Patel, N. H. (1994). The evolution of arthropod segmentation: insights from comparisons of gene expression patterns. *Development*, *207*, 201–207.

Patel, N. H., Martin-Blanco, E., Coleman, K. G., Poole, S. J., Ellis, M. C., Kornberg, T. B., & Goodman, C. S. (1989). Expression of engrailed proteins in arthropods, annelids, and chordates. *Cell*, *58*(5), 955–968. https://doi.org/10.1016/0092-8674(89)90947-1

Peel, A. D., Chipman, A. D., & Akam, M. (2005). Arthropod segmentation: beyond the Drosophila paradigm. *Nature Reviews Genetics*, *6*(12), 905–16. https://doi.org/10.1038/nrg1724

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: five essential questions. *Nature Reviews Genetics*, *14*(4), 288–95. https://doi.org/10.1038/nrg3458

Perry, M. W., Boettiger, A. N., & Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 1–12. https://doi.org/10.1073/pnas.1109873108

Pevny, L. H., & Lovell-Badge, R. (1997). Sox genes find their feet. *Current Opinion in Genetics & Development*, *7*(3), 338–344. https://doi.org/10.1016/S0959-437X(97)80147-5

Pevny, L., & Placzek, M. (2005). SOX genes and neural progenitor identity. *Current Opinion in Neurobiology*, *15*(1), 7–13. https://doi.org/10.1016/j.conb.2005.01.016

Phochanukul, N., & Russell, S. (2010). No backbone but lots of Sox: Invertebrate Sox genes. *International Journal of Biochemistry and Cell Biology*, *42*(3), 453–464. https://doi.org/10.1016/j.biocel.2009.06.013

Popovic, J., Stanisavljevic, D., Schwirtlich, M., Klajn, A., Marjanovic, J., & Stevanovic, M. (2014). Expression analysis of SOX14 during retinoic acid induced neural differentiation of embryonal carcinoma cells and assessment of the effect of its ectopic expression on SOXB members in HeLa cells. *PLoS ONE*, *9*(3), 1–12. https://doi.org/10.1371/journal.pone.0091852

Posnien, N., Schinko, J., Grossmann, D., Shippy, T. D., Konopova, B., & Bucher, G. (2009). RNAi in the Red Flour Beetle (Tribolium). *Cold Spring Harbor Protocols*, *2009*(8), pdb.prot5256. https://doi.org/10.1101/pdb.prot5256

Prasad, N., Tarikere, S., Khanale, D., Habib, F., & Shashidhara, L. S. (2016). A comparative genomic analysis of targets of Hox protein Ultrabithorax amongst distant insect species. *Scientific Reports*, *6*(1), 27885. https://doi.org/10.1038/srep27885

Pueyo, J. I., Lanfear, R., & Couso, J. P. (2008). Ancestral Notch-mediated segmentation revealed in the cockroach Periplaneta americana. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(43), 16614–16619. https://doi.org/10.1073/pnas.0804093105

Qian, W., Liao, B. Y., Chang, A. Y. F., & Zhang, J. (2010). Maintenance of duplicate genes and their functional redundancy by reduced expression. *Trends in Genetics*, *26*(10), 425–430. https://doi.org/10.1016/j.tig.2010.07.002

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

Raper, J. A., Bastiani, M., & Goodman, C. S. (1983). Pathfinding by neuronal growth cones in grasshopper embryos. II. Selective fasciculation onto specific axonal pathways. *The Journal of Neuroscience*, *3*(1), 31–41.

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., & Lieb, J. D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biology*, *12*(7), R67. https://doi.org/10.1186/gb-2011-12-7-r67

Regier, J. C., Shultz, J. W., Zwick, A., Hussey, A., Ball, B., Wetzer, R., … Cunningham, C. W. (2010). Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. *Nature*, *463*(7284), 1079–1083. https://doi.org/10.1038/nature08742

Reiprich, S., & Wegner, M. (2015). From CNS stem cells to neurons and glia: Sox for everyone. *Cell & Tissue Research*, *359*(1), 111–124. https://doi.org/10.1007/s00441-014-1909-6

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 and Development*, *17*(16), 2048–2059. https://doi.org/10.1101/gad.269303

Renn, S. C. P., Armstrong, J. D., Yang, M., Wang, Z., An, X., Kaiser, K., & Taghert, P. H. (1999). Genetic analysis of the Drosophila ellipsoid body neuropil: Organization and development of the central complex. *Journal of Neurobiology*, *41*(2), 189–207. https://doi.org/10.1002/(SICI)1097-4695(19991105)41:2<189::AID-NEU3>3.0.CO;2-Q

Rhee, H. S., & Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell*, *147*(6), 1408–1419. https://doi.org/10.1016/j.cell.2011.11.013

Richards, G. S., & Rentzsch, F. (2015). Regulation of Nematostella neural progenitors by SoxB, Notch and bHLH genes. *Development*, *142*(19), 3332–3342. https://doi.org/10.1242/dev.123745

Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., … Bucher, G. (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, *452*(7190), 949–955. https://doi.org/10.1038/nature06784

Rosenberg, I. M. (2006). *Protein Analysis and Purification*. *Protein Analysis and Purification: Benchtop Techniques: Second Edition*. Boston, MA: Birkhäuser Boston. https://doi.org/10.1007/b138330

Roth, S., & Hartenstein, V. (2008). Development of Tribolium castaneum. *Development Genes and Evolution*, *218*(3–4), 115–118. https://doi.org/10.1007/s00427-008-0215-2

Russell, S. (2000). The Drosophila dominant wing mutation Dichaete results from ectopic expression of a Sox-domain gene. *Molecular & General Genetics*, *263*(4), 690–701. https://doi.org/10.1007/s004380051218

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.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, *16*(10), 944–945. https://doi.org/10.1093/bioinformatics/16.10.944

Saeed, I., Bachir, D. G., Chen, L., & Hu, Y. G. (2016). The Expression of TaRca2-α Gene associated with net photosynthesis rate, biomass and grain yield in bread wheat (Triticum aestivum L) under field conditions. *PLoS ONE*, *11*(8), 1–19. https://doi.org/10.1371/journal.pone.0161308

Sakai, D., Suzuki, T., Osumi, N., & Wakamatsu, Y. (2006). Cooperative action of Sox9, Snail2 and PKA signaling in early neural crest development. *Development*, *133*(7), 1323–1333. https://doi.org/10.1242/dev.02297

Sallam, M. N. (2008). *Insect Damage: Post-Harvest Operations*. *Post-harvest Compendium, Food and Agriculture Organization of the United Nations*.

Sánchez-Soriano, N., & Russell, S. (1998). The Drosophila SOX-domain protein Dichaete is required for the development of the central nervous system midline. *Development*, *125*(20), 3989–96.

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

Sanguinetti, G., Lawrence, N. D., & Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, *22*(22), 2775–2781. https://doi.org/10.1093/bioinformatics/btl473

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. https://doi.org/10.1016/j.stem.2012.12.007

Sarrazin, A. F., Peel, A. D., & Averof, M. (2012). A segmentation clock with two-segment periodicity in insects. *Science*, *336*(6079), 338–341. https://doi.org/10.1126/science.1218256

Schauer, T., Schwalie, P. C., Handley, A., Margulies, C. E., Flicek, P., & Ladurner, A. G. (2013). CAST-ChIP Maps Cell-Type-Specific Chromatin States in the Drosophila Central Nervous System. *Cell Reports*, *5*(1), 271–282. https://doi.org/10.1016/j.celrep.2013.09.001

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. https://doi.org/10.1016/S1534-5807(02)00223-X

Schinko, J. B., Weber, M., Viktorinova, I., Kiupakis, A., Averof, M., Klingler, M., … Bucher, G. (2010). Functionality of the GAL4/UAS system in Tribolium requires the use of endogenous core promoters. *BMC Developmental Biology*, *10*(53), 1–12. https://doi.org/10.1186/1471-213X-10-53

Schmid, A., Chiba, A., & Doe, C. Q. (1999). Clonal analysis of Drosophila embryonic neuroblasts: neural cell types, axon projections and muscle targets. *Development*, *126*(21), 4653–4689.

Schmidt, D. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science*, *328*(5981), 1036–1040.

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., … Odom, D. T. (2010). Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*, *328*(5981), 1036–1040. https://doi.org/10.1126/science.1186176

Schmidt, H., Rickert, C., Bossing, T., Vef, O., Urban, J., & Technau, G. M. (1997). The Embryonic Central Nervous System Lineages ofDrosophila melanogaster. *Developmental Biology*, *189*(2), 186–204. https://doi.org/10.1006/dbio.1997.8660

Schmitt-Engel, C., Schultheis, D., Schwirz, J., Ströhlein, N., Troelenberg, N., Majumdar, U., … Bucher, G. (2015). The iBeetle large-scale RNAi screen reveals gene functions for insect development and physiology. *Nature Communications*, *6*, 7822. https://doi.org/10.1038/ncomms8822

Schnitzler, C. E., Simmons, D. K., Pang, K., Martindale, M. Q., & Baxevanis, A. D. (2014). Expression of multiple Sox genes through embryonic development in the ctenophore Mnemiopsis leidyi is spatially restricted to zones of cell proliferation. *EvoDevo*, *5*(1), 15. https://doi.org/10.1186/2041-9139-5-15

Scholtz, G. (1990). The Formation, Differentiation and Segmentation of the Post-Naupliar Germ Band of the Amphipod Gammarus pulex L. (Crustacea, Malacostraca, Peracarida). *Proceedings of the Royal Society B: Biological Sciences*, *239*(1295), 163–211. https://doi.org/10.1098/rspb.1990.0013

Scholtz, G. (1992). Cell lineage studies in the crayfish Cherax destructor (Crustacea, Decapoda): Germ band formation, segmentation, and early neurogenesis. *Roux's Archives of Developmental Biology*, *202*(1), 36–48. https://doi.org/10.1007/BF00364595

Schoppmeier, M., & Damen, W. G. M. (2001). Double-stranded RNA interference in the spider Cupiennius salei: The role of Distal-less is evolutionarily conserved in arthropod appendage formation. *Development Genes and Evolution*, *211*(2), 76–82. https://doi.org/10.1007/s004270000121

Schröder, R. (2003). The genes orthodenticle and hunchback substitute for bicoid in the beetle Tribolium. *Nature*, *422*(April), 621–625. https://doi.org/10.1038/nature01518.1.

Schröder, R., Beermann, A., Wittkopp, N., & Lutz, R. (2008). From development to biodiversity - Tribolium castaneum, an insect model organism for short germband development. *Development Genes and Evolution*, *218*(3–4), 119–126. https://doi.org/10.1007/s00427-008-0214-3

Schuster, E., McElwee, J. J., Tullet, J. M. a, Doonan, R., Matthijssens, F., Reece-Hoyes, J. S., … Gems, D. (2010). DamID in C. elegans reveals longevity-associated targets of DAF-16/FoxO. *Molecular Systems Biology*, *6*(399), 1–6. https://doi.org/10.1038/msb.2010.54

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Serfling, E., Jasin, M., & Schaffner, W. (1985). Enhancers and eukaryotic gene transcription. *Trends in Genetics*, *1*(C), 224–230. https://doi.org/10.1016/0168-9525(85)90088-5

Shahzad Asif, H. M., Rolfe, M. D., Green, J., Lawrence, N. D., Rattray, M., & Sanguinetti, G. (2010). TFInfer: A tool for probabilistic inference of transcription factor activities. *Bioinformatics*, *26*(20), 2635–2636. https://doi.org/10.1093/bioinformatics/btq469

Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, *26*(10), 1135–1145. https://doi.org/10.1038/nbt1486

Shinzato, C., Iguchi, A., Hayward, D. C., Technau, U., Ball, E. E., & Miller, D. J. (2008). Sox genes in the coral Acropora millepora: divergent expression patterns reflect differences in developmental mechanisms within the Anthozoa. *BMC Evolutionary Biology*, *8*, 311. https://doi.org/10.1186/1471-2148-8-311

Siebert, K. S., Lorenzen, M. D., Brown, S. J., Park, Y., & Beeman, R. W. (2008). Tubulin superfamily genes in Tribolium castaneum and the use of a Tubulin promoter to drive transgene expression. *Insect Biochemistry and Molecular Biology*, *38*(8), 749–755. https://doi.org/10.1016/j.ibmb.2008.04.007

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. https://doi.org/10.1038/346240a0

Skeath, J. B. (1999). At the nexus between pattern formation and cell-type specification: The generation of individual neuroblast fates in the drosophila embryonic central nervous system. *BioEssays*, *21*(11), 922–931. https://doi.org/10.1002/(SICI)1521-1878(199911)21:11<922::AID-BIES4>3.0.CO;2-T

Skeath, J. B., & Carroll, S. B. (1992). Regulation of proneural gene expression and cell fate during neuroblast segregation in the Drosophila embryo. *Development*, *114*(4), 939–946.

Skeath, J. B., Panganiban, G. F., & Carroll, S. B. (1994). The ventral nervous system defective gene controls proneural gene expression at two distinct steps during neuroblast formation in Drosophila. *Development*, *120*(6), 1517–1524.

Skeath, J. B., & Thor, S. (2003). Genetic control of Drosophila nerve cord development. *Current Opinion in Neurobiology*, *13*(1), 8–15. https://doi.org/10.1016/S0959-4388(03)00007-2

Smits, P., Li, P., Mandel, J., Zhang, Z., Deng, J. M., Behringer, R. R., … Lefebvre, V. (2001). The Transcription Factors L-Sox5 and Sox6 Are Essential for Cartilage Formation. *Developmental Cell*, *1*(2), 277–290. https://doi.org/10.1016/S1534-5807(01)00003-X

Sock, E., Rettig, S. D., Enderich, J., Bosl, M. R., Tamm, E. R., & Wegner, M. (2004). Gene targeting reveals a widespread role for the high-mobility-group transcription factor Sox11 in tissue remodeling. *Mol Cell Biol*, *24*(15), 6635–6644. https://doi.org/10.1128/mcb.24.15.6635-6644.2004

Sokal, R. R., & Sokoloff, A. (1973). The Biology of Tribolium with Special Emphasis on Genetic Aspects. Volume 1. *The Quarterly Review of Biology*, *48*(3), 500–501. https://doi.org/10.1086/407731

Song, X., Huang, F., Liu, J., Li, C., Gao, S., Wu, W., … Li, B. (2017). Genome-wide DNA methylomes from discrete developmental stages reveal the predominance of non-CpG methylation in Tribolium castaneum. *DNA Research*, *0*(0), 1–14. https://doi.org/10.1093/dnares/dsx016

Soullier, S., Jay, P., Poulat, F., Vanacker, J. M., Berta, P., & Laudet, V. (1999). Diversification pattern of the HMG and SOX family members during evolution. *Journal of Molecular Evolution*, *48*(5), 517–527. https://doi.org/10.1007/PL00006495

Southall, T. D., Gold, K. S., Egger, B., Davidson, C. M., Caygill, E. E., Marshall, O. J., & Brand, A. H. (2013). Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: Assaying RNA pol II occupancy in neural stem cells. *Developmental Cell*, *26*(1), 101–112. https://doi.org/10.1016/j.devcel.2013.05.020

Steensel, B., & Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. *Nature Biotechnology*, *18*(4), 424–428. https://doi.org/10.1038/74487

Stent, G. S., & Weisblat, D. A. (1985). Cell lineage in the development of invertebrate nervous systems. *Annual Review of Neuroscience*, *8*, 45–70. https://doi.org/10.1146/annurev.ne.08.030185.000401

Stollewerk, A. (2016). A flexible genetic toolkit for arthropod neurogenesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1685), 20150044. https://doi.org/http://dx.doi.org/10.1098/rstb.2015.0044

Stollewerk, A., & Simpson, P. (2005). Evolution of early development of the nervous system: A comparison between arthropods. *BioEssays*, *27*(9), 874–883. https://doi.org/10.1002/bies.20276

Stolt, C. C., Lommes, P., Friedrich, R. P., & Wegner, M. (2004). Transcription factors Sox8 and Sox10 perform non-equivalent roles during oligodendrocyte development despite functional redundancy. *Development*, *131*(10), 2349–2358. https://doi.org/10.1242/dev.01114

Stolt, C. C., Lommes, P., Sock, E., Chaboissier, M. C., Schedl, A., & Wegner, M. (2003). The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes and Development*, *17*(13), 1677–1689. https://doi.org/10.1101/gad.259003

Stolt, C. C., Schmitt, S., Lommes, P., Sock, E., & Wegner, M. (2005). Impact of transcription factor Sox8 on oligodendrocyte specification in the mouse embryonic spinal cord. *Developmental Biology*, *281*(2), 309–317. https://doi.org/10.1016/j.ydbio.2005.03.010

Stork, N. E. (1988). Insect diversity: facts, fiction and speculation. *Biological Journal of the Linnean Society*, *35*(4), 321–337. https://doi.org/10.1111/j.1095-8312.1988.tb00474.x

Sulston, I. A., & Anderson, K. V. (1996). Embryonic patterning mutants of Tribolium castaneum. *Development*, *122*(3), 805–814.

Taguchi, S., Tagawa, K., Humphreys, T., & Satoh, N. (2002). Group B sox genes that contribute to specification of the vertebrate brain are expressed in the apical organ and ciliary bands of hemichordate larvae. *Zoological Science*, *19*(1), 57–66. https://doi.org/10.2108/zsj.19.57

Takayama, S., Dhahbi, J., Roberts, A., Mao, G., Heo, S. J., Pachter, L., … Boffelli, D. (2014). Genome methylation in D. melanogaster is found at specific short motifs and is independent of DNMT2 activity. *Genome Research*, *24*(5), 821–830. https://doi.org/10.1101/gr.162412.113

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. *Molecular and Cellular Biology*, *24*(20), 8834–8846. https://doi.org/10.1128/MCB.24.20.8834-8846.2004

Terpe, K. (2006). Overview of bacterial expression systems for heterologous protein production: From molecular and biochemical fundamentals to commercial systems. *Applied Microbiology and Biotechnology*, *72*(2), 211–222. https://doi.org/10.1007/s00253-006-0465-8

The ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Thomas, J. B., Bastiani, M. J., Bate, M., & Goodman, C. S. (1984). From grasshopper to Drosophila: a common plan for neuronal development. *Nature*, *310*(19 July), 203–207. https://doi.org/10.1038/310203a0

Trauner, J., Schinko, J., Lorenzen, M. D., Shippy, T. D., Wimmer, E. a, Beeman, R. W., … Brown, S. J. (2009). Large-scale insertional mutagenesis of a coleopteran stored grain pest, the red flour beetle Tribolium castaneum, identifies embryonic lethal mutations and enhancer traps. *BMC Biology*, *7*(73), 1–12. https://doi.org/10.1186/1741-7007-7-73

Tuch, B. B., Galgoczy, D. J., Hernday, A. D., Li, H., & Johnson, A. D. (2008). The evolution of combinatorial gene regulation in fungi. *PLoS Biology*, *6*(2), 0352–0364. https://doi.org/10.1371/journal.pbio.0060038

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. *Mechanisms of Development*, *84*(1–2), 103–120. https://doi.org/10.1016/S0925-4773(99)00083-0

Ungerer, P., Geppert, M., & Wolff, C. (2011). Axogenesis in the central and peripheral nervous system of the amphipod crustacean Orchestia cavimana. *Integrative Zoology*, *6*(1), 28–44. https://doi.org/10.1111/j.1749-4877.2010.00227.x

Ungerer, P., & Scholtz, G. (2008). Filling the gap between identified neuroblasts and neurons in crustaceans adds new support for Tetraconata. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1633), 369–376. https://doi.org/10.1098/rspb.2007.1391

Uwanogho, D., Rex, M., Cartwright, E. J., Pearl, G., Healy, C., Scotting, P. J., & Sharpe, P. T. (1995). Embryonic expression of the chicken Sox2, Sox3 and Sox11 genes suggests an interactive role in neuronal development. *Mechanisms of Development*, *49*(1–2), 23–36. https://doi.org/10.1016/0925-4773(94)00299-3

van de Wetering, M., & Clevers, H. (1993). Sox 15, a novel member of the murine Sox family of HMG box transcription factors. *Nucleic Acids Research*, *21*(7), 1669–1669. https://doi.org/10.1093/nar/21.7.1669

van de Wetering, M., Oosterwegel, M., van Norren, K., & Clevers, H. (1993). Sox-4, an Sry-like HMG box protein, is a transcriptional activator in lymphocytes. *The EMBO Journal*, *12*(10), 3847–54.

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554–566. https://doi.org/10.1016/j.cell.2015.01.006

Villar, D., Flicek, P., & Odom, D. T. (2014). Evolution of transcription factor binding in metazoans-mechanisms and functional implications. *Nature Reviews Genetics*, *15*(4), 221–233. https://doi.org/10.1038/nrg3481

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. a, Holt, A., … Pennacchio, L. a. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. https://doi.org/10.1038/nature07730

Visel, A., Bristow, J., & Pennacchio, L. A. (2007). Enhancer identification through comparative genomics. *Seminars in Cell and Developmental Biology*, *18*(1), 140–152. https://doi.org/10.1016/j.semcdb.2006.12.014

Vogel, M. J., Peric-Hupkes, D., & van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nature Protocols*, *2*(6), 1467–78. https://doi.org/10.1038/nprot.2007.148

Walldorf, U., Binner, P., & Fleig, R. (2000). Hox genes in the honey bee Apis mellifera. *Development Genes and Evolution*, *210*(10), 483–492. https://doi.org/10.1007/s004270050337

Wegerhoff, R., & Breidbach, O. (1992). Structure and development of the larval central complex in a holometabolous insect, the beetle Tenebrio molitor. *Cell & Tissue Research*, *268*(2), 341–358. https://doi.org/10.1007/BF00318803

Wegerhoff, R., Breidbach, O., & Lobemeier, M. (1996). Development of locustatachykinin immunopositive neurons in the central complex of the beetle Tenebrio molitor. *Journal of Comparative Neurology*, *375*(1), 157–166. https://doi.org/10.1002/(SICI)1096-9861(19961104)375:1<157::AID-CNE10>3.0.CO;2-S

Wegner, M. (1999). From head to toes: The multiple facets of Sox proteins. *Nucleic Acids Research*, *27*(6), 1409–1420. https://doi.org/10.1093/nar/27.6.1409

Wei, L., Cheng, D., Li, D., Meng, M., Peng, L., Tang, L., … Lu, C. (2011). Identification and characterization of Sox genes in the silkworm, Bombyx mori. *Molecular Biology Reports*, *38*(5), 3573–3584. https://doi.org/10.1007/s11033-010-0468-5

Wheeler, S. R., Carrico, M. L., Wilson, B. A., & Skeath, J. B. (2005). The Tribolium columnar genes reveal conservation and plasticity in neural precursor patterning along the embryonic dorsal-ventral axis. *Developmental Biology*, *279*(2), 491–500. https://doi.org/10.1016/j.ydbio.2004.12.031

Wheeler, S. R., Carrico, M. L., Wilson, B. a, Brown, S. J., & Skeath, J. B. (2003). The expression and function of the achaete-scute genes in Tribolium castaneum reveals conservation and variation in neural pattern formation and cell fate specification. *Development*, *130*(18), 4373–4381. https://doi.org/10.1242/dev.00646

Whelan, S., & Goldman, N. (2001). A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*, *18*(5), 691–699. https://doi.org/10.1093/oxfordjournals.molbev.a003851

Whitfield, L. S., Lovell-Badge, R., & Goodfellow, P. N. (1993). Rapid sequence evolution of the mammalian sex-determining gene SRY. *Nature*, *364*(6439), 713–5. https://doi.org/10.1038/364713a0

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., … Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, *153*(2), 307–319. https://doi.org/10.1016/j.cell.2013.03.035

Wilson, M. E., Yang, K. Y., Kalousova, A., Lau, J., Kosaka, Y., Lynn, F. C., … German, M. S. (2005). The HMG box transcription factor Sox4 contributes to the development of the endocrine pancreas. *Diabetes*, *54*(12), 3402–3409. https://doi.org/10.2337/diabetes.54.12.3402

Wilson, M. J., & Dearden, P. K. (2008). Evolution of the insect Sox genes. *BMC Evolutionary Biology*, *8*(120), 1–13. https://doi.org/10.1186/1471-2148-8-120

Wilson, M. J., McKelvey, B. H., van der Heide, S., & Dearden, P. K. (2010). Notch signaling does not regulate segmentation in the honeybee, Apis mellifera. *Development Genes and Evolution*, *220*(7–8), 179–190. https://doi.org/10.1007/s00427-010-0340-6

Wolff, C., Sommer, R., Schröder, R., Glaser, G., & Tautz, D. (1995). Conserved and divergent expression aspects of the Drosophila segmentation gene hunchback in the short germ band embryo of the flour beetle Tribolium. *Development*, *121*(12), 4227–4236.

Wolpert, L., Tickle, C., & Martinez Arias, A. (2005). *Principles of Development* (Fifth). Oxford University Press.

Wood, H. B., & Episkopou, V. (1999). Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mechanisms of Development*, *86*(1–2), 197–201. https://doi.org/10.1016/S0925-4773(99)00116-1

Wright, E., Hargrave, M. R., Christiansen, J., Cooper, L., Kun, J., Evans, T., … Koopman, P. (1995). The Sry-related gene Sox9 is expressed during chondrogenesis in mouse embryos. *Nature Genetics*, *9*(1), 15–20. https://doi.org/10.1038/ng0195-15

Wright, E. M., Snopek, B., & Koopman, P. (1993, February 11). Seven new members of the Sox gene family expressed during mouse development. *Nucleic Acids Research*. Oxford University Press. https://doi.org/10.1093/nar/21.3.744

Wu, T. P., Wang, T., Seetin, M. G., Lai, Y., Zhu, S., Lin, K., … Xiao, A. Z. (2016). DNA methylation on N6-adenine in mammalian embryonic stem cells. *Nature*, *532*(7599), 1–18. https://doi.org/10.1038/nature17640

Yan, Y.-L., Willoughby, J., Liu, D., Crump, J. G., Wilson, C., Miller, C. T., … Postlethwait, J. H. (2005). A pair of Sox: distinct and overlapping functions of zebrafish sox9 co-orthologs in craniofacial and pectoral fin development. *Development*, *132*(5), 1069–1083. https://doi.org/10.1242/dev.01674

Yan, Y. L., Miller, C. T., Nissen, R. M., Singer, A., Liu, D., Kirn, A., … Postlethwait, J. H. (2002). A zebrafish sox9 gene required for cartilage morphogenesis. *Development*, *129*(21), 5065–5079.

Yang, Z. (1998). On the Best Evolutionary Rate for Phylogenetic Analysis. *Systematic Biology*, *47*(1), 125–133. https://doi.org/10.1080/106351598261067

Yin, R., Mao, S. Q., Zhao, B., Chong, Z., Yang, Y., Zhao, C., … Wang, H. (2013). Ascorbic acid enhances tet-mediated 5-methylcytosine oxidation and promotes DNA demethylation in mammals. *Journal of the American Chemical Society*, *135*(28), 10396–10403. https://doi.org/10.1021/ja4028346

Yu, G. (2014). ChIPseeker: an R package for ChIP peak Annotation, Comparision and Visualization ChIP profiling. *Bioconductor*.

Yu, J., Zhang, L., Li, Y., Li, R., Zhang, M., Li, W., … Bao, Z. (2017). Genome-wide identification and expression profiling of the SOX gene family in a bivalve mollusc Patinopecten yessoensis. *Gene*, *627*(June), 530–537. https://doi.org/10.1016/j.gene.2017.07.013

Zhang, G., Huang, H., Liu, D., Cheng, Y., Liu, X., Zhang, W., … Chen, D. (2015). N6-methyladenine DNA modification in Drosophila. *Cell*, *161*(4), 893–906. https://doi.org/10.1016/j.cell.2015.04.018

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. *International Journal of Developmental Biology*, *51*(2), 107–115. https://doi.org/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. https://doi.org/10.1371/journal.pone.0016570

Zhu, L. J., Gazin, C., Lawson, N. D., Pagès, H., Lin, S. M., Lapointe, D. S., & Green, M. R. (2010). ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, *11*(237), 1–10. https://doi.org/10.1186/1471-2105-11-237