Profiling Doublesex-DNA interactions with targeted DamID in *Drosophila melanogaster*

Luis A. D'Souza

Lincoln College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2019

For my family.

DECLARATION

This thesis and the work presented herein are my own unless otherwise specified. All work was carried out at the Centre for Neural Circuits and Behaviour, University of Oxford. This thesis is submitted for the degree of DPhil in Physiology, Anatomy and Genetics and has not been submitted for any other degree.

The following experiments were conducted in collaboration with others:

i. Library preparation and NGS of TaDa samples were completed by Jérôme Nicod (DSX-brain) and Sarah Lamble (DSX-head) at the Oxford Genomics Centre.

ii. Bettina Fischer (University of Cambridge) assisted with implementing damidseq_pipeline and find_peaks R scripts.

iii. Imaging *dsx*/ neuropeptide^{AD} neuronal co-expression was completed by Tetsuya Nojima. I generated transgenic lines, raised stocks, and set up crosses for imaging.

iv. Yeast-one hybrid data of *Drosophila* transcription factors kindly provided by Korneel Hens (unpublished, 2017).

This work was funded by the University of Oxford, DPAG and the MRC.

ABSTRACT

The transcriptional regulators *doublesex* (*dsx*) and *fruitless* (*fru*) function together to specify sexual physiology and behaviour in *Drosophila melanogaster*. Whilst *fru* has only been located in insect lineages, *dsx* is structurally and functionally conserved throughout the animal kingdom. The role of FRU proteins as transcriptional regulators has been studied extensively with DamID – a method that uses DNA adenine methylation to create 'fingerprints' in the genome that correspond to where Dam-tagged proteins have interacted with DNA. Comparatively little research has focused on DSX, and there are few direct *dsx* target genes in the CNS. In this study, we implement the novel targeted DamID (TaDa) approach for profiling Dsx^M and Dsx^F expressing neurons in the *Drosophila* adult CNS, marking the first time these neurons will be profiled in a cell-specific manner. TaDa harnesses the phenomenon of low frequency ribosome re-initiation, and the introduction of a primary open reading frame (ORF) encoded before the Dam-fusion, to enable low level expression of the Dam-fusion in a cell-type specific manner when driven using Gal4/UAS. We generate a rich list of putative *dsx* targets in the CNS. Comparisons with a published DSX-fat body DamID dataset, an organ in which *dsx* is known to be widely expressed, reveal a tendency towards tissue-specificity. Our Gene Ontology analyses reveal DSX targets are involved in nervous system development, confirming the known role attributed to *dsx*, and motif analyses identify novel putative GATA cofactors. Alongside independent protein-DNA interaction screens (ChIP-seq and yeast-one hybrid), we identify the neuropeptides Diuretic Hormone 31 (*Dh31*), Tachykinin 1 (*Tk1*) and Neuropeptide-like precursor 1 (*Nplp1*) for further analysis. *dsx*/ *Nplp1* expression analyses reveal a single cluster of neurons in the female brain, likely pMN1, which we speculate is involved in post-mating behaviour. The wealth of putative *dsx* target genes generated here pave the way for further characterisation of the *dsx* machinery and thus a better understanding of its role in specifying sexual physiology and behaviour.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank Prof Lars Chittka for igniting a passion in studying insect behaviour, and Prof Robert Wilkins for leading me to this foray into fly neurobiology. It's been a truly remarkable experience.

To my supervisors, Prof Stephen F. Goodwin and Dr Megan Neville, thank you for allowing me the freedom to pursue my scientific interests. I am grateful for your guidance, support, and advice over the past four years. I have learnt the virtue and importance of scientific independence. It's been wonderful to get to know members of the Goodwin lab. Carolina, Hania and Tetsuya, thank you for your encouragement and kindness from back in the heady Sherrington days. Ammerins, I appreciate your thorough technical support when I was just getting started in the wet lab, and for your advice out of it, thank you. Thank you Bettina Fischer for overseeing my training in Cambridge, and for the whirlwind personal bioinformatics course. Your guidance has been invaluable in shaping my understanding of a field completely new to me. Indeed, your technical direction has been profound in shaping the latter part of this project.

Thanks Jess for the solace of sanity and wisdom in our CNCB tea breaks. Thanks for the encouragement this summer Max, writing up in Lincoln library without you would have been significantly harder. And thank you, Johanna, for being there through it all. Some of my best memories in life have been made in Oxford. It will be sorely missed. I feel privileged to have been given the chance to pursue a DPhil at Oxford by the Department of Physiology, Anatomy and Genetics, and I am thankful to DPAG, the MRC, and Lincoln College who funded this research.

Most of all, thank you to my family for your unrelenting love and unwavering support on this journey. To my parents, you have made this possible. This is for you.

CONTENTS

List of Abbreviations

Abg	abdominal ganglion
AD	Activation domain
ChIP	Chromatin Immunoprecipitation
CNS	Central Nervous System
DamID	DNA adenine methyltransferase Identification
DBD	DNA binding domain
dsx	*doublesex*
FB	Fat body
FBE	Fat body enhancer
fru	*fruitless*
gDNA	genomic Deoxyribonucleic Acid
OGC	Oxford Genomics Centre
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
POI	Protein of Interest
RE	Regulatory Element
SDH	Sex Determination Hierarchy
TaDa	Targeted DamID
TE	Transposable Element
TF	Transcription Factor
TFBS	Transcription Factor Binding Site
VNC	Ventral Nerve Cord

1 INTRODUCTION

1.1 GENES, THE BRAIN AND BEHAVIOUR

Neuroscientists have worked tirelessly to unravel the relationship between genes and behaviour since the turn of the twentieth century. Indeed, Sir Francis Galton conducted the first systematic studies investigating the hereditability of human behaviour over a century ago (Galton, 1874). However, decoding connections between genotype and complex behavioural traits is difficult, because it necessitates an understanding of the physiological functions of the underlying neural substrate as well as how genes regulate their development. The vinegar fly, *Drosophila melanogaster*, is an excellent model organism for decoding these connections. The availability of its well characterised sequenced genome, small genome size (~180 Mb), the simplicity of its chromosome complement (four easily identifiable pairs of chromosomes) (Adams et al., 2000), plus a significant genetic toolkit, makes *D. melanogaster* a powerful genetic system. This genetic toolkit has developed to allow sophisticated genetic manipulations, to single neuron resolution, to aid understanding of the complexity of the nervous system to

unprecedented detail. Also, practically *D. melanogaster* has many benefits for genetic research. For instance, rearing *D. melanogaster* in laboratory conditions is relatively easy and inexpensive, they have a short generation time (10 days at 25 °C), they have a high tolerance of inbreeding, and they produce large numbers of externally laid embryos (every female can lay over 100 eggs). The versatility of *D. melanogaster* has enabled its effective use for over a century to study a diverse range of biological processes, contributing to our understanding of, for example, nervous system development (Doe, 2008; Finkelstein et al., 1990; Hartenstein et al., 2008), exocytosis and endocytosis at synapses (Bellen et al., 2010; Richmond and Broadie, 2002; Südhof, 2004), the neural circuits underlying behaviours such as diurnal rhythms and sleep (Crocker and Sehgal, 2010; Donlea et al., 2014; Vosshall et al., 1994; Zehring et al., 1984), aggression (Asahina et al., 2014; Koganezawa et al., 2016; Kravitz and Huber, 2003), learning and memory (Dudai et al., 1976; Krashes et al., 2007; McGuire et al., 2005), and courtship (Manoli et al., 2013; Pavlou and Goodwin, 2013; Rezával et al., 2012; Villella and Hall, 2008).

1.2 COURTSHIP BEHAVIOUR IN *DROSOPHILA*

The behavioural courtship ritual performed by *D. melanogaster* males towards a conspecific female is a classic example of an animal exhibiting a sex-specific, innate behaviour (Baker et al., 2001; Billeter et al., 2006; Hall, 1979). The sequence of behaviours that comprise the courtship ritual are robust and highly stereotyped and largely in the domain of the male: he will follow the female (Cook, 1979; Manning, 1967; Markow, 1987; Tompkins et al., 1983), tap her abdomen with his forelegs (Amrein and Thorne, 2005; Greenspan and Ferveur, 2000; Manning, 1967), sing a

species-specific courtship song by extending and vibrating one wing (Bennet-Clark and Ewing, 1969; von Schilcher, 1976b, 1976a), use his mouthparts to contact her genitalia (Tompkins et al., 1983), then bend his abdomen to attempt to mount and copulate. *D. melanogaster* females assess the courting male for species type and fitness before sanctioning mating. Females unwilling to mate display rejection behaviours by kicking or extruding their ovipositor, an organ that facilitates the laying of eggs (Cook and Connolly, 1973; Ejima et al., 2001; Rideout et al., 2007; Spieth and Ringo, 1983; Villella et al., 2008). If she accepts his advances, she will slow down and open her vaginal plates to facilitate copulation (Figure 1; Villella et al., 2008). A mixture of sperm and seminal fluid is transferred to fertilise her eggs, triggering a change in her physiological state and behaviour. This post-mating switch causes mated females to be temporarily sexually unreceptive to further copulatory events, increase their display of rejection behaviours, and increase egg laying (Chapman et al., 2003; Kubli, 2003; Liu and Kubli, 2003; Rezával et al., 2012).

The behaviour of *D. melanogaster* males during the courtship ritual was first described by Sturtevant (1915), but the complete genetic basis of these behaviours is still unknown. Most studies decoding courtship behaviour have focused on two pivotal transcription factors of the sex-determination hierarchy (described later), *doublesex* (*dsx*) and *fruitless* (*fru*), which act in concert to specify sex-specific physiology and neural circuitry (Clough et al., 2014; Dauwalder, 2011; Kimura et al., 2008; Neville et al., 2014; Pavlou et al., 2013; Yamamoto and Koganezawa, 2013). Whilst *dsx* has been studied for over fifty years, there are still few defined direct target genes, and even those cannot fully explain the full repertoire of behaviours that comprise the behavioural courtship ritual in *D. melanogaster*.

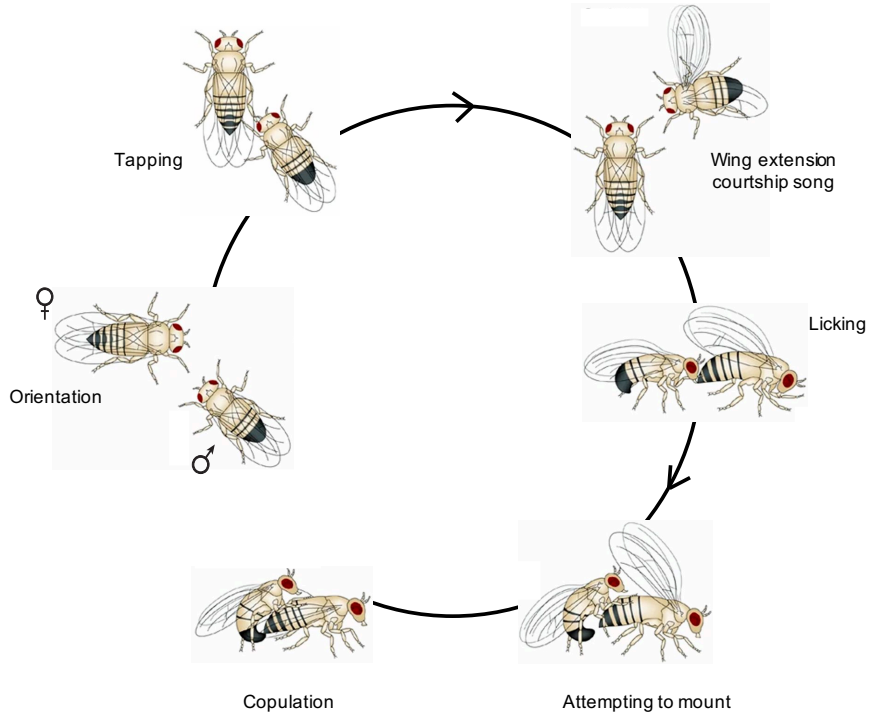


Figure 1 *The courtship ritual in Drosophila melanogaster (adapted from Sokolowski et al., 2001).*

1.3 THE HISTORY OF *DROSOPHILA* BEHAVIOURAL GENETICS

Thomas Hunt Morgan and his students, Sturtevant, Bridges, and Muller greatly refined the theory of inheritance, first proposed by Gregor Mendel, using *Drosophila* in the early 1910s. Their seminal work provided important evidence for the chromosome theory of heredity through correlating the inheritance of red- and white- eyed variants with the segregation of the X chromosome (Bridges, 1916; Morgan, 1910). Morgan's group were the first to delineate linkage of genes in a linear order on a chromosome (Sturtevant, 1913), and to dissect chromosomal rearrangements, beginning with genic deficiencies (Bridges, 1917). These major scientific breakthroughs in *Drosophila* research founded the field of modern genetics. Bridges further established so-called polytene chromosomes as a key tool for chromosome mapping (Bridges, 1935),

following their discovery in other insects in the 1930s, and their initial description in *D. melanogaster* salivary glands by Painter (1933). Polytene chromosomes consist of many replicated but not separated DNA molecules which are easily visualised by staining, resulting in a unique banding pattern for each chromosome.

While Morgan and his students had established a correlation between genotype and phenotype, in-depth causational studies required the ability to manipulate genetic data. Muller first discovered gene mutations are inducible by X-rays in a dose-dependent manner through examining offspring of irradiated flies (Muller, 1927). This seminal work on the nature and origin of mutations paved the way for scientists to generate (using radiation and later chemicals) vast collections of mutant strains for dissecting countless biological processes. Lewis and Bacher (1968) later described an efficient method to induce point mutations using ethyl methanesulphonate (EMS). EMS mutagenesis facilitates an important tool of *D. melanogaster* research, namely the power of forward genetic screens to dissect the genes that affect a specific phenotype. Lewis' later work using radiation-induced and spontaneous mutants focused on development of the body plan in *D. melanogaster* (Lewis, 1978). Soon after, Nüsslein-Volhard and Wieschaus (1980) conducted large-scale mutagenesis screens to identify and describe mutations affecting the early patterning steps of embryonic development. These seminal studies enabled the in-depth molecular characterisation of conserved genes that govern the basic body plan of most metazoans. Lewis, Hogness and colleagues later completed the first cloning and genomic analysis of a gene, Ubx, within the Bithorax complex (Bender et al., 1983), setting the stage for the broad use of molecular biology in *Drosophila* genetics.

Mutational analyses also led to breakthrough findings in behavioural genetics, with Benzer identifying the first known genes associated with circadian rhythms, learning and memory (Dudai et al., 1976; Konopka and Benzer, 1971). Benzer and Konopka identified *D. melanogaster* flies with defective 24 h rhythms, which all had a mutation in the same gene. In 1984, Hall and Rosbash molecularly characterised this gene called *period* (*per*) (Zehring et al., 1984), finding *per* protein levels oscillated with a 24 h pattern. A decade later, Young identified a second core clock protein, *tim*, encoded by the *timeless* gene (Vosshall et al., 1994). Alongside additional clock proteins, they concluded that a self-sustaining transcription-translation feedback loop governs the 24 h rhythms observed. Indeed, the 1980s and 1990s saw significant growth in the molecular characterisation of many *D. melanogaster* genes and showed how conserved many developmental pathways were between flies and mammals. For instance, Gehring's work showed that the role of the Pax6 gene as a master regulator of eye development was shared between flies and humans (Gehring, 1996). This rise in the molecular understanding of *Drosophila* genes was enabled, in part, by the Gal4/UAS system developed for use in *D. melanogaster* by Brand and Perrimon (1993), allowing spatiotemporally controlled transgene expression. Functioning as a binary expression system, the yeast transcription factor Gal4 could be expressed in a tissue specific and/or time sensitive manner, activating reporter transgenes under the control of a UAS promoter.

Further facilitating the study of *Drosophila* genetics, the fly genome was sequenced and published in 2000, just eleven months ahead of the human genome (Adams et al., 2000), and knock-out technology was discovered around the turn of the 21st century (Rong and Golic, 2000, 2001). The 2010s have seen the development of novel targeted

gene knock-down and knock-out approaches using RNA interference (RNAi) and the CRISPR/Cas9 system. The latter has quickly and dramatically expanded in popularity and use (Bassett and Liu, 2014; Gasiunas et al., 2012; Gratz et al., 2013; Jinek et al., 2012, 2013; Zhang et al., 2013).

1.4 CHROMOSOMAL SEX DETERMINATION IN *DROSOPHILA*

Sex determination in *D. melanogaster* was one of the first major developmental processes shown to be under genetic control (Bridges, 1921). Indeed, both chromosomal and genic components appear to be involved in this regulation. At the chromosomal level, Bridges' seminal studies (1916, 1921, 1922, 1925, 1932, 1939) of flies with abnormal chromosome complements indicated the Y chromosome is not involved in determining sex. Instead, sexual fate is governed by the balance of female determinants on the X chromosome (X) and male determinants on the autosomes (A). Flies have either one or two X chromosomes and two sets of autosomes. Accordingly, if there is one X chromosome in a diploid cell (1X:2A) the fly is male; and if there are two X chromosomes in a diploid cell (2X:2A), the fly is female (Bridges, 1921, 1925; Steinmann-Zwicky and Nöthiger, 1985). This ratio, or balance, between female determinants on the X chromosome and male determinants on the autosomes governs which sex-specific pattern of transcription will be initiated. Gynandromorphs, animals which are composed of both male and female cells, are observed in *Drosophila*. This could occur if one of the two X chromosomes of a *Drosophila* embryo is lost, then the cells that descend from that cell are XO (male) instead of being XX (female). Every cell lineage makes its own sexual 'decision' independent of its neighbours. XO cells display male characteristics whilst XX cells display female characteristics.

1.4.1 Dosage compensation

Although originally homologous to the X chromosome, the Y chromosome has degenerated over time, creating an imbalance in X-linked gene products between males and females. The chromosomal basis of sex determination thus poses problems given that homogametic females (XX) have twice as many X-linked genes as the heterogametic males (XY). The imbalance in X chromosomes is dealt with by dosage compensation mechanisms that equalise X-linked gene expression between the two sexes, ensuring the appropriate balance of X-chromosomal and autosomal gene products in each sex (Charlesworth, 1996). Muller first described the occurrence of dosage compensation in *D. melanogaster* almost ninety years ago, whilst studying eye pigment levels of individuals carrying partial loss-of-function X-linked mutations (Muller, 1932). The X:A ratio controls both dosage compensation and sex determination by regulating a critical binary switch, *Sex lethal* (*Sxl*).

1.4.2 The sex determination hierarchy

The sex determination hierarchy (SDH) in *D. melanogaster* portrays the coordination of the development and differentiation of sex-specific tissues, physiology and neural circuitry. Atop the SDH, *Sxl* functions as the female or male master switch of fly sex determination (Cline and Meyer, 1996; Cronmiller and Salz, 1994), transcribed when the X:A ratio equals or is greater than one. Hence, *Sxl* activity is 'on' in XX animals whereby the programme of female development follows. *Sxl* expression in XX females additionally prevents activation of the male-specific dosage compensation system. *Sxl* remains 'off' in XY animals, dosage compensation is activated, and the programme of male development follows (Figure 2) (Bashaw and Baker, 1997; Kelley et al., 1997).

Sxl controls the male specific lethal (MSL) complex, increasing transcript levels on the single male X chromosome to equal transcript levels in XX females (Baker et al., 1994). The MSL is a ribonucleoprotein complex enriched on the X chromosome and made up of at least five proteins, the *male-specific lethal 1* (*msl*-1) scaffolding protein, the *male-specific lethal 2* (*msl*-2) RING finger protein, the *male-specific lethal 3* (*msl*-3) chromodomain protein, *males absent on the first* (MOF) histone acetyltransferase, *maleless* (MLE) RNA helicase, and two functionally redundant long non-coding RNAs: *RNA on the X 1* (roX1) and *RNA on the X 2* (roX2). By virtue of sitting at the top of a regulatory cascade that includes dosage compensation, loss of *Sxl* function in XX animals results in female specific lethality, and inappropriate *Sxl* expression in XY animals leads to male-specific lethality (Salz and Erickson, 2010).

Differential splicing occurs in XX and XY animals when *Sxl* binds its downstream target, the *transformer* (*tra*) gene, generating protein-coding *tra* mRNA in females only (Figure 2; Yamamoto et al., 2013). Expression of female-specific *tra* alongside non-sex-specific *transformer-2* (*tra 2*) creates a sex-specific splice in the pivotal downstream transcription factors (TFs) *doublesex (dsx)* and *fruitless* (*fru*). In females, *dsx* transcripts are spliced to generate the female-specific isoform, Dsx^F, whilst a stop codon introduced into *fru* mRNAs truncates their translation to generate a non-functional peptide. In males, *dsx* and *fru* transcripts undergo default splicing to generate the male specific isoforms Dsx^M and Fru^M, respectively (Burtis and Baker, 1989; Nagoshi et al., 1988). *dsx* and *fru*, found at the bottom of the sex-determination hierarchy, specify the ability to display these sexually dimorphic courtship behaviours discussed above (Baker et al., 2001; Billeter et al., 2006; Burtis, 1993; Christiansen et

al., 2002; Cline et al., 1996; Rideout et al., 2007; Shirangi and McKeown, 2007; Villella et al., 2008).

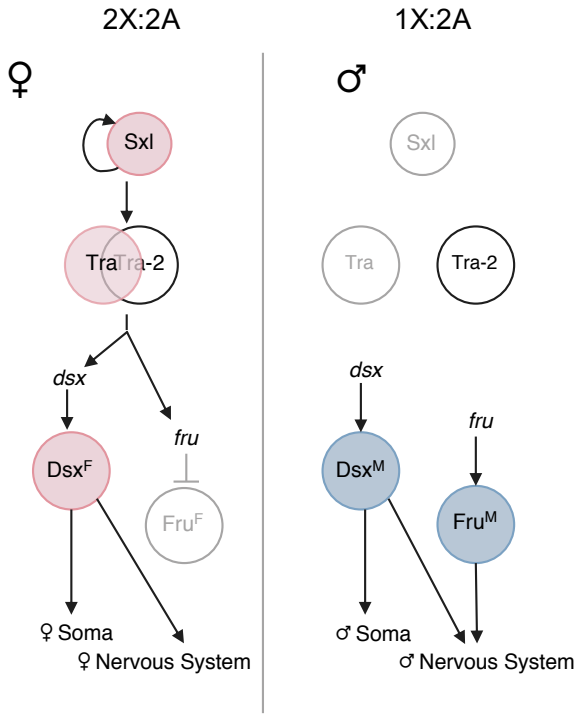


Figure 2 *The SDH in Drosophila melanogaster: sex-specific alternative cascades leading to generation of XX females (left) and XY males (right) (adapted from Billeter et al., 2006).*

1.4.3 DOUBLESEX

The *Drosophila dsx* gene is conserved in species as diverse as worms (mab3 in *Caenorhabditis elegans*) and mammals (*doublesex* and *mab-3* related transcription factor 1, DMRT1). Along with other *mab-3* related TFs (DMRTs), *dsx* has been experimentally shown to control sex determination and differentiation (Zarkower, 2013). For example, in male XY humans, deletions of three DMRT genes have

proposed to be involved in sex reversal (Hong et al., 2007; Raymond et al., 1999). Structurally, Dsx^F and Dsx^M isoforms share the same DNA-binding and dimerisation domains, but differ in their C-termini (Bayrer et al., 2005; Zhang et al., 2006). Intersex (IX) binds the C terminus of Dsx^F and is essential for its function (Yang et al., 2008). It is likely therefore that the DSX C-terminus modulates gene expression by acting as an effector domain which interacts with cofactors.

There are 400-700 *dsx*-expressing neurons in the male CNS (Lee et al., 2002; Pavlou et al., 2016; Rideout et al., 2010) and 300-400 *dsx*-expressing neurons in the female CNS (Figure 3; Pavlou et al., 2016; Rideout et al., 2010). Generation of the dsx^{Gal4} allele – whereby the yeast transcription factor *Gal4* was targeted to the *dsx* locus (Rideout et al., 2010) – has helped towards a better understanding of anatomy and function of *dsx*-expressing cells by specifying the *dsx* expression pattern in the CNS. The dsx^{Gal4} allele revealed *dsx* expression in the fat body, oenocytes and reproductive organs – tissues fundamental to specifying sex and physiological state (Rideout et al., 2010). Taken together then, *dsx* plays a fundamental role in the development and differentiation of sex-specific neural circuits and tissues, including the establishment and maintenance of sex-specific physiology.

Dsx^F and Dsx^M are necessary for proper sexual development in *Drosophila* (Coschigano and Wensink, 1993). Expression of both isoforms in the same fly generates an intersexual phenotype (Nagoshi and Baker, 1990). Activation of all *dsx* neurons in Fru^M-null males stimulates robust courtship behaviours (Pan et al., 2011), whilst their inhibition prevents the display of all courtship behaviours (Rideout et al., 2010). *dsx*

plays a key role in coordinating male copulation: sexually dimorphic *dsx*/glutamatergic neurons control genital coupling, and *dsx*/GABAergic neurons in the abdominal ganglion are involved in the termination of copulation (Pavlou et al., 2016). Further, thermogenetic activation of *dsx* neurons in the female brain induces female animals to perform male-like courtship behaviours towards conspecific males or females, and other *Drosophila* species (Rezával et al., 2016).

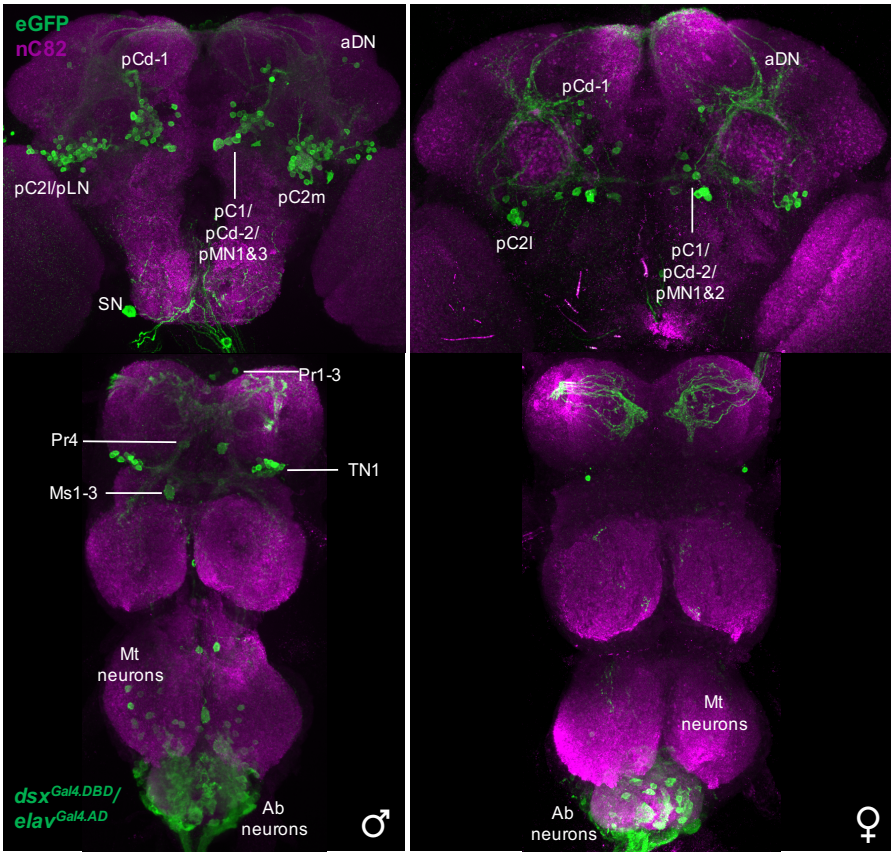


Figure 3 *Sexually dimorphic expression of dsx^{Gal4} in the CNS of male and female adult Drosophila melanogaster. 5-7 day-old male and female brain and VNC, dsx^{Gal4}∩elav^{AD}>mCD8::eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta) (Nojima unpublished, 2017).*

1.4.4 KNOWN DSX TARGET GENES

Presently, there are just few defined and validated direct *dsx* target genes. These are the Yolk Proteins (*yp-1* and *yp-2*), *bric-à-brac 1 (bab1)*, and *desatF* (also known as *Fad2*) (Burtis et al., 1991; Coschigano et al., 1993; Luo and Baker, 2015; Shirangi et al., 2009; Williams et al., 2008). At the molecular level, regulation of yolk protein gene expression by DSX is well characterised. Dsx^M and Dsx^F bind directly to a number of sites in an enhancer sequence designated the fat body enhancer (FBE) (Burtis et al., 1991; Coschigano and Wensink, 1993). Binding of Dsx^M to the FBE significantly reduces expression of the yolk protein genes. Dsx^F, however, cooperates with additional factors, such that binding of the same FBE sequences activates transcription (An and Wensink, 1995a, 1995b). In this way, DSX can act as either a transcriptional activator or repressor (Suzuki and Shimada., 2013). The *Drosophila* fat body (FB) is a multifunctional tissue involved in energy storage, immune response, and nutritional sensing, where it is able to release energy according to demands (Bownes and Hames, 1978; Haunerland, 1996; Meister et al., 1997; Yongmei Xi, 2015). The yolk proteins, synthesised in the FB (Bownes and Hames, 1978), provide a nutritional supply and bind conjugated hormones necessary for embryonic development (Bownes et al., 1988; Butterworth et al., 1992).

The gene *bab1* is part of the conserved proximal-distal gene regulatory network module. It is involved in the development of a number of sexually dimorphic features including pattern formation and ovary morphogenesis (Chen et al., 1995; Godt et al., 1993; Williams et al., 2008). *bab1* has been shown to be directly regulated by DSX in the gonad (Camara et al., 2019), and directing sex-specific pigmentation in the

abdomen (Williams et al., 2008). *Fad2* encodes a female-specific transmembrane fatty acid desaturase involved in sex pheromone signalling (Shirangi et al., 2009), and is directly regulated by DSX in oenocytes (Shirangi et al., 2009) and indirectly in the fat body (Clough et al., 2014).

1.4.5 FRUITLESS

Whilst the *Drosophila fru* gene does not have an obvious mammalian homolog like *dsx* does, it is known to function in sex determination in other species such as the mosquito *Anopheles gambiae* (Gailey et al., 2006). Structurally, all Fru proteins comprise a common BTB (protein-protein interaction) N-terminal domain, and one of three C-terminal zinc-finger (Zn-finger) DNA-binding domains through alternative splicing (Ito et al., 1996; Ryner et al., 1996; Usui-Aoki et al., 2000). A number of BTB-Zn-finger proteins function as transcriptional regulators, some with key roles in development (Siggs and Beutler, 2012). Fru^M proteins are generated from *fru* transcripts whose expression initiates from the most distal promoter, P1, under SDH control. Male-specific Fru^M proteins share a 101 amino acid male-specific sequence and one of three (A, B, or C) alternative C_2H_2 Zn-finger domains (Billeter et al., 2006). Expression of Fru^M proteins are noted in the nervous system at the origin of metamorphosis when the CNS remodels from larval to adult form (Lee et al., 2000).

There are ~1,700 *fru*-expressing neurons in the adult male CNS (Lee et al., 2000; Stockinger et al., 2005), and ~3,000 *fru*-expressing neurons in the female CNS (Pavlou et al., 2016; Rideout et al., 2010). *fru* functions in the control of male sexual behaviour (Villella et al., 2008). *fru* alleles that perturb male sexual behaviour include

chromosomal insertions, deletions, or rearrangements disrupting P1 transcripts specifically (Anand et al., 2001; Goodwin et al., 2000). Therefore male-specific splicing of *fru* P1 transcripts gives rise to male sexual behaviour and orientation (Baker et al., 2001).

In its regulation of sexual behaviour, *fru* functions by encoding putative transcription factors that control development of sexually dimorphic neuronal circuitry. It has been shown that *fru* acts with the transcriptional cofactor Bonus (Bon) generating a complex that then interacts with key chromatin regulators (Ito et al., 2012). These chromatin regulators establish male-specific neurite projections and dendritic branching – Histone deacetylase 1 (HDAC1) masculinises individual sexually dimorphic neurons, and conversely Heterochromatin protein 1a (HP1a) demasculinises them. *fru* can act therefore as both a transcriptional activator or repressor through forming antagonistic chromatin regulating complexes that masculinise or demasculinise specific neuronal subtypes (Ito et al., 2012). Males that lack Fru^M expression display courtship behaviours towards other males, but exhibit no behaviour that would indicate sexual interest towards females (Demir and Dickson, 2005). Females constitutively expressing Fru^M can only perform some steps of the courtship ritual and at subnormal thresholds – initiation, orientation, following, tapping and wing extension – towards *wild type* females, but cannot generate recognisable song nor attempt copulation (Demir et al., 2005; Rideout et al., 2007). Optogenetic activation of all *fru*-expressing neurons in intact and decapitated flies elicits unilateral wing extension and courtship song in males and females (Clyne and Miesenböck, 2008). However, to generate *bona fide* courtship song, both Fru^M and Dsx^M must be expressed in the CNS along with direction from male-specific higher order command neurons (Clyne et al., 2008; Rideout et al., 2007).

1.4.6 THE *FRU*/*DSX*-EXPRESSING P1 NEURONAL COMMAND CENTRE

The P1 neuronal cluster in the dorsal posterior brain in *D. melanogaster* near the mushroom body is regarded as a fundamental higher-order command centre that functions to initiate male-type courtship behaviour (Kimura et al., 2008). P1 is a male-specific *fru*/ *dsx*-coexpressing cluster composed of 20 interneurons, each of which has a primary transversal neurite with extensive ramifications in the bilateral protocerebrum. In females, P1 is fated to die through the action of the feminising protein Dsx^F. In males, the masculinising protein, Fru, is necessary in the male brain for correct positioning of the terminals of the P1 neurites (Kimura et al., 2008; Lee et al., 2000, 2002; Ren et al., 2016; Zhou et al., 2014). Therefore, the coordinated function of *dsx* and *fru* is required to confer the ability to initiate male-typical sexual behaviour on P1 neurons. In males, neuronal silencing of P1 neurons impairs courtship song production amongst other courtship elements (Crickmore and Vosshall, 2013; Pan and Baker, 2014). Their thermogenetic activation initiates the courtship ritual and triggers pulse song with or without another female present (Crickmore et al., 2013; Pan et al., 2014). Artificially inducing a P1 clone in females initiates male-typical behaviour in such females even when other parts of the brain are not masculinised (Kimura et al., 2008).

1.5 GENOMICS APPROACHES TO ASSESS GLOBAL PROTEIN-DNA INTERACTIONS

Protein-DNA interactions are pivotal for key biological processes in living cells, including transcriptional regulation and DNA modification. To study these complex interactions, a number of experimental methods have been developed. Each of these

methods has varying utility depending on the experimental question, and with relative advantages and disadvantages. Here, we discuss two complementary techniques, Chromatin immunoprecipitation (ChIP) and DNA adenine methyltransferase (DamID), that have gained significant momentum as a means of generating global maps of protein-DNA interactions.

1.5.1 CHROMATIN IMMUNOPRECIPITATION (CHIP)

The first attempts to assay protein occupancy in a living cell, that is to determine where proteins bind to chromatin *in vivo*, began more than thirty years ago with the development of Chromatin immunoprecipitation (ChIP) (O'Neill and Turner, 1996; Orlando and Paro, 1993; Solomon et al., 1988; Turner and O'Neill, 1995). Chromatin structure is dynamic, made up of DNA, proteins and RNA (Bernstein, 2005; Felsenfeld and Groudine, 2003; Schübeler and Elgin, 2005), responding to intra- and extracellular signals and controlling gene expression, DNA replication and repair (Felsenfeld et al., 2003; Sims, 2004; Thiriet and Hayes, 2005). In ChIP, antibodies against a protein of interest (POI) or nucleosome enrich DNA fragments at loci that bind specifically to the POI or nucleosome. The DNA-binding-POI is immunoprecipitated along with its cognate DNA, and DNA binding at specific sites determined (Nelson et al., 2006; Ostrowski et al., 2003). The introduction of microarrays enabled fragments located from ChIP to be identified by hybridisation to a microarray (ChIP-chip). Microarrays allowed a global view of DNA-protein interactions at specified resolutions (Blat and Kleckner, 1999; Ren et al., 2000). On high density tiling arrays, oligonucleotide probes from across the entire genome, or certain regions thereof, could be placed on a chip.

Significant advances in Next Generation Sequencing (NGS) technologies have transformed the toolbox of genomic assays through the ability to sequence tens or hundreds of millions of short DNA fragments in a single run (Bentley, 2006; Mardis, 2008; Shendure and Ji, 2008). Chromatin immunoprecipitation followed by sequencing (ChIP-seq) was an early application of NGS, with the first studies to employ it published in 2007 (Barski et al., 2007; Johnson et al., 2007; Mikkelsen et al., 2007; Robertson et al., 2007). With ChIP-seq, the immunoprecipitated DNA fragments are directly sequenced. Sequences are then mapped back to the genome to determine where these regions, or reads, originate (Figure 4). ChIP-seq data has higher resolution, fewer artefacts, greater coverage and a larger dynamic range versus ChIP-chip generating a significantly improved dataset. Whilst the short reads (~35 bp) generated by NGS platforms are problematic for applications such as *de novo* genome assembly, they are acceptable for ChIP-seq. Indeed, given the higher precision of mapped protein-DNA binding sites using ChIP-seq, the method enables the more accurate generation of a list of targets for TFs and enhancers, as well as the better identification of sequence motifs (Johnson et al., 2007; Visel et al., 2009).

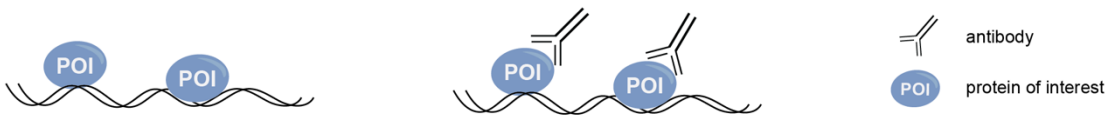


Figure 4 *ChIP analyses identify genome-wide DNA binding sites for TFs and other proteins. In ChIP-seq experimentation, DNA-bound protein (DNA-POI interactions) are immunoprecipitated using a specific antibody. The bound DNA is co-precipitated, purified and sequenced.*

1.5.1 ChIP-exo

In ChIP analyses of DNA-binding proteins, DNA fragments associated with a specific protein are enriched. Typically, the DNA-binding proteins are crosslinked to DNA *in vivo* by treating cells with formaldehyde, chromatin is sheared by sonication into small fragments (200-600 bp), and an antibody specific to the POI is employed to precipitate the DNA-POI complex. Lastly, the crosslinks are reversed and the released DNA is identified by either PCR, microarrays (Blat et al., 1999), or NGS (Deplancke et al., 2004; Johnson et al., 2007). Adaptations to steps in the ChIP method have been developed that bring further precision to the method. In ChIP-exo, exonucleases are used to degrade strands of the protein-bound DNA in the 5' to 3' direction to within a small number of nucleotides of the protein binding site. These exonuclease-treated nucleotide ends are determined using a combination of DNA sequencing, microarrays, and PCR. Sequences are mapped back to the genome to identify locations where the protein binds. Compared to ChIP-seq, ChIP-exo improves the resolution of binding sites from hundreds of nucleotides to almost single nucleotide resolution (Rhee and Pugh, 2011, 2012).

1.5.2 DNA adenine methyltransferase Identification (DamID)

The chief alternative to ChIP for assessing chromatin binding on a global scale is DNA adenine methyltransferase Identification (DamID). First developed two decades ago, DamID serves as a robust and reproducible method aimed at locating putative binding sites *in vivo* (van Steensel et al., 2001; van Steensel and Henikoff, 2000). DamID is a chromatin profiling technique utilising a fusion protein consisting of the *Escherichia*

coli DNA adenine methyltransferase (Dam) and a protein of interest (POI) (van Steensel and Henikoff, 2000). *E. coli* Dam methylates adenines specifically at the N^6-position in the DNA sequence GATC (Brooks et al., 1983). Dam tags the native binding sites of the POI upon expression of the fusion protein in a transgenic organism or in cultured cells. This results in localised methylation of adenine, $G^{6m}ATC$, in sequences close to accessible chromosomal binding sites. These sites are later recognised with methylation-sensitive restriction enzymes and mapped to the genome. GATC sites are present at a high frequency in eukaryotic genomes, on average every 200-400 bp, so the technique allows good coverage of the entire genome. DNA adenine methylation creates 'footprints' in the genome that relate to positions where Dam-fusion proteins have interacted with DNA. A 'Dam-only' control sample, i.e. an expressed untethered Dam methyltransferase, is run alongside experimental samples as a measure of background Dam-only methylation given the promiscuity of Dam. During data analysis, the Dam-POI fusion is normalised against the background methylation obtained from the Dam-only control sample (Figure 5).

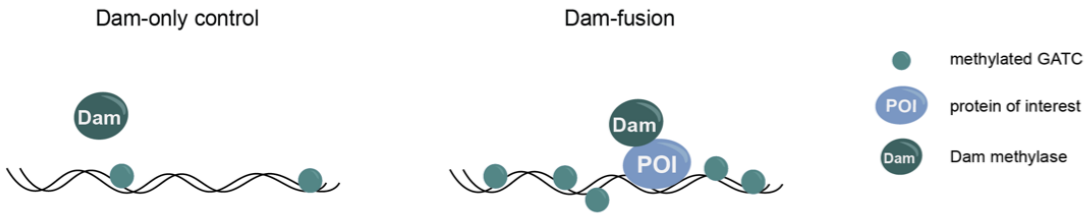


Figure 5 *DamID-seq analyses are an alternative technique to identify genome-wide DNA binding sites for TFs and other proteins. An untethered Dam methyltransferase 'Dam-only' control is run alongside the experimental Dam-POI fusion in DamID experiments to account for background methylation. Dam is a promiscuous enzyme!*

DNA adenine methylation is widespread across many bacterial phyla, and is observed at very low levels, fewer than one in one million deoxyadenines (Koziol et al., 2016), in the genomes of various eukaryotic species (Parashar et al., 2018). These levels are so low that they do not interfere with DamID; this is crucial, because having no background DNA adenine methylation allows for the generation of very sensitive chromatin profiles. Initially implemented in *Drosophila* cell cultures and in whole flies (van Steensel et al., 2000), DamID has been used to assess chromatin states in plants (Germann et al., 2006; Germann and Gaudin, 2011; Zhang et al., 2007), yeast (Lebrun et al., 2003; Steglich et al., 2012; Woolcock et al., 2011), *C. elegans* (González-Aguilera et al., 2014; Schuster et al., 2010; Towbin et al., 2012), and mammalian cells (Tosti et al., 2018; Vogel et al., 2007). Theoretically, the technique can be implemented in any organism in which generating transgenic animals or cell lines is possible. Like with ChIP, DamID is coupled with hybridisation to microarrays (DamID-chip), and recently, to high throughput massively parallel sequencing (DamID-seq).

1.5.3 TARGETED DAMID (TADA)

DamID experiments have traditionally relied on low-level expression from basal, non-induced promoters such as *heat shock protein 70* (*hsp70*), since high expression levels of Dam methylase is toxic to cells (Bianchi-Frias et al., 2004; Neville et al., 2014; van Steensel et al., 2000; Vogel et al., 2007). This dose-dependent toxicity is due to the disruption caused to gene regulatory and DNA replication mechanisms through saturation of GATC adenine methylation (Southall et al., 2013). A downside to the DamID approach is that regulating the Dam-fusion protein using non-induced *hsp70* means that it is expressed constitutively in all cell types. Using cell type-specific promoters or attempting to drive the Dam-fusion cell-specifically using expression

systems like Gal4/UAS (Brand and Perrimon, 1993) results in excessively high levels of Dam methylase and associated toxicity (Southall et al., 2013). A number of attempts to round this problem, aiming to regulate the levels of the Dam-fusion protein expression, have been published: for instance, inserting recombination cassettes (Kozhevnikova et al., 2018; Pindyurin et al., 2016), or adding any of a hormone-responsive element (Hass et al., 2015), a protein degradation signal (Kind et al., 2015), or a tamoxifen-inducible intein (Pindyurin et al., 2016). The first successful approach to achieve cell-type specific, inducible expression of Dam-fusion proteins was developed in 2013 (Southall et al., 2013). Termed targeted DamID or *TaDa*, the technology found a way to utilise the Gal4/UAS system to enable conditional, targeted expression of the Dam-fusion but avoid problems with toxicity or potential artefacts from overexpression as seen previously. Introducing a primary open reading frame (ORF) before the Dam-fusion, and taking advantage of an unusual feature of eukaryotic ribosome translation where rates of translation re-initiation are low enabled low level expression of the Dam-fusion in a cell-type specific manner when driven using Gal4/UAS (Brand et al., 1993) or any other targeted expression system. Coupling the DamID approach with the Gal4/UAS system in this way in *Drosophila* allows neuronal profiling at spatial resolution.

TaDa is a powerful tool that enables the identification of chromatin binding sites on a global scale *in vivo*. It is a highly sensitive, robust and reproducible approach that could profile small populations of cells in complex tissues. TaDa-fusion proteins can be driven in the *Drosophila* CNS through all stages of development and adulthood without previously seen toxicity associated with the Dam methylase (Marshall et al., 2016). Crucially, particularly so for the studies described in this thesis, TaDa allows TFs to be

expressed cell-specifically at extremely low levels, enabling profiling of neurons without altering cell fate (Marshall et al., 2016). Whilst first developed for use in *Drosophila* (Southall et al., 2013), TaDa has been broadly implemented (Albert et al., 2018; Doupé et al., 2018; Korzelius et al., 2014; Loza-Coll et al., 2014; Marshall and Brand, 2017; Otsuki and Brand, 2018; Sen et al., 2019; Spéder and Brand, 2018; Tamirisa et al., 2018; Vissers et al., 2018; Widmer et al., 2018), and more recently modified for inducible expression in mammalian systems – mammalian targeted DamID, *MaTaDa* (Cheetham et al., 2018; Tosti et al., 2018).

1.5.4 ChIP Vs. DamID

Whilst ChIP and DamID both generate a map of global DNA-binding, each technique has drawbacks, and each technique could potentially be more suitable for one application versus the other. Arguably, the most significant difference between ChIP and DamID is a temporal one: whilst ChIP generates a snapshot of protein occupancy at a specific time point, DamID generates a cumulative picture of all interactions that have occurred between the Dam-POI and DNA. Given DamID is reliant on DNA methylation over a period of several hours (van Steensel et al., 2001; van Steensel and Henikoff, 2000), ChIP may thus be theoretically better suited to an experiment requiring profiling for a short time interval. Considering the expected abundance of POI-DNA interactions when choosing between the suitability of both approaches is also important. If POI-DNA interactions are expected to be low, obtaining sufficient antibody binding and purification as required in ChIP may be technically difficult. The resolution of ChIP is based on the amount of POI. Antibodies capture each POI on the DNA creating a direct correlation between the amount of POI, the amount of POI-DNA crosslinks, and the number of locations ultimately detected. With DamID however,

there is no such correlation as the methylation pattern is cumulative and several interactions can be captured from one POI molecule. Thus low levels of Dam-POI are enough to detect a wide range of interactions, which is particularly useful as inducing low levels of Dam is important to avoid Dam methylase toxicity.

In addition to these conceptual differences, one must also consider the practical differences between ChIP and DamID. In contrast to ChIP, DamID does not require antibodies that bind with high affinity and specificity to the POI (Goens et al., 2009). Such antibodies are expensive and difficult to generate; time-consuming thorough optimisations are therefore necessary for every ChIP experiment to make sure the binding is specific, and the precipitation of DNA sequences is targeted. DamID instead necessitates transgenic cells or organisms expressing the Dam-POI fusion. Although both techniques can be used to analyse chromatin interactions in individual cells or tissue types within a whole organism, the picture is more complicated with ChIP. Here, cells or nuclei of interest must be dissociated from the starting material, for instance via dissection, a technically difficult and time-consuming practice. With DamID, genetic techniques can be implemented to target Dam methylation specifically in tissues of interest, eliminating the need for cell isolation practices (Southall et al., 2013).

DamID, as an alternate experimental methodology to chromatin purification-based methods, is a powerful complementary approach to ChIP, and as such can be used to garner higher confidence targets through independent verification. A number of recent studies have used both techniques to validate *bona fide* chromatin interactions (Cheetham et al., 2018; Moorman et al., 2006; Nègre et al., 2006; Shimbo et al., 2013;

Southall et al., 2013), and found strong similarity in the binding profiles generated by the two methods. This is particularly important as there is a growing body of work that shows a proportion of identified ChIP targets are experimental artefacts. These so-called *phantom peaks* (or false positive hits) are chromatin-associated proteins that are wrongly seen to associate with unrelated highly expressed gene promoters (Jain et al., 2015; Park et al., 2013; Teytelman et al., 2013). These targets are extremely difficult to separate from *bona fide* binding sites in ChIP-seq experiments, may still appear if using an alternative antibody, and even remain once a target gene is ablated (Jain et al., 2015).

1.6 PREVIOUS STUDIES PROFILING DSX WITH DAMID-SEQ

Whilst there have been many genome-wide analyses on the sexes (Lebo et al., 2009), there have been considerably fewer efforts to link these expression analyses directly to DSX. For example, a study focussing on genital development in *Drosophila* identified *dsx*-dependent and sex-biased gene expression but did not identify whether these genes were directly or indirectly regulated by *dsx* (Chatterjee et al., 2011). At the same time, a study investigating Dsx^F occupancy on a genome-wide scale using a modified DamID approach located 650 Dsx^F binding locations in the genome, extracted an optimal palindromic 13 bp putative *dsx* binding sequence and forecasted 23 direct target genes (Luo et al., 2011). This study however did not capture known DSX targets and is thus unlikely to be complete.

To delineate whether Dsx^M and Dsx^F bind different targets and how DSX proteins direct different outcomes in diverse tissues, the Goodwin group performed ChIP-seq in Schneider 2 (S2) cells expressing tagged Dsx^M or Dsx^F, as well as Dsx^M or Dsx^F DamID

in adult female and male fat body tissue in transgenic flies coupled to sequencing or hybridisation to microarrays (Clough et al., 2014). These organs were chosen because *dsx* plays a role in maintaining sexually dimorphic gene expression there. The group found DSX bound thousands of the same targets in multiple tissues in males and females, although the targets had sex- and tissue-specific functions. Interestingly, there was a striking conservation of DSX targets identified in the independent screens in the *Drosophila* genus that are orthologs of mouse DMRT1 targets (Murphy et al., 2010) suggesting control of sexual dimorphism may be similar in diverse species across the animal kingdom (Clough et al., 2014). The types of target genes predicted by the analyses illustrated potentially why DSX influences such a diverse set of developmental processes. Predicted target genes included those involved in paracrine (e.g. WNT and DPP) and endocrine (e.g. insulin and ecdysone) signalling, suggesting DSX expression can have far-reaching effects on the development of surrounding cells and beyond. Transcriptional regulators, a majority of which had sex-specific expression (Barmina et al., 2005; Chatterjee et al., 2011; Williams et al., 2008), were another major class of potential DSX targets, alongside signalling molecules including neuropeptides.

1.6.1 Neuropeptides modulate sexual behaviour in *Drosophila*

Neuropeptides form the largest group of signalling molecules in animals and are important regulators for a range of physiological processes. 119 neuropeptide precursor genes have been predicted from the *Drosophila* genome bioinformatically (Clynen et al., 2010). They transmit and regulate biological information in the circulatory and neuronal systems, acting mostly on G-protein coupled receptors (GPCRs) (Brody and Cravchik, 2000; Hauser et al., 2006; Hewes and Taghert, 2001; Vanden Broeck, 2001). Neuropeptides have long been known to modulate sexual behaviour in *Drosophila*. The

neuropeptide SIFamide, encoded in the *Drosophila* genome (Vanden Broeck, 2001), is a classic example: male SIFamide knockout flies perform vigorous and indiscriminate courtship towards either *wild type* males or females (Terhzaz et al., 2007). Further, the female 'post-mating switch' in behaviour described above is regulated predominantly by a single peptide in sperm called *sex peptide* (SP) (Liu et al., 2003). Experiments pairing wild type females with SP-deficient males result in higher lifetime reproductive success and fitness in females (Wigby and Chapman, 2005). It is highly probable therefore that both neuropeptides and their receptors are expressed in the neural circuits involved in the regulation of sexual behaviour (Anderson, 2016; Aranha and Vasconcelos, 2018; Ávila et al., 2011; Billeter and Wolfner, 2018; Carmel et al., 2016; Castellanos et al., 2013; Dickson, 2008; Jang et al., 2017; Kim et al., 2016; Li et al., 2011; Manoli et al., 2013; Neville and Goodwin, 2012; Pavlou et al., 2013; Sellami and Veenstra, 2015; Yamamoto and Koganezawa, 2013).

1.7 PREVIOUS STUDIES PROFILING FRU WITH DAMID-SEQ

Whilst both *dsx* and *fru* orchestrate to specify sexual physiology and behaviour in *D. melanogaster*, most work has focussed on delineating Fru function. For example, to elucidate how FRU proteins function at the level of transcriptional regulation, and to investigate the role of Fru isoform diversity in the establishment and development of a male-specific nervous system, the Goodwin group performed Fru^{M} isoform (Fru^{MA}, Fru^{MB} and Fru^{MC}) DamID in nervous system tissue in larval, pupal and adult transgenic flies coupled to hybridisation to microarrays (Neville et al., 2014). The DamID analyses revealed all Fru^{M} isoforms directly targeted genes involved in the specification of the nervous system including neuronal morphogenesis pathways, albeit each isoform

exhibiting unique binding specificities. Neville et al., localised a putative Fru-DNA binding motif and found genomic loci that contain the motif exhibited sexually dimorphic expression in *fru* neurons. Neville et al. identified an overrepresentation of identified DSX target genes in the Fru^M datasets, suggesting that Fru^M and Dsx^M act together, either in a physical complex or through co-regulation of genomic targets, to specify the male-specific nervous system (2014). Indeed, profiling DSX neural cells with DamID could similarly reveal an overrepresentation of Fru^M target genes thereby adding support to this conclusion.

1.8 BIOINFORMATICS TO PROCESS DAMID-SEQ DATASETS

Interpreting data from DamID, ChIP or indeed any global protein-DNA interaction screen requires a careful application of bioinformatic techniques tailored to the experimental question. Whereas both techniques generate similar data, one needs to consider, biologically, what each technique is de facto assessing and thus apply appropriate downstream computational bioinformatic manipulations. The computational tools to analyse ChIP data are substantially more developed than those currently available for DamID, because ChIP was established a decade earlier than DamID. Compounding this difference further, most DamID studies initially utilised microarray-based approaches, and DamID-seq has just relatively recently been introduced. However, a number of pipelines have now been published that analyse DamID-seq data (Li et al., 2015; Marshall and Brand, 2015). Most notably, the same group that engineered the TaDa approach released a computational bioinformatic pipeline at the same time as their methods paper (Marshall et al., 2015). This "damidseq_pipeline" is a single script that handles a number of processes: aligning sequenced reads, extending

reads, binned counts, normalisation, pseudocount addition, and generating a final ratio file. The script uses a FASTQ or BAM file as input, and outputs the final log2 ratio files in bedGraph or GFF format. The group also developed a "find_peaks" computational pipeline (Marshall and Brand, 2015) as a means of directly assessing the output from the former tool. Peak calling is a computational method that identifies regions in a genome that have been enriched with aligned reads as a result of a genome scale sequencing experiment. These enriched regions are recognised as potential protein-DNA interaction loci. In find_peaks, binding intensity thresholds are identified in a sequencing dataset, the dataset is randomly shuffled, and the frequency of consecutive regions (i.e. GATC fragments or bins) with a score higher than the specified threshold calculated.

Numerous nuanced experimental modifications to the DamID approach have been developed to improve specificity and sensitivity. For instance, in iDamIDseq the methylation-sensitive restriction endonuclease digestion steps are inverted and additional steps involving a phosphatase and exonuclease are introduced (Gutierrez-Triana et al., 2016). These new techniques are coupled with their own standalone computational pipelines including, iDEAR (iDamID Enrichment Analysis with R), an efficient pipeline to generate reliable profiles of transcription factor binding sites (TFBS). Further, bioinformatic refinements address the idea that induced adenine methylation signals at TFBS are not symmetrically distributed. A lately developed non-parametric method for peak calling (NPPC) rounds this problem (Li et al., 2015). A major challenge in peak calling is understanding the levels of background signal from sequence data. In NPCC, a bootstrap resampling method for short sequence reads is implemented in the Dam-only control. This change affects reads resampling,

normalisation, computing signal-to-noise fold changes, and filtering. Downstream called peaks would hence differ from those called by processing with damidseq_pipeline and find_peaks (Li et al., 2015; Marshall and Brand, 2015). Comparing data from homologous protein-DNA interaction screening methods is only possible if both the experimental and bioinformatic pipelines align. It is fundamental that this is taken into consideration when interpreting data across experiments or studies.

1.9 OUTLINE OF THESIS

In this thesis, we present a series of experiments to elucidate the DSX transcriptional network, utilising novel tools to identify putative target genes. We use the recently developed TaDa approach (Southall et al., 2013) that enables the unprecedented rapid genome-wide analysis of cell-type-specific protein binding *in vivo*. As discussed here, TaDa escapes problems associated with Dam methylase toxicity experienced with the DamID system. Through coupling the introduction of a primary ORF before the Dam-fusion with the low rates of translation re-initiation, low level expression of the Dam-fusion in a cell-type specific manner are achieved when driven using Gal4/UAS (Brand et al., 1993; Southall et al., 2013). Here, we specifically target *dsx* neuronal populations in the adult CNS, marking the first time these neuronal cells will be profiled cell-specifically using TaDa. This population is particularly important given that there are still few identified direct *dsx* targets in the CNS. We complete two alternative TaDa screens profiling Doublesex-Dam interactions in both brains and whole heads (including peripheral fat body cells) in male and female adults. We conduct a thorough

bioinformatic analysis of DSX occupancy in the CNS – from assessing called peak characteristics, assigning peaks to genes, to downstream analysis including a Gene Ontology study, predicting regulatory features using an integrative genomics approach, and MOTIF analyses, comparing biological replicates within and between screens.

In addition, we take our bioinformatics analyses one step further in completing a detailed meta-analysis comparing our TaDa-seq DSX-CNS experimental datasets with published DamID-seq DSX-fat body experiments (Clough et al., 2014) and draw meaningful biological conclusions with our analyses pointing towards nuanced *dsx* function. Indeed, our analyses reveal DSX binding targets show a tendency towards tissue specificity. The binding targets identified exemplify the broad mode of DSX function, with *de novo* motif analyses consistently identifying GATA TFs across datasets, that we propose could function alongside *dsx* as co-factors. We also compare these results with a published ChIP-seq S2 cell-DSX dataset (Clough et al., 2014), and a y1H screen involving 722 *Drosophila* transcription factors including Dsx^F (Hens et al., 2011). We independently investigate three putative *dsx* target genes that appear in one or more of these screens – the neuropeptides Diuretic hormone 31 (*Dh31*), Neuropeptide-like precursor 1 (*Nplp1*) and Tachykinin 1 (*Tk1*). We generate a series of lines for our neuropeptides of interest and conduct expression analyses to investigate their physiology in relation to *dsx*. The wealth of Dsx^M target genes and putative co-factors identified in this study pave the way for a better delineation of the *dsx* machinery, and how it functions in the control of sex differentiation and behaviour.

2 METHODOLOGY

2.1 *DROSOPHILA MELANOGASTER* STOCKS

All flies described in this study are of the species *Drosophila melanogaster*. The *wild type* strain was *Canton-S* (CS). Flies were raised at 25 °C at 40-50 % relative humidity on a 12 h light/ 12 h dark cycle unless otherwise stated and flipped every ten days. Fly stocks not in use were kept at 18 °C and transferred to new vials or bottles every 21 days. All flies were raised on standard cornmeal agar food supplemented with dried yeast in plastic vials or bottles.

For our TaDa experiments, we received the pUAST-attB-LT3-Dam donor vector from T. Southall, and generated UAS-Dam-dsx^{M} (pUAST-attB-LT3-Dam + dsx^{M}) and UAS-Dam-dsx^{F} (pUAST-attB-LT3-Dam + dsx^{F}). BDSC 8622 ($y^{1}w^{67c23}$; P{CaryP}attP2) donor attP landing site flies were used for microinjection. We used dsx^{Gal4} (Rideout et al., 2010) to express DSX-Dam in all *dsx* cells, and generated a *Bar*; +; dsx^{Gal4} driver for sex bias assays. For w^{+}/w^{-} ; + ; dsx^{DBD},UAS-Dam-X/TM3,Sb recombinant line generation attempts we used w^{+}/Y; + ; dsx^{DBD} and w^{-}/Y ; + ; Dr/TM3,Sb (Bloomington

Drosophila Stock Center). Flies for TaDa were raised at 25 °C for their entire life-cycle. Prior to gDNA extraction, flies were kept at 29 °C for at least 24 h as per Southall et al., 2013.

We generated p65-*Dh31*^{AD} (pDSPp65AD-*Dh31*), p65-*Tk1*^{AD} (pDSPp65AD-*Tk1*) and p65-*Nplp1*^{AD} (pDSPp65AD-*Nplp1*) transgenic constructs. These were microinjected into BDSC 25709 (y^1v^1 P{nos-phiC31\int.NLS}X; P{CaryP}attP40), a 2nd chromosome attP docking site for phiC31 integrase-mediated transformation, and crossed to *UAS-2eGFP; elav*$^{Gal4\text{-}DBD}$ (Bloomington *Drosophila* Stock Center) or *UAS-2eGFP; dsx*$^{Gal4\text{-}DBD}$ for imaging expression patterns.

2.2 GENERAL MOLECULAR BIOLOGY PROTOCOLS

Method	Commercial protocol
Polymerase Chain Reaction (PCR) amplification	Bioline MyTaq™ HS Red Mix (Standard MyTaq HS Mix Red Product Manual, PI-50154 v6)
PCR purification	QIAGEN QIAquick® PCR Purification Kit (Quick-Start Protocol, July 2018 version)
Restriction digestion	New England Biolabs (NEB®) (Restriction Digest Protocol, October 2014 version)
Gel extraction	QIAGEN QIAquick® Gel Extraction Kit (Quick-Start Protocol, July 2018 version)
Ligation for PCR cloning	NEB® Quick Ligation™ Kit (Quick Ligation Protocol, M2200 version 2.0, August 2017)
Transformation for PCR cloning	NEB® Transformation Protocol (January 2015 version)
Transformation with electro-competent DH10ß *E.coli* cells	ElectroMAX™ DH10 ß™ Competent Cells (Invitrogen™ Transformation Procedure, March 2003 version)

Transformation with chemically competent DH5α *E.coli* cells	MAX Efficiency™ DH5α ™ Competent Cells (Invitrogen™ Transformation Procedure, October 2006 version)
Transformation with chemically competent DH10ß *E.coli* cells	MAX Efficiency™ DH10 ß ™ Competent Cells (Invitrogen™ Transformation Procedure, October 2006 version)
Plating for PCR cloning	NEB® PCR Cloning Kit (Plating Protocol, Instruction manual Version 3.0, 2017)
Colony PCR insert screening	NEB® Colony PCR (NEB® PCR using Hot Start Taq DNA Polymerase, M0495, October 2012)
Minipreparation of plasmid DNA	QIAGEN QIAprep® Spin Miniprep (QIAGEN® Plasmid Mini, Midi and Maxi Kits Quick-Start Protocol, March 2016 version)
Maxipreparation of plasmid DNA	QIAGEN QIAprep® Spin Maxiprep (QIAGEN® Plasmid Mini, Midi and Maxi Kits Quick-Start Protocol, March 2016 version)

Table 1 *General molecular biology methods used in this study alongside commercial protocols*.

2.2.1 PREPARATION OF AGAR PLATES

To prepare LB medium, 1000 ml ultrapure water was added to 25 g LB broth powder (Sigma-Aldrich) and sterilised by autoclaving. For preparation of agar plates with LB medium and ampicillin, 1000 ml LB liquid medium was mixed with 1.5% (w/w) bacteriological (bacto) agar and autoclaved. Ampicillin was added at 5% (w/v) and mixtures poured into petri dishes. This recipe is enough for 30-35 plates.

2.2.2 THERMOCYCLING CONDITIONS FOR ROUTINE PCR

Step	Temperature	Time
Initial Denaturation	95 °C	30 s
30 cycles	95 °C 45-68 °C 68 °C	15-30 s 15-60 s 1 min/Kb
Final Extension	68 °C	5 min
Hold	4-10 °C	∞

Table 2 *Thermocycling conditions used in routine PCR reactions such as for genotyping experiments.*

2.2.3 DNA GEL ELECTROPHORESIS

DNA fragments were resolved on agarose (Boehringer Mannheim) gels. To generate a standard 1% gel, 1 g agarose powder was dissolved in 100 ml 1x TAE buffer (40 mM Tris/ AcOH (pH 8.2), 20 mM NaOAc, 1 mM EDTA) microwaved and mixed with ethidium bromide (EtBr) to a final concentration of 0.2-0.5 μg/ ml. EtBr binds the DNA and allows visualisation under ultraviolet (UV) light. DNA marker ladders used were either 100 bp (NEB), 1 Kb (NEB), or 1 Kb Plus (NEB and Invitrogen) depending on the expected fragment size. Agarose gels were generated at 0.8-2% (w/v) concentrations depending on the size of the fragment to be resolved. Higher concentrations (1.5-2%) were used to separate out smaller fragments (<500 bp) (Sambrook et al., 2001). DNA samples were mixed with Gel Loading Dye, purple 6X (NEB) and loaded into agarose gels. Gel trays were submerged in 1 x Tris-acetate-EDTA (TAE) buffer. The gels were run at varying conditions, most typically at 70 V for 1 hr, 100 V for 45 min, or 120 V

for 30 min. Shorter run times were used for diagnostic experimentation. Longer run times were used to separate out similar sized fragments.

DNA was visualised using either a short wave (254 nm) or long wave (365 nm) ultraviolet (UV) transilluminator after staining the gels with EtBr (0.2-0.5 μg/ ml) or GelRed® (Biotium) (Lee et al., 2012).

2.3 TARGETED DAMID

The published DamID protocol (Vogel et al., 2007), targeted DamID protocol (Marshall et al., 2016), and the Marshall lab updated wet-lab experimental protocols (v2016-Sep-29 and v2017-Oct-9) were used for the basis of TaDa optimisation experiments (see chapter 4, optimising TaDa). Table 2 describes the oligos, their specific nucleotide sequence, and purpose, required for the DamID protocol.

Oligo name	Sequence (5′–3′)	Purpose
AdRt oligo (unmodified)	CTAATACGACTCACTATAGGGCAG CGTGGTCGCGGCCGAGGA	DamID adaptor (top strand) for amplification
AdRb oligo (unmodified)	TCCTCGGCCG	DamID adaptor (bottom strand) for amplification
DamID_PCR oligo (unmodified)	GGTCGCGGCCGAGGATC	Primer for amplifying dsAdR adaptor-ligated DNA

Table 3 *Oligos required for TaDa.*

2.4 BIOINFORMATICS

Method	Protocol
Raw sequence data quality control	FastQC version 0.11.5 Babraham (Andrews, 2010) https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ MultiQC version 0.9 (Ewels et al., 2016) https://multiqc.info cutadapt version 1.13 (Martin, 2011) https://cutadapt.readthedocs.io/en/stable/
DamID-seq computational pipeline	'damidseq_pipeline' v1.4 Oct 2018 (Marshall and Brand, 2015) https://owenjm.github.io/damidseq_pipeline/
Parametric peak calling	'find_peaks' v1 Jan 2016 (Marshall, 2015) https://github.com/owenjm/find_peaks
Non-parametric peak calling	Non-parametric method for peak calling, NPPC (Li et al., 2015) R script for NPPC algorithm; requires SPP software to run
Generating gene lists	Peak Annotation and Visualisation, PAVIS (Huang et al., 2013) https://manticore.niehs.nih.gov/pavis2
Gene Ontology (GO)	FlyMine GO analysis v46.1 Nov 2018 (Lyne et al., 2007) http://www.flymine.org Protein ANalysis ThrougH Evolutionary Relationships, PANTHER version 14.1 (Mi et al., 2019) http://www.pantherdb.org
Searching for DSX motif	MEME Suite v4.12.0 (Grant et al., 2011) FIMO (Find Individual Motif Occurrences) version 5.0.5 http://meme-suite.org/tools/fimo
Predicting regulatory features	i-*cis*Target (Herrmann et al., 2012; Imrichová et al., 2015) https://gbiomed.kuleuven.be/apps/lcb/i-cisTarget/
de novo motif analysis	HOMER (Hypergeometric Optimisation of Motif EnRichment) version 4.10 May 2018 (Heinz et al., 2010)

Published data acquisition	NCBI Gene Expression Omnibus https://www.ncbi.nlm.nih.gov/geo/

Table 4 *Computational bioinformatics methods implemented in this study with their accompanying location.*

For PAVIS (Huang et al., 2013), we used the *D. melanogaster* flybase annotation version R6.01 for gene annotations. For i-*cis*Target (Herrmann et al., 2012; Imrichová et al., 2015), peaks were searched using the 'full analyses' option containing the entire motif collection: PWMs, modERN, TF binding sites, Non TF binding sites, Histone modifications and RNA polymerase (version 5.0 of the database). Genomic coordinate .BED peaks were supplied, and the *Drosophila melanogaster* Ensembl BDGP6 reference genome specified.

2.5 VISUALISATION

Flies were reared at 25 °C and aged for 4-8 days after eclosion prior to dissection and staining. Dissected brains and VNCs were double-stained with primary antibodies of anti-rabbit GFP (1:1000, A6455, Thermo Fisher Scientific) and anti-mouse Brp (1:10, Developmental Studies Hybridoma Bank). The tissues were subsequently stained with secondary antibodies of anti-rabbit Alexa 488 (1:500, A11034, Thermo Fisher Scientific) and anti-mouse Alexa 633 (1:500, A21050, Thermo Fisher Scientific). Images were taken on a Leica TCS SP5 confocal microscope at 1024 x 1024 pixel resolution with the slice size of ≥1 μm. Images taken were subsequently processed on an image-editing software, Fiji.

3 TRANSGENICS

GENERATING DOUBLESEX^{X}-DAM TRANSGENIC CONSTRUCTS FOR TARGETED DAMID (TADA)

3.1 INTRODUCTION

3.1.1 TRANSGENESIS IN *DROSOPHILA MELANOGASTER*

Broadly speaking, transgenesis refers to an assembly of methods and technologies that enable foreign DNA to be introduced into a host cell or organism (Venken and Bellen, 2005, 2007). Since the 1980s, significant developments in this assembly have effectively enabled defined cloned DNA sequences to be introduced into animal germ lines (Brinster and Palmiter, 1986.; Daniels et al., 1990; Ivics et al., 2014; Kroll and Kirschner, 1999). These inserted sequences, transgenes, can stably pass from one generation to the next. The expression of transgenes may be regulated tissue-specifically, developmentally or physiologically. It is possible to analyse the role and regulation of specific cloned genes in a whole living organism – a transgenic organism (David Murphy and Carter, 1993). *D. melanogaster* has gained substantial popularity as a model organism over the past number of years for numerous reasons as discussed in the introduction to this thesis. Alongside the availability of the published, annotated genome sequence (Adams et al., 2000), the *Drosophila* genome unquestionably allows the most sophisticated manipulations of any known eukaryote. The genetic tools available for its dissection are immense (Greenspan, 2004). Numerous *Drosophila* resources are also readily accessible for researchers including the online database, FlyBase, and fly stocks from international fly stock centres, such as the Bloomington Drosophila Stock Centre (BDSC) in Indiana (Matthews et al., 2005), amongst others. Indeed, improvements in transgenesis technology are important because they help in both the identification of novel genes and their functional characterisation (Venken et al., 2005). Furthermore, given that the majority of genes involved in human disease have a counterpart in the *Drosophila* genome, the fly is a popular model system to study

the molecular underpinnings of human pathophysiology. This includes genes involved in genetic disorders and cancer (Bier, 2005; Vidal and Cagan, 2006). Efficient and reliable transgenesis systems are paramount for research of this nature.

3.1.2 PAST VS. PRESENT METHODS OF INTRODUCING TRANSGENES

The majority of transgenesis techniques in *D. melanogaster* require direct microinjection of transgenes into fly embryos, and the following selective genetic crossing of surviving adults (Ringrose, 2009). Screening for a successful transgenesis event by crossing single flies is a time-consuming process. Practically speaking, this reduces the number of transgenesis techniques considered worthwhile to those associated with higher efficiency frequencies: those where one in ten to all surviving adult flies carry the transgene. Previously, only transposon-mediated integration (Handler and James, 2000) fulfilled these criteria and dominated insect transgenesis more broadly. Indeed, in the fly, initial transgenesis techniques were dependent on *P* element transposable elements discussed below (Handler et al., 2000). However, more lately, significant improvements in homologous recombination and site-specific integration of transgenic DNA at specific genomic docking sites using various recombinases and integrases (Bateman et al., 2006; Bischof et al., 2007; Groth et al., 2004; Horn and Handler, 2005; Oberstein et al., 2005; Venken et al., 2006) have eclipsed these previous levels of integration efficiency. The use of the powerful CRISPR/Cas9 targeted gene editing approach has revolutionised the means of inactivating, tagging, and overexpressing any gene precisely and rapidly in recent times (Doudna and Charpentier, 2014; Ewen-Campen et al., 2017). These technological advances have facilitated structure-function analyses of *Drosophila* genes to a higher resolution, increasing the precision, speed and ease in which flies can be manipulated

and assessed. Microinjection of embryos remains pivotal to the majority of transgenesis techniques.

3.1.3 TRANSPOSON-MEDIATED TRANSGENESIS IN *DROSOPHILA*

Transposons, or 'jumping genes', are DNA sequences that are able to change their position in the genome, or jump, first discovered in the 1940s (McClintock, 1950). There are numerous types of transposable elements (TEs) and methods for their characterisation, although the main divide is between the TEs that require reverse transcription to be able to transpose (retrotransposons) and those that do not (DNA transposons) (Baltimore, 1970; Feschotte and Pritham, 2007; Temin and Mizutani, 1970). *P* elements are DNA transposons (Finnegan, 1992), originally identified in the fly genome (de Castro and Carareto, 2004), and discovered as a result of studying hybrid dysgenesis (HD). HD occurs when specific *D. melanogaster* strains are crossed, resulting in sterility, mutation, chromosome breakage, male recombination, and nondisjunction. When males of strains established from natural populations (P strains) were mated with laboratory female strains (M strains), HD would occur; reciprocal crosses with P strain females and M strain males, however, did not induce HD. Hence, the interaction between maternal cytoplasm and elements residing on the paternal chromosomes controlled HD (Kidwell et al., 1977). *P* elements were first used in the early 1980s in *Drosophila* germ line transgenesis (Rubin and Spradling, 1982), and many transgenesis techniques that followed were heavily *P* element-mediated. Structurally, transposons including *P* elements possess two terminal repeats: an inverted repeat sequence and an internally located sequence *motif* necessary for their transposition (Beall and Rio, 1997). Mobile *P* element transposons encode *P*

transposase, an enzymatic protein that catalyses transposition through both terminal repeats of the transposon.

Transgenesis mediated by *P* elements is dependent on the separation of the *P* transposase from the *P* element transposon backbone (Rubin et al., 1982). In the binary transformation system (Figure 6), one plasmid encoding the *P* transposase (the 'helper plasmid') is combined in *trans* with a second plasmid (the transgene) which contains the transposon backbone, the sequence of interest and a marker (Karess and Rubin, 1984). *In vitro* synthesised mRNA encoding the transposase or the purified transposase protein itself (Kaufman and Rio, 1991) could alternatively be co-injected with modified *P* elements. Rates of transposition differ. For example, co-injections could reduce transposase activity, or utilising a hyperactive form of *P* transposase could increase transposition (Beall et al., 2002). Another method would be to express the transposase from a genomic source (Cooley et al., 1988), thus enabling injection of a *P* element in the absence of a helper plasmid. Including an inducible heat shock promoter, such as the heat shock protein 70 (*hsp70*) promoter, allows further refined control of transgene expression. To ensure stable integration and maintenance of the injected transgene, transposons are traditionally injected into fly strains that lack the same transposons to prevent unwanted mobilisation events of transposons present in the genome. Transposition events are marked by the excision or replication of the transposon from the injected plasmid and its insertion into the host genome. Recognising integration events relies on incorporating dominant markers identified through screening or selection.

However, *P*-element transgenesis is limited in two measures: insertion sites are random, and the size of the DNA to be integrated is limited. Random insertion sites vary according to different transposons. For instance, with *P* elements, integration is biased towards the 5' end of genes and there are specific insertion sites that attract *P* elements at significantly higher frequency than others – so called hot spots (Bellen et al., 2004; Spradling et al., 1995; Venken et al., 2011, 2007).

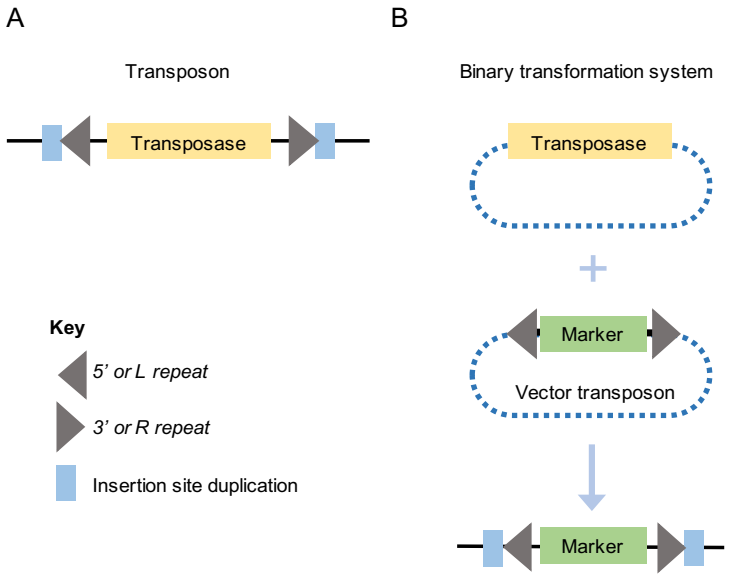


Figure 6 *Transposons are mobile elements containing two inverted terminal repeats (grey) flanking an Open Reading Frame (ORF) that encodes a transposase. Inverted repeats are termed 5' or Left (L) and 3' or Right (R) (A). The binary vector/ helper transposon transformation system composed of transposon and transposase, enables regulated transposition of transgenes into the genome. Transposition events are marked with dominant markers (green), and causes duplication of the insertion site (blue) (B).*

3.1.4 SITE-SPECIFIC TRANSGENESIS USING PHIC31 INTEGRASE

The major limitation with transposon-mediated transgenesis is the random nature of where the transgene is inserted in the genome, and potentially the subsequent positional effects that arise from this (Jaenisch et al., 1981; Rossant et al., 2011; Wilson et al.,

1990). A pioneering strategy to achieve site-specific and efficient transgene integration in the fly was developed in the early 2000s using the bacteriophage PhiC31 integrase (Groth et al., 2004). PhiC31 integrase catalyses the recombination between a phage attachment site (*attP*) present in its own bacteriophage genome, and a bacterial attachment site (*attB*) in a bacterial host genome (Thorpe and Smith, 1998). Early research in the strategy showed PhiC31 integrase catalyses the site-specific integration of attB-containing plasmids into attP-containing docking or landing sites introduced into mammalian cell lines (Groth et al., 2000; Thyagarajan et al., 2001). This was later confirmed in *Drosophila* (Nimmo et al., 2006). Here, recombination mediated by PhiC31 integrase from an mRNA source occurs between an attP docking site integrated into the fly genome using transposons, and an attB site from an injected plasmid (Groth et al., 2004) (Figure 7). The recombination products, attR and attL, are not substrates for the PhiC31 integrase and thus the reaction is irreversible. This feature makes site-specific transgenesis with PhiC31 integrase particularly useful. Three attP docking sites have so far been identified in the *Drosophila* genome, one of which is located in the endogenous transposable element *copia*. The true number of these pseudo-attP sites may be much higher (Kaminker et al., 2002). These sites were not receptive to attB plasmids however, and integration events were identified at the desired specified attP sites (Groth et al., 2004). Non-specific integrations have been noted in *Drosophila*, albeit rarely (Bischof et al., 2007; Nimmo et al., 2006; Venken et al., 2006). PhiC31 integrase-mediated transgenesis allows the insertion of transgenes at specific docking sites, and enables far larger DNA fragments to be integrated in the fly genome than could be allowed by *P* element-mediated integration (Venken et al., 2006, 2007).

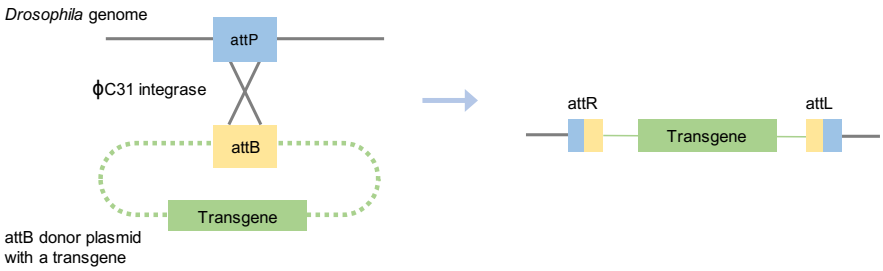


Figure 7 *The PhiC31 site-specific recombinase strategy enables the targeted insertion of cloned DNA sequences to specified locations in the genome. PhiC31 integrase catalyses the site-specific integration of attB-containing plasmids into attP-containing docking sites in the Drosophila genome, forming attR and attL sites respectively.*

3.1.5 DAMID VS. TARGETED DAMID TRANSGENIC CONSTRUCTS

As will be discussed in detail in the following chapter, briefly DamID is an *in vivo* chromatin profiling technique founded on the creation of a fusion protein consisting of the *Escherichia coli* DNA adenine methyltransferase (Dam) and a protein of interest (POI) (van Steensel et al., 2001; van Steensel and Henikoff, 2000). Dam methylates adenines in the sequence GATC and is selectively targeted to the native binding sites of the POI once the fusion protein is expressed in a transgenic organism. This results in localised methylation of adenine in sequences close to an accessible chromosomal binding site.

To achieve the *in vivo* targeting of Dam to specific DNA sequences, the founding group (van Steensel and Henikoff, 2000) used the well-characterised budding yeast protein, Gal4, as the targeting protein (Fischer et al., 1988). They generated a fly line expressing both the fusion protein containing the full-length Dam ('GalDam') and the DNA-binding domain of Gal4 ($Gal4^{DBD}$), 'GalDam1'. To genetically introduce the binding

sequence of Gal4, the GalDam1 flies were crossed to a fly line containing a *P* element with 14 tandem binding sites for Gal4 (UAS^{14}) (Rørth, 1996) inserted into a sequenced region of chromosome 2 - fly line EP(2)0750. As a control condition, van Steensel and Henikoff crossed EP(2)0750 to a fly line expressing Dam alone - Me4 (Simpson, 1999). The progeny of these crosses were used to test whether $Gal4^{DBD}$ is able to target Dam directly to the GATC sequences in the vicinity of the UAS^{14} array. van Steensel and Henikoff used a sensitive assay based on quantitative real-time PCR (Lie and Petropoulos, 1998) to delineate the methylation frequencies of individual GATC sequences. A number of GATCs at various distances from the UAS^{14} array were tested at the resolution of single flies. Given that local differences in chromatin accessibility affect the methylation frequency of individual GATC sequences (Gottschling, 1992; Kladde and Simpson, 1994; Singh and Klar, 1992; Wines et al., 1996), the ratio of methylation by GALDam and methylation by Dam was calculated for each GATC sequence. To ascertain the level of non-targeted background methylation by the freely diffusing GALDam protein, the methylation levels of two remote loci were measured: GATCs in the pericentric Bari-1 element located >10 Mb away, and GATCs in the blastopia element – present in ~10 copies throughout the genome. The methylation ratio (GalDam1: Me4) in the Bari-1 and blastopia elements were ~0.2-0.3, reflecting the methyltransferase activity and nuclear concentrations of both proteins. The methylation ratios for the GATC sequences in the vicinity of the UAS^{14} array were five- to ten-fold higher in the range of ~1-3, showing that methylation by GALDam is clearly targeted to the vicinity of the UAS^{14} array. Indeed, all six GATC sequences tested within ~2.5 Kb of the UAS^{14} array showed significant targeted methylation when compared to the Bari-1 GATC ($p<0.05$, t-test for ratios (Goldstein, 1964)). This set of pioneering experiments

proved that methylation by Dam can be targeted to the vicinity of a specific DNA sequence, *in vivo*.

Targeted DamID couples the established DamID technique with the Gal4 system, enabling expression of the Dam-fusion protein at low levels in a cell- or tissue-specific manner with spatial and temporal control. To establish the technique, the founding group (Southall et al., 2013) harnessed the phenomenon of low-frequency ribosome reinitiation (Child et al., 1999; Luukkonen et al., 1995; van Blokland et al., 2011) and designed a series of transgenic constructs whereby the Dam-fusion is encoded on a secondary ORF where its expression is attenuated, and varying length primary ORFs. The group generated three transgenic lines which contained a primary ORF (ORF1) followed by two stop codons and a secondary ORF (ORF2) encoding Dam methylase. In UAS-LT1-Dam (low level translation version 1), the ORF1 encoded six amino acids (AAs), in UAS-LT2-Dam, the ORF1 encoded the first 80 AAs of mGFP6 (Schuldt et al., 1998), and in UAS-LT3-Dam ORF1 encoded the full-length mCherry, 246 AAs (Shaner et al., 2004).

The group first assayed toxicity. They crossed each transgenic line to *ase-Gal4* which drove neural stem cell expression specifically. As a control, the group also crossed *ase-Gal4* with UAS-Dam (Choksi et al., 2006) in which ORF1 encodes Dam methylase, resulting in 100% embryonic lethality. The same was observed for UAS-LT1-Dam, suggesting toxicity from Dam methylase expression from ORF2 was too high. 100% lethality in the pupal stage was seen in UAS-LT2-Dam. However, the group found that expression of UAS-LT3-Dam driven by *ase-Gal4* was not toxic at embryonic, pupal or

adult developmental stages, with the added advantage of encoding the full-length mCherry to mark expressing cells. The group therefore focused their efforts on the UAS-LT3-Dam construct, and assessing whether there is expression in uninduced cells. To do this, they crossed UAS-LT3-Dam to *insc-Gal4* ($Gal4^{MZ1407}$, a line expressing neural stem cells from the embryonic stage onwards) and to the control wild-type w^{1118}. gDNA from brains of third instar larvae was extracted from experimental and control animals and digested with the methylation-sensitive endonuclease DpnI, cutting solely at methylated GATC sites. Methylated material was subsequently amplified by PCR. The group found methylated gDNA from *insc-Gal4* x UAS-LT3-Dam brains had been amplified, while DNA concentrations were almost undetectable in w^{1118} x UAS-LT3-Dam brains, comparable to control w^{1118} brains. Their results suggested Dam methylase was expressed, active, and able to methylate gDNA in the Gal4-driven UAS-LT3-Dam brains, whilst there was no detectable methylated gDNA in the absence of Gal4. By attenuating the level of translation of the Dam methylase, the group generated a system whereby Dam-fusion proteins can be targeted temporally and spatially using the Gal4 targeted expression system.

3.2 AIMS

1. Genetically engineer Dsx^{M} and Dsx^{F} bicistronic targeted DamID transgenic expression constructs for microinjection into *D. melanogaster* embryos.

2. Test the functionality of the DSX TaDa constructs by two independent methods:

    i. DNA sequencing to characterise recombinant plasmid DNA.

    ii. Delineate the phenotypic penetrance of DSX TaDa constructs in overexpression analyses using dsx^{Gal4}.

3.3 METHODS

3.3.1 DSX TADA CONSTRUCT GENERATION

All primers were designed, mapped and cloning planned in Geneious 10.2.3 (Build 2017.07.10). For amplification of *dsx*^{M} and *dsx*^{F}, sex-specific forward and reverse primers were designed. The sex-specific forward PCR primers were *dsx*^{M}.F1 (ACA GAA ACT CAT CTC TGA AGA GGA TCT GGC CGG CGC **AGA TCT** CAT GGT TTC GGA GGA GAA CTG G), and *dsx*^{F}.F1 (ACA GAA ACT CAT CTC TGA AGA GGA TCT GGC CGG CGC **AGA TCT** CAT GGT TTC GGA GGA GAA CTG), both containing a BglII site (bold). The sex-specific reverse primers were *dsx*^{M}.R1 (AGT AAG GTT CCT TCA CAA AGA TCC **TCT AGA** CTA CGT GGC AGC CGT GGA) and *dsx*^{F}.R1 (AGT AAG GTT CCT TCA CAA AGA TCC **TCT AGA** TCA TCC ACA TTG CCG CGT), both containing an XbaI site (bold).

3.3.2 ELECTRICAL VS. CHEMICAL TRANSFORMATION OF DSX TADA CONSTRUCTS

For transformation, 20 µl *E. coli* electrocompetent (ElectroMAX™ DH10ß™, Invitrogen™) or chemically competent cells (MAX Efficiency™ DH10ß™, Invitrogen™) were added to 2 µl (~5 ng) of the chilled ligation reaction mixture from the previous step (UAS-Dam-*dsx*^{M} and UAS-Dam-*dsx*^{F}). Contents were mixed by flicking the 1.5 ml Eppendorf tube four to five times and placed on ice for 30 minutes. Samples transformed in *E. coli* electrocompetent cells were electroporated using a BioRad GenePulser II electroporator, 2.0 kV, 200 Ω, 25 µF. Samples transformed into *E. coli* chemically competent cells were heat shocked at 42 °C for 45 s in a water bath. No mixing after either of these steps. 950 µl SOC medium (Super Optimal broth with

Catabolite repression, (Hanahan, 1983)) were added to the tube at room temperature. Tubes were placed at 37 °C for 60 minutes shaking vigorously (250 revolutions per minute, rpm). We transformed the pUC19 (Yanisch-Perron et al., 1985) construct, which conveys ampicillin resistance, alongside experimental samples as a positive control. Cells transformed with the pUC19 control plasmid DNA were diluted 1:100 with SOC medium. 50 µl or 100 µl of the dilutions were spread on the pre-warmed plates containing 100 ug/ ml ampicillin. Experimental controls were diluted as necessary (1:10) and 100 µl or 200 µl of this dilution were spread on pre-warmed plates. Plates were incubated overnight from 12 h to 24 h at 37 °C. LB-ampicillin agar plates were controlled for positive and negative control transformants: pUC19 and a construct containing tetracycline resistance that should not generate colonies. Optimal colony growth spread referred to bacterial cells being amply separated to be able to easily pick individual colonies, and hence no associated overgrowth.

3.3.3 Characterisation of DSX TaDa constructs

Assessment with high-throughput colony PCR

Primers for colony PCR and subsequent DNA Sanger sequencing were ordered through Sigma-Aldrich, Haverhill, UK. We requested DNA Oligos in tubes for applications including PCR, sequencing and cloning, providing Sigma-Aldrich with our user defined Oligo name, and the sequence (5' to 3'). We requested the primers at 0.05 µM concentration with the recommended Desalt purification. Primers were designed to target three regions in both UAS-Dam-dsx^M and UAS-Dam-dsx^F constructs across insert and plasmid backbone. In UAS-Dam-dsx^M, forward primer dsx^M-Dam.F1 (AGG GCG AAT GTC TGG TTG AG) and reverse primer, dsx^M-Dam.R1 (CTT TAG GCC ATG

ATT GCG GC) were used for region one. The forward primer *dsx*^{M}-Dam.F2 (CGG CTG TCC ACA ACC AAA AG) and the reverse primer *dsx*^{M}-Dam.R2 (TCA TCA GTT CCA TAG GTT GG) were used for region two. The forward primer *dsx*^{M}-Dam.F2 was again used for region three with the *dsx*^{M}-Dam.R3 (CGA GAC CGT GAC CTA CAT CG) reverse primer. In UAS-Dam-*dsx*^{F}, forward primer *dsx*^{F}-Dam.F1 (CAG CTC CCA GGT GCT TAC AA) and reverse primer *dsx*^{F}-Dam.R1 (CTT TAG GCC ATG ATT GCG GC) were used for region one. The forward primer *dsx*^{F}-Dam.F2 (GAT CAC CAC ATG ACC ACG GT) and the reverse primer *dsx*^{F}-Dam.R2 (CGA GAC CGT GAC CTA CAT CG) were used for region two. The forward primer *dsx*^{F}-Dam.F3 (GTG ATC ACT AGC GCC GAT CA) was used with the *dsx*^{F}-Dam.R2 reverse primer, from region two, for region three.

For colony PCR, 50 µl reactions were set up using 5 µl 10X standard Taq reaction buffer, 1 µl 10 mM dNTPs, 1 µl 10 µM forward primer, 1 µl 10 µM reverse primer, 0.25 µl Hot Start *Taq* DNA polymerase, up to 50 µl nuclease-free water, and variable template DNA: picked single bacterial colonies using 20 µl filter pipette tips (STAR labs) and dipped into each reaction mixture. We picked six single colonies from the LB-ampicillin agar plates: two for each UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} and pUAST-attB-LT3-NDam ('UAS-Dam') control. We followed the thermocycling conditions for routine PCR (see Methodology, section 2.2.2 for thermocycling conditions) except the initial denaturation is increased from 30 s to 5 min at 95 °C to lyse cells. PCR products were assessed by gel electrophoresis. Predicted size of DNA product band(s) were modelled using Geneious 10.2.3.

Assessment with DNA Sanger sequencing

Sequencing was conducted by the facility at Source Bioscience, Oxford, UK. We provided the facility with the recombinant plasmid DNA (at least 100 ng/ µl per reaction) and our own specifically designed forward and reverse primers (at least 3.2 pmol/ µl per reaction) preparing both samples in 5 µl volumes per reaction as specified by the facility.

3.3.4 Microinjection of DSX TaDa constructs in *Drosophila* embryos

We completed maxipreparation of our recombinant plasmid DNA, each UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} and UAS-Dam, to obtain the amount required by the facility at BestGene Inc., Chino Hills CA, U.S.A. for microinjection. We followed the QIAGEN® Plasmid Maxi Kit Quick-Start Protocol, March 2016 version (see Methodology, section 2.2) to extract the recombinant plasmid DNA. For each sample, we prepared 200 ml Lysogeny broth (LB), a nutritionally rich medium primarily used for the growth of bacteria (Bertani, 1951) and autoclaved in a 500 ml flask. Once cooled, we added 200 µl 50 mg/ ml ampicillin. Bacterial colonies confirmed to contain the verified recombinant plasmid DNA were picked using a 10 µl graduated pipette tip (STAR labs), dropped directly into each flask, and sealed with aluminium foil. Cultures were incubated for 12 h at 37 °C and 200 rpm. We provided 20 to 30 µg plasmid DNA (each UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} and UAS-Dam) diluted to 0.5-1.5 µg/µl concentration. DNA provided was dissolved in QIAGEN® Buffer EB.

3.3.5 DSX-DAM OVEREXPRESSION ASSAYS

For UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F}, and UAS-Dam overexpression assays using *dsx*^{Gal4}, 5-7 day-old adult flies were anaesthesised using CO_2 before surgical decapitation. Flies were mounted onto a white-backed dissection dish using dissection pins for imaging the genitalia, and turned over for imaging the abdominal pigmentation. Forelegs were removed using a scalpel and placed flat on the dish for imaging sex combs. Flies were imaged using a Sony Cyber-shot DSC-V3 digital camera connected to Zeiss Axio optical light microscope via a Sony Alpha 7 III phototube adapter.

3.3.6 *DSX*^{DBD}, UAS-DAM-X* RECOMBINANT LINE GENERATION

DNA Sanger sequencing was employed to assess potential recombination events. We designed three sets of forward/ reverse primer pairs specific to *dsx*^{DBD} and UAS-Dam-X*. Forward primers, 21 to 24 bp, originating in the DSX coding region or the Gal4-DBD coding gene, *dsx*^{DBD}.F1 (GCA CGG TTA CAT AAA CTT AGA CGC), *dsx*^{DBD}.F2 (TAG GAT ATC AGA AGC TGG AAT C), and *dsx*^{DBD}.F3 (GAG GTG AGC CAG TAC GAG AC). Reverse primers, 20-28 bp, all located within the Gal4-DBD coding gene were used in tandem: *dsx*^{DBD}.R1 (TAC AGT CAA CTG TCT TTG ACC TTT GTT A), *dsx*^{DBD}.R2 (CTT ATT CTA TCG TGT CTC AAT GTT AG), and *dsx*^{DBD}.R3 (GGC GCA CTT CGG TTT TTC TT). The sequencing primers were used in consecutive pairs, for example, *dsx*^{DBD}.F1 with *dsx*^{DBD}.R1 and span regions ranging between approximately 650 bp to 130 bp.

3.4 RESULTS

3.4.1 GENERATING DSX MALE AND FEMALE TADA CONSTRUCTS

As a first step to generate TaDa DSX-Dam constructs, *dsx* male (dsx^M) and *dsx* female (dsx^F) complementary DNA (cDNA) sequences were amplified by polymerase chain reaction (PCR) using Q5® Hot Start High-Fidelity 2x Master Mix (New England Biolabs), and sex-specific *dsx* cDNA plasmid templates (a gift from G. Lee). The sex-specific forward PCR primers dsx^M.F1 and dsx^F.F1 both contained a BglII site, and the sex-specific reverse primers dsx^M.R1 and dsx^F.R1 both contained an XbaI site, and were used to amplify dsx^M and dsx^F respectively. The primers were designed to flank the DSX coding peptides: the Dsx^M coding peptide is 1,650 base pairs (bp) and Dsx^F is 1,284 bp.

The TaDa donor plasmid, pUAST-attB-LT3-NDam vector (a gift from T. Southall, Figure 8) encodes mCherry in the primary ORF followed by two stop codons. We aimed to insert the DSX peptide as a fusion with Dam as the secondary ORF. Directly after Dam in the sequence, the Myc (human c-Myc oncogene) epitope follows along with the Multiple Cloning Site (MCS). The pUAST-attB-LT3-NDam vector was prepared for insertion of the amplified $dsx^{M/F}$ fragments by digestion with restriction enzymes BglII and XbaI. These restriction enzymes were selected because they flank the dsx^M and dsx^F inserts but do not cut within the insert, and whilst located in the MCS are not located anywhere else in the TaDa donor construct. The BglII/ XbaI digested vector was purified via gel electrophoresis and gel extraction to remove residual nicked and supercoiled vector DNA and the small (22 bp) DNA excision segment. This

reduced background of non-recombinants due to efficient transformation of undigested vector. *dsx*^{M} and *dsx*^{F} PCR products were cloned and sequenced in full prior to ligation as BglII/XbaI fragments into the BglII and XbaI sites of pUAST-attB-LT3-NDam, generating plasmids pUAST-attB-LT3-NDam+*dsx*^{M} ('UAS-Dam-*dsx*^{M}') and pUAST-attB-LT3-NDam+*dsx*^{F} ('UAS-Dam-*dsx*^{F}'), respectively.

Figure 8 *Empty donor targeted DamID construct, pUAST-attB-LT3-NDam, 10, 106 bp total length. mCherry is encoded on the primary ORF, and the Dam-fusion on the secondary ORF.*

3.4.2 Bacterial transformation of DSX TaDa constructs

In the previous step of DNA cloning, ligation, the DSX coding peptide for males or females was inserted into the pUAST-attB-LT3-NDam plasmid using restriction enzymes and DNA ligase. Bacterial transformation is a vital step which transfers the

newly made recombinant plasmids to bacteria. Bacterial competence depends on the ability of a given cell to take up recombinant DNA and survive the introduction of the foreign DNA into the cell. Microorganisms vary in their ability to do this, but the cell membrane/ cell wall must be damaged either electrically or chemically. The 'uptake' efficiency is measured in colony forming units per microgram of DNA (cfu/ μg). In *E. coli*, uptake efficiencies vary depending on electrical or chemical competence: electrocompetent cells have a higher transformation efficiency (1.0×10^8 to 2.0×10^{10}) compared to chemically competent cells (1.0×10^6 to 5.0×10^9).

However, achieving successful transformation proved tricky, perhaps due to the size of the recombinant plasmids generated (>10 Kb). Table 5 details a brief overview of the combinations of incubation time, competent cell types, and whether colonies subsequently formed. Notably, we found our transformation reactions only formed colonies with electrocompetent cells, incubating for either 16 h or 24 h at 37 °C. Theoretically, both competent cell types had an uptake efficiency easily high enough to allow uptake of the recombinant plasmid. However, electrocompetent cells have a higher transformation efficiency for larger constructs > 10 Kb (NEB, manufacturer's troubleshooting information). Diluting experimental samples 1:10 and plating 100 μl had optimal colony growth spread. Size differences in individual bacterial cell colonies following transformation were reported by the authors of the TaDa technology (Southall et al., 2013; Marshall et al., 2016). Indeed, bacterial cells carrying mutated Dam seemed to have a growth advantage, and the authors report specifically selecting for the smaller colonies after transformation of the recombinant plasmid. We similarly observed this phenomenon, and also specifically picked the smaller colonies. As a side note, incubating for either 16 or 24 h did not generate microsatellite colonies, but

increasing the incubation time to 36 h did, suggesting 36 h is associated with overgrowth and hence too long an incubation time.

	Colony formation (Yes/ No)		
	12 h	16 h	24 h
Electrocompetent ElectroMAX™ DH10ß *E.coli* cells	No	Yes	Yes
Chemically competent MAX Efficiency™ DH10ß *E.coli* cells	No	No	No

Table 5 *Comparing transformation rate in electrocompetent and chemically competent cells. Incubation at 37 ℃. n≥5 experimental trials for all 'No' options. n≥2 single colonies for all 'Yes' options.*

3.4.3 Characterising DSX TaDa constructs

Assessment with high-throughput colony PCR

Following the transformation stage, a subset of single bacterial cells was selected from the LB-ampicillin agar plates. Colonies were assessed to identify whether they had the correct recombinant plasmid DNA using the Colony PCR method (see Methodology, section 2.2, for NEB® Colony PCR protocol using Hot Start Taq DNA Polymerase). We designed primers targeting the plasmid backbone DNA (pUAST-attB-LT3-NDam), and the insert itself (*dsx^{M} or dsx^{F}*). Three regions were targeted in both the UAS-Dam-*dsx^{M}* and UAS-Dam-*dsx^{F}* TaDa constructs for colony PCR and to be subsequently used for DNA Sanger sequencing. This design allowed us to determine the presence, specificity, size and orientation of the insert. For UAS-Dam-*dsx^{M}*, region one spans 922 ungapped bases from the 80th bp of Dam on secondary ORF (840 bp total), through the

entire Myc epitope tag, and 141 bp into the Dsx^{M} coding peptide (1,650 bp total length), using *dsx^{M}*-Dam.F1 forward and *dsx^{M}*-Dam.R1 reverse primers. See black arrows in these regions described in the schematic in Figure 9A. Region two spans 489 ungapped bases from the 1,310th bp of the Dsx^{M} coding peptide, through 366 bp of the SV40 polyA signal (700 bp total length) using *dsx^{M}*-Dam.F2 forward and *dsx^{M}*-Dam.R2 reverse primers. Region three spans 934 ungapped bases starting from the 1,430th bp of the Dsx^{M} coding peptide, through the entire SV40 polyA signal, and 5 bp of the attB site (285 bp total length). As with region two, common forward primer *dsx^{M}*-Dam.F2 was used with *dsx^{M}*-Dam.R3 reverse primer. See orange arrows in regions described in the schematic in Figure 9A.

For UAS-Dam-*dsx^{F}*, region one spans 1,157 ungapped bases starting from the 678th bp of mCherry on the primary ORF (738 bp total), through the entire Dam region, Myc epitope tag, and 139 bp into Dsx^{F} coding peptide (1,284 bp total length), using *dsx^{F}*-Dam.F1 forward and *dsx^{F}*-Dam.R1 reverse primer pairs. See grey arrows in regions described in the schematic in Figure 9B. Region two spans 1,003 ungapped bases starting from the 904th bp of the Dsx^{F} coding peptide, through the entire SV40 polyA signal and 4 bp into the attB site (285 bp total), using *dsx^{F}*-Dam.F2 forward and *dsx^{F}*-Dam.R2 reverse primer pairs. Region three spans 1,408 ungapped bases from the 1,188th bp of the Dsx^{F} coding peptide, through the entire SV40 polyA signal and 4 bp into the attB site. The *dsx^{F}*-Dam.F3 forward primer was used with the *dsx^{F}*-Dam.R2 common reverse primer, also used for region two. See blue arrows in regions described in the schematic in Figure 9B.

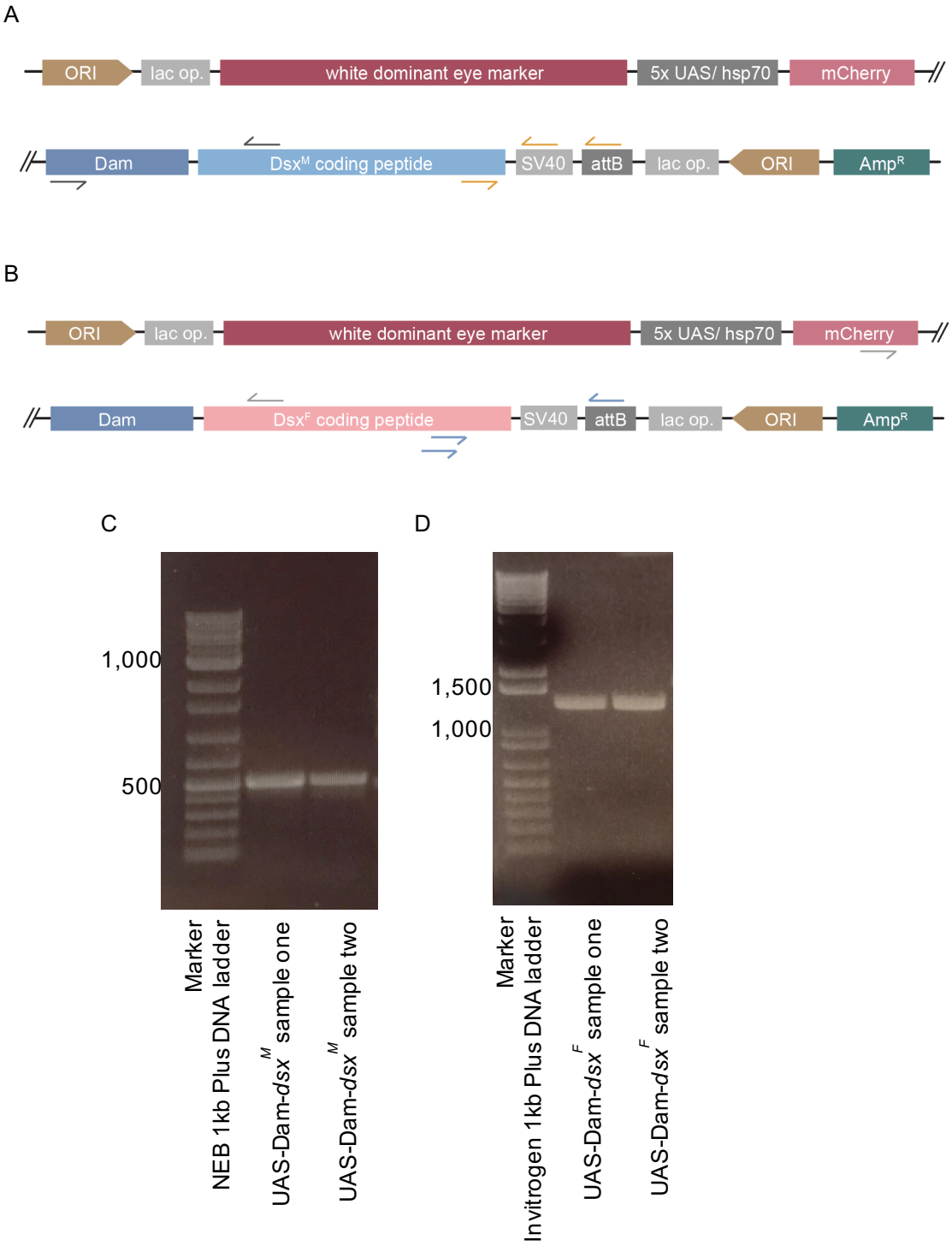


Figure 9 *Linear plasmid map of UAS-Dam-dsx^{M} (A) and UAS-Dam-dsx^{F} (B). For A, black arrows represent location of sequencing primers in region one, and orange arrows for regions two and three. For B, grey arrows represent location of sequencing primers in region one, and blue arrows for regions two and three. Forward primers for regions two and three are shared, respectively. Plasmid maps visualised in Geneious 10.2.3. Gel electrophoresis confirmation of UAS-Dam-dsx^{M} (C) and UAS-Dam-dsx^{F} TaDa constructs (D). Gel electrophoresis of PCR reactions using primers targeting region two in UAS-Dam-dsx^{M} and region three in UAS-Dam-dsx^{F} constructs. Bands identified at ~500 bp in UAS-Dam-dsx^{M} and ~1400 in UAS-Dam-dsx^{F} in two independent samples as modelled in Geneious 10.2.3. NEB 1 Kb Plus DNA ladder used in C, Invitrogen 1 Kb Plus ladder used in D.*

ASSESSMENT WITH DNA SANGER SEQUENCING

Having ascertained which bacterial colonies had the correct recombinant plasmid DNA, we picked these colonies and used the commercially available QIAGEN® kit for the mini-preparation, i.e. isolation and purification, of small quantities of plasmid DNA from bacterial cells (QIAprep® Spin Miniprep kit). It is necessary to isolate the recombinant plasmid DNA in a highly purified form for downstream assessments including DNA Sanger sequencing. DNA Sanger sequencing of the DSX-TaDa constructs enables the most thorough characterisation of the recombinant plasmid DNA. This is particularly important as the authors of the landmark TaDa protocol reported occasional spontaneous mutations arising in the Dam sequence after passage through bacteria (Marshall et al., 2016). This is potentially a result of the function of adenine methylation in mismatch repair in *E. coli* (Modrich and Lahue, 1996).

The size of the DNA insert dictates how many and which primers are needed to determine the complete sequence of the recombinant plasmid DNA. We used the primer pairs described above for colony PCR to sequence three regions within both the UAS-Dam-dsx^M and UAS-Dam-dsx^F TaDa constructs. Each of the regions assessed in both constructs were verified by DNA Sanger sequencing.

3.4.4 MICROINJECTION OF DSX TADA CONSTRUCTS IN *DROSOPHILA* EMBRYOS

SITE-SPECIFIC INTEGRATION USING PHIC31

Site-specific integration using PhiC31 was used to minimise expression differences and thus allow direct comparison of TaDa UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} constructs with UAS-Dam control. The UAS-Dam-*dsx*^{M} and UAS-Dam-*dsx*^{F} constructs containing the DSX transgene and donor sequence (attB) are co-injected with C31 integrase mRNA into attP2-containing recipient embryos (BDSC 8622), resulting in the site-specific insertion of the transgene into the attP site (BestGene Inc., Chino Hills CA, U.S.A.). The attP2 transposable element insertion site is located on chromosome arm 3L (3L:11,070,538..11,070,538). Southall successfully previously used the attP2 site in DamID experiments as it yielded low UAS background expression (Southall et al., 2013). Hybrid sites (attL and attR) are formed during this process, which prevents further integrase-catalysed movement of the integrated transgene (Figure 7).

We completed a maxipreparation of our recombinant plasmid DNA to obtain 20 to 30 μg plasmid DNA (each UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} and UAS-Dam) required by the facility at BestGene Inc. for microinjection. Approximately 200 *Drosophila* embryos were initially injected with *attB* DNA sample, with a survival rate of ~40% to larvae. Positively, this rate was consistent between UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} and Dam-only control samples (Table 6).

n	UAS-Dam	UAS-Dam-*dsx*^{M}	UAS-Dam-*dsx*^{F}
Injected embryos	~200	~200	~200
Surviving larvae	~80	~80	~80
Crossing G_0 adult to *yw*	~45	~45	~45
Red eye G_1 adult	4	5	5
1st round balancing cross	4	5	5
Balanced line	4	5	5

Table 6 *Microinjection of TaDa constructs in D. melanogaster embryos. Survival rates from embryos to larvae, and subsequent crossing scheme for balancing.*

3.4.5 BALANCING DSX TADA TRANSGENIC FLIES

In addition to the initial microinjection of the TaDa constructs into *Drosophila* embryos, flies were balanced at the facility in BestGene Inc. As a first step, BestGene Inc back-crossed selected survival G_0 adults to *yw* – the *yellow white* genetic background – to generate stable transformants. Screening was conducted for the loss of w^+ transformants. The screening marker, w^+, is the normal white gene in *Drosophila* required for the production of red pigment in the eyes. All transformants were picked from individually injected G_0. The G_1 adult transformants were expanded (red eye) by crossing to *yw* again. G_2 transformant flies were balanced with FM7i for X, CyO for the 2nd, and TM3 (stubble, sb) on the 3rd chromosome. The final genotype of the balanced stocks received from BestGene Inc was:

w^{-}/ FM7i; CyO; UAS-Dam-X*/ TM3 (sb)

where UAS-Dam-X* is UAS-Dam, UAS-Dam-*dsx*^{M}, or UAS-Dam-*dsx*^{F}.

3.4.6 Functional analysis of DSX TaDa transgenic flies

Effect of overexpression of Dsx^M-Dam or Dsx^F-Dam on sexual morphology

In the *dsx*^{Gal4} allele, the Gal4 coding sequence was inserted into the first, non-sex specific coding exon of *dsx* using ends-in homologous recombination creating a tandem duplication at the locus (Rideout et al., 2010). To determine whether *dsx*^{Gal4}-expressing cells were able to function in the direction of a sex-specific programme of development, Rideout et al., restrictively manipulated the sex of *dsx*^{Gal4}-expressing cells and assayed the result of overexpression of Dsx^M or Dsx^F on sexual morphology (Figure 10). They found Dsx^M overexpression masculinised females: abdominal pigmentation and sex combs appeared *wild type* male-typical. Sex combs in *Drosophila* are confined to males only (Kopp and True, 2002; Lakovaara and Saura, 1982; Lemeunier et al., 1986) providing better grasping of the female during copulation, hence increasing the selective fitness of the male (Devi et al., 2013; Markow et al., 1996; Polak et al., 2004; Spieth, 1952). The structure is composed of an array of one or a few rows of chitinised teeth in the proximal tarsal segments of the prothoracic leg (Figure 10D). Females instead have a basic row of bristles on the tibia and first tarsal segments (T1) of the prothoracic legs (Figure 10C) (Devi et al., 2013). Dsx^M overexpression resulted in the female genitalia, although masculinised, to appear rotated and malformed, perhaps an inability to overcome the effects of endogenous Dsx^F. With Dsx^F overexpression, males were feminised with *wild type* female-typical abdominal pigmentation, genitalia and no sex combs. Dsx^F appeared sufficient to direct a female-specific programme of development even whilst competing with endogenous Dsx^M production. Given the number of independent *dsx*^{Gal4} injected stocks available (for example, Goodwin lab and BDSC), in this study we recapitulated these findings to ensure the *dsx*^{Gal4} allele was

functional, prior to our TaDa functionality testing described later. Our overexpression findings were very similar to those described by Rideout et al., 2010 (Figure 10).

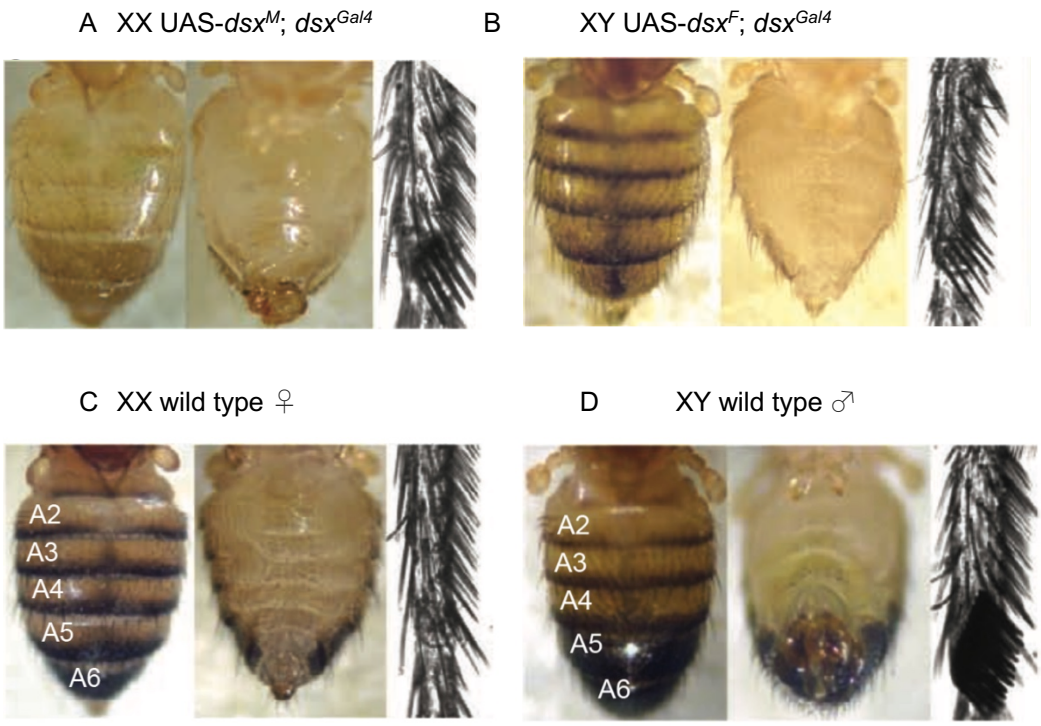


Figure 10 *Light microscope images of dsx^{Gal4} transformation of secondary sexual characteristics, dorsal abdominal cuticular pigmentation, external genitalia and T1 leg basitarsal detail. Shown are masculinised XX UAS-dsx^{M}/dsx^{Gal4} female ('pseudo-male') (A), XY feminised UAS-dsx^{F}; dsx^{Gal4} male ('pseudo-female') (B) adults as compared to wild type XX female (C) and XY male (D) (images from Rideout et al., 2010).*

In addition to Sanger sequencing, we assessed the functionality of the TaDa DSX-Dam constructs by assessing the phenotypic result of DSX overexpression in line with the Rideout et al., 2010 overexpression assay. Using the dsx^{Gal4} allele described in the experiment above (Rideout et al., 2010) with our TaDa DSX-Dam constructs, the assay assessed the function, phenotypic penetrance of DSX-Dam given it is encoded on the secondary ORF, and hence whether DSX-Dam will be able to direct a programme of

development. To do this, we crossed five male homozygous UAS-Dam-*dsx*^{M}, UAS-Dam-*dsx*^{F} or UAS-Dam adults (5-7 days after eclosion) with ten female homozygous *dsx*^{Gal4} virgin adults (Rideout et al., 2010). DSX-Dam will therefore be driven in all *dsx* tissues and neuronal populations (Figure 11). Three experimental replicate crosses were set up for each genotype, where each cross generated ≥ 30 F1 progeny indicating a healthy/ viable cross. Twelve crosses total, including three independent *dsx*^{Gal4} x *wild type*. For phenotypic comparisons, 5-7 day-old adult progeny were compared to *wild type* flies.

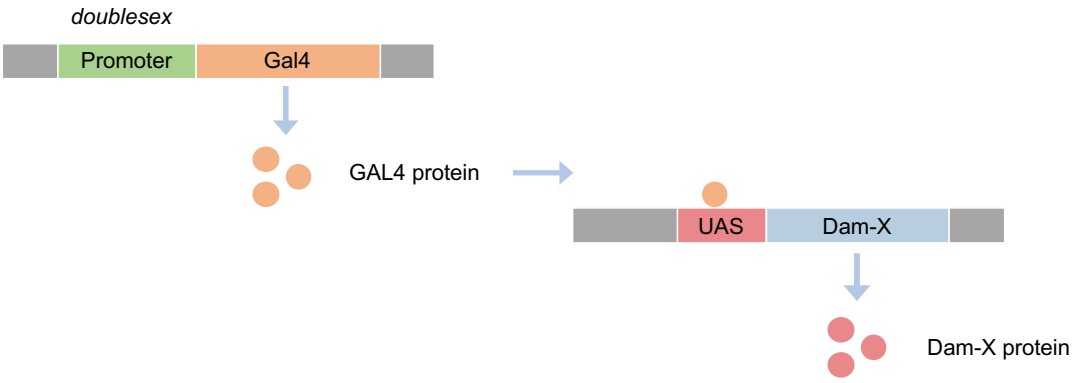


Figure 11 *For overexpression analyses, we used the Gal4/UAS system to drive DSX-Dam expression in all dsx tissues and neuronal populations with the dsx*^{Gal4} *allele (Rideout et al., 2010).*

In our overexpression analyses, we noted overexpression of both Dsx^{M}-Dam and Dsx^{F}-Dam had profound effects on phenotype. Specifically, overexpression of Dsx^{M}-Dam in female flies (UAS-Dam-*dsx*^{M}/ *dsx*^{Gal4} females) resulted in the phenotypic masculinisation of certain features (Figure 12A and B). Overexpression of Dsx^{M}-Dam in females resulted in the masculinisation of the dorsal abdominal cuticular

pigmentation. Pigmentation resembled that typical in *wild type* male adults: darker pigmentation clustered towards the end of the abdomen in abdominal segments A5 and A6. External genitalia in a number of animals were phenotypically transformed. Female genitalia appeared malformed, extended at the vaginal plates and broader, less pointed than typical in *wild type* females (Figure 12B). Despite the additional copy of Dsx^{M}-Dam, females did not develop sex-specific male sex combs, perhaps an inability to overcome the effects of endogenous Dsx^{F} (Figure 12B). Contrastingly, overexpression of Dsx^{F}-Dam in male flies (UAS-Dam-dsx^{F}/ dsx^{Gal4} males) resulted in the phenotypic feminisation of some male-typical features (Figure 12C and D). A significant proportion of male progeny external genitalia appeared similar to *wild type* female external genitalia (Figure 12D). Their dorsal abdominal cuticular pigmentation appeared banded through alternative segments (from A3 to A6) in the abdomen rather than clustered pigmentation towards the end of the abdomen (A5 and A6) as is typical in *wild type* males. Basitarsal sex combs appeared feminised in a proportion of these males, appearing in a reduced form (Figure 12D). No phenotypic change was noted on the tibia or T1 of the prothoracic legs in the female. These findings are positive, in line with the UAS-dsx^{M}/ dsx^{Gal4} and UAS-dsx^{F}; dsx^{Gal4} penetrance assays described above (Figure 10, Rideout et al., 2010), they suggest that the lines generated here are functional. Overexpression of an extra copy of the male or female DSX isoform results in corresponding masculinisation or feminisation, albeit to different levels. This difference is expected as in our assay the DSX coding peptide is fused to Dam and encoded on a secondary ORF.

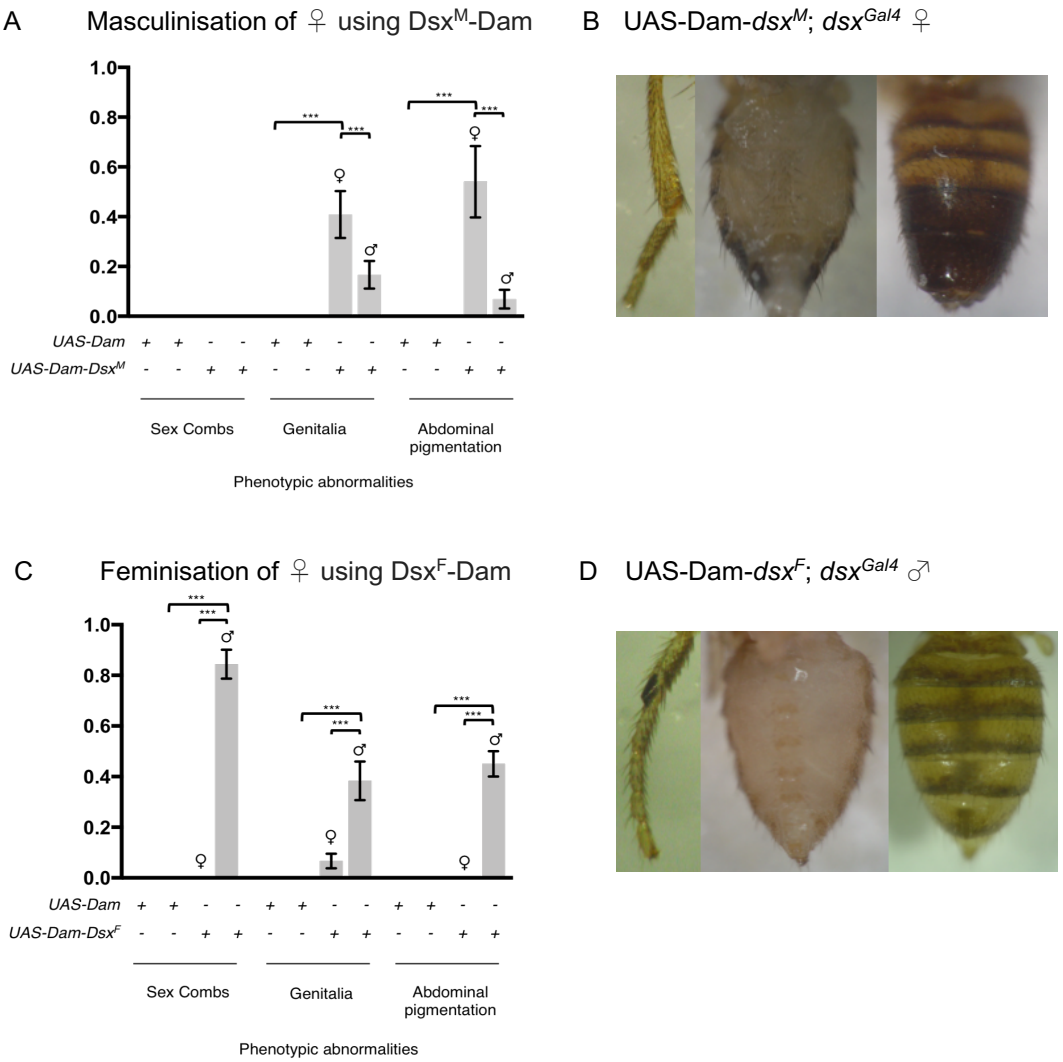


Figure 12 *DSX overexpression assays in TaDa UAS-Dam-dsx^{M} (A), and UAS-Dam-dsx^{F} (C) constructs using dsx^{Gal4}. Transformation of secondary sexual characteristics, sex combs, external genitalia, and dorsal abdominal cuticular pigmentation assessed. Phenotypic transformation compared to UAS-Dam/ dsx^{Gal4} control ♀ (left) and males ♂ (right) for each feature. UAS-Dam/ dsx^{Gal4} male and female controls all appeared phenotypically wild type. Average ratio of progeny transformed from three independent crosses for each UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} and UAS-Dam controls. Significant differences ***p<0.0001 (unpaired t-test).*

Light microscope images of masculinised UAS-Dam-dsx^{M}/ dsx^{Gal4} adult females (B), and feminised UAS-Dam-dsx^{F}/ dsx^{Gal4} adult males (D). Adult flies imaged at 5-7 days-old.

3.4.7 Test for sex-bias in DSX TaDa transgenic flies

To delineate whether there was a difference in offspring sex ratios from *wild type*, we generated a *Bar*; +; *dsx^{Gal4}* driver and completed a series of genetic crosses with our

UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} transgenic flies. In the X-linked dominant mutation, *Bar*, the ommatidia which make up the compound eye in *Drosophila*, are reduced in number and restricted to an area shaped like a vertical bar or band, the edges of which are often irregular in shape (Tice, 1914). *Bar* causes the eyes to be small and slit-like in males and homozygous females compared to round in *wild type* flies. The vertical narrow bar-shape is composed of ~90 facets in the male and ~70 facets in the female compared to ~740 and ~780 facets in *wild type* males and females respectively (Sturtevant, 1925). Female heterozygotes have kidney-bean shaped eyes. Given the sexual transformations described above, particularly transformations of external genitalia, the *Bar*; +; dsx^{Gal4} driver in combination with our UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} flies enables precision in sexing F1 progeny. Three crosses were set up with five *Bar*; +; dsx^{Gal4} homozygous XX virgin females (5-7 days after eclosure) and three +; +; UAS-Dam-dsx^{M}, or +; +; UAS-Dam-dsx^{F}, or +; +; UAS-Dam control homozygous XY males. F1 progeny were counted five days after offspring had begun to eclose. Figure 13 details the number of male and female progeny from each experimental and control cross. Interestingly, the *Bar*; +; dsx^{Gal4} x +; +;UAS-Dam-dsx^{M} cross produced statistically more females than males (unpaired t-test, $p<0.05$). This was surprising -we did not expect this male:female offspring ratio to differ from 50:50. Given that sex-determination typically occurs upstream of *dsx* expression, perhaps this finding could be explained by our experimental set-up where we used 5-8 day-old males. Indeed, it has been shown previously that offspring sex ratios can potentially vary depending on the age of the male. Namely, females mated to older males produce more daughters (Mange, 1970). Statistically, both the *Bar*; +; dsx^{Gal4} x +; +; UAS-Dam-dsx^{F} experimental cross and the *Bar*; +; dsx^{Gal4} x +; +; UAS-Dam control cross produced similar numbers of male and female progeny (unpaired t-test, $p\geq0.05$). We observed a

significant amount of variation in the numbers of male and female flies amongst the *Bar*; +; dsx^{Gal} x +; +; UAS-Dam-dsx^{F} experimental independent crosses (Figure 13).

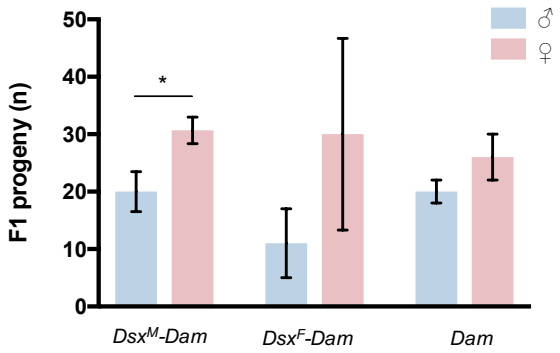


Figure 13 *Number of male and female F1 progeny from Bar; +; dsx^{Gal4} x +; +; UAS-Dam(-$dsx^{M/F}$) crosses. Significant differences were noted in numbers of male and female progeny in Bar; +; UAS-Dam-dsx^{M}/ dsx^{Gal4}, *p<0.05 (unpaired t-test). There was no significant difference in numbers of male and female progeny in either Bar; +; UAS-Dam-dsx^{F}/ dsx^{Gal4} (p>0.05) or Bar; +; UAS-Dam/ dsx^{Gal4} control flies (p>0.05). Crosses set-up with five Bar; +; dsx^{Gal4} homozygous XX virgin females, and three homozygous XY UAS-Dam(-$dsx^{M/F}$) males. Three independent crosses set up for each genotype. Number of offspring counted at midday, five days after the flies started eclosing. Standard deviation shown.*

3.4.8 GENETIC STRATEGY TO TARGET *DOUBLESEX* NEURONS IN THE CNS

While there are several methods to spatially target neurons, there is unfortunately no single strategy that is a panacea when it comes to dissecting neural circuits (Simpson, 2009). *dsx* expression is present in both the nervous system and non-neuronal adult tissues (Rideout et al., 2010). Genetic crosses involving our UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} TaDa constructs with the dsx^{Gal4} allele (Rideout et al., 2010) would drive Dsx^{M}-Dam or Dsx^{F}-Dam expression in all *dsx*-expressing cells. Here therefore, we attempt to use the Split-Gal4 system to restrict *dsx* expression to the nervous system. Given there are only 400-700 *dsx*-expressing neurons in the male CNS (Lee et al., 2002; Pavlou et al., 2016; Rideout et al., 2010) and 300-400 *dsx*-expressing neurons in the

female CNS (Pavlou et al., 2016; Rideout et al., 2010), the proposed approach described below will allow for accurately delineating the appropriate neurons. Split-Gal4 makes use of the fact that the two functional domains of Gal4 are separable: the DNA binding domain (DBD), and the transcriptional activation domain (AD) (Luan et al., 2006). AD and DBD are separately fused to a hetero-dimerising leucine zipper, and when expressed in the same cell, associate, and reconstitute transcriptional activity (Figure 14).

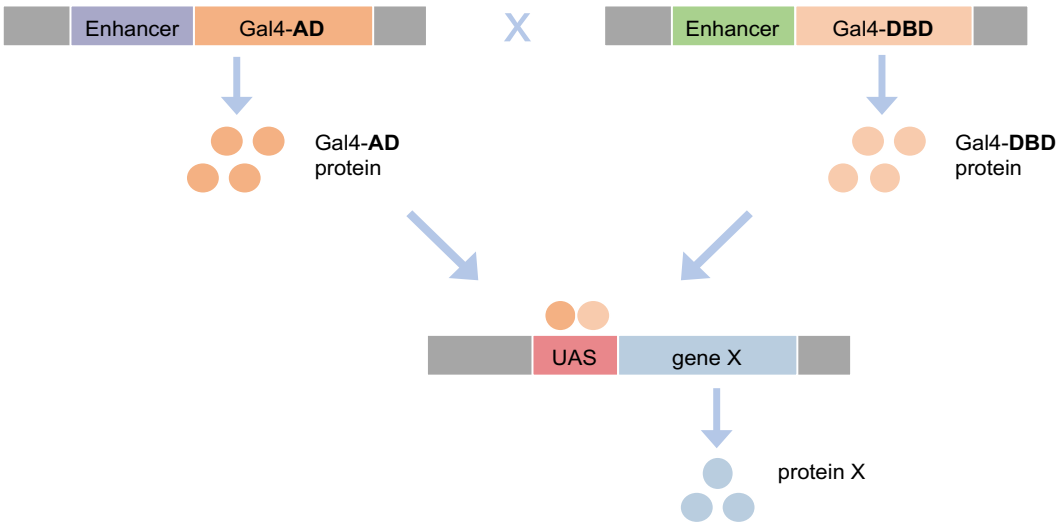


Figure 14 *The Split-Gal4 system can be implemented to enable finer genetic intersections. The system makes use of the fact the two functional domains of Gal4 are separable: the DNA binding domain (DBD), and the transcriptional activation domain (AD) (Luan et al., 2006).*

The genetic mating scheme in Figure 15 details the three steps required to generate an established recombinant line between UAS-Dam-X* and dsx^{DBD} on the third chromosome, which can later be crossed to a neuronal specific AD line, such as the neuron-specific driver using the regulatory sequences of the *embryonic lethal abnormal vision* (ELAV) gene.

Genetic cross 1: w^- ; + ; UAS-Dam-X* **x** w^+/Y ; + ; dsx^{DBD}

↓

Genetic cross 2: w^-/w^+ ; + ; dsx^{DBD}/UAS-Dam-X* **x** w^-/Y ; + ; Dr/TM3,Sb

↓

Genetic cross 3: w^+/w^+ ; + ; TM3,Sb/TM6 **x** w^-/Y ; + ; dsx^{DBD},UAS-Dam-X*/TM3,Sb

↓

w^+/w^- ; + ; dsx^{DBD},UAS-Dam-X*/TM3,Sb

Figure 15 *Generating w^+/w^-;+; dsx^{DBD},UAS-Dam-X/TM3,Sb recombinant line for targeted intersection of dsx neurons in the CNS. w^-/w^+;+; dsx^{DBD}/UAS-Dam-X* females are isolated from the F1 progeny in cross one, and required for cross two as recombination only occurs in females (Hiraizumi et al., 1971). Where UAS-Dam-X* is UAS-Dam, UAS-Dam-dsx^M or UAS-Dam-dsx^F.*

We utilised DNA Sanger sequencing using primers specific to dsx^{DBD} and UAS-Dam-X* to test for recombination events using the genomic DNA extracted from single flies (progeny from genetic cross three, Figure 15). The *dsx* Split-Gal4 allele (dsx^{DBD}) was generated in the Goodwin lab, Oxford (Pavlou et al., 2016) by homologous recombination at the *dsx* locus. Three forward/ reverse sequencing pairs assessing similar but independent genomic regions were designed. Forward primers, dsx^{DBD}.F1, dsx^{DBD}.F2, and dsx^{DBD}.F3, length 21 to 24 bp, originated in the DSX coding region or the Gal4-DBD coding gene. Reverse primers, dsx^{DBD}.R1, dsx^{DBD}.R2, and dsx^{DBD}.R3, length 20 to 28 bp, all located within the Gal4-DBD coding gene were used alongside. Primer pairs span regions ranging between approximately 650 bp to 130 bp (Figure 16). The sequencing primers detailed in sub-section 3.3 of this chapter specify primers used for sequencing UAS-Dam-X*.

Despite three separate attempts to generate the w^{+}/w^{-};+; dsx^{DBD},UAS-Dam-X/*TM3,Sb* recombinant line, we were unsuccessful. At genetic cross two (Figure 15), > 70 individual male/ female crosses are set up owing to the low recombination rate (~6%) between dsx^{DBD} and UAS-Dam-X*. This percentage was computed using a recombination calculator tool that uses the 'Marey map' approach (Fiston-Lavier et al., 2010). This method takes into account genetic and physical maps to infer recombination rates along major chromosomes of the *D. melanogaster* genome. The low recombination rate is due to the close physical location of UAS-Dam-X* (3L:11,070,538..11,070,538) and dsx^{DBD} (3R:7,924,323..7,967,408) near the centromere on the 3rd chromosome, physically 12 centiMorgans. To combat this low recombination rate, we could generate new lines whereby the physical distance between the two transgenes is extended. For example, we could inject the dsx^{DBD} transgene on the same chromosome arm as UAS-Dam-X* and hence increase the likelihood of recombination.

Although our attempts to develop the recombinant line were unsuccessful, using our UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} TaDa constructs with the dsx^{Gal4} allele (Rideout et al., 2010) would still drive Dsx^{M}-Dam or Dsx^{F}-Dam expression in all *dsx*-expressing cells. Precisely, we would use UAS-Dam-dsx^{M}/ dsx^{Gal4} male progeny and UAS-Dam-dsx^{F}/ dsx^{Gal4} female progeny for these experiments. Albeit more labour-intensive, we could couple this genetic approach with manually dissecting neural tissue (brains or whole heads) to ask the same research questions in profiling *dsx* neural cells tissue-specifically.

Figure 16 *Linear plasmid map of dsx^{DBD} (Pavlou et al., 2016). Brown, grey and blue arrows denote forward and reverse sequencing primer pair locations used in DNA Sanger sequencing verification.*

3.5 DISCUSSION

Here, we describe a series of experiments for the generation and validation of novel DSX-Dam targeted DamID lines (UAS-Dam-*dsx*^{M} and UAS-Dam-*dsx*^{F}). These will be used for the profiling of DSX neuronal populations in the CNS of adult *D. melanogaster* discussed in the following chapters. To insert the DSX *male* and *female* coding peptides into the secondary ORF of the TaDa vector, pUAST-attB-LT3-NDam (Southall et al., 2013), we implemented a PCR/ restriction enzyme-based cloning strategy. As described in the introduction to this chapter, attenuating levels of translation of Dam methylase by encoding the Dam-POI on the secondary ORF allows the Dam-POI to be targeted spatially and temporally using the Gal4 system (Southall et al., 2013). We utilised electrocompetent *E. coli* cells for the transformation stage given their increased transformation efficiency and the length of the generated recombinant plasmids (>10 Kb, Figure 8 and Table 5). Site-specific integration using PhiC31 integrase was utilised for the microinjection of our DSX-Dam TaDa constructs into *Drosophila* embryos. The approach allows for direct comparisons of UAS-Dam-*dsx*^{M} and UAS-Dam-*dsx*^{F} and UAS-Dam control transgenic flies as expression differences are minimised.

DSX-Dam TaDa constructs were verified independently by sequencing both prior to and following microinjection – using the same primer sequences as previously and the gDNA extracted from single flies (Figure 9). Functionality of DSX-Dam was assayed by overexpressing DSX using the *dsx*^{Gal4} allele (Rideout et al., 2010). This is in line with a landmark study that showed restrictively manipulating the sex of *dsx*^{Gal4}-expressing cells using Dsx^{M} or Dsx^{F} has an effect on secondary sexual characteristics

(Rideout et al., 2010, Figure 10). Indeed, mutant analyses and ubiquitous overexpression of DSX isoforms have shown *dsx*'s ability to determine secondary sexual characteristics (Camara et al., 2008). Here, we aimed to determine the phenotypic penetrance of DSX-Dam, testing with dsx^{Gal4} (Rideout et al., 2010) whether DSX-Dam will be able to direct a programme of development given its position encoded on the secondary ORF in the TaDa construct. We found the overexpression of either Dsx^{M}-Dam or Dsx^{F}-Dam resulted in the transformation of some secondary sexual characteristics. This is a positive sign as it suggests DSX binds the appropriate targets involved in specifying secondary sexual characteristics. Overexpression of Dsx^{M}-Dam in UAS-Dam-dsx^{M}/ dsx^{Gal4} female flies resulted in the masculinisation of dorsal abdominal cuticular pigmentation in a significant proportion of the female progeny (~50%, Figure 12A). Despite the additional copy of Dsx^{M}-Dam, females did not develop male-typical sex combs, nor did the *wild type* typical expression of sex combs change in males (Figure 12A and B). Indeed, female genitalia appeared masculinised to an extent, at least overexpression of Dsx^{M}-Dam resulted in their genitalia appearing broader and malformed. This could be the result of the incapability of Dsx^{M}-Dam to overcome the effects of endogenous Dsx^{F}, similar to what has been reported previously in the literature (Rideout et al., 2010). Overexpression of Dsx^{F}-Dam in UAS-Dam-dsx^{F}/ dsx^{Gal4} male flies resulted in the feminisation of their sex combs, external genitalia and dorsal abdominal cuticular pigmentation (~82%, ~38%, and ~44% respectively, Figure 12C and D). The finding corroborates Rideout et al., 2010 overexpression analyses (Figure 10) showing Dsx^{F} was sufficient to direct a female-specific programme of development even when competing with endogenous Dsx^{M} production, however here in a smaller subset of female progeny. Overall, the finding suggests that Dsx^{F}-Dam competes better with the effects of endogenous Dsx^{M}, than Dsx^{M}-Dam does with

endogenous Dsx^F. Indeed, all three secondary sexual characteristics assayed for transformation, showed changes with Dsx^F-Dam. We note sexual transformations, both for masculinisation and feminisation experiments, only occur in a subset of progeny. Despite all flies raised at 25 °C, we could speculate that the number of DSX-Dam transcripts available in individual flies varies, hence expression differences and subsequently transformations occur in some progeny but not all.

In our assays testing sex-bias in UAS-Dam-dsx^M and UAS-Dam-dsx^F transgenic flies, we generated a *Bar*; +; dsx^{Gal4} driver line to more accurately phenotype offspring sex given the transformation of secondary sexual characteristics observed when overexpressing Dsx^M-Dam or Dsx^F-Dam. Across all experimental crosses (and indeed each of the three replicates for each cross) we observed slightly higher numbers of female progeny. Of these, our assays showed the *Bar*; +; dsx^{Gal4} x +; +;UAS-Dam-dsx^M cross produced statistically more females than males (unpaired t-test, $p<0.05$) (Figure 13), whilst crosses with UAS-Dam-dsx^F and Dam control produced no significant difference in numbers of male and female progeny (both unpaired t-tests, $p\geq0.05$). The result for Dsx^M-Dam was unexpected, as we predicted a 50:50 male:female sex ratio as seen in *wild type* animals. Indeed, sex-determination is achieved by a balance of female determinants on the X chromosome and male determinants on the autosomes (Bridges, 1921, 1925), the ratio of which can be affected by genes such as *Sxl*, *tra*, and *tra2* (Baker and Ridge, 1980; Belote et al., 1985), upstream of *dsx*. Perhaps then this finding could be explained by our experimental set-up. Using males that vary by four days in age for our experimental crosses, for example, could introduce variation in the sex-ratios given females mated to older males are known to produce more daughters (Mange, 1970). Secondly, our offspring count was taken in the morning, and since

Drosophila are photoperiod-sensitive, females tend to eclose early in the morning, perhaps further biasing towards more female progeny.

Our repeated attempts to generate a Split-Gal4 DSX^{DBD}/ UAS-Dam-X* (w^+/w^-;+; dsx^{DBD},UAS-Dam-X*/*TM3,Sb*) recombinant line were unsuccessful (Figure 15), likely a result of the low recombination rate between the two alleles physically located on either arm on the 3rd chromosome but crucially close to the centromere. If our attempts to generate the recombinant line had been successful, we could have potentially genetically targeted certain *dsx* neuronal populations of interest in a cell-type specific manner, and winnowed these subsets of neuronal populations. Importantly, given *dsx* is expressed broadly in neuronal and non-neuronal tissues, the generation of a recombinant line would have allowed us to separate domains of expression by carefully choosing specific Split-Gal4 driver lines. By combining our DSX^{DBD}/ UAS-Dam-X* recombinant line with appropriate AD lines, we could have accurately targeted for instance, solely inhibitory *dsx* neurons using glutamic acid decarboxylase 1 (Gad1), or *dsx* motor neurons using vesicular glutamate transporter (vGlut), for use in TaDa. These neuronal populations are of interest because both have been found to have a pivotal function in coordinating *Drosophila* male copulation. *dsx* GABAergic inhibitory neurons mediate genital uncoupling by inhibiting key motor neurons, and *dsx* glutamatergic motor neurons innervate muscles of the genitalia and enable genital attachment (Pavlou et al., 2016).

Here, we have produced novel, and functional, DSX-Dam targeted DamID transgenic reagents. Genetic crosses coupling these lines with dsx^{Gal4} (Rideout et al., 2010) and downstream manual dissection of neural tissues (brains or heads) will allow the targeted

DamID whole genome profiling of these DSX-expressing neural cells in an unprecedented manner to a resolution not previously undertaken.

4 OPTIMISING TADA

OPTIMISING TADA TO PROFILE DOUBLESEX-DNA INTERACTIONS IN THE *DROSOPHILA* BRAIN AND HEAD

4.1 INTRODUCTION

4.1.1 STRATEGIES TO ANALYSE PROTEIN-DNA INTERACTIONS

Interactions between protein complexes and DNA are at the heart of vital cellular processes including transcription, DNA replication and genome maintenance. The activation of genes by DNA-binding proteins is a key regulatory mechanism initiating RNA synthesis. The mechanism involves chromatin modifications and transcription complexes (Bulyk et al., 1999). Indeed, genes must be tightly regulated in a temporal and spatial manner throughout development. This is achieved through the action of chromatin-binding proteins, including transcription factors (TFs), histone modifiers, and nucleosome remodelers (Bulyk, 2006). It is of utmost importance to develop techniques to identify DNA loci that interact with specific proteins. One early method, the electrophoretic mobility shift assay (EMSA), is an *in vitro* method to identify protein-nucleic acid associations based on differences in DNA mobility during electrophoresis. Protein-bound DNA in general moves slower than free DNA, resulting in a band shift in agarose or native polyacrylamide gels (Garner and Revzin, 1981). EMSA has been modified into numerous variants using a plethora of DNA labelling methods (Hellman and Fried, 2007). Another early method, the DNase footprinting assay, was the gold standard for detecting protein-DNA interactions and identifying a protein's core binding sequence at single-nucleotide resolution. Based on DNA cleavage, this assay utilises an enzyme deoxyribonuclease (DNaseI) to cut DNA randomly. Protein bound to DNA protects the bound DNA regions from enzymatic cleavage. Hence, patterns of DNA fragments with and without a bound protein can be compared using gel electrophoresis to reveal the footprint of the binding protein (Galas and Schmitz, 1978).

Since the advent of these techniques, a handful of independent methods have become well-established in identifying protein-DNA interactions. One such technique is ChIP-Sequencing (ChIP-Seq). As discussed in the introduction to this thesis, in ChIP-Seq, DNA and DNA-bound proteins are cross-linked using formaldehyde. The chromatin is then isolated from nuclei and fragmented either mechanically by sonication or enzymatically by digestion with micrococcal nuclease (MNase). DNA-TF complexes are specifically immunoprecipitated using an antibody against a specific TF or DNA-binding protein (Snyder et al., 2010). The applicability of this technology for understanding gene regulation is vast: at its inception, both Barski et al. 2007 and Mikkelsen et al. 2007 applied the technology to profile histone modifications in mammalian cells. Histones are proteins located in eukaryotic cell nuclei that package and order DNA into nucleosomes. They are a chief component of chromatin and play a role in gene regulation. Johnson et al. 2007 used the technique to map *in vivo* binding of the neuron-restrictive silencer factor in the human genome. Since its inception, numerous improvements of the technique have been introduced, such as reducing the required number of cells and increasing resolution (Furey, 2012).

4.1.2 DamID experimental method

In the early 2000s, DamID emerged as a promising technique to effectively map protein-DNA interactions globally, *in vivo* (van Steensel and Henikoff, 2000). Experimentally, methylated genomic DNA (gDNA) from transgenic cells is extracted and digested with the methylation sensitive restriction endonuclease DpnI, which cuts specifically at adenine-methylated GATC sites to fragment the gDNA. Adaptors for

Polymerase Chain Reaction (PCR) amplification are ligated to DpnI-cut fragments, and subsequently digested with DpnII, which cuts only non-methylated GATC sites to stop unmethylated regions of DNA being aberrantly amplified. These steps allow the methylated sequences to be selectively amplified by PCR (Figure 17). Sequencing or array-based methods follow to generate a chromatin profile of loci that were in close proximity to the POI for the duration of the Dam-fusion expression.

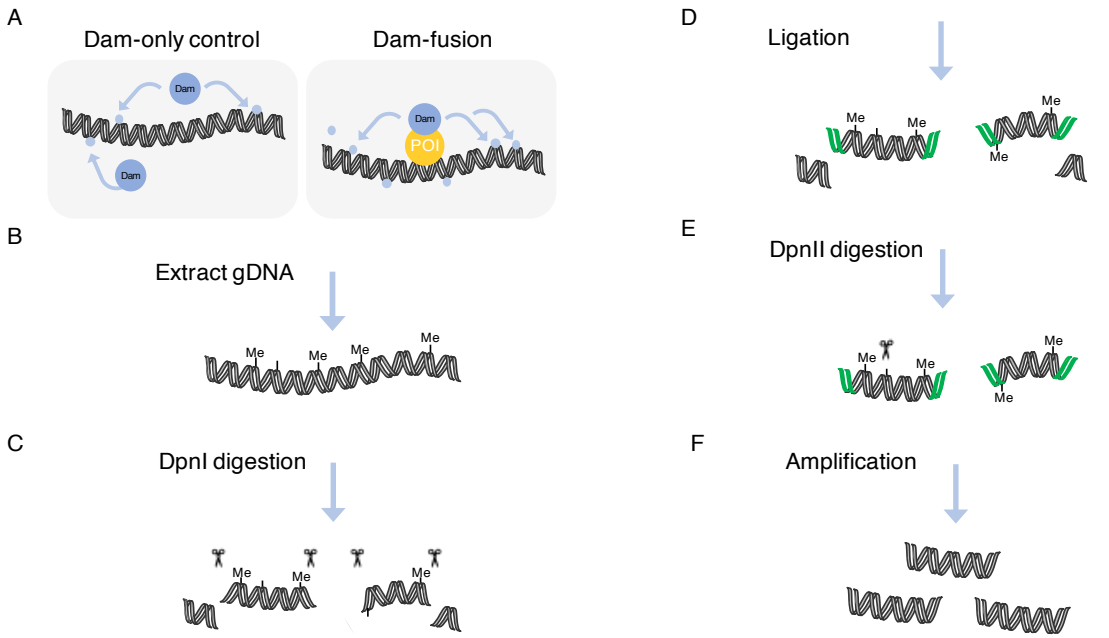


Figure 17 *DamID schematic pipeline. Dam (DNA adenine methyltransferase)-only control or Dam-protein of interest (POI) fusion expressed in transgenic flies (A). gDNA containing methylation sites (Me) extracted from Dam-fusion and Dam-only control in parallel (B). gDNA digested with methylation-sensitive restriction enzyme, DpnI (C). Double-stranded adaptors are ligated to DpnI-cut fragments (D). Ligated fragments are digested with non-methylation sensitive restriction enzyme, DpnII (E). Digested fragments are amplified with PCR (F).*

4.1.3 ATTENUATING DAM EXPRESSION IN DAMID AND TADA

It had previously been shown that DNA cytosine methyltransferase could be targeted to specific DNA sequences *in vitro* by means of tethering to a DNA-binding protein (Xu and Bestor, 1997). DamID was born out of an adaptation to this technique: testing whether Dam from *E. coli* could be targeted to a specific locus in *D. melanogaster in vivo*. van Steensel and Henikoff (2000) chose Dam for several reasons: firstly, endogenous methylation of adenine does not occur in the DNA of most eukaryotes; secondly, Dam is active when expressed in yeast (Gottschling, 1992; Kladde et al., 1994; Singh et al., 1992) and *Drosophila* (Wines et al., 1996); and thirdly, low-level Dam is not known to have any detectable effects on *Drosophila* development or viability (Wines et al., 1996). Adenine methylation has only minor effects on DNA topology (Barras and Marinus, 1989); even though ~50% of all GATCs in the fly genome are methylated, development is not affected (Boivin and Dura, 1998; Wines et al., 1996).

When designing DamID experiments, it is imperative to consider two facets regarding the inherent architecture and function of Dam. Firstly, at high concentrations, methylation levels are saturated, making it difficult to determine *bona fide* binding sites. And secondly, ectopic expression of Dam is toxic in most cell types, giving rise to experimental artefacts (Southall et al., 2013; van Steensel et al., 2000). Choosing the appropriate promoter to express the Dam-fusion protein at very low concentrations is thus critical to an effective experiment. As discussed in the introduction to this thesis, suitable levels of Dam expression are achieved through use of minimal 'leaky' expression from inducible promoters. Basal heat shock promoters, such as the *hsp70*

promoter, are commonly used in *Drosophila* without inducing the promoter in transgenic flies or cultured cells. These yield almost undetectable levels of Dam with reproducible methylation profiles. In mammalian cells, 'leaky' expression from uninduced inducible ecdysone promoters have been used for DamID (Aughey and Southall, 2016).

In 2013, Southall et al., published an adaptation to the now established DamID technique, facilitating expression of the Dam-fusion protein at low levels in a cell- or tissue- specific manner with spatial and temporal control, TaDa. Southall et al., harnessed the biological phenomenon that at low frequency, eukaryotic ribosomes reinitiate translation on bicistronic messages that lack an obvious internal ribosome entry segment (IRES) (Child et al., 1999; Luukkonen et al., 1995; van Blokland et al., 2011). Specifically, given that ribosome reinitiation is dependent on the size of the primary Open Reading Frame (ORF), the group generated a series of constructs containing primary ORFs of varying lengths, followed by a secondary ORF encoding the Dam methylase *(TaDa construct generation is discussed in depth in the previous chapter)*. With the Dam-fusion encoded as a secondary ORF in a bicistronic expression construct downstream of a UAS enhancer, when no transcriptional activation occurs, basal transcription causes low level translation of the upstream ORF and negligible translation of the Dam-fusion from the secondary ORF. Coupled to a Gal4 driver, levels of translation are increased from the primary ORF, and Dam-fusion translation occurs at low levels, causing tissue-specific methylation of target sequences without associated toxicity (Figure 18; Marshall et al., 2016). To detect appropriate methylation levels, whereby the Dam methylase is functional but non-toxic, the protein is undetectable by western blotting and immunofluorescence (Vogel et al., 2007). TaDa requires less than

10,000 cells, as compared to 4-6 million required for other approaches such as ChIP-seq (Southall et al., 2013).

Here, we employ the TaDa method to profile the small number of DSX-expressing neurons in the *Drosophila* CNS of adult male and female animals in a cell-specific manner. This is important as DSX has a significant sex-specific role in the development of physiology and behaviour in *Drosophila*, functioning as a TF at the bottom of the SDH. TaDa theoretically probes where DSX interacts with the genome *in vivo*, hence identifying putative TFBS and genes involved in how *dsx* functions. Implementing this technique will allow for a comprehensive comparison of *dsx* function in male and female adults through assessing similarities and differences in putative target genes identified. Experimentally, implementing TaDa in *Drosophila* requires genetic crosses driving fly lines containing DSX-Dam TaDa transgenic constructs specifically in *dsx* neuronal populations and thorough optimisation of the TaDa protocol. This chapter details the optimisation experiments for each of the main five steps in the TaDa protocol required to profile DSX neurons in the CNS. These optimisations are necessary to adapt the protocol for profiling small cell numbers in *Drosophila*, and to generate experimental reproducibility across biological replicates.

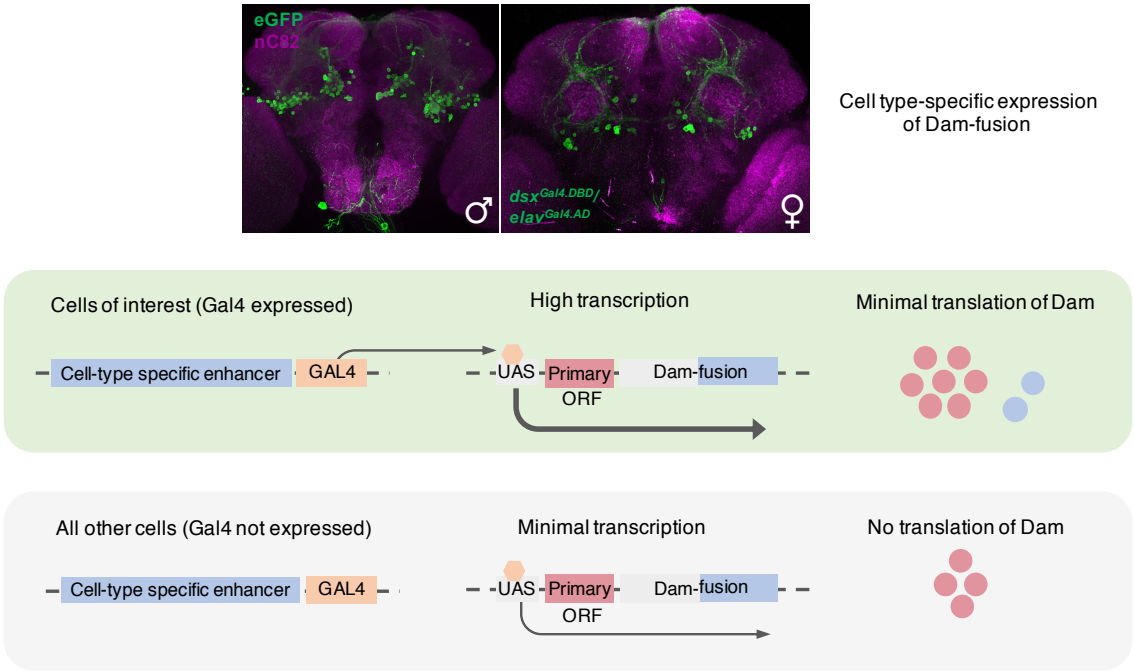


Figure 18 *TaDa schematic pipeline. Expression of Dam in distinct cell populations, here in dsx-expressing neuronal populations (green) in the male and female adult brain (magenta). In cells where Gal4 is expressed (green), the bicistronic transcript is expressed at high levels, resulting in low-level translation of the secondary ORF containing Dam-fusion. There is minimal expression of the bicistronic transcript where Gal4 is not expressed (magenta) (adapted from Aughey and Southall, 2016).*

4.2 AIMS

1. Generate animals driving the TaDa DSX-Dam bicistronic expression construct in all *dsx* neuronal populations in *Drosophila*.

2. Optimise the five main steps of the published DamID (Vogel et al., 2007) and targeted DamID (Marshall et al., 2016) protocols for targeting DSX-expressing neural cells in the *Drosophila* CNS.

3. Prepare animal material for library preparation and next generation sequencing for two experimental rounds of TaDa independently profiling dsx^{M} neurons and dsx^{F} neurons in the *Drosophila* CNS.

    i. DSX brain TaDa: targeting *dsx* neuronal populations in dissected brains.

    ii. DSX head TaDa: targeting *dsx* neuronal populations in whole heads.

4.3 METHODS

4.3.1 DNA SANGER SEQUENCING DSX TADA TRANSGENIC FLIES

For DNA Sanger sequencing, we verified the presence of *dsx*^{Gal4} using two sets of forward and reverse sequencing primer pairs targeting the junction between the *dsx* coding gene and the Gal4 exon. The forward sequencing primers, 22-24 bp, *dsx*^{Gal4}.F1 (GCA CGG TTA CAT AAA CTT AGA CGC) and *dsx*^{Gal4}.F2 (TAG GAT ATC AGA AGC TGG AAT C), were used in conjunction with the reverse sequencing primers, 26-28 bp, *dsx*^{Gal4}.R1 (TAC AGT CAA CTG TCT TTG ACC TTT GTT A) and *dsx*^{Gal4}.R2 (CTT ATT CTA TGC TGT CTC AAT GTT AG). To verify the presence of UAS-Dam-*dsx*^{M} or UAS-Dam-*dsx*^{F}, primers were designed to target three regions across the *dsx*^{M} or *dsx*^{F} coding region and pUAST-attB-LT3-NDam plasmid backbone. In UAS-Dam-*dsx*^{M}, forward primer *dsx*^{M}-Dam.F1 and reverse primer *dsx*^{M}-Dam.R1 were used for region one. The forward primer *dsx*^{M}-Dam.F2 and the reverse primer *dsx*^{M}-Dam.R2 were used for region two. The forward primer *dsx*^{M}-Dam.F2 was again used for region three with the *dsx*^{M}-Dam.R3 reverse primer. In UAS-Dam-*dsx*^{F}, forward primer *dsx*^{F}-Dam.F1 and reverse primer *dsx*^{F}-Dam.R1 were used for region one. The forward primer *dsx*^{F}-Dam.F2 and the reverse primer *dsx*^{F}-Dam.R2 were used for region two. The forward primer *dsx*^{F}-Dam.F3 was used with the *dsx*^{F}-Dam.R2 reverse primer, from region two, for region three. gDNA was isolated from five whole UAS-Dam-*dsx*^{M}/ *dsx*^{Gal4} male animals, and five whole UAS-Dam-*dsx*^{F}/ *dsx*^{Gal4} female animals using the Qiagen DNeasy Blood and Tissue Kit, twice replicated (see Methodology section 2.2 for protocol). DNA Sanger sequencing was completed by the facility at Source Bioscience, Oxford, UK.

4.3.2 PCR GENOTYPING DSX TADA TRANSGENIC FLIES

For PCR genotyping, 25 µl 2x Hot Start MyTaq™ Red Mix (Bioline, contains MyTaq reaction buffer and MyTaq DNA Polymerase), 1 µl forward primer, 1 µl reverse primer, 2 µl gDNA, and 21 µl dH_2O was used per reaction. Followed by thermocycling conditions: denaturation for 3 min at 95°C, 30 cycles of 30 s at 95°C, 30 s at 58°C, and 1 min at 72°C; and one cycle of 5 min at 72°C. Reactions were run on a Biometra T1 Thermoblock PCR thermal cycler. For electrophoresis, 2 µl of each PCR reaction was run on a 1% EtBr agarose gel submerged in TAE for 1 h at 70 Volts (see Methodology for agarose gel recipe, electrophoresis conditions, and visualisation).

4.3.3 OPTIMISATION OF PUBLISHED DAMID/ TADA PROTOCOLS

In this study, we followed the Vogel et al., 2007 DamID and Marshall et al., 2016 TaDa protocols for DSX brain and head TaDa. Each of the five major steps of these protocols, however, required thorough optimisation for profiling small numbers of *D. melanogaster* neural tissue. The following describes how we veer from the published protocols to effectively generate material for library preparation and NGS. Step one, gDNA isolation, is descrbed in detail below for DSX brain and head. To assess whether the DpnI restriction digestion of DNA was successful (step two), we introduced an additional measure whereby 2 µl of experimental samples were run with 2 µl gel loading dye (New England Biolabs) on 0.8% agarose gel at 70 V for 1 h. A successful digestion should produce a smear of the DpnI-cut blunt ended DNA fragments of varying lengths (200 to 10,000 bp). For ligation of DamID adaptors to DpnI-digested DNA (step three), we used 750 ng DNA diluted using MiliQ water to 15 µl. Here, we assessed DNA amount using two measures – NanoDrop Spectrophotometer (Thermo

Fisher Scientific) and QUBIT® Flurometer (QuantiFluor™ dsDNA System, Promega) – given this step is most efficient with the 750 ng stipulated in both protocols. Each sample was transferred to 0.2 ml PCR tubes, 4 µl of premade adaptor ligation buffer (2 µl 10× T4 DNA ligase buffer, 0.8 µl dsAdR (see Methodology section 2.3 for sequence), and 1.2 µl dH_2O) was added with 1 µl (400 U) T4 DNA ligase enzyme (NEB) and mixed well. The ligation reaction was incubated for 2 h at 16 °C followed by 10 min at 65 °C, to inactivate the ligase, in a Biometra T1 Thermoblock PCR thermal cycler. For efficient DpnII digestion (step four), the reaction was made up to a total volume of 40 µl with 20 µl ligated-sample from the previous step, 1 µl DpnII restriction enzyme, and 19 µl TaDa DpnII digestion buffer (4 µl 10× DpnII buffer, and 15 µl MiliQ water). This was digested for 2 h at 37 °C, and heat-inactivated for 20 min at 65 °C. We followed protocol for the PCR amplification reaction (step five), where the entire DpnII digested-DNA product from the previous step was used (40 µl). 118 µl premade DamID PCR buffer (16 µl 10× cDNA PCR buffer, 2.5 µl (50 µM) DamID_PCR primer (see Methodology section 2.3 for sequence), 3.2 µl 10 mM dNTPs, and 96.3 µl MiliQ water) were added to the DpnII digested-DNA, with 2 µl Advantage 2 cDNA polymerase enzyme (Clontech) and mixed well. For efficient PCR amplification, the reaction was split into 4 × 40 µl reactions in 0.2 ml PCR tubes.

4.3.4 gDNA isolation for Brain and Head TaDa

For DSX brain TaDa, adult animals were 5-7 day-old, and housed with male and female conspecifics. UAS-Dam-X*/ dsx^{Gal4} transgenic animals were anesthetised on ice for up to 15 minutes prior to dissection. A brief ethanol wash step was introduced by dipping the animal into 70% ethanol for 3-5 s to dewax the cuticle, preventing air bubbles

adhering to the cuticle when the animal is submerged in solution in the next step (Tito et al., 2016). Dissection pins were used to secure the animal. Dissection of animal material was completed in dissection dishes/ plates brimmed with DamID PCR buffer (Marshall et al., 2016). Following dissection, brains were grouped into batches of approximately twenty and snap frozen using liquid nitrogen. Excess DamID PCR buffer was carefully removed from the Eppendorf tubes containing the dissected brain tissue, and using metallic tongs submerged in liquid nitrogen for 30 s before being transferred to the -80 °C freezer. This method of freezing slows the action of proteases and nucleases, importantly inhibiting the degradation of RNA or proteins. gDNA was isolated from manually dissected brains using the QIAamp DNA Micro kit (Qiagen) for tissues, with modifications including a 10 h incubation with lysis buffer prior to the addition of Proteinase K, addition of 400 µl of Buffer AL and 300 µl of ≥99.5% ethanol, two rounds of both AW1 and AW2 buffer, heating the elution buffer to 56 °C for 5 min prior to elution, and an incubation with the elution buffer for 30 min prior to two rounds of elution.

For DSX head TaDa, as before, adult animals were 5-7 day-old, and housed with male and female conspecifics. UAS-Dam-X*/ dsx^{Gal4} transgenic animals were anesthetised on ice for up to 15 minutes before *Drosophila* whole heads were removed surgically in dissection dishes using a scalpel, and snap frozen using liquid nitrogen in batches of twenty. Alternatively, whole flies were snap frozen in Eppendorf tubes in batches of fifty, vortexed for 30 s at maximum power (Scientific Industries, Vortex Genie 2), and passed through 710 µm and 425 µm sieves (Hogentogler and Co, No. 25 mesh and No. 40 mesh) to separate heads, bodies, and appendices. In both cases, heads were then transferred to -80 °C for storage.

4.3.5 Library Preparation and Next Generation Sequencing (NGS)

Library preparation of all samples and subsequent NGS of DamID amplicons was completed at the Wellcome Trust Centre for Human Genetics, Oxford Genomics Centre, Oxford UK (OGC). For both DSX brain and head TaDa, each lane generated between 36-46 GB of data. Data was provided by an .FTP link.

4.4 RESULTS

4.4.1 DRIVING DSX-DAM IN *DSX* CELLS

We utilise our newly generated, and functionally tested, novel DSX-Dam TaDa transgenic flies to profile DSX neurons in the *Drosophila* CNS in a cell-specific manner using targeted DamID. To do this, we must drive expression of DSX-Dam in *dsx*-expressing cells. We hence implemented the Gal4/UAS system with dsx^{Gal4} (Figure 19, Rideout et al., 2010) driving expression of UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} in all *dsx* neuronal cells and tissue populations. The dsx^{Gal4} allele was generated using ends-in homologous recombination to insert the *Gal4* sequence into the first, non-sex specific coding exon of *dsx*, creating a tandem duplication at the locus. dsx^{Gal4} generated *wild type dsx* transcripts and *Gal4*-containing transcripts in both sexes (Rideout et al., 2010). Genetically, homozygous dsx^{Gal4} virgin females were crossed to homozygous UAS-Dam-dsx^{M}, UAS-Dam-dsx^{F} or UAS-Dam control adult male flies, generating heterozygous UAS-Dam-X*/ dsx^{Gal4} driven flies. To ascertain both DSX-Dam and dsx^{Gal4} were present in all heterozygous UAS-Dam-X*/ dsx^{Gal4} driven transgenic animals, to be used in both brain and head TaDa experiments, we used two methods of molecular genotyping: PCR genotyping, and DNA Sanger sequencing.

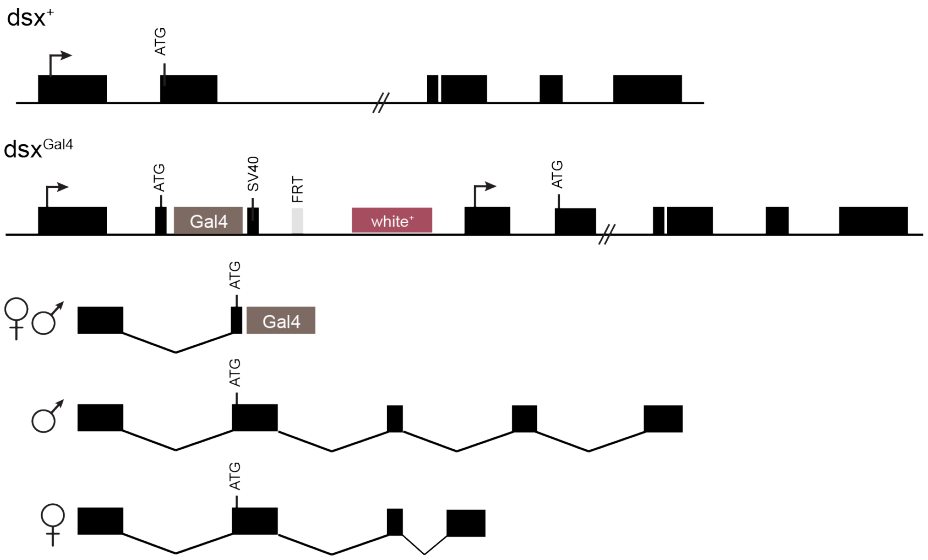


Figure 19 *The dsx^{Gal4} allele recapitulates dsx endogenous expression. The top schematic shows dsx (dsx^{+}) and the Gal4 knock-in allele (dsx^{Gal4}). Arrows indicate transcriptional start sites and black boxes indicate exons. dsx^{Gal4} produces wild type dsx transcripts and Gal4-containing transcripts in males and females. The schematic shows the dsx^{Gal4} locus predicted transcripts in both sexes (schematic adapted from Rideout et al., 2010).*

4.4.2 Assessment of DSX TaDa transgenic flies with DNA Sanger sequencing

We genotyped 5-7 day-old UAS-Dam-*dsx^{M}*/ *dsx^{Gal4}* and UAS-Dam-*dsx^{F}*/ *dsx^{Gal4}* transgenic animals for use in TaDa-seq experiments prior to their use. For *dsx^{Gal4}*, we designed two sets of forward and reverse sequencing primer pairs targeting the junction between the *dsx* coding gene and the Gal4 exon (Figure 20A). The same primers used to sequence verify UAS-Dam-*dsx^{M}* or UAS-Dam-*dsx*^{F} recombinant plasmid DNA were used here for transgenic animals. Three primer pairs target the *dsx^{M}* or *dsx^{F}* coding region and pUAST-attB-LT3-NDam plasmid backbone (Figure 20B and C). Sequencing reactions, with sequencing pairs for both UAS-Dam-X* and *dsx^{Gal4}*, each

showed both transgenes were present in the gDNA samples provided, across a number of independent gDNA extraction samples tested.

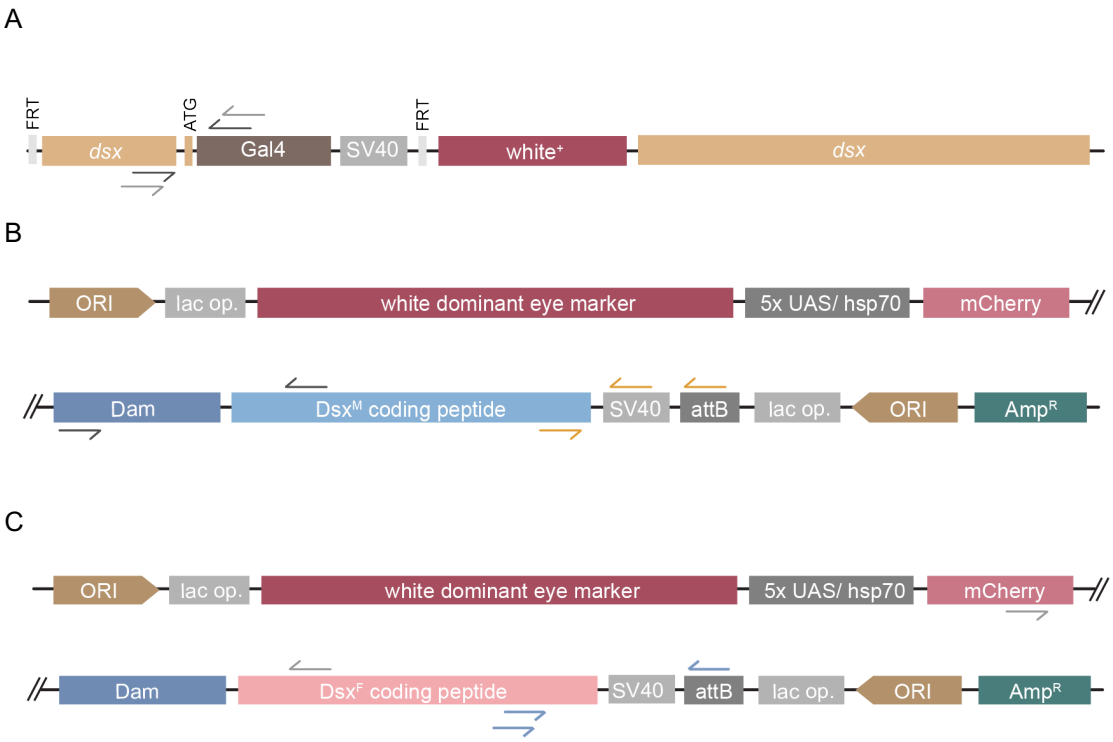


Figure 20 *Linear plasmid map of dsx^{Gal4} (Rideout et al., 2010) (A), UAS-Dam-dsx^{M} (B), and UAS-Dam-dsx^{F} (C). For A, grey arrows represent location of sequencing primers in region one, and black arrows for region two. For B, black arrows represent location of sequencing primers in region one, and orange arrows for regions two and three. For C, grey arrows represent location of sequencing primers in region one, and blue arrows for regions two and three. Plasmid maps visualised in Geneious 10.2.3.*

4.4.3 Assessment of DSX TaDa transgenic flies with PCR genotyping

As a second complimentary means to assess the genotype of the UAS-Dam-X*/ *dsx^{Gal4}* fly lines, verifying the presence of both transgenes, we used a combination of PCR and gel electrophoresis. For consistency and for increasing the likelihood that the above sequencing reactions were successful, we used the same forward primers as above

targeting the *dsx* gene and reverse primers targeting the Gal4 exon as previously for verifying dsx^{Gal4}. To verify the presence of UAS-Dam-X* we used the same primer pairs as above. Namely, two sets of forward and reverse primer pairs targeting both the DSX coding peptide itself and the vector backbone (either Dam or mCherry). We set up genotyping PCR reactions for UAS-Dam-dsx^{M}/ dsx^{Gal4} male animals, UAS-Dam-dsx^{F}/ dsx^{Gal4} female animals as well as for control animals, UAS-Dam/ dsx^{Gal4} males and UAS-Dam/ dsx^{Gal4} females. gDNA was isolated from five whole animals of each genotype using the Qiagen DNeasy Blood and Tissue Kit (see Methodology section 2.2 for protocol). Following electrophoresis and visualisation, for dsx^{Gal4} we identified bands at ~300 bp and ~450 bp, bands at ~900 bp and ~1000 bp for UAS-Dam-dsx^{M} and bands at ~950 bp and ~1000 bp for UAS-Dam-dsx^{F} (Figure 21). In each case, these bands were the expected band sizes. The result suggests amplification of the appropriate specific regions by PCR, and we therefore concluded that our experimental fly line for DSX brain and head TaDa expressed both UAS-Dam-X* and dsx^{Gal4} transgenes.

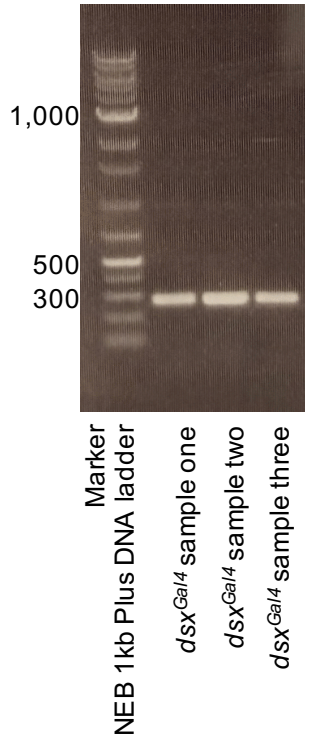


Figure 21 *Gel electrophoresis confirmation of presence of dsx^{Gal4} in transgenic fly lines. Gel electrophoresis of PCR reactions using primers targeting region two shown, band identified at ~300 bp in three independent samples, as modelled in Geneious 10.2.3. Gel electrophoresis of PCR reactions targeting regions two and three in UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} respectively completed as described in Transgenics chapter.*

4.4.4 Complete optimisation of established DamID/ TaDa protocols

Published experimental DamID and TaDa protocols (Vogel et al., 2007 and Marshall et al., 2016)

The original DamID protocol was first released in 2007 (Vogel et al., 2007) and its extension, the targeted DamID protocol, in 2016 (Marshall et al., 2016). With regards to the latter, the Marshall lab has recently made available their updated wet-lab experimental protocols on request (v2016-Sep-29 and v2017-Oct-9). Whilst both protocols are largely similar, there are small intricacies that differ at each of the five major stages. The TaDa protocol, established in profiling neurons in *D. melanogaster*,

can generate binding profiles from as few as 10,000 total induced cells. In the series of experiments discussed in this thesis, TaDa screens were conducted on >10,000 cells profiling DSX neuronal populations in dissected brain tissue or whole heads in *Drosophila*. Figure 22 below details the 'methylation smears', which the final step of the protocol demands prior to Next Generation Sequencing (NGS). There, experimental sample replicates produce visually similar methylation smears in the expected range (200 bp to 2,000 bp) (Vogel et al., 2007).

We found it necessary to optimise each step of the protocol to generate experimental reproducibility within and between sample replicates. Thorough optimisation of the DamID protocol was completed at each of the five major steps, one) isolation of genomic DNA, two) DpnI digestion of DNA, three) ligation of DamID adaptors, four) DpnII digestion of DNA, and five) PCR cycling conditions. References to the complete published protocols are located in the Methodology. The following sub-chapters detail the variation we introduced in experimental conditions at each step, as well as the additional sub-steps to assay the effectivity of experimental conditions, to attain the required experimental reproducibility.

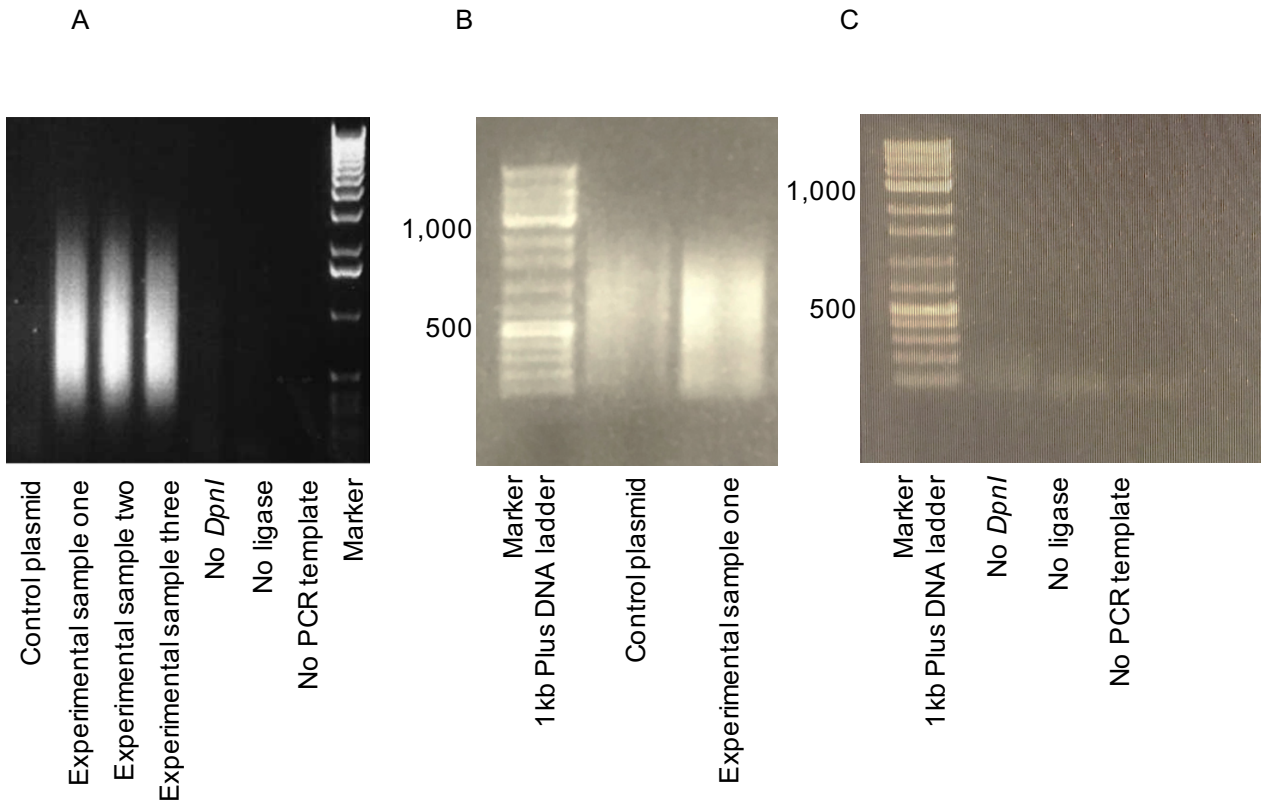


Figure 22 *Experimental methylation smears for TaDa protocol. Three lanes containing experimental Dam-fusion methylation smears (PCR product amplified from cells that expressed Dam (-fusion) protein). Alongside DNA marker, no smears are visualised for control plasmid or experimental control steps of the TaDa protocol: no DpnI, no ligase, and no PCR template (gel image taken from Vogel et al., 2007) (A). Example of TaDa methylation smears for UAS-Dam-X*/ dsx^{Gal4} experimental animals in this study (B) and associated experimental controls (C).*

4.4.5 TaDa Step 1 of 5: isolation of genomic DNA from tissues

gDNA isolation from animal material is required as the first step in the TaDa protocol (Marshall et al., 2016). We hence first optimised the starting animal material by conducting a series of gDNA isolation experiments. We used transgenic flies driving DSX-Dam in all *dsx* expressing tissues and neuronal populations (UAS-Dam-X*/ *dsx*^{Gal4}). The animal material used in DSX brain and head TaDa differs, and therefore initial surgical dissection of animals in the two rounds is substantially different. In brain

TaDa, we dissected fly brains, paying particular attention to remove the encapsulating fat body tissue that surrounds the brain (Aguila et al., 2007; Benes et al., 1990; Lazareva et al., 2007). This is a time-consuming, labour intensive method requiring fine technical skill. Adult animals were 5-7 day-old at the point of dissection. In contrast to the TaDa protocol which specifies dissection in PBS (Marshall et al., 2016), dissection of animal material was completed in dissection dishes/ plates brimmed with DamID PCR buffer (Marshall et al., 2016). The latter was similarly used by experimenters in Neville et al., 2014 for comparable animal dissections for DamID. In head TaDa, we used whole animal heads, which we removed with a scalpel, or snap-froze, vortexed and sieved to separate heads from other body appendages and debris.

For our gDNA isolation optimisation experiments, we used primarily the QIAamp DNA Micro Kit as specified in the TaDa protocol (Marshall et al., 2016), but compared its efficiency to the DNeasy Blood and Tissue Kit – both commercially available from Qiagen. gDNA isolation using the latter kit has been successfully used in published DamID experiments (Luo et al., 2011). The QIAamp DNA Micro Kit is designed for isolation of gDNA from less than 10 mg tissue, and the DNeasy Blood and Tissue Kit up to 25 mg tissue. Table 7 describes the experimental optimisations assayed including varied starting animal material (whole animals, whole heads, or dissected brains), varied tissue lysis times (1 h, 5 h, 10 h, or 24 h), and gDNA extraction kit (QIAamp DNA Micro Kit, or DNeasy Blood and Tissue Kit). In each sample, gDNA was visible as a single band at the top of a 0.8% agarose gel (and not a smear). This is important because a single band denotes a clean gDNA extraction (Figure 23). We used UAS-Dam-dsx^{M}/ dsx^{Gal4} male animals or UAS-Dam-dsx^{F}/ dsx^{Gal4} female animals for all experimental trials. Overall, we found the QIAamp DNA Micro Kit more efficient for

Drosophila gDNA extractions, and that 10 h tissue lysis at 56 °C was optimal for final gDNA extraction.

Qiagen QIAamp DNA Micro Kit			
Tissue lysis time	Animal material		
	10 whole animals *n≥5*, ng/µl mean +/- SD	30 whole heads *n≥5*, ng/µl mean +/- SD	50 dissected brains *n≥1*, ng/µl mean +/- SD
1 h	24.3 +/- 4.4	20.0 +/- 3.9	20.5
5 h	72.0 +/- 14.7	42.7 +/- 5.2	77.0
10 h	142.1 +/- 12.3	102.2 +/- 9.8	74.4 +/- 16.4
24 h	139.4 +/- 5.4	98.8 +/- 6.7	71.7 +/- 11.2

Qiagen DNeasy Blood and Tissue Kit			
Tissue lysis time	Animal material		
	10 whole animals *n≥5*, ng/µl mean +/- SD	30 whole heads *n≥5*, ng/µl mean +/- SD	30 dissected brains *n≥1*, ng/µl mean +/- SD
1 h	10.4 +/- 3.3	17.8 +/- 4.5	14.5
5 h	45.5 +/- 8.2	37.7 +/- 6.2	103.4
10 h	120.4 +/- 21.4	86.3 +/- 12.9	61.1 +/- 24.4
24 h	121.3 +/- 14.6	89.4 +/- 13.7	63.9 +/- 12.8

Table 7 *Optimising gDNA isolation. Varying tissue lysis time, amount of animal starting material, or Qiagen gDNA isolation kits. Values in table refer to final gDNA isolation concentration (ng/ µl) as measured on NanoDrop™ 2000 Spectrophotometer (Thermo Fisher Scientific) using 2µl sample. Final spin-down elution in 40 µl Buffer EB. For each experimental condition for whole animals and whole heads, mean +/- SD reported for n≥5 independent gDNA extractions. For dissected brain samples, 1 replicate for 1 h and 5 h lysis, and mean +/- SD reported for n=3 for others.*

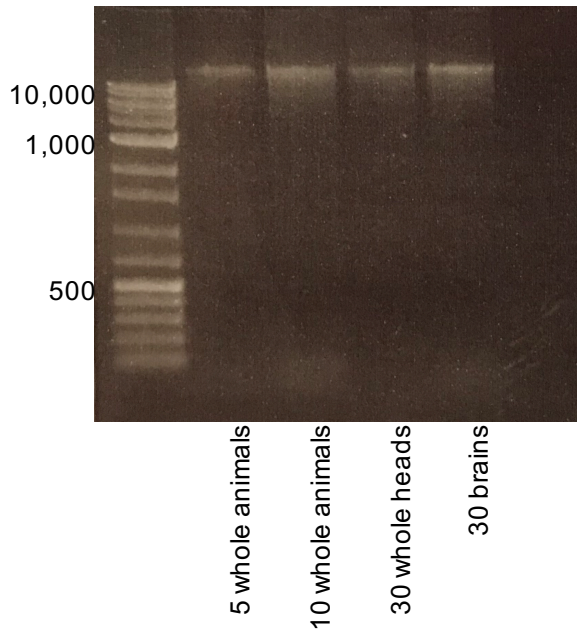


Figure 23 *Example of 'clean' gDNA extraction from 5 whole animals, 10 whole animals, 30 whole heads, and 30 dissected brains using Qiagen QIAamp DNA Micro Kit, tissue lysis for 5 h. 1 Kb Plus DNA ladder (NEB).*

4.4.6 TaDa Step 2 of 5: DpnI restriction digestion of DNA

The methylation-sensitive restriction enzyme DpnI cuts methylated GATCs between GA and TC nucleotides (GA^{me}). Digesting with DpnI therefore results in a pool of blunt-ended DNA fragments with a 5' TC and 3' GA^{me} sequence. Across protocols, the exact time for DpnI digestion at 37 °C varies. The original DamID protocol stipulates preferable digestion "overnight" but 4 h is "also possible" (Vogel et al., 2007). The TaDa protocol stipulates "overnight" digestion for at least 12 h (Marshall et al., 2016). However, 2 h digestion was suggested in the Marshall wet lab protocol (v2017-Oct-9). Across protocols, DpnI is always heat inactivated at 80 °C for 20 minutes. In our optimisation experiments, we trialled incrementally increasing DpnI digestion times: 2 h, 4 h, 12 h and 24 h. Table 8 details the relationship between amount of starting animal material, DpnI digestion time, and whether the animal material was successfully

DpnI-digested. Very low or undetectable DNA concentrations at this point are expected as uncut gDNA (which should be the majority of DNA in most cases) should not pass through the spin column. The purified DNA should come solely from induced cells. Measuring DNA concentration at this stage using the NanoDrop Spectrophotometer would not be useful in determining whether the digestion was successful.

Instead, we assayed whether the digestion was successful by electrophoresis: a successful digestion should produce a smear of the DpnI-cut blunt ended DNA fragments of varying lengths (Figure 24). Consistency across all animal tissue types (whole animals, whole heads, and dissected brains) was observed when samples were digested overnight for 12 h. Experimentally therefore, we used gDNA isolated with the QIAamp DNA Micro Kit eluted in 43.5 µl AE buffer, digested overnight (12 h) at 37°C with 1.5 µl DpnI (NEB) and 5 µl NEB CutSmart Buffer (NEB) as stipulated by Marshall et al., 2016. Whilst the DpnI restriction enzyme could have been heat-inactivated for 20 min at 80 °C, this is deemed unnecessary as the enzyme is removed during PCR clean-up prior to ligation. DpnI-digested DNA was cleaned up with the Qiagen PCR Purification Kit, and eluted in 32 µl of MiliQ water.

	Animal material		
DpnI digestion time	10 whole animals *n=12*	30 whole heads *n=15*	50 dissected brains *n≥2*
2 h	No smear 0/12	No smear 2/15	No smear 0/2
4 h	No smear 2/12	No smear 1/15	Smear 3/4
12 h	Smear 9/12	Smear 15/15	Smear 4/4
24 h	Smear 9/12	Smear 15/15	Smear 2/2

Table 8 *Various DpnI digestion times were trialled with various starting animal material to delineate the optimal digestion time. We considered a DpnI digestion event to have worked if an appropriate methylation smear was seen (200 bp to 10,000 bp) when 2 µl of the gDNA product was run on a gel via electrophoresis. 'Smear/ no smear' refers to the presence of methylation smear. For each experimental condition n refers to the number of independent sample replicates that produced the smear.*

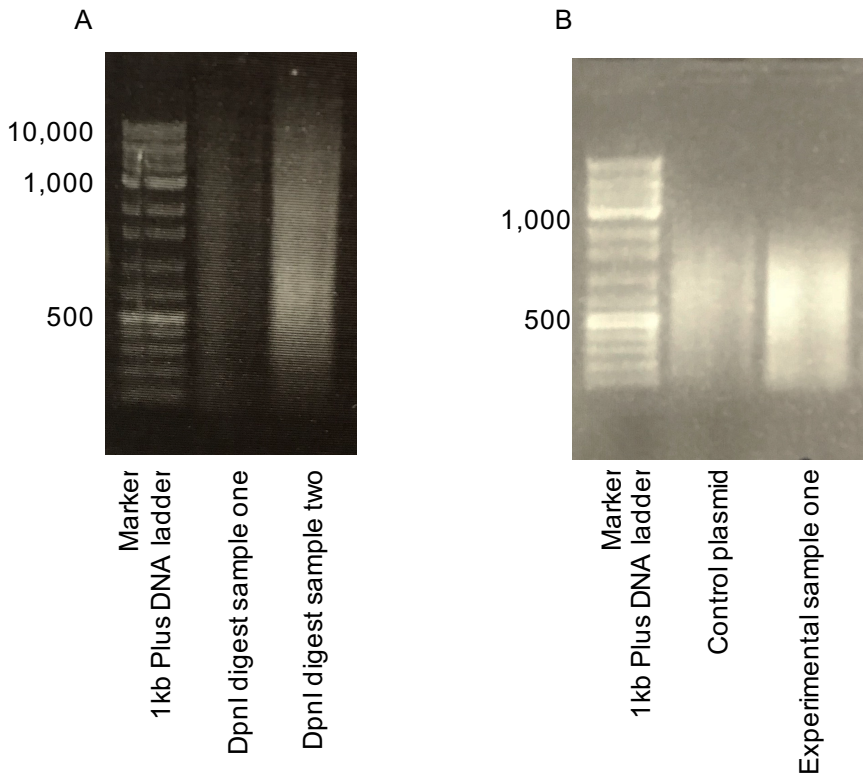


Figure 24 *Example of 'successful' DpnI digestion gel electrophoresis smears in two experimental samples (200 to 10,000 bp) (A). DpnI digestion for 12 h with thirty whole heads, and fifty dissected brains as starting material in lanes two and three consecutively. For comparison, PCR Dam-fusion methylation smears in control plasmid and experimental sample (B). Methylation smears for PCR amplified material cover a smaller range: 200 to 2000 bp (part B same as Figure 22B).*

4.4.7 TaDa Step 3 of 5: Ligation of DamID adaptors to DpnI-digested DNA

Having optimised DpnI digestion conditions in the previous step, we moved on to trialling ligation reaction conditions. In the ligation step, a double stranded adaptor oligonucleotide with a 32 bp 5' overhang ('dsAdR', comprising oligos AdRt and AdRb) is ligated to the pool of blunt ended DNA fragments. Double-stranded adaptor oligos ligate more efficiently compared to single-stranded oligonucleotides. The 5' overhang ensures directional ligation. Using the right DNA concentration and amount is key for the efficiency of the ligation step. 750 ng of total gDNA in 15 µl is optimal for the following DpnII digestion step. For this step, we simply ensured that our ligated

samples strictly fell within this concentration range in the required volume. To do this, we measured DNA concentration independently using both the NanoDrop Spectrophotometer (Thermo Fisher Scientific) and QUBIT® Fluorometer (QuantiFluor™ dsDNA System, Promega). Given the starting material and proportion of induced cells, very low to undetectable DNA concentrations at this step are normal. For ligation, 4 µl of premade adaptor ligation buffer and 1 µl (400 U) T4 DNA ligase enzyme was added to each 15 µl sample. The ligation reaction was incubated for 2 h at 16 °C followed by 10 min at 65 °C to inactivate the ligase.

4.4.8 TaDa Step 4 of 5: DpnII restriction digestion of DNA

Digestion with the methylation-sensitive restriction enzyme DpnII follows the ligation step, where unmethylated GATC sequences are selectively cut. This prevents amplification of DNA fragments containing unmethylated GATCs in the following step. Again, across protocols, conditions for DpnII reactions vary. The DamID protocol (Vogel et al., 2007) stipulates DpnII digestion at 37 °C for 1 h, whilst the TaDa protocol suggests digestion for at least 2 h with a "longer digestion acceptable" (Marshall et al., 2016). The Marshall lab's experimental protocol (Marshall lab, v2017-Oct-9) has an additional step of 20 min at 65 °C to heat-inactivate the restriction enzyme. While there is no discernible method to check whether the DpnII digestion step has been successful, we can assess following the next and final step of the TaDa protocol: PCR amplification. Here, we would hope to see an appropriate methylation smear (200 to 2000 bp) in post-PCR amplified samples via electrophoresis. We trialled DpnII digestions for the various times stipulated in the protocols. Reported PCR cycling conditions (Marshall et al., 2016) were used for the following step. We found DpnII

digestion from the shortest time point (1 h) to the longest (4 h) were successful in at least a proportion of replicates. However, we found DpnII digestion for 2 h was sufficient, and indeed at this time point the DpnII digestion step seemed to have worked for the vast majority of independent sample replicates across animal materials (Table 9). For efficient DpnII digestion, the reaction was made up to a total volume of 40 µl with 20 µl ligated-sample from the previous step, 1 µl DpnII restriction enzyme, and 19 µl TaDa DpnII digestion buffer. Our optimisation experiments for this step suggested prime conditions for digestion would be 2 h at 37 °C followed by heat-inactivating the enzyme for 20 min at 65 °C (Figure 25). To this point, our optimisation experiments have focused on carefully isolating specifically DSX-Dam methylated sequences targeted to neural cells in *Drosophila*. The following step refers to amplifying these sequencing via PCR.

	Animal material		
DpnII digestion time	10 whole animals *n≥8*	30 whole heads *n≥8*	50 dissected brains *n≥2*
1 h	3/8	4/8	1/2
2 h	8/9*	15/15*	4/4*
4 h	5/8	7/8	2/2

Table 9 *Various DpnII digestion times were trialled with various starting animal material to delineate the optimal digestion time. Proportion of 'successful' DpnII digestion reactions across different starting material and digestion times. Reaction deemed 'successful' if appropriate methylation smear is generated with sample post PCR-amplification. Excluding one sample replicate, all samples marked with an asterisk (*) digested for 2 h produced an appropriate PCR-amplified methylation smear. This time point was considered the optimal digestion time.*

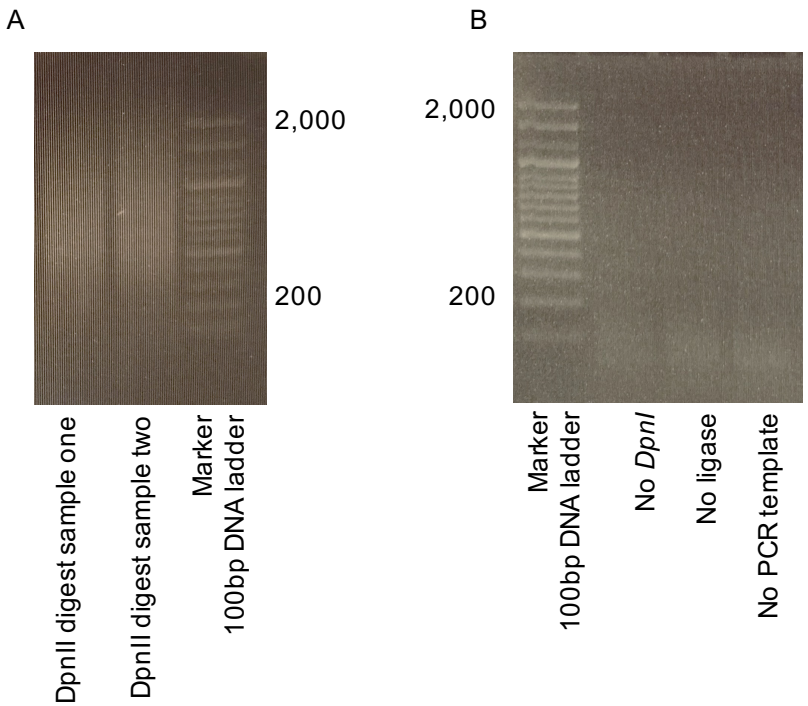


Figure 25 *PCR amplified DpnII-digested animal material gel electrophoresis smears in two samples (A). Methylation smears cover 200 to 2000 bp range. 2 h DpnII digestion with 30 whole heads starting material for both samples. Appropriate TaDa experimental controls for preparing material for NGS (B) including no DpnI, no ligase, and no PCR template. 100 bp DNA ladder (ThermoFisher) shown alongside both.*

4.4.9 TaDa Step 5 of 5: PCR cycling conditions

In the final stage of the experimental TaDa protocol, the DSX-Dam-methylated DNA is selectively amplified using PCR. The PCR cycling conditions specified in Vogel et al., 2007, Marshall et al., 2016, and the Marshall lab experimental protocols (v2016-Sep-29 and v2017-Oct-9) are identical (Table 10). The initial long extension for 10 min and the subsequent long cycles "appear to be required" (Marshall lab experimental protocol v2017-Oct-9). The 3 min extension time, present in the final 17 cycles, should amplify all fragments below 5 Kb. These represent 99.97% of all GATC fragments in the *D. melanogaster* genome. The entire product from the previous step, the DpnII digested-DNA, was used for the PCR amplification reactions, ~300 ng (20 µl). 118 µl

premade DamID PCR buffer and 2 µl Advantage 2 cDNA polymerase enzyme (Clontech) was added to the DpnII digested-DNA. For efficient PCR amplification, the reaction is split into 4 × 40 µl reactions.

Cycle number	Denature	Anneal	Extend
0			68 °C, 10 min
1	94 °C, 30 s	65 °C, 5 min	68 °C, 15 min
2–4	94 °C, 30 s	65 °C, 1 min	68 °C, 10 min
5–21	94 °C, 30 s	65 °C, 1 min	68 °C, 2 min
22			68 °C, 5 min

Table 10 *Cycling conditions for efficient PCR amplification (from Vogel et al., 2007, Marshall et al., 2016, and Marshall lab experimental protocols v2016-Sep-29 and v2017-Oct-9).*

In our repeated PCR assays, we noticed a sex difference in the PCR-amplified samples dependent on the amount of starting animal material used in each reaction. In brain TaDa we found that 50 dissected brains were sufficient to generate a methylation smear and hence successful amplification in UAS-Dam-dsx^{M}/ dsx^{Gal4} males. However, with UAS-Dam-dsx^{F}/ dsx^{Gal4} females, 50 dissected brains produced a methylation smear in just under half of our experimental replicates, but increasing this number to 70 brains produced methylation smears in all replicates. In head TaDa, 30 whole heads were enough to generate consistent methylation smears in experimental replicates for both UAS-Dam-dsx^{M}/ dsx^{Gal4} males and UAS-Dam-dsx^{F}/ dsx^{Gal4} females (Table 11). In head TaDa, we raised experimental and control flies at 25 °C as compared to 21 °C for the entire duration of their life cycle, not just for the 24 h suggested prior to extraction of

gDNA (Southall et al., 2013). Theoretically, this would increase the number of *dsx*^{Gal4} transcripts available, potentially boosting the driving of UAS-Dam-*dsx*$^{M/F}$ in *dsx* cells.

RT-qPCR was carried out to ascertain the specific PCR cycle number where amplification massively increases in our UAS-Dam-*dsx*^{M}/ *dsx*^{Gal4} male experimental samples but proved inconclusive across biological replicates. We hence added up to ten additional amplification cycles (cycle number 5-21 became 5-31, Table 10) in our RT-qPCR analyses to see whether we could produce the expected methylation smears in the UAS-Dam-*dsx*^{F}/ *dsx*^{Gal4} female experimental samples. Whilst this invariably did generate the desired methylation smears, the introduction of additional PCR cycles increases the risk of a size selected bias. We thus aimed to optimise the amount of starting material in the TaDa protocol instead. This was particularly important to generate biological reproducibility in PCR-amplified methylation smears in the UAS-Dam-*dsx*^{F}/ *dsx*^{Gal4} female experimental samples (Table 11).

A

Brain TaDa-seq						
	Experimental samples				Control samples	
	UAS-Dam-*dsx*^{M}/ *dsx*^{Gal4} ♂ *n=8*	UAS-Dam-*dsx*^{F}/ *dsx*^{Gal4} ♀ *n=8*			UAS-Dam/ *dsx*^{Gal4} ♂ *n=4*	UAS-Dam/ *dsx*^{Gal4} ♀ *n=3*
Brains (n)	50	30	50	70	50	70
PCR result	Smear 8/8	No smear 2/8	No smear 3/8	Smear 8/8	Smear 4/4	Smear 3/3

B

Head TaDa-seq				
	Experimental samples		Control samples	
	UAS-Dam-*dsx*^{M}/ *dsx*^{Gal4} ♂ *n=7*	UAS-Dam-*dsx*^{F}/ *dsx*^{Gal4} ♀ *n=8*	UAS-Dam/ *dsx*^{Gal4} ♂ *n=3*	UAS-Dam/ *dsx*^{Gal4} ♀ *n=4*
Heads (n)	30	30	30	30
PCR result	Smear 7/7	Smear 8/8	Smear 3/3	Smear 4/4

Table 11 *Comparing amount of starting animal material in experimental and control samples in DSX-brain TaDa (A) and head TaDa (B) with subsequent PCR amplification. 'Smear/ no smear' refers to whether a methylation smear was visualised when 2 µl of product was run on a gel via electrophoresis (200 to 2,000 bp). n refers to number of experimental trials for each condition. A successful PCR amplification step seemed dependent on starting animal material (gDNA) in brain TaDa. 20 more brains were required in experimental UAS-Dam-dsx*^{F}*/ dsx*^{Gal4} *females compared to experimental UAS-Dam-dsx*^{M}*/ dsx*^{Gal4} *males.*

4.4.10 Optimised experimental conditions for TaDa protocol (Marshall et al., 2016)

Bringing together the five major steps of the TaDa protocol, Table 12 highlights the main optimisation conditions trialled for each step as well as the peak experimental condition. This optimised version of the Marshall et al., 2016 TaDa protocol will be used for the TaDa experiments described in this thesis profiling DSX-expressing neural cells in the *Drosophila* CNS.

TaDa Step		**Conditions**			
1	gDNA extraction (tissue lyse time)	1 h	5 h	**10 h**	24 h
2	DpnI digestion time	2 h	4 h	**12 h**	24 h
3	Ligation	**2 h**			
4	DpnII digestion time	1 h	**2 h**	4 h	
5	PCR	Marshall et al., 2016 cycling conditions			

Table 12 *Final TaDa experimental conditions for DSX brain and head TaDa within overall optimisation experimental combinations. Starting material for DSX brain TaDa was 50 dissected brains for UAS Dam-dsx^{M}/ dsx^{Gal4} and 70 dissected brains for UAS-Dam-dsx^{F}/ dsx^{Gal4}; starting material for DSX head TaDa was 30 whole heads.*

4.4.11 Dam mutation in DSX TaDa transgenic flies

Given the elapsed time during experimental optimisation of the TaDa protocol, and given advice from another group working with DamID in *Drosophila* (Southall lab, Imperial College London), we re-sequenced a subset of our UAS-Dam, UAS-Dam-*dsx^{M}*, and UAS-Dam-*dsx^{F}* transgenic fly stocks every six months. To do this, we used

the DNA Sanger sequencing primer pairs described previously (sub-section 4.1.1) for UAS-Dam-dsx^M and UAS-Dam-dsx^F transgenic flies. For UAS-Dam, we used two primer pairs targeting a 400 and 700 bp expanse in the Dam region (~800 bp total) encoded on the secondary ORF (Figure 26). Whereas the subset of our UAS-Dam-dsx^M male and UAS-Dam-dsx^F female transgenic animals tested all returned expected sequences in three assayed occasions, we noted a point mutation located in the Dam region in one of the UAS-Dam stocks at one of the sequencing time points. Five individual flies were checked from three separate UAS-Dam stocks. In one stock, we identified the point mutation in each of the five flies tested. This suggests the mutation must be in the germline to be identified in multiple flies. This point mutation rendered the Dam not functional. These stocks were removed from the collection.

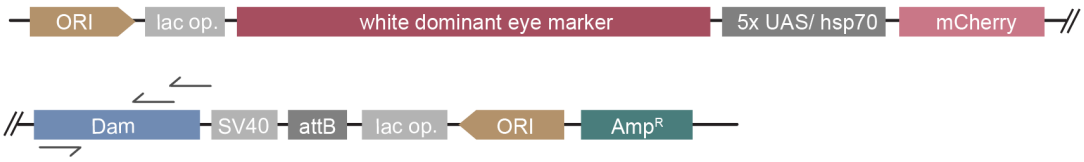


Figure 26 *Linear plasmid map of UAS-Dam. Grey arrows denote forward and reverse sequencing primer pair locations used in DNA Sanger sequencing verification.*

4.4.12 Profiling DSX neurons in the *Drosophila* CNS using TaDa

Having completed the set of optimisation experiments described here, we implemented our modified, enhanced version of the Marshall et al., 2016 TaDa protocol across two TaDa rounds both aiming to profile DSX neurons, cell-specifically, in the *Drosophila* CNS with two slightly different approaches. In DSX brain TaDa, we profiled DSX-

expressing neural cells in the brain of *Drosophila* in a cell-type specific manner. To do this, we set up genetic crosses between our TaDa UAS-Dam-X* transgenic animals with dsx^{Gal4} (Rideout et al., 2010). dsx^{Gal4} drives expression in all *dsx* cells, nervous system and non-neuronal adult tissues (Rideout et al., 2010). For brain TaDa, we hence manually dissected brains from UAS-Dam-X*/ dsx^{Gal4} transgenic animals, completing the TaDa profiling on this tissue, specifically. For dissections, we were particularly careful to remove *dsx*-expressing fat body tissue that encapsulates the brain (Aguila et al., 2007; Benes et al., 1990; Lazareva et al., 2007). Brains from fifty 5-7 day-old adults were dissected from each UAS-Dam-dsx^{M}/ dsx^{Gal4} males and UAS-Dam/ dsx^{Gal4} control males, and seventy UAS-Dam-dsx^{F}/ dsx^{Gal4} females and UAS-Dam/ dsx^{Gal4} control females. Three biological replicates were completed for both experimental samples as well as both control samples (720 brains total).

In DSX head TaDa, we also used our TaDa UAS-Dam-X* transgenic animals driven with dsx^{Gal4} (Rideout et al., 2010). For head TaDa, however, we profiled all DSX-expressing cells in the whole head. This profiling would include both DSX-expressing neural cells in the brain as well as DSX-expressing fat body and subcuticular tissue present in the *Drosophila* head. We completed the TaDa profiling from heads of thirty 5-7 day-old adults that were surgically removed from each UAS-Dam-dsx^{M}/ dsx^{Gal4} males, UAS-Dam-dsx^{F}/ dsx^{Gal4} females, and UAS-Dam/ dsx^{Gal4} control males and females. Two biological replicates were completed for both experimental samples as well as both control samples (240 heads total).

4.4.13 LIBRARY PREPARATION AND NGS

DNA was fragmented, end-repaired, A-tailed, and adapter-ligated before size selection and amplification at OGC. In brain TaDa, ten of the twelve sample library preparations passed the OGC quality control (QC) steps, and were fragmented by sonication using Covaris with expected insert size of ~200 bp. In head TaDa, all eight samples passed library preparation QC prior to sonication. Following library preparation QC, NGS was completed at 75 bp paired-end (PE) resolution using Illumina HiSeq4000 for both brain and head TaDa. Each lane generated a minimum of 240 M reads. One lane covered the four samples (UAS-Dam-dsx^{M}/ dsx^{Gal4} males, UAS-Dam-dsx^{F}/ dsx^{Gal4} females, and UAS-Dam/ dsx^{Gal4} control males and females) replicated three times each in brain TaDa. One lane also therefore covered the four samples replicated twice in head TaDa. OGC aligned data to the reference genome specified and assessed quality, delivering standard data files .fastq and .bam. Figure 27 details the overarching three step process for conducting a TaDa experiment.

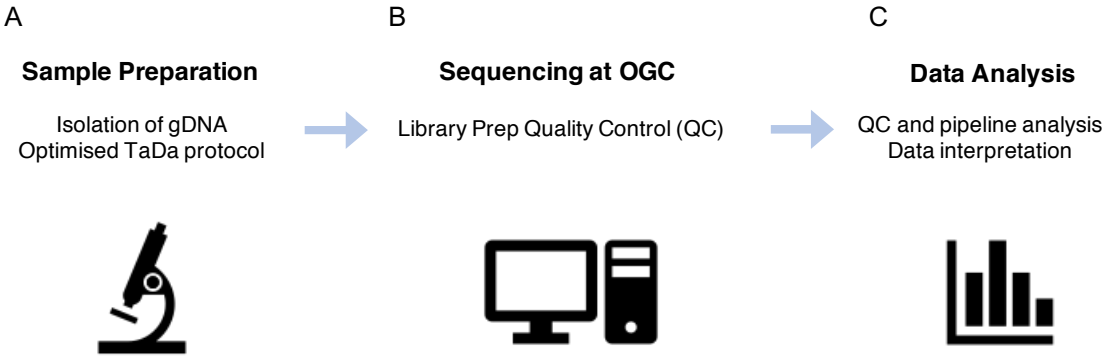


Figure 27 *The targeted DamID workflow is split in three main steps. Optimisation of the TaDa experimental protocol (described here) is essential for sample preparation (A). Library preparation and NGS of prepared samples are completed externally at OGC (B). Assessments of the quality of the NGS datasets and interpreting the biological meaning (C) are discussed in the following chapter.*

4.5 DISCUSSION

Here, we have conducted a thorough functional comparison and optimisation of the five major steps of the experimental DamID and TaDa published protocols (Marshall et al., 2016; Vogel et al., 2007). These optimisations were necessary to adapt the protocol for profiling small numbers of induced cells in the brain and head of *D. melanogaster* adult transgenic tissue, as well as to generate experimental reproducibility across biological replicates.

Although DSX has been studied extensively in *Drosophila* for over 50 years, there are still few defined DSX target genes. Downstream analyses of those previously identified cannot explain the plethora of sexually dimorphic morphologies and behaviours regulated by this sex determination gene. Employing TaDa to profile *dsx* neuronal populations in adult *Drosophila* aims to further delineate putative target genes and thus try to better explain DSX function. The novel approach taken here involves expressing *Drosophila* dsx^{M} and dsx^{F} coding peptides fused to *E. coli* Dam on the secondary ORF in a TaDa construct, microinjected into *Drosophila* embryos, driven in *dsx* neuronal populations using a series of genetic crosses. This approach potentially allows us to compare Dsx^{M} and Dsx^{F} function by way of comparing the putative genes revealed in the analysis across the two screens.

Following confirmation that our DSX-Dam TaDa constructs are functional, we conducted a series of genetic crosses to drive these constructs in *dsx* neuronal populations using the published dsx^{Gal4} allele (Rideout et al., 2010, Figure 19). We assessed if transgenic animals contained both transgenes and hence TaDa DSX-Dam

was driven in *dsx* neuronal populations by two methods: DNA Sanger sequencing and PCR/ gel electrophoresis genotyping (Figure 20 and Figure 21). Complete optimisation of the DamID/ TaDa method was necessary to adapt the experimental protocol to profile firstly *Drosophila* adult transgenic tissue, and secondly the small number of DSX-Dam induced neural cells in the CNS. We hence aimed to ascertain the peak conditions at each of the five major experimental stages to appropriately isolate gDNA, and selectively cut and amplify only the DSX-Dam methylated regions. Most extensively, complete optimisation of each step of the protocol as discussed in this chapter was also necessary to achieve reproducibility within and between experimental biological replicates.

In step one, as recommended in the TaDa protocol (Marshall et al., 2016), our analyses revealed gDNA extraction using the Qiagen QIAamp DNA Micro Kit was the most effective, with tissue lysis for 10 h. Interestingly, increasing lysis time above 10 h for any tissue type (whole animals, whole heads, or dissected brains) did not significantly increase gDNA concentrations (Table 7). In step two, the range of stipulated DpnI digestion times across protocols were broad, assumingly to account for the range of gDNA amounts being used in this step. 'Overnight' digestion for 12 h was optimal for the 30 whole *Drosophila* heads or 50 dissected brains (Table 8). In step three, the ligation stage, conditions stipulated across both protocols were identical and strictly followed in this study. In step four, DpnII digestion times were trialled between one and four hours (Table 9). Digestion for 2 h at 37 °C appeared sufficient, plus an additional step, heat-inactivation of the enzyme for 20 min at 65 °C not stipulated as imperative in the TaDa protocol (Marshall et al., 2016). For the final step, as with the ligation stage –

experimental PCR conditions were constant across DamID and TaDa protocols, and thus kept consistent in this study as well (Table 10).

Achieving reproducibility within and between experimental biological replicates proved difficult. We saw a significant level of inconsistency in the generated methylation smears in our UAS-Dam-dsx^{M}/ dsx^{Gal4} male and UAS-Dam-dsx^{F}/ dsx^{Gal4} female post-PCR amplified samples. We triple checked the genetics of all experimental crosses were correct, genetically and phenotypically. We raised flies at 25 °C and kept flies at 29 °C for 72 h prior to isolation of gDNA, and trialled using more starting animal material. Raising flies and keeping flies at higher temperature prior to isolation of gDNA theoretically boosts the number of dsx^{Gal4} transcripts available to drive the TaDa DSX-Dam construct, hence boosting expression levels. We found experimental UAS-Dam-dsx^{F}/ dsx^{Gal4} female samples required higher amounts of starting material (gDNA) compared to UAS-Dam-dsx^{M}/ dsx^{Gal4} males: starting with 70 brains in females compared to 50 in males generated consistent methylation smears (Table 11). This could be explained by the fact that more cells are induced in UAS-Dam-dsx^{M}/ dsx^{Gal4} males overall for TaDa profiling compared to UAS-Dam-dsx^{F}/ dsx^{Gal4} females (Pavlou et al., 2016; Rezával et al., 2016; Rideout et al., 2010).

Overall, we repeated iterations of the TaDa protocol over fifty times in both brain and head TaDa before we generated biological reproducibility and hence sent samples for library preparation and NGS. We considered the biological variation in expression levels of individual UAS-Dam-dsx^{M}/ dsx^{Gal4} male and UAS-Dam-dsx^{F}/ dsx^{Gal4} female transgenic animals could point towards the variation we see in the method output,

despite all animals being raised at 25 °C. Taking samples forward with varying DSX-Dam methylation footprints will carry this variation through the protocol or even amplify it, and indeed at each step there are potential further opportunities for variation to be introduced. For example, perhaps batch snap-freezing dissected brains or whole heads caused protein degradation in a subset of animal material. Snap-freezing dissected brains or individual heads immediately as they are dissected could potentially round this problem. For the steps involving the selective cutting and amplification of DSX-Dam methylated material, efficiency of each step is in part dependent on the ratio of induced to non-induced cells. Thus, small biological differences in DSX-Dam methylated regions can potentially be exacerbated as the protocol progresses. Our repeated TaDa optimisation assays to generate reproducible DSX-Dam methylation smears in both UAS-Dam-dsx^{M}/ dsx^{Gal4} male and UAS-Dam-dsx^{F}/ dsx^{Gal4} female samples garner confidence in the biological relevance, meaning, and reproducibility of these TaDa screens.

Brain and head TaDa probe slightly different experimental questions. In brain TaDa, we profile *dsx*-expressing neural cells using dissected brains from animals expressing UAS-Dam-dsx^{F}/ dsx^{Gal4} or UAS-Dam-dsx^{M}/ dsx^{Gal4}. In head TaDa, we use the whole heads of animals of the same genotype – therefore also assaying signal from *dsx*-expressing fat body tissue that encapsulates the brain. In the following chapters, we directly compare and contrast the output of this TaDa screening process, and independently compare the results from previously published DSX-fat body DamID experiments (Clough et al., 2014).

As discussed in the previous chapter, if our attempts to generate the Split-Gal4 *dsx*^{DBD}/ UAS-Dam-X* recombinant line were successful, we could have combined this line with, for example, a neuron specific driver such as the ELAV gene. This would drive TaDa DSX-Dam specifically in *dsx* neural cells, hence negating the need to manually dissect brains as we did in brain TaDa. Further, this method would remove risk of contamination of profiling *dsx*-expressing subcuticular tissue or fat body that could potentially be profiled if not carefully removed from every single brain dissected. Background signal from *dsx*-expressing subcuticular tissue or fat body could interfere with the sample signal when profiling *dsx* neuronal populations. However, we were particularly vigilant when dissecting individual brains in brain TaDa to carefully remove all encapsulating fat body tissue from the brain. Our screen in head TaDa embraces the *dsx*-expressing fat body and subcuticular tissue fully in profiling all these cells in *Drosophila* heads alongside the *dsx*-expressing neural cells in the brain. Subsequently, our comparisons with previously published DSX-fat body DamID experiments (Clough et al., 2014) allow us to later draw biological conclusions about *dsx*'s mode of action across tissues.

5 DOUBLESEX TADA-SEQ

BIOINFORMATIC ANALYSIS OF DSX BRAIN AND WHOLE HEAD TADA-SEQUENCING DATASETS

5.1 INTRODUCTION

5.1.1 PRACTICAL CONSIDERATIONS WHEN PROCESSING NGS DATASETS

In any genome-scale experiment involving the generation of sequencing data, arguably the appropriate assessment of the data obtained is the most important step. Indeed, a carefully considered implementation of bioinformatic techniques tailored specifically to the type of experiment in question is crucial (Aughey et al., 2016). The techniques discussed in this thesis, TaDa and ChIP, are no exception to this. Whilst ChIP as a technique has been established earlier, and therefore computational tools and pipelines to analyse ChIP-seq data are better developed, bioinformatic tools to analyse both ChIP and DamID are available (Li et al., 2015; Marshall and Brand, 2015). The sequencing considerations that follow are fundamental to ensure a successful sequencing experiment and allow for the appropriate interpretation of the data.

SAMPLE PREPARATION AND SEQUENCING

During sample preparation, several considerations are necessary to accommodate for repetitive regions in the DNA sequence, and chromatin accessibility. Fragmented DNAs, ~150-500 bp, from whole genome analyses including DamID-seq and ChIP-seq are sequenced in smaller ~36-100 bp units, or reads. Single-end reads are typically used for ChIP-seq and DamID-seq analyses. Paired-end reads could enhance library complexity, and increase mapping efficiency at repetitive regions in the genome (Chen et al., 2012; Nakato and Shirahige, 2017). Whilst paired-end reads can also be used to garner information on fragment size-distribution, a number of methods exist to estimate this distribution from single-end mapped data (Hansen et al., 2015; Kharchenko et al., 2008; Yang et al., 2008). An important potential confounding factor to be considered is

the accessibility of chromatin. Chromatin accessibility varies across genomic loci during fragmentation: a bias exists in certain open-chromatin regions where DNA is amenable to fragmentation and thus preferentially represented in a fragmented sample. This causes false-positive read enrichment (Auerbach et al., 2009). Such regions include actively transcribed promoter regions. Conversely, regions that are tightly packed such as heterochromatin are sheared to a smaller extent by DNA fragmentation confounding weak enrichment of true binding sites for heterochromatin markers (Chen et al., 2012). Fragmentation biases are potentially mitigated by shearing longer DNA fragments (i.e. 350-800 bp) enzymatically or mechanically using sonication (Deardorff et al., 2012). Including longer fragments widens the obtained 'peaks' (explained below) whilst not affecting the resolution of the peak-summit (Chen et al., 2012). In our TaDa-seq analyses we attempt to mitigate this fragmentation bias by employing enzymatic shearing of longer DNA fragments in the library preparation stage of sample preparation.

READ MAPPING

Sequenced reads are mapped to a specified version of a genome using mapping tools such as Bowtie 2 (Langmead et al., 2009; Langmead and Salzberg, 2012; Li and Durbin, 2009). The majority of DamID-seq and ChIP-seq experiments do not require gapped alignments which consider insertions and deletions, 'indels'. This is because the sequenced reads do not contain them unlike in exon junctions in RNA-seq analyses (Furey, 2012; Nakato et al., 2017). An important consideration in read mapping is what to do with multiple mapped reads, i.e. single reads mapped to multiple loci in a given genome. If multiple mapped reads are included, the number of usable reads is hence increased, which heightens the sensitivity of peak detection. However, this method

increases the number of false positive read maps, given all but one of the several maps of the one read must be incorrect. Subsequently, it may increase the number of false positive called peaks, if the read count exceeds the specified statistical threshold (Chung et al., 2011). To analyse the function of TFs, singularly mapped reads are sufficient (Day et al., 2010). Assessing the proportion of mapped reads, the 'mapping ratio', is important for understanding whether a sequencing experiment has potentially worked or not. The mapping ratio varies depending on species and read lengths.

Library complexity

The primary measure of library complexity is the non-redundant fraction (NRF), that is the ratio of non-redundant reads to the total number of mapped reads. Non-redundant reads are reads mapped to the same genomic positions T times or less, where T is the threshold for redundant reads. Optimally, redundant reads should be omitted from further analysis. If $T = 1$, then the expected number of mapped reads per base pair is less than or 1 (the 'sequencing depth') as seen in humans. For smaller genome sizes as for instance in yeast, relaxing the T threshold to $T > 1$ is important because stringent filtering has a minimal effect on peak detection (Chen et al., 2012). Given the NRF is dependent on the total number of mapped reads, read sampling is pivotal for comparing NRF scores across samples. Low library complexity arises when samples are generated from a small amount of starting material. Even if the number of sequenced reads is theoretically sufficient following PCR amplification, the substantial read power would theoretically be small resulting in a small significance.

Peak calling

'Peak calling' is a computational technique used to discover locations in a given genome enriched with aligned reads as a result of an NGS experiment such as DamID-

seq or ChIP-seq. A peak is called if a number of reads mapped to the genome at a particular location either exceeds the number of reads according to a specified threshold, or if there is minimum enrichment compared to background signal. The positions, or peaks, are thought to be loci where a protein-DNA interaction occurs. If the protein of interest (POI) is a TF, the enriched areas are potentially close to transcription factor binding sites (TFBS). Numerous tools are publicly available to process NGS datasets, some of which apply both methods, which differ in the algorithms used to identify and call peaks, and of which parameters can be adjusted. This leads to significant variation in the number of peaks called within and between peak calling algorithms (Nakato and Shirahige, 2017). A major task for peak calling programs is distinguishing the signal of interest from molecular or experimental noise, and taking into consideration loci that are systematically more accessible by the experiment (Thomas et al., 2017).

Peak shape varies amongst proteins and is categorised into different modes. 'Sharp (or narrow) mode' peaks are located at small, specific loci in the genome, 'broad mode' peaks are associated with large genomic domains, and 'mixed mode' involving both of these modes (Figure 28A; Nakato et al., 2017; Park, 2009). The majority of TFs have sharp modes, and therefore the majority of peak-calling algorithms are designed for this. Mixed mode is observed for RNA polymerase II (Pol II), and transcription elongation factors (Lin et al., 2011). In a recent review of thirty methods for peak calling, six methods were specified that performed best across 300 simulated and three real ChIP-seq datasets (Thomas et al., 2017). Performance was based on the ability to identify candidate peaks and testing candidate peaks for statistical significance. Methods included Model based Analysis for ChIP-seq version 2 (MACS2) (Zhang et

al., 2008), MUltiScale enrIchment Calling for ChIP-seq (MUSIC) (Harmanci et al., 2014) and Genome wide Event finding and Motif discovery (GEM) (Guo et al., 2012). MACS2, a refinement to the original MACS published over a decade ago, models the distance between the paired forward and reverse strand peaks in ChIP-seq data (Zhang et al., 2008). MACS2 slides a window across the genome to locate enriched regions as compared to background. The expected background is the number of reads multiplied by their length, divided by the mappable genome size. MACS2 extends the reads in the 3' direction to the fragment length obtained from modelling, then locating candidate peaks by scanning the genome again using a sliding window size twice the fragment length. To account for local bias in background read levels, MACS2 calculates a significance value for each peak using a dynamic Poisson distribution. If a control sample is available, this is used to calculate for local background.

The MACS2 method of peak calling, developed for ChIP-seq datasets, is extremely inefficient and generates high levels of false positives if applied to DamID-seq datasets. This is because MACS2 is based on the idea reads tagging both DNA strands are symmetrically distributed at a binding site (Gaspar, 2018). Hence, the middle point of the estimated distance between two distributions of tags within a called peak is a potential TFBS (Figure 28B; Zhang et al., 2008). In DamID-seq datasets, expression of Dam-POI results in specific methylation of adenines in the GATC sites surrounding the binding sites. GATC sequences are present in the genome every 200 bp to 2.5 Kb on average, however not at regular intervals. Isolated fragments are likely to contain regions nearby or within genes in addition to the binding site itself. Therefore, induced methylation signals are not guaranteed to be symmetrically distributed at TFBS (Figure 28C). This is impossible for MACS2 to model. Peak callers have been recently

developed for DamID-seq to account for this (Li et al., 2015; Marshall and Brand, 2015).

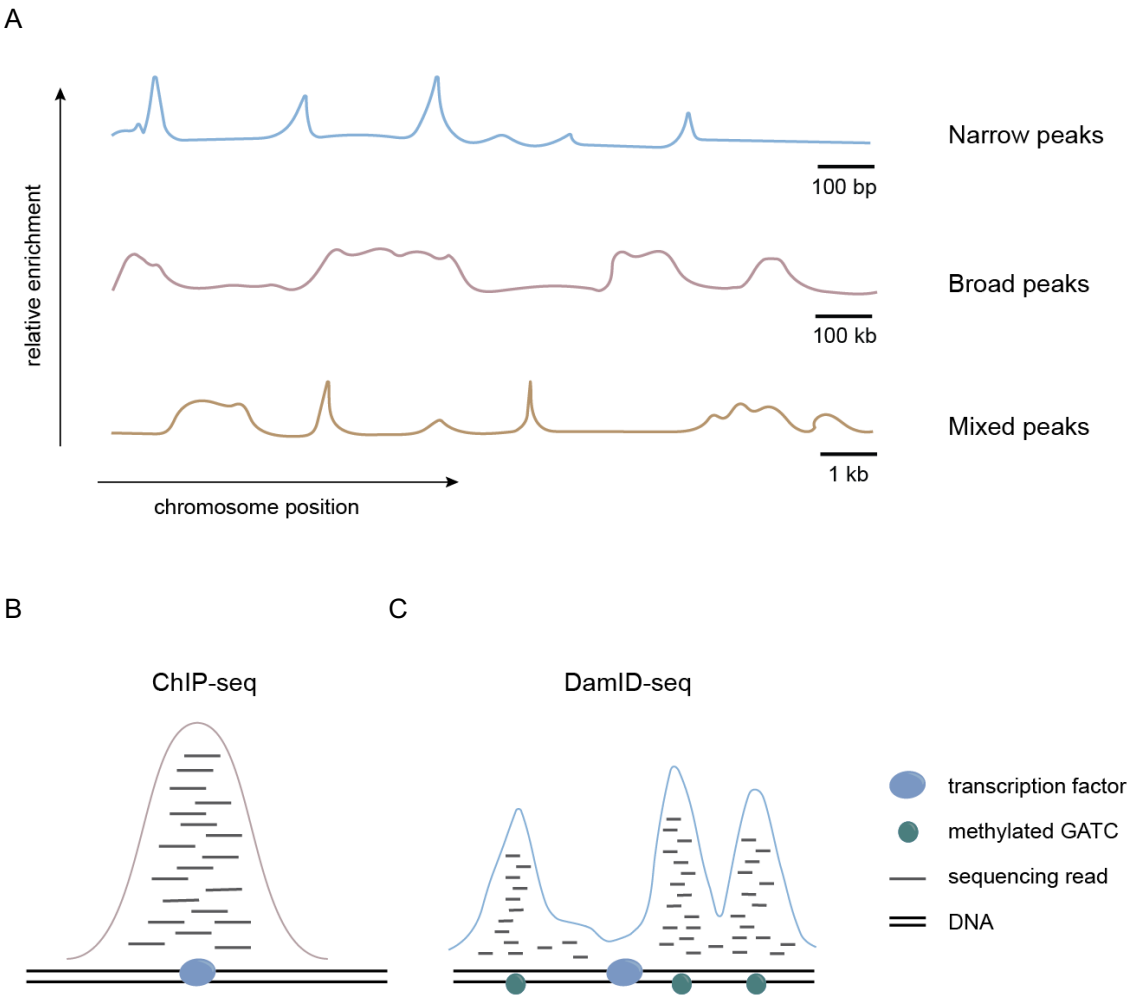


Figure 28 *Peak shape differs amongst proteins and is divided into three categories: 'sharp/ narrow', 'broad' or 'mixed' (A, top middle and bottom respectively). NGS read alignment characteristics differ in ChIP-seq (B) and DamID-seq (C) datasets. For ChIP-seq, TF binding sites are predicted to be centrally located between two distributions of tags in called peaks. For DamID-seq, induced methylation signals are not guaranteed to be symmetrically distributed at TFBS. If samples are fragmented enzymatically or by sonication, more reads are visible between GATC sites. Peak calling strategies therefore differ in ChIP-seq and DamID-seq experiments.*

SEQUENCING DEPTH

The relationship between sequencing depth and number of called peaks is linear: when sequencing depth increases so does the peak number. This is because weaker sites become significantly enriched with a higher number of reads (Sims et al., 2014). 'Saturation analyses' are used to determine sufficient sequencing depths. These analyses subsample the original read set in a stepwise manner, and calculate the number of identified peaks overlapping the original peaks at each depth (Park, 2009). Where the proportion is saturated, the depth is determined to be sufficient.

5.1.2 ESTABLISHED PIPELINES TO PROCESS DAMID-SEQ DATASETS

In 2015, the same group that developed the targeted DamID technology published the software pipeline "damidseq_pipeline" to process the raw sequencing data generated from a DamID-seq experiment. As previously discussed, bacterial Dam protein uniquely methylates adenine in the sequence GATC. Higher eukaryotes lack native adenine methylation. Hence, the unique DNA-binding POI 'footprint' can be identified through selectively extracting sequences flanked by adenine methylated GATC sites. Normalisation based on read-counts alone in DamID-seq datasets potentially results in high background and loss of bound signal – DamID-seq therefore presents novel challenges. Whilst DamID-seq data can be aligned and binned in the same way as other NGS data, the damidseq_pipeline takes a novel approach to the issues of normalisation and high background noise. Several studies have coupled NGS with DamID (Carl and Russell, 2015; Clough et al., 2014; Lie-A-Ling et al., 2014; Wu and Yao, 2013), comparing peak binding intensities between read-count-normalised Dam-POI and Dam samples. This method has caveats, however, and dependent on the characteristics of the

Dam-POI could result in loss of 'real' signal. Correctly normalising the datasets is imperative to detect all binding interactions by many Dam-POI's.

MARSHALL AND BRAND (2015) DAMID_SEQ PIPELINE

In many Dam-POI datasets, there is a large contribution from non-specific methylation of accessible genomic regions. This needs to be accommodated for in the data processing pipeline. The mean correlation between analogous Dam-only controls and Dam-POI datasets is 0.70 (*n=4*, Spearman's correlation; Marshall and Brand, 2015), and thus presenting the data as a Dam-POI/ Dam ratio is a good means to test non-specific methylation. However, the presence of strong methylation signals at highly bound regions in the Dam-POI dataset will relatively reduce the numbers of reads present at accessible genomic regions in the dataset, and if normalisation is based on read counts alone, a strong negative bias would be introduced to the ratio file. Depending on the characteristics of the POI, this negative bias could result in 'real' signal being lost. To round the negative bias, Marshall and Brand used the read counts from accessible genomic regions identified from the Dam-only dataset as the basis for normalisation. Their method theoretically avoids regions likely to contain 'real' Dam-POI signal. Marshall and Brand developed the following algorithm to adapt the Dam-POI dataset, based on the GATC fragment resolution of DamID. The algorithm was designed to divide the read counts into GATC fragments, remove all GATC fragments without read counts, and divide the remaining fragments into deciles. Fragments in the top decile were removed because of the high likelihood that these Dam-POI read counts represent 'real' signal. Further, in the Dam control dataset, the group found the first three deciles generally generated inconsistent normalisation values, and were thus excluded. The distribution of the $\log_2$(Dam-POI: Dam) ratio for the remaining GATC

fragments (x_1, x_2... x_n) was established through the following Gaussian kernel density estimate:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp(-\frac{(x_i-x)^2}{2h^2})$$

where h = bandwidth, estimated via the method of Silverman (1986)

$$h = 0.9 \frac{\min(\sigma, IQR)}{1.34} n^{-1/5}$$

Σ = standard deviation of sample, IQR = interquartile range

The maximum point on the Gaussian kernel density marks the maximum point of correspondence between Dam-POI and Dam values. If both datasets were correctly normalised, this value should equal 0. The group estimated the kernel density over three hundred equally spaced points within the interval [max (-5, min(x)), min (5, max(x))].

Dam-POI values were normalised by $1/(2^{\arg\max(\hat{f}_h(x))})$.

Genomic regions containing non-specific methylation with randomly distributed background counts generate large amounts of noise when calculating the Dam-POI: Dam ratio. High levels of background noise potentially cause problems with downstream peak detection and peak calling. As a means of lessening the noise, Marshall and Brand added pseudocounts to both the Dam-POI and Dam control datasets. To ensure equality in replicates containing a different number of reads, the number of pseudocounts added is directly proportional to the sequencing coverage, c

(reads/ bins), where c is a constant. This works on the assumption that $genome_{bound} \leq genome_{unbound}$. The addition of pseudocounts increases the signal: noise ratio, and hence the genomic coverage and total number of detected and called peaks.

MARSHALL AND BRAND (2015) FIND_PEAKS PIPELINE

Marshall and Brand (2015) released a "find_peaks" pipeline alongside their damidseq_pipeline. find_peaks uses an algorithm developed by Southall for peak detection and calling (Wolfram et al., 2014) based on a False Discovery Rate model (FDR). The binding intensity thresholds in the dataset are identified, the dataset is shuffled randomly, and the frequency of consecutive regions (i.e. GATC fragments or bins) with a score higher than a specified threshold (FDR) are calculated. Given the relationship between the number of consecutive fragments and the frequency of observations in a random dataset is logarithmic, using linear regression one can model effectively, for any number of fragments. The FDR is therefore the observed/expected ratio for a number of consecutive fragments above a certain threshold. This find_peaks software is designed to process output from the damidseq_pipeline specifically, but can handle any DamID-seq processed dataset, i.e. any DNA binding track in .bedgraph or .GFF format. The script, implemented in Perl, is fully adjustable depending on the expected genome coverage of the POI. One can amend the resolution of peak detection at the expense of speed, or detect peaks of lower heights. The pipeline outputs a .GFF file of all detected and called peaks with an FDR below a certain value, where the default value is FDR = 0.01. Information on mean binding intensity of each peak, their exact FDR and genomic coordinates, amongst others, are generated.

Here, we present a series of bioinformatics analyses of two targeted DamID experiments profiling DSX neuronal populations in the brain and whole head. Whilst no one bioinformatic method is a panacea for analysis, our considered approach takes into account the practicalities of processing NGS datasets described here. This significantly aids in the interpretation of the findings from the experiments, and helps to draw biologically meaningful conclusions.

5.2 AIMS

1. Assess TaDa-sequencing NGS quality from experiments profiling Dsx^{M}-Dam and Dsx^{F}-Dam in *D. melanogaster* brains and whole heads.

2. Process TaDa-seq data using published computational pipelines (Marshall and Brand, 2015), call peaks, and generate genome-wide gene lists.

3. Conduct downstream analyses on both DSX brain and head datasets, including Gene Ontology and MOTIF analyses, and make functional comparisons.

5.3 RESULTS

5.3.1 OVERVIEW OF PROCESSING TADA-SEQ DATASETS

Appropriately processing NGS datasets including TaDa-seq, for accurate biological interpretation, involves the careful selection and implementation of a number of bioinformatics tools. The steps depicted here describe the handling, rationale and subsequent interpretation of such large datasets. In both DSX brain and head TaDa, the Oxford Genomics Centre (OGC) aligned data to the *Drosophila melanogaster* Ensembl BDGP6 reference genome specified and assessed quality, delivering data files in .fastq and .bam formats. To process DamID-seq datasets of this nature, we firstly completed pre-processing whereby sequencing reads were assessed for quality and biases using the FastQC tool (Andrews, 2010). Following this, the automated damidseq_pipeline (Marshall and Brand, 2015) was implemented to process both datasets, and the accompanying find_peaks (Marshall and Brand, 2015) to call significant peaks. These identified areas of enrichment are putative Doublesex-DNA interactions. Enriched regions were mapped to genes. Downstream of this, we assessed the global similarity within and between replicates in both datasets, and conducted nuanced bioinformatic analyses including Gene Ontology (GO) and motif searching drawing biological interpretation of how DSX functions in the CNS (Figure 29).

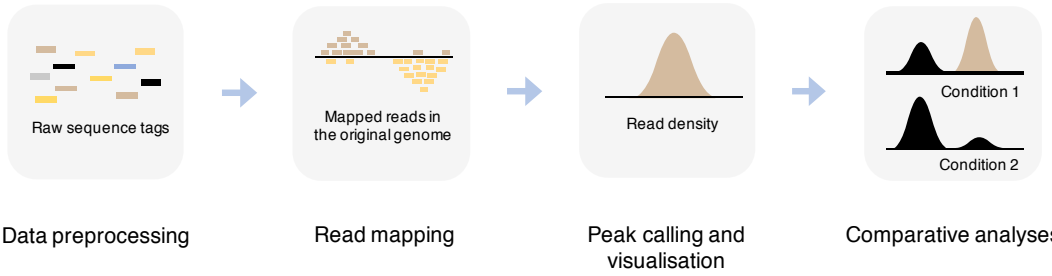


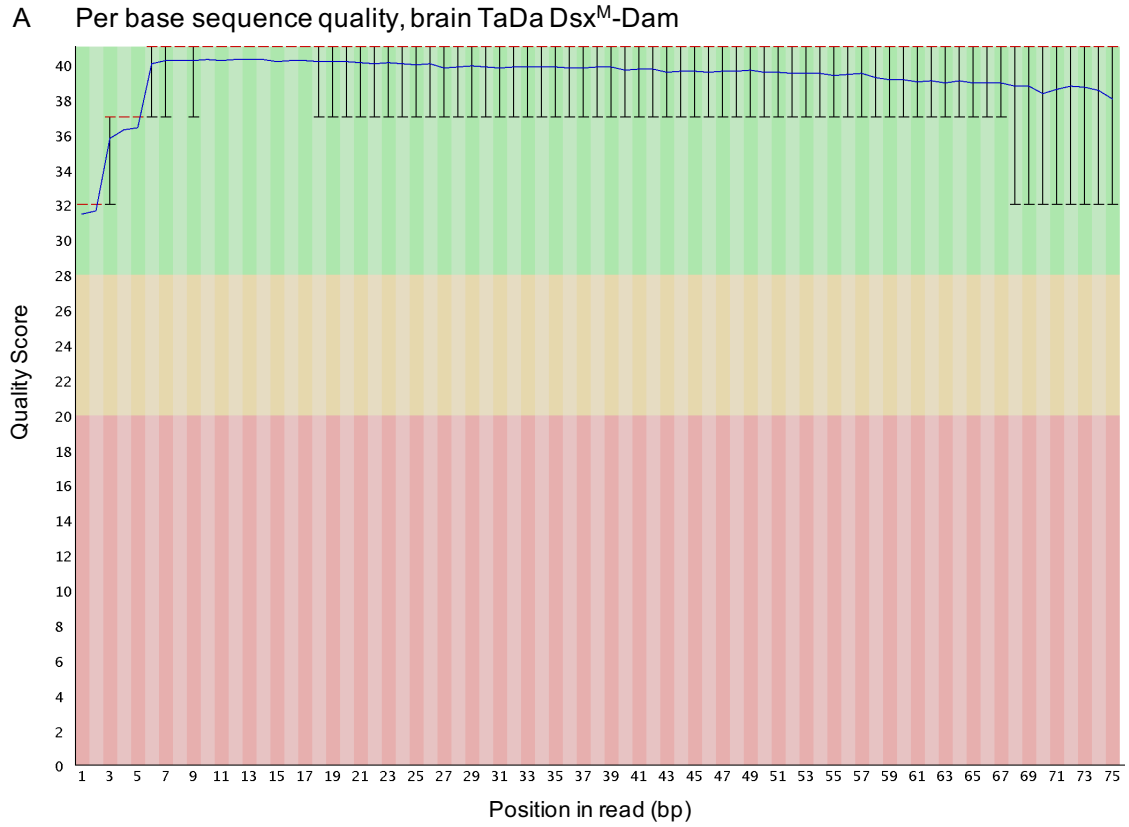
Figure 29 *Processing raw NGS datasets involves a number of steps. Our approach involves pre-processing the datasets to assess sequencing read quality of DamID amplicons, processing data using the automated damidseq_pipeline (Marshall and Brand, 2015), calling peaks and visualising using find_peaks (Marshall and Brand, 2015), and conducting thorough downstream comparative analyses.*

The aim of both brain and head TaDa was to identify putative Doublesex targets specific to the nervous system. Experimentally, the TaDa-sequencing experiment in brain TaDa profiles cell-specifically dissected brains driving Dsx^{M}-Dam or Dsx^{F}-Dam in *dsx* neural cells in adults; in head TaDa we profile Dsx^{M}-Dam or Dsx^{F}-Dam driven in *dsx* in the whole heads of adult animals. Functionally therefore we are asking slightly different questions across experimental rounds, whereby head TaDa also takes into account the profiling of *dsx*-expressing somatic and fat body tissue expressed in the adult head in addition to the brain (Jiang et al., 2005; Lazareva et al., 2007).

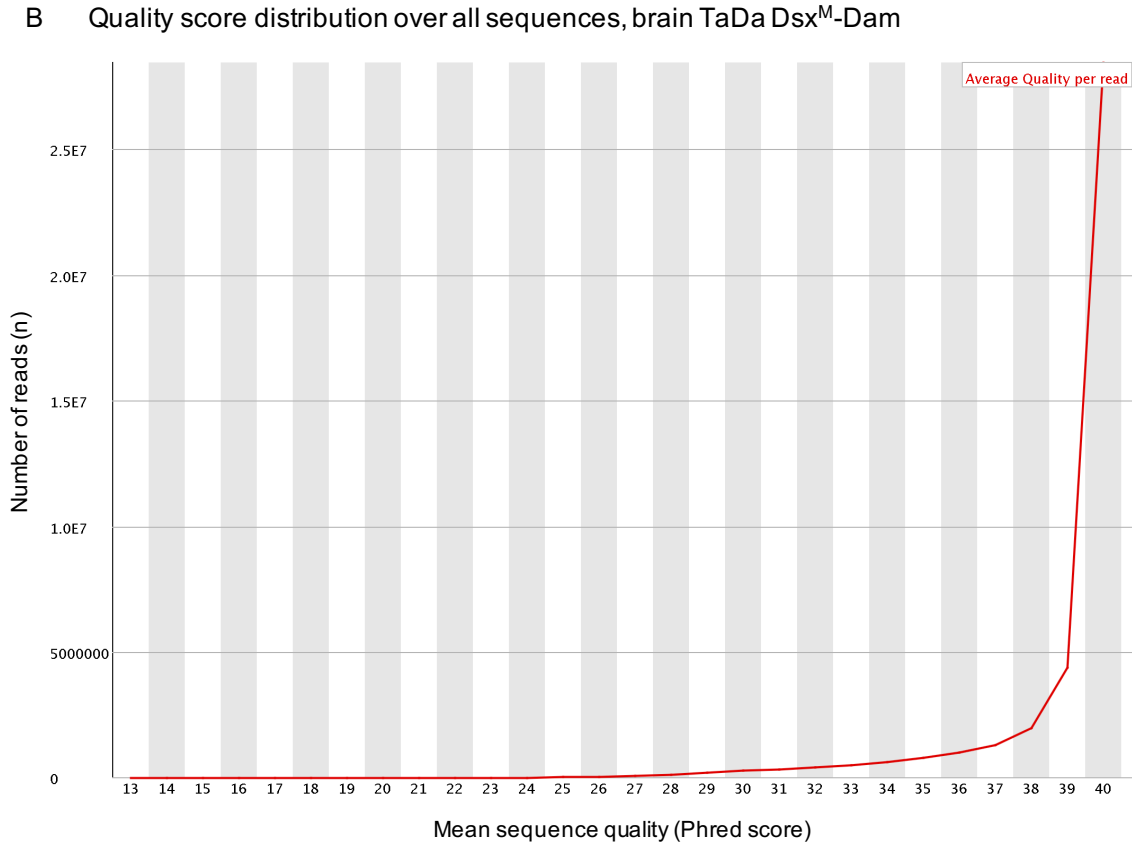
5.3.2 Assessing TaDa-seq datasets with FastQC (Andrews, 2010)

We used the computational tool FastQC version 0.11.5 (Andrews, 2010) to assess data quality (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The program completes a set of analyses designed to locate potential problems in high throughput sequencing datasets. We assessed the .fastq files from each replicate from brain and head TaDa. Read quality was visualised using charts which were subsequently aggregated and visualised using MultiQC version 0.9 (Ewels et al., 2016). We saw the

"per base sequence quality" had reliable base calls across all reads in each biological replicate in brain and head TaDa. The average sequence quality per read was high (Phred scores: 30-40), and the "sequence length distribution" confirmed the majority of sequencing reads were of similar length. Residual adaptor sequences and low quality bases were removed with cutadapt version 1.13 (Martin, 2011). Figure 30A-C shows FastQC plots for the above parameters (Andrews, 2010), for brain TaDa Dsx^{M}-Dam as an example. See Appendix for FastQC plots for brain TaDa Dsx^{F}-Dam, and head TaDa Dsx^{M}-Dam and Dsx^{F}-Dam.



(figure continued on next page)



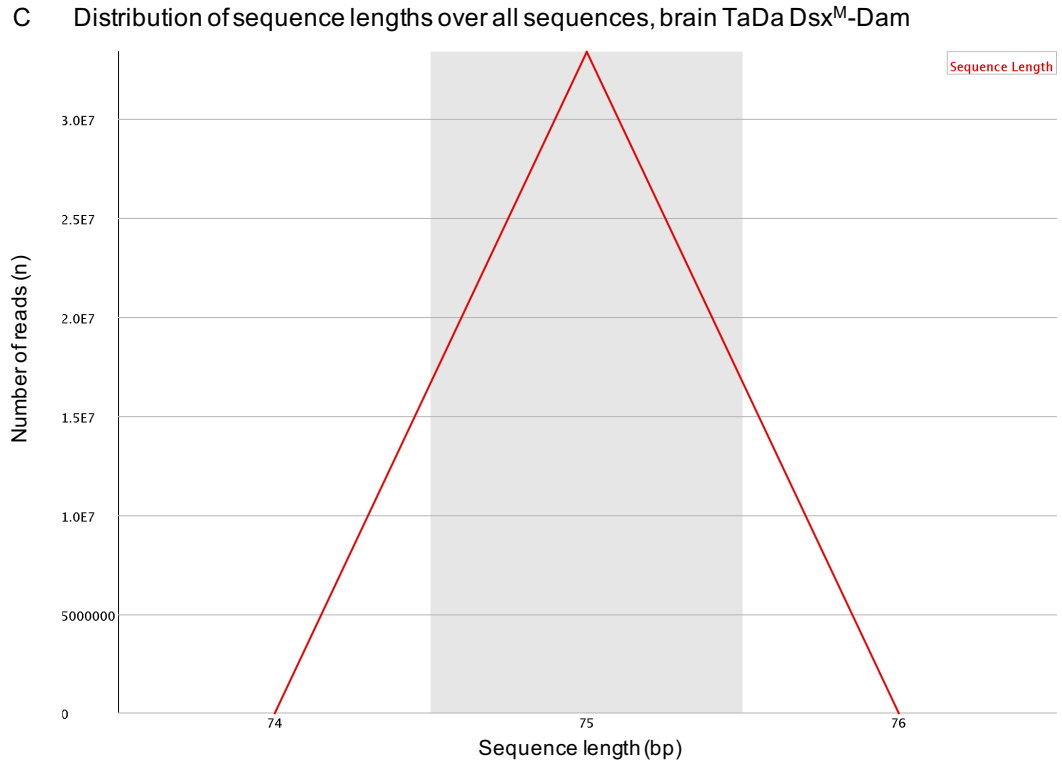


Figure 30 *Assessing Dsx^{M}-Dam brain TaDa read quality using FastQC (Andrews, 2010). Per base sequence quality (A) has reliable base calls across every base in 75 bp reads. The graph background divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of*

poor quality (red). The blue line represents the mean quality. The central line is the median value. The upper and lower whiskers represent the 10% and 90% points. The quality score distribution over all sequences (B) was high (Phred scores: 30-40). The distribution of sequence lengths over all sequences (C) confirmed the majority of all sequencing reads were of similar length.

5.3.3 PROCESSING TADA-SEQ DATASETS WITH DAMIDSEQ_PIPELINE (MARSHALL AND BRAND, 2015)

TaDa-seq identifies areas in the genome which are enriched for aligned reads, indicating a potential site of protein-DNA interaction. Here, TaDa-seq aims to identify putative DSX TFBS in the CNS of *D. melanogaster*. Processing DamID-seq data involves extending single-end reads, aligning the reads to the genome and determining the coverage. Analysing single-end read datasets can be difficult because the region which is sequenced is not directly over the interaction site but is rather one end of a fragment that covers the interaction site. Single-end reads were hence extended by a specified length so they more closely represent the sequenced region, reflecting more appropriately the relative size of DamID enriched regions. This made the datasets easier to interpret. The damidseq_pipeline (Marshall and Brand, 2015) handled sequence alignment, read extension, binned counts, normalisation, pseudocount addition and final ratio file generation as discussed above. The final ratio file is represented as a $\log_2$ ratio of Dam-*dsx*: Dam. Both the Dam-*dsx* and Dam-only controls are normalised, and pseudocounts added to mitigate the effect of background count.

damidseq_pipleine (v1.4, October 2018) was implemented via the Mac OSX Terminal and Perl v5.28.2. Installation of Bowtie 2 v2.3.3.1 and SAMtools v0.1.9 were required for damidseq_pipeline function. Bowtie 2 (Langmead et al., 2012) is a tool for aligning sequencing reads to long reference sequences. Bowtie 2 index files (.bt files) were

downloaded from Illumina's iGenome (*Drosophila melanogaster* Ensembl BDGP6). SAM (Sequence Alignment/Map) manipulates alignments, including sorting, merging, indexing and generating alignments in a per-position format. A prebuilt .gff GATC fragment file containing all GATC sites in the genome for *Drosophila melanogaster* r6 (Dmel BDGP6) was also available for download from the Marshall damidseq_pipeline GitHub page. We followed installation instructions available on the Marshall damidseq_pipeline GitHub page for damidseq_pipeline and associated files. The following command was used to implement the pipeline, location of Bowtie 2 index files and GATC fragment .gff files were specified:

damidseq_pipeline--gatc_frag_file=path/to/Dmel_r6.GATC.gff.gz--
bowtie2_genome_dir=path/to/dmel_r6/dmel_r6

The damidseq_pipeline was designed to process single-end sequencing data. In brain and head TaDa, we sequenced to a 75 bp paired-end resolution. We received one .bam file (sequencing reads pre-aligned to genome) and two .fastq files from OGC per sample referring to the forward and reverse ends of the sequencing orientation. To process the data to paired-end resolution the script could simply be implemented twice to account for both sequencing orientations given the pipeline conventionally handles sequencing at single-end resolution.

5.3.4 Peak Calling with find_peaks (Marshall and Brand, 2015)

In DamID-seq data, induced methylation signals are not certainly symmetrically distributed at TFBSs, meaning existing peak calling algorithms used for ChIP-seq such as MACS and MACS2 are inappropriate. We therefore conducted peak calling using

the novel find_peaks pipeline software designed specifically for DamID-seq datasets (Marshall and Brand, 2015). Peaks were called at two defined False Discovery Rates: 0.01 and 0.05. Table 13 summarises peak numbers for 'Dsx^{M}-Dam' (UAS-Dam-*dsx^{M}*/ *dsx^{Gal4}*) male and 'Dsx^{F}-Dam' (UAS-Dam-*dsx^{F}*/ *dsx^{Gal4}*) female replicates in brain and head TaDa at FDR 0.01 and FDR 0.05. Our analyses revealed that called peak numbers were very similar across both sequencing orientations for each biological replicate across brain and head TaDa (>95% similarity for each). See Appendix for peak numbers for other sequencing orientation. We hence took the downstream analysis forward using the reads from one orientation only for each biological replicate. We found good levels of reproducibility in terms of statistically significant similar numbers of peaks called amongst the Dsx^{M}-Dam male replicates 1 and 2 in brain TaDa (unpaired t-test, $p>0.05$), as well as both the Dsx^{F}-Dam female replicates 1 and 2 (unpaired t-test, $p>0.05$) and Dsx^{M}-Dam male replicates 1 and 2 in head TaDa (unpaired t-test, $p>0.05$). Overall, peak numbers were three- to four-fold higher in head TaDa, perhaps directly as a result of the higher number and biological/ functional diversity of cells profiled in this round.

Brain TaDa-seq				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	449	538	473	113
FDR 0.05	821	1255	1276	306
Head TaDa-seq				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	784	890	1203	1720
FDR 0.05	2104	2551	2067	1838

Table 13 *Summary of called peak numbers in Dsx^{M}-Dam male and Dsx^{F}-Dam female replicates in DSX-brain TaDa-seq (top) and DSX-head TaDa-seq (bottom). Peaks called at FDR 0.01 and 0.05.*

In addition to defining the number of peaks called across Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates in the brain and head datasets, we mapped the physical chromosomal location of each of the called peaks alongside their relative fold enrichment (Figure 31A-H). Overall, across all brain and head biological replicates, the spread of called peaks across both arms of chromosome 2 and 3 appears equal as well as the X chromosome. Significantly smaller numbers of peaks are called on chromosome 4, expected given the significantly smaller size of chromosome 4. These findings are in line with previous DSX DamID studies that have reported a similar chromosomal spread of called peaks (Clough et al., 2014; Li et al., 2015). This is positive as it suggests Dsx^{M}-Dam, globally, binds appropriate *dsx* target genes. Dsx^{M}-Dam brain biological replicates one and two peak locations appear very similar (Figure 31A and B). The sparsity of called peaks in Dsx^{F}-Dam brain biological replicate two is visualised in plot Figure 31D. Interestingly, despite statistically similar called peak numbers in the Dsx^{M}-Dam head biological replicates, the fold enrichment of biological

replicate one peaks appear far higher compared to replicate two (Figure 31E and F). This is in line with the observed higher read depth in biological replicate one compared to two (Table 14, page 201).

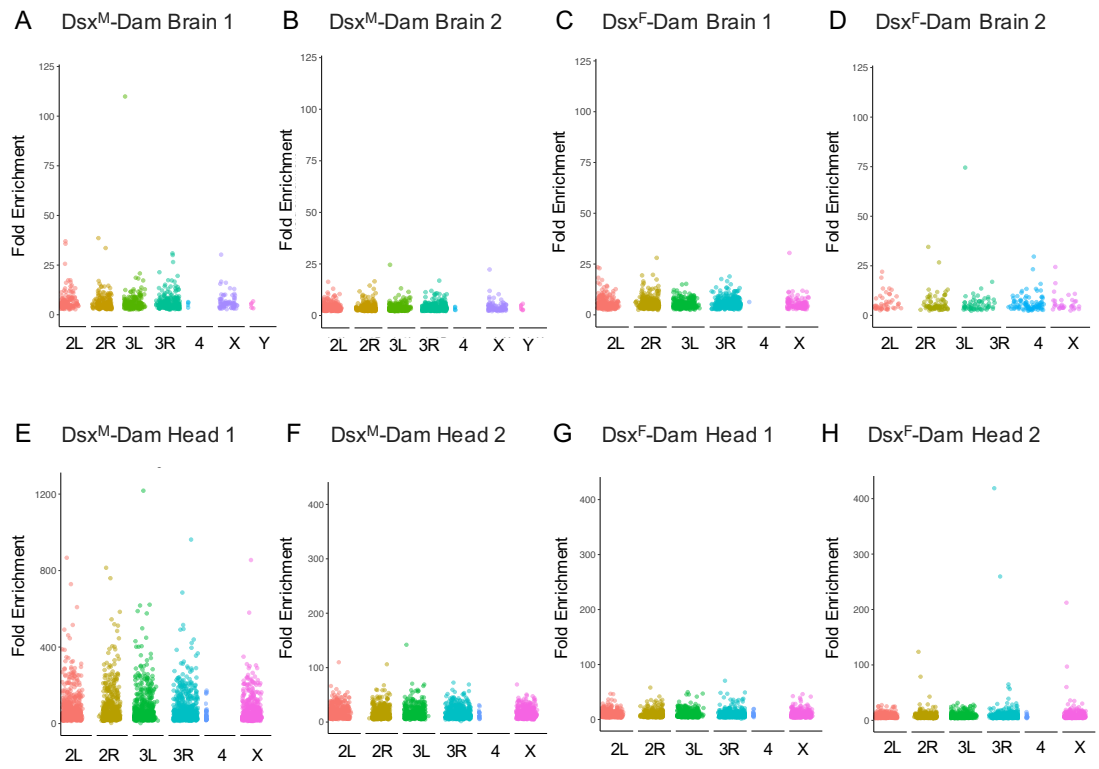


Figure 31 *Plots showing spread of fold enrichment of called peaks (find_peaks, FDR 0.05) and their chromosomal location in individual sample replicates. Dsx^{M}-Dam brain one (A) and two (B), Dsx^{F}-Dam brain one (C) and two (D), and Dsx^{M}-Dam head one (E) and two (F), Dsx^{F}-Dam head one (G) and two (H). Note fold enrichment axes differ between samples. Relative chromosomal distances mapped on x-axes. Plots generated in Bioconductor in R.*

We employed area-quantitative Venn diagrams as a means to visually inspect the overlap between called peaks in sex-specific biological replicates within and between DSX brain and head TaDa. We saw a statistically significant level of reproducibility in the overlap between the two Dsx^{M}-Dam biological replicates in brain TaDa (unpaired t-test, $p>0.05$; Figure 32A) and the same reproducibility between both Dsx^{M}-Dam

biological replicates in head TaDa (unpaired t-test, $p>0.05$; Figure 32C). However, in both TaDa Rounds there was substantial variation in called peak numbers within Dsx^F-Dam biological replicates (unpaired t-test, $p<0.05$ and unpaired t-test, $p<0.05$, Figure 32B and D). Furthermore, our comparisons of called peaks in individual Dsx^M-Dam biological replicates in brain and head TaDa (Figure 32E) and called peaks in individual Dsx^F-Dam biological replicates in brain and head TaDa (Figure 32F) exemplified the biological variability in our DSX brain and head TaDa experimental datasets. Indeed, whilst Dsx^M-Dam peaks show a tendency towards tissue-specificity, the variability within Dsx^F-Dam biological replicates means it is impossible to draw biological conclusions about the sex-specificity DSX binding.

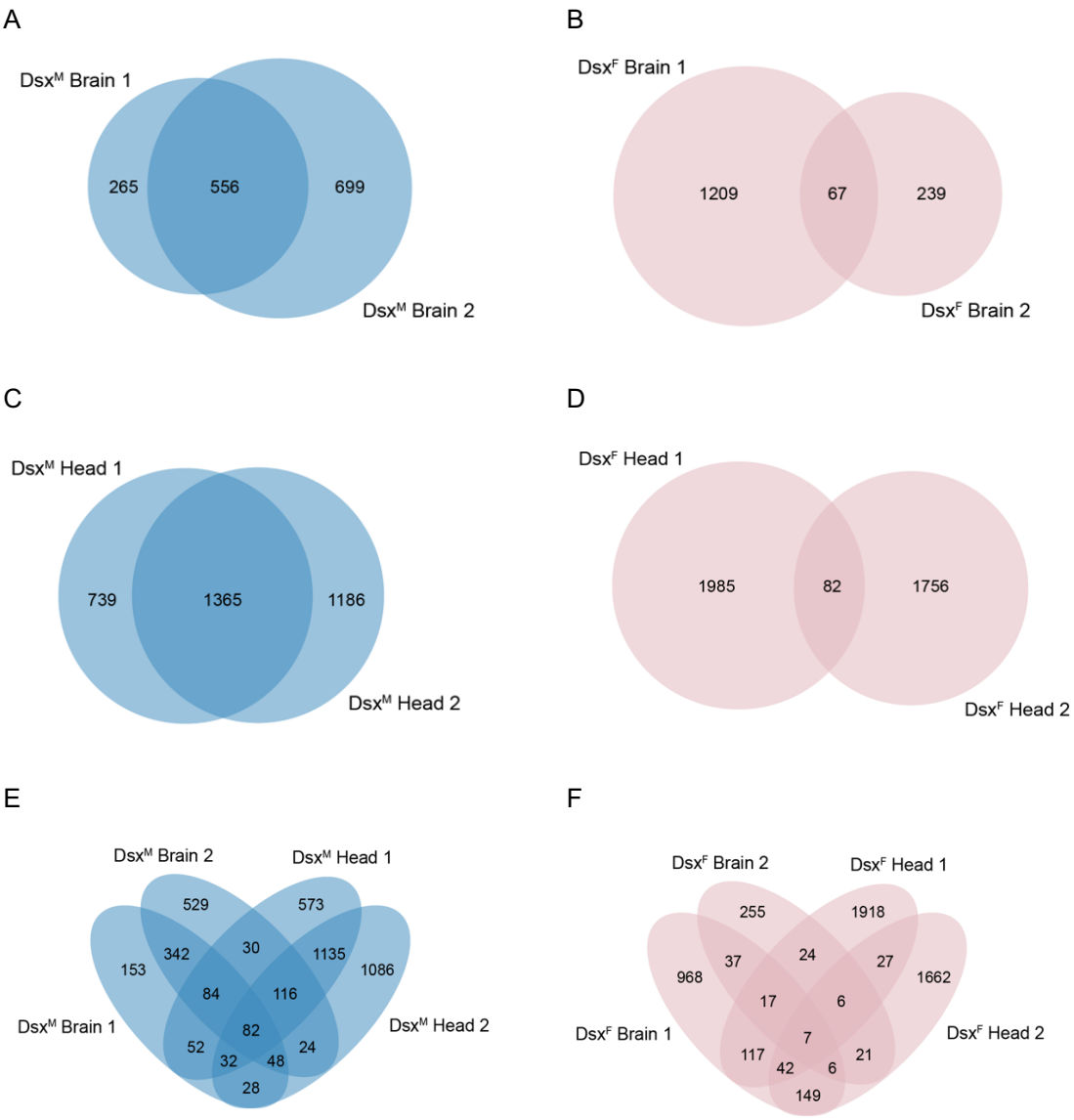


Figure 32 *Area-quantitative Venn diagram of overlapping called peaks in Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates within and between brain and head TaDa. Overlapping called peaks in brain TaDa Dsx^{M}-Dam biological replicate one and two (A), brain TaDa Dsx^{F}-Dam biological replicate one and two (B), head TaDa Dsx^{M}-Dam biological replicate one and two (C) head TaDa Dsx^{F}-Dam biological replicate one and two (D), brain TaDa Dsx^{M}-Dam replicate one and two, and head TaDa Dsx^{M}-Dam replicate one and two (E), and brain TaDa Dsx^{F}-Dam replicate one and two, and head TaDa Dsx^{F}-Dam replicate one and two (F). Peaks called to FDR 0.05 (find_peaks, Marshall and Brand, 2015). Two-way Venn diagrams generated in BioVenn (Hulsen et al., 2008). Four-way Venn diagrams generated using Venny 2.1.0.*

5.3.5 Overview of downstream bioinformatics analyses

Figure 33 describes an overview of the bioinformatics analyses we employed to assess our TaDa-seq data profiling DSX neurons in the CNS of *Drosophila*. Given the variability we noted in our Dsx^{F}-Dam datasets in both brain and head TaDa, we separated our downstream analyses by sex following damidseq_pipeline processing, peak calling and peak-to-gene assignment. For our Dsx^{M}-Dam datasets in both brain and head TaDa we split our analyses into functional and sequence categories and, as discussed in the following chapter, compare the dataset to alternative tissue. More specifically, we looked for relationships in gene lists attempting to group genes more broadly according to function (Gene Ontology assays). Various MOTIF analyses looked to identify significant sequences of biological importance in the enriched regions, and analyses to compare these to known DSX binding sequences derived from various experimental methods (Erdman et al., 1996; Oliphant et al., 1989; Yi and Zarkower, 1999). These analyses help to winnow candidate gene lists, and alongside independent screens discussed in the following chapter, generate a sensible number of putative targets representing 'low-hanging fruit'. For our Dsx^{F}-Dam datasets in both brain and head TaDa we take a more nuanced approach to tease apart the dataset and attempt to explain the variability we see.

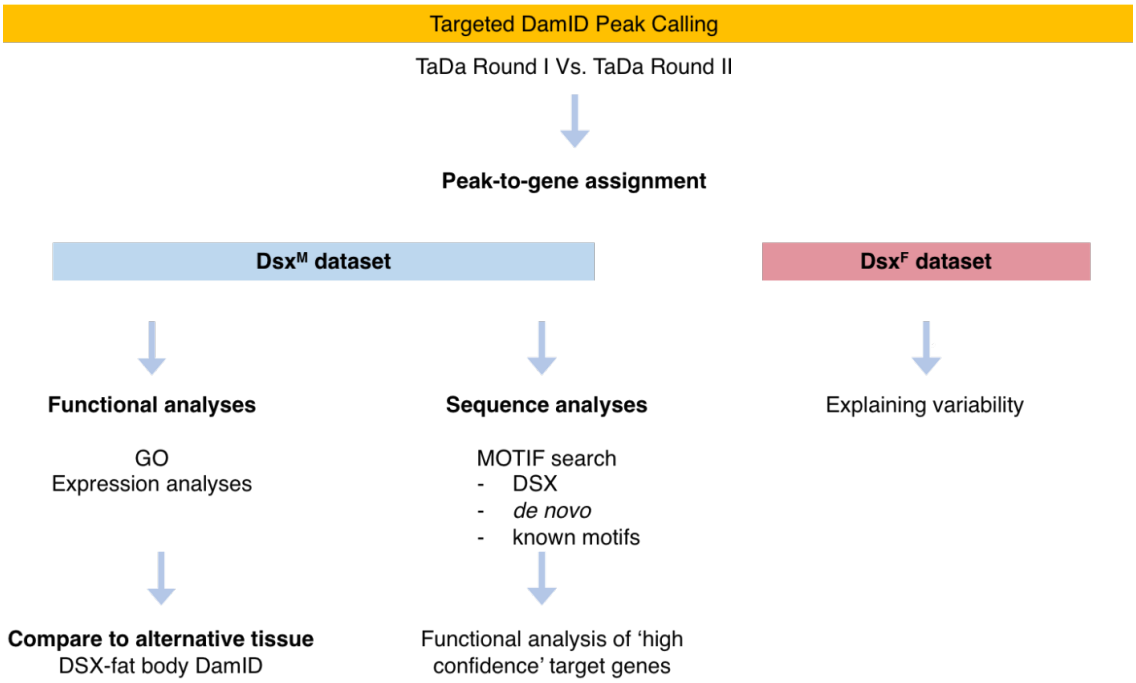


Figure 33 *Overview of downstream bioinformatics analyses in DSX brain and head TaDa. Given the variability within the Dsx^{F}-Dam biological replicates in both brain and head TaDa, we assess the datasets sex-specifically. For Dsx^{M}-Dam we take two approaches: functional analyses, and sequence analyses. For Dsx^{F}-Dam we try a number of computational techniques to assess the variability across biological replicates.*

5.3.6 Generating genome-scale candidate gene lists

Following peak detection and calling at various FDR thresholds across biological replicates, we aimed to associate identified called peaks with *Drosophila* genes. This is a challenging process, and previous *Drosophila* DamID papers differ in their approaches to this task (for example, Clough et al., 2014; Li et al., 2015; Neville et al., 2014). Here, we trialled and compared two methods of gene annotation. Firstly, we assigned peaks to genes by using a 2 Kb window centred on the annotated Transcription Start Site (TSS). This fixed-range method limits artificial contributions of nearby upstream genes and uncouples gene length from occupancy, however misses binding at

many intronic enhancers. The second method assigned DSX peaks to genes occurring in the gene body + 1 Kb upstream of TSS and + 1 Kb downstream of Transcription Termination Site (TTS). Importantly, this method takes into consideration the variation in gene body lengths and captures intronic enhancers, however it biases towards longer genes (Figure 34, Method B). The known DSX binding target gene *bab1* was annotated in both approaches, and both approaches were used for peak to gene annotation in similar DSX DamID experiments completed in the fat body tissue (Clough et al., 2014). We hence employed the gene body + 1 Kb upstream of TSS + 1 Kb downstream of TTS peak to gene annotation method in this study across biological replicates in brain and head TaDa.

We use PAVIS, a tool for Peak Annotation and Visualisation (Huang et al., 2013) for annotating and visualising DamID-seq data, version 02-05-2018. For the gene annotation of query peaks, PAVIS identifies the closest gene to each peak and its relative location: upstream of the gene TSS, intron, exon, 5'/3'- untranslated region, or downstream of TTS. When there are multiple genes nearby for a peak, PAVIS associates the peak with genes whose TSS is closest to the peak if the peak is within the gene region. If the peak is not within the gene region, then it is associated with the genes closest to the TSS, unless the peak is closer to the TTS of another gene. The called peaks .bed output from find_peaks (Marshall and Brand, 2015) at FDR 0.05 were utilised as input files for PAVIS, and generated gene lists with the parameters described above.

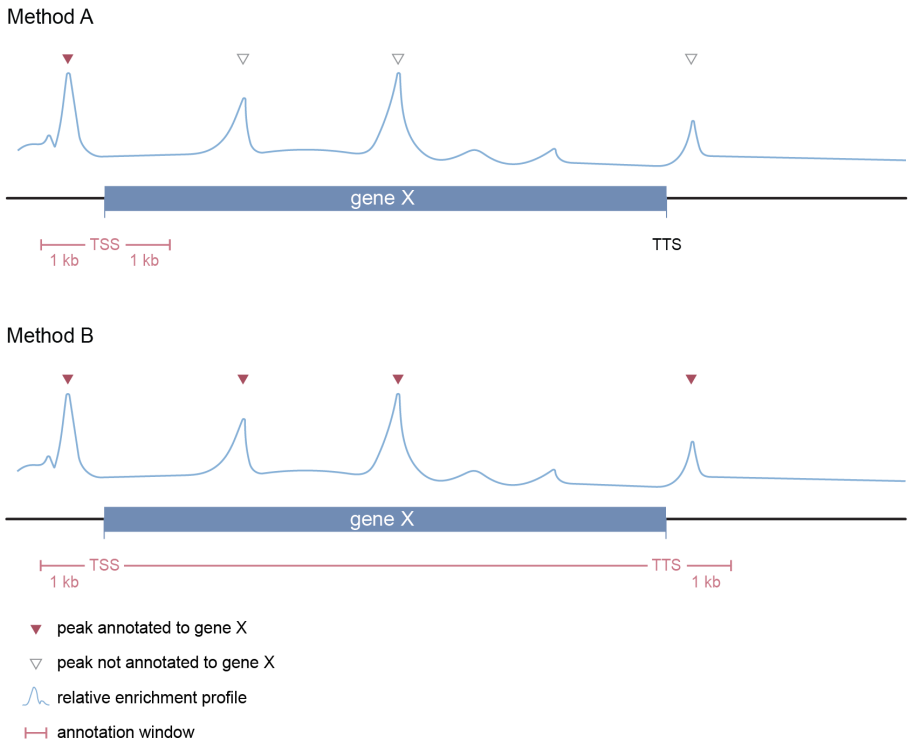


Figure 34 *Methods to associate peaks to genes vary. In this study we trialled two approaches: assigning peaks to genes using a 2 Kb window centred on the annotated TSS (Method A), or assigning peaks to genes occurring in the gene body + 1 Kb upstream of the TSS and + 1 Kb downstream of the TTS (Method B).*

Peak to gene annotation is the first step to adding biological meaning to peaks. Once this is completed, as a first step to assess the global binding characteristics of the TF in question, we delineate whether genes annotated from certain genomic categories are over-represented. To do this, we implement PAVIS (Huang et al., 2013) to sort called genes into annotations from peaks in intronic, exonic, 5' UTR, 3' UTR and upstream and downstream regions as denoted by our gene annotation parameters (Method B, Figure 33).

In brain TaDa, we selectively associated peaks to genes common to both Dsx^{M}-Dam biological replicates, given their statistically significant overlap (Figure 32A). In the

overlapping peaks across both biological replicates, we found that 83% of the loci were associated with genes, with two thirds of the peaks associated with intronic or exonic gene regions. We observed very similar findings in each biological replicate separately (Figure 35A). Given the variation in the Dsx^{F}-Dam biological replicates, we generated separate gene lists from both replicates. In replicate one, 90% of the loci were associated with genes, and over half of all peaks associated with exonic regions as compared to the third in Dsx^{M}-Dam (Figure 35A). In replicate two, 92% of the loci were associated with genes, and as with Dsx^{M}-Dam around two thirds of the peaks associated with intronic or exonic gene regions (Figure 35A).

In head TaDa, again given the statistically significant overlap between the Dsx^{M}-Dam replicates one and two, we looked specifically at the peaks common to both replicates, only associating those said peaks to genes. In the overlapping peaks across both biological replicates, we found that 92% of the loci were associated with genes, and significantly almost two thirds with intronic regions, and just ~5% with exonic regions. As before, we observed similar findings in each biological replicate separately (Figure 35A). We generated separate gene lists for head TaDa Dsx^{F}-Dam biological replicates given their variation. In replicate one, 83% of the loci were associated with genes, and 88% of the loci were associated with genes in replicate two. We noted some variation in these replicates whereby in the first replicate there was a preference towards binding intronic regions (~42%), and in the second a preference towards exonic regions (~47%) (Figure 35A).

The spread of called peak lengths across the biological replicates are described in Figure 35B. The variation we observe in gene occupancy preferences across biological replicates in brain and head TaDa are aligned with the differences we observe in called peak lengths. Indeed, for head TaDa Dsx^{M}-Dam, not only is the spread of the called peak lengths far higher than any other replicate, both the median (~1,750 bp) and Interquartile Range (IQR) of the data (~1,100-2,700 bp) are higher than the other replicates. This could describe the skew we note for the head TaDa Dsx^{M}-Dam replicate's preference towards intronic regions. Moreover, biological replicate one of Dsx^{F}-Dam in both brain and head TaDa which show similar gene occupancy preferences, also show similar peak length spreads (median ~600 bp in both, and IQR ~500-800 bp and ~500-900 bp consecutively). Peak lengths resembled 'narrow' and 'mixed mode' peaks across all biological replicates (see Figure 28A), with median peak length between ~500-600 bp in three biological replicates, and ~1,000 in two others, plus ~1,750 bp in the head TaDa Dsx^{M}-Dam replicate. These peak lengths are in line with DamID TF peak length binding characteristics. Across both brain and head TaDa we saw consistent rates in peak to gene associations (Figure 35C). All occupancy analyses completed in PAVIS (Huang et al., 2013).

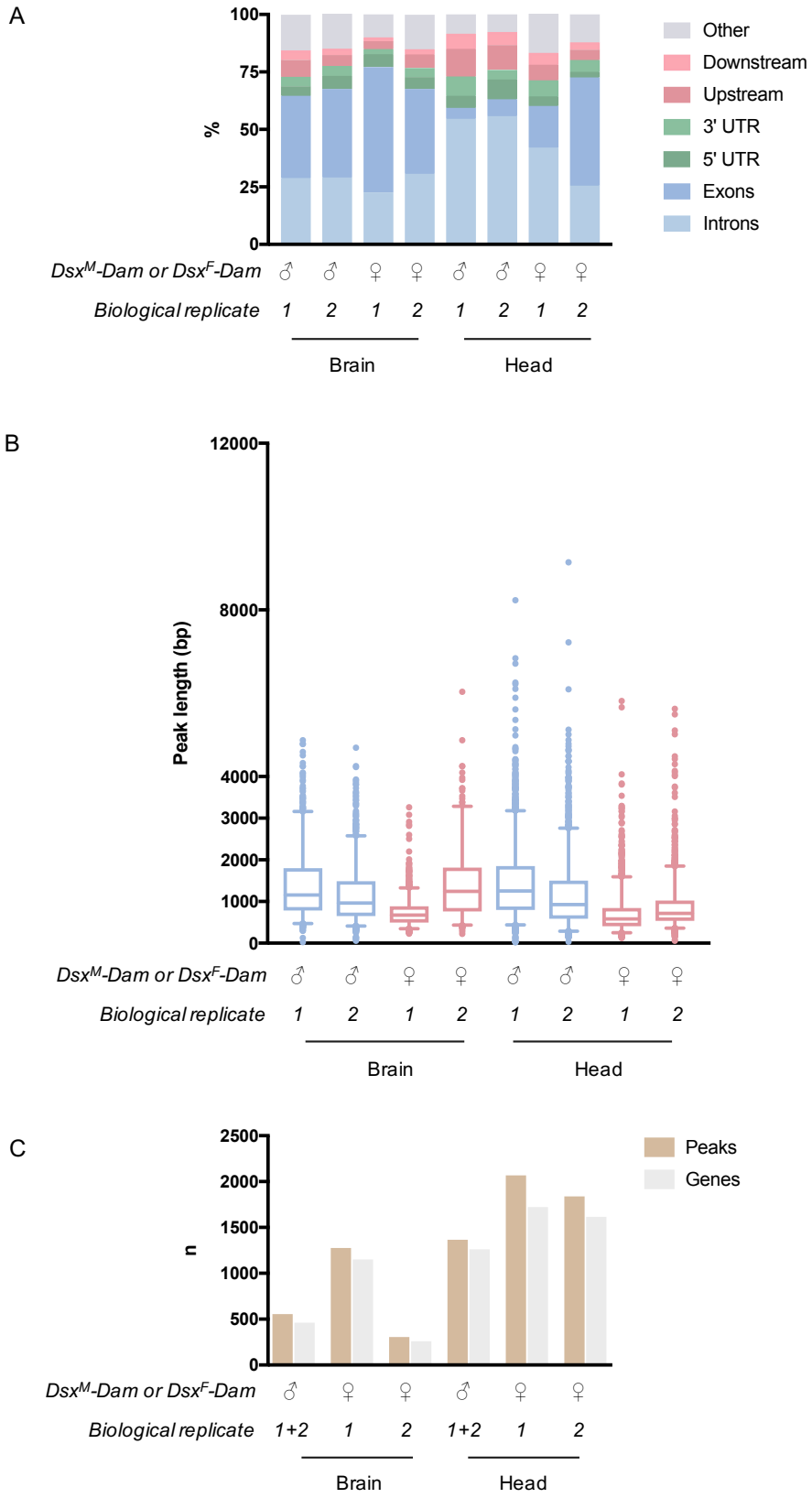


Figure 35 *Comparing gene occupancy preferences in Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates in brain and head TaDa (A). Occupancy analyses for individual biological replicates for both Dsx^{M}-Dam and Dsx^{F}-Dam in brain and head TaDa are shown. In all samples, occupancy was significantly overrepresented in exonic and intronic regions as compared to 5'/ 3' UTR or other upstream or*

downstream elements. Very similar occupancy is noted between Dsx^{M}-Dam biological replicate one and two in both brain and head TaDa. In brain TaDa, Dsx^{F}-Dam replicate one occupancy was significantly overrepresented in exonic regions, where over half of all peaks were associated to exonic gene regions. In replicate two, occupancy was overrepresented in exonic and intronic regions in similar proportions to both Dsx^{M}-Dam replicates. In head TaDa, Dsx^{M}-Dam occupancy was significantly overrepresented in intronic regions, over half of bound genes in both replicates one and two. For Dsx^{F}-Dam, we saw variation in occupancy across biological replicates: in replicate one there is bias towards intronic regions and in replicate two, bias towards exonic regions. Analysis completed in PAVIS (Huang et al., 2013). Peaks called at FDR 0.05. Peaks to genes assigned to genes occurring in the gene body + 1 Kb upstream of TSS and + 1 Kb downstream of the TTS (Method B).

Box plots compare called peak lengths (bp) in Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates in brain and head TaDa (B). Box represents the Interquartile Range, middle line is the median value, and the whiskers represent the 5 to 95 percentiles of the data.

Peak to gene annotations in brain and head TaDa across Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates (C). The overlap of Dsx^{M}-Dam peaks/ genes are shown given these are specifically taken forward for further analysis given their biological reproducibility. The difference between number of loci and gene associations were statistically similar across all biological replicates (one-way ANOVA, $p<0.05$).

5.3.7 DSX^{M} TARGET GENES HAVE A TENDENCY TOWARDS TISSUE-SPECIFICITY

Analysis of the generated gene lists across Dsx^{M}-Dam biological replicates in both brain and head TaDa showed a high degree of similarity. As previously described, we saw a three- to four- fold higher number of peaks called in head TaDa compared to brain TaDa. We found this difference in called peak numbers was proportionally transferred to gene numbers. Comparing the common peaks/ genes within both Dsx^{M}-Dam biological replicates across brain and head TaDa, we saw a small overlap (5%) in gene lists. Theoretically, this subtractive comparison delineates the difference in profiling DSX-neural cells with DSX-neural and DSX-fat body cells (Figure 36A). As we expected, we identified DSX targets showed a tendency towards tissue specificity. A number of targets identified in the head TaDa dataset were fat body-specific, such as *Desat1*. Figure 36B compares the separate Dsx^{F}-Dam biological replicates across both brain and head TaDa, highlighting the variation in biological replicates between all four biological replicates. Given this variability, it is unfortunately impossible to directly

compare Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates to draw biological conclusions about sex-specificity of annotated DSX target genes.

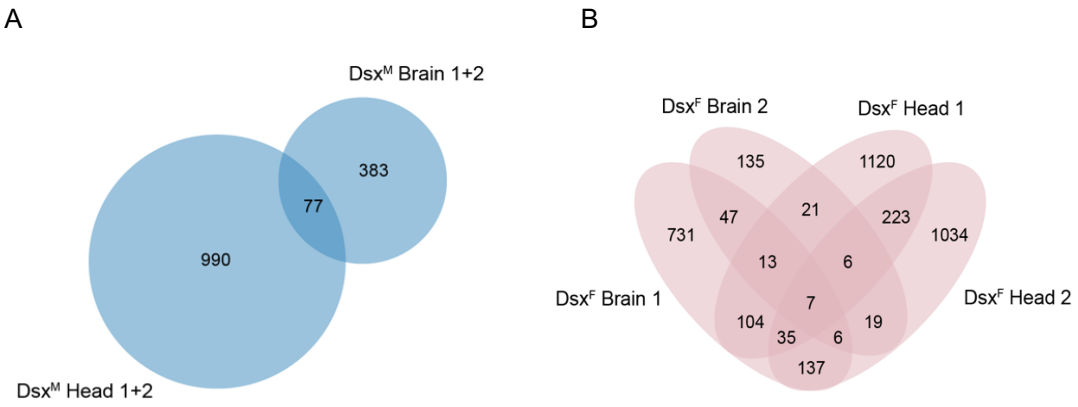
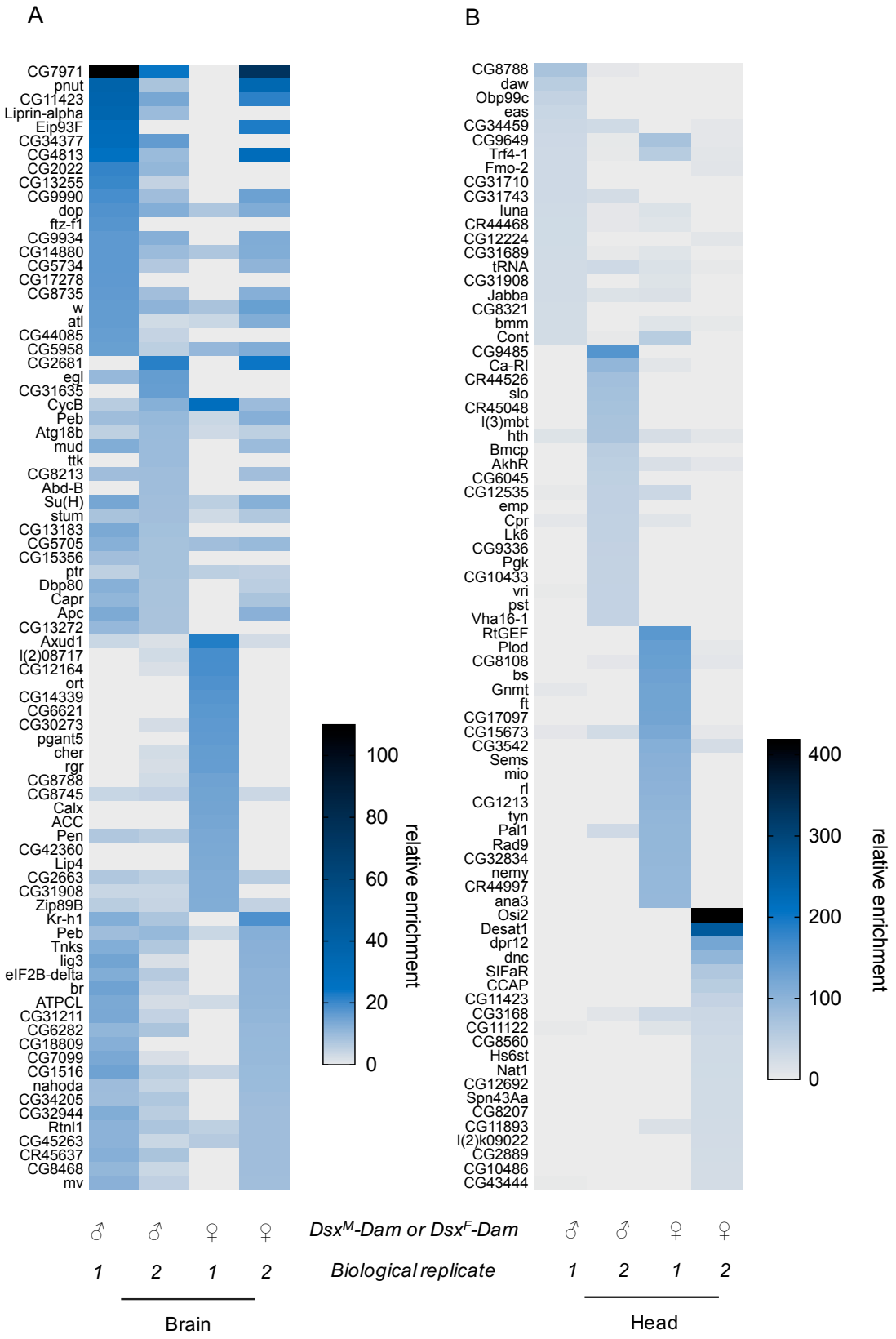


Figure 36 *Assessing generated gene lists for Dsx^{M}-Dam and Dsx^{F}-Dam brain and head TaDa. Overlapping gene associations in brain TaDa Dsx^{M}-Dam merged replicates and head TaDa Dsx^{M}-Dam merged replicates (A), and Dsx^{F}-Dam replicates one and two in both brain and head TaDa (D). Two-way area-quantitative Venn diagrams generated in BioVenn (Hulsen et al., 2008). Four-way Venn diagram generated using Venny 2.1.0.*

So far, we processed DSX brain and head TaDa-seq male and female samples using the damidseq_pipeline (Marshall and Brand, 2015), called peaks, and annotated peaks to genes. Through these means, we were able to ascertain the reproducibility between biological replicates, and compare biologically reproducible Dsx^{M}-Dam global binding architecture in brain and head samples. We next compared enrichment scores for each gene annotation across Dsx^{M}-Dam and Dsx^{F}-Dam brain and head samples. Shown in Figure 37A and B are the ranked top twenty most highly enriched genes in each Dsx^{M}-Dam and Dsx^{F}-Dam biological replicate in the brain and head datasets individually. For ranked genes appearing within the top twenty in multiple replicates, enrichment scores are shown alongside the first biological replicate the gene was located, and additional genes included up to twenty. Brain TaDa Dsx^{M}-Dam biological replicates show a high

level of cogency in genes with the highest enrichment scores. In head TaDa, whilst Dsx^{M}-Dam biological replicates bind a largely overlapping set of genes, enrichment scores of these individual genes vary (Figure 37B). Figure 37C compares the most highly enriched genes in the Dsx^{M}-Dam brain and head biological replicates. Here, we observed just a small number of genes (~20%) common to both datasets within the top twenty most highly enriched genes. Given the head dataset also profiles DSX-expressing fat body tissue surrounding the brain, the finding suggests Dsx^{M}-Dam has a tendency towards binding tissue-specifically. Notably, Dsx^{M}-Dam TaDa head binding target genes had significantly higher enrichment scores compared to Dsx^{M}-Dam TaDa brain samples.



(figure continued on next page)

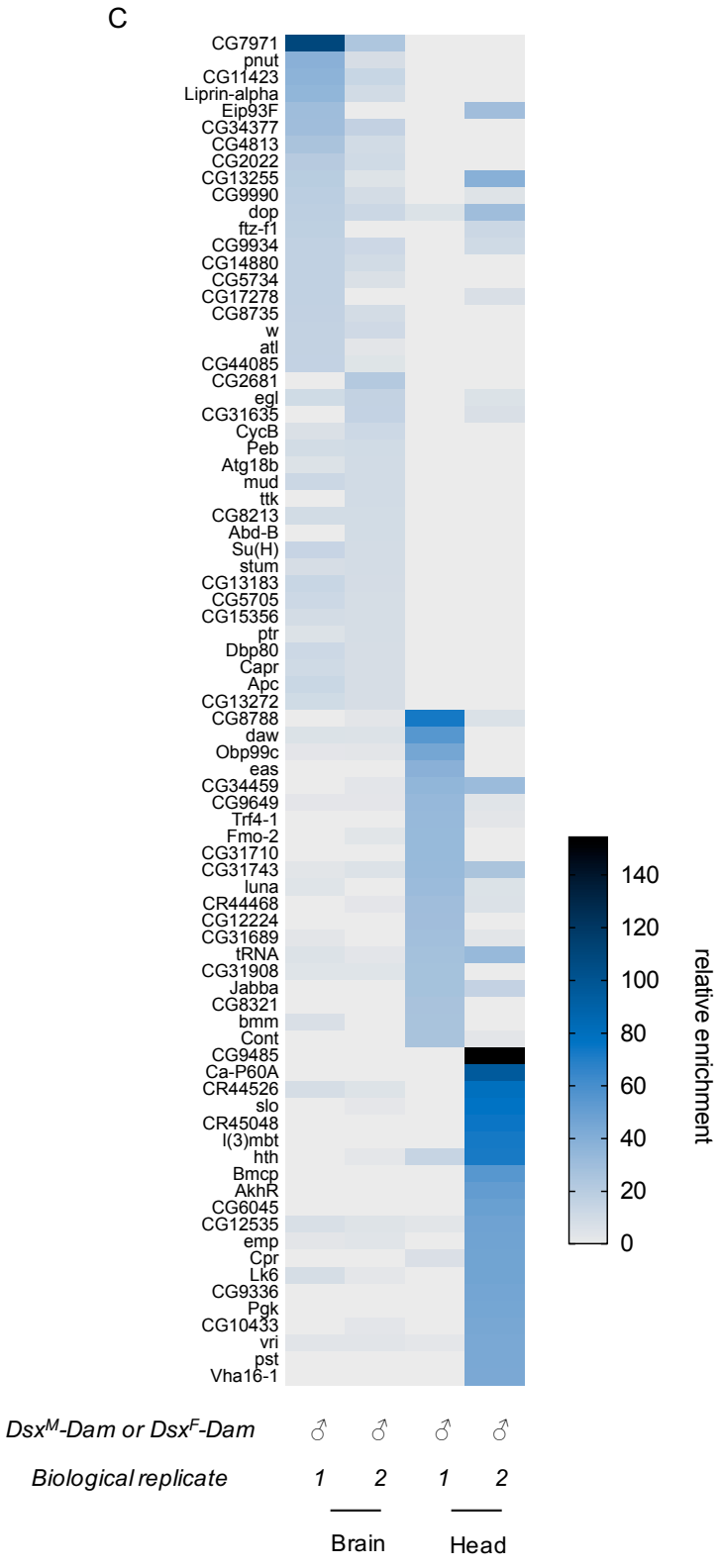


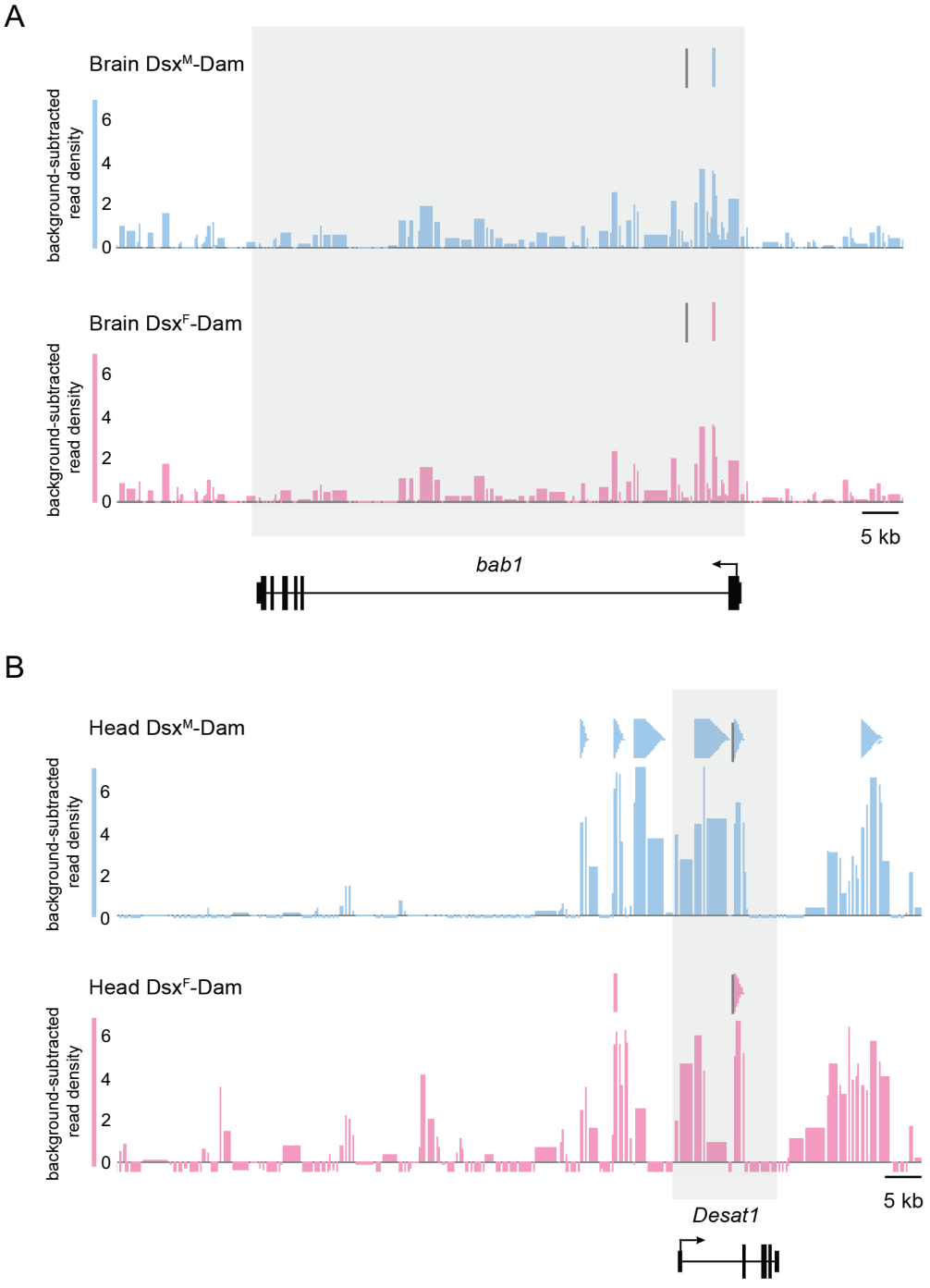
Figure 37 *Heat maps showing the top 20 enriched genes across Dsx^{M}-Dam and Dsx^{F}-Dam brain (A) and head (B) datasets, and directly comparing the top 40 enriched genes in Dsx^{M}-Dam brain and head samples (C). Genes are ranked according to fold relative enrichment, note scale difference in brain and head datasets in A and B. Overall gene enrichment in head datasets appears significantly higher than in the brain. The heat maps suggest Dsx^{M}-Dam and Dsx^{F}-Dam bind largely dissimilar genes in both the brain and head. Peaks/ genes common to both Dsx^{M}-Dam biological replicates were assessed in the brain*

and head datasets. For Dsx^F-Dam, both biological replicates were assessed separately given their variability in both rounds. Heat maps made in PRISM v7.0d.

5.3.8 Binding of known DSX targets in Brain and Head TaDa

Next, we add an alternative means to assess the integrity and quality of the TaDa-seq datasets by identifying specific binding events of known DSX target genes. Three true direct targets of *dsx* have been identified to date: *bric a brac 1* (*bab1*) (Williams et al., 2008), *fatty acid desaturase 2* (*Fad2*) (or *desaturaseF* (*desatF*)) (Shirangi et al., 2009) and *Yolk protein 1 (Yp 1) and 2* (*Yp 2*) (Burtis et al., 1991; Coschigano et al., 1993; Hutson and Bownes, 2003). Most other targets have been found to be indirectly regulated by *dsx* (Arbeitman et al., 2004; Chapman and Wolfner, 1988; Wolfner, 1988). We therefore began by looking for binding events of these targets in our DSX brain and head TaDa-seq datasets. We saw a level of variation in these results as previously reported in similar published datasets (Clough et al., 2014; Li et al., 2015). Namely, only some Dsx^M-Dam and Dsx^F-Dam replicates both brain and head TaDa exhibited binding to these loci. The *bab1* gene was called in both Dsx^M-Dam biological replicates, as well as Dsx^F-Dam replicate two in brain TaDa (Figure 38A). *bab1* did not appear in any head TaDa biological replicate. The Fatty acid desaturase (*Desat1*) gene is known to modulate courtship behaviour in *D. melanogaster* through reducing the production of courtship-stimulatory pheromones (Bousquet et al., 2012; Grillet et al., 2006; Ueyama et al., 2005). *Desat1* peaks were identified in both Dsx^M-Dam biological replicates as well as both Dsx^F-Dam biological replicates in head TaDa, and one Dsx^M-Dam biological replicate from brain TaDa (Figure 38C). We further searched for *fru* enrichment, expected given its pivotal role alongside *dsx* in the SDH. We located peaks in both Dsx^M-Dam biological replicates and both Dsx^F-Dam biological replicates from

brain TaDa, as well as both Dsx^{M}-Dam biological replicates and one Dsx^{F}-Dam biological replicate from head TaDa (Figure 38B).



(figure continued on next page)

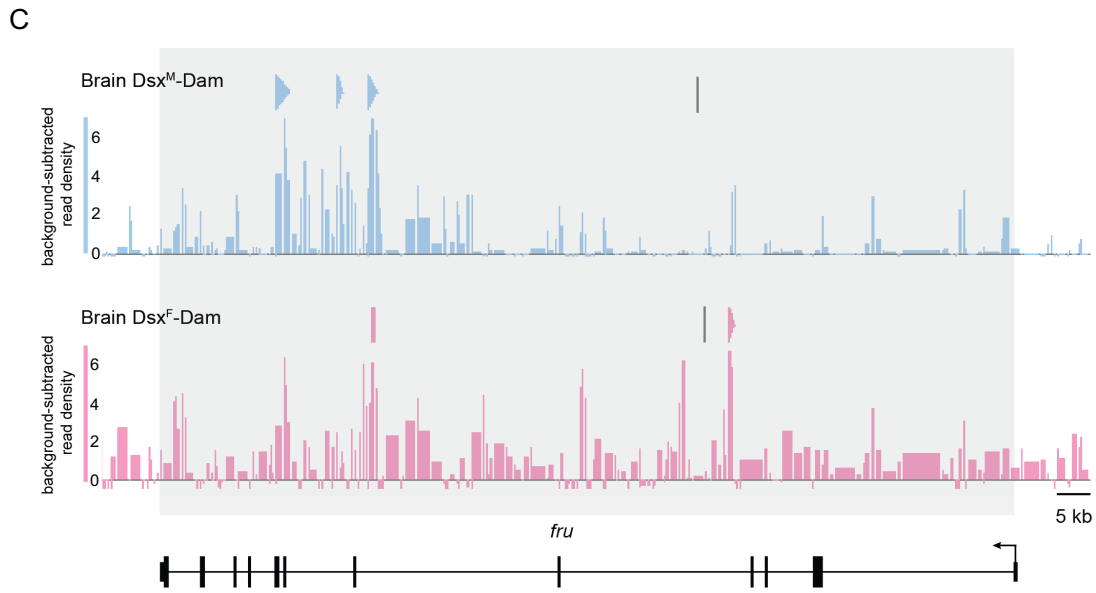


Figure 38 *Enrichment binding profiles of known DSX binding target genes bab1 (A), desat1 (B), and fellow sex-determination gene fru (C). Binding of bab1 is seen in both Dsx^{M}-Dam replicates and one Dsx^{F}-Dam replicate in brain TaDa. Binding of desat1 is seen in both Dsx^{M}-Dam and both Dsx^{F}-Dam biological replicates from head TaDa and one Dsx^{M}-Dam biological replicate from brain TaDa. Binding of fru is seen in both Dsx^{M}-Dam and both Dsx^{F}-Dam biological replicates from brain TaDa, and both Dsx^{M}-Dam and one Dsx^{F}-Dam biological replicate from head TaDa. Peaks called at known dsx binding genes at FDR 0.05. Grey boxes on called peaks binding track show locations of known DSX binding motifs. Scaled read-density plots are background subtracted, arbitrary scale. FlyBase gene models showing coding exons (thick rectangles), noncoding regions (thin rectangles), and introns (lines). Enrichment tracks visualised in Integrated Genome Browser v9.0.0.*

5.3.9 Assessing gene function using Gene Ontology (GO) analyses

NGS studies have repeatedly shown that a large proportion of genes specifying core biological functions are shared by all eukaryotes. Using knowledge of the biological function of a given protein in one organism, one can infer this knowledge to another organism. The Gene Ontology (GO) consortium (Ashburner et al., 2000) generates a dynamic vocabulary to define concepts used to describe biological function of gene products in all eukaryotes, classifying these and delineating the relationships between these concepts. This is an active vocabularic database that develops as knowledge of

gene and protein roles in eukaryotic cells is changing and accumulating. GO classifies functions along three ontologies: *biological process*, *molecular function* and *cellular component* domains (Ashburner et al., 2000). The *biological process* refers to the pathways and larger processes made up of the activities of multiple gene products, molecular events key to the function of cells, tissues, organs or organisms. The *molecular function* refers to elemental activities of a gene product at the molecular level, for instance binding. The *cellular component* refers to parts of a cell or its extracellular environment where gene products are active. In this study, we associate the called peaks to genes, and conduct GO analyses on sets of genes to ascertain putative function (Figure 39). GO analyses were completed on specific gene lists uploaded to FlyMine v45.1 (Lyne et al., 2007) and PANTHER 14.1 (Mi et al., 2019). With DamID-seq genome-wide binding experiments, binding intervals are more likely to be associated with longer genes than shorter ones, hence normalisation by gene length. The Holm-Bonferroni correction was applied, maximum p-value <0.05.

We completed enrichment analyses using GO for Dsx^{M}-Dam in both brain and head TaDa. For both rounds the enrichment analysis was performed using the peaks/ genes common to both replicates. Given the biological variation we see in Dsx^{F}-Dam replicates within and between both rounds, an enrichment analysis of this nature would not be biologically nor statistically meaningful. Alternative downstream analyses are discussed later in this chapter.

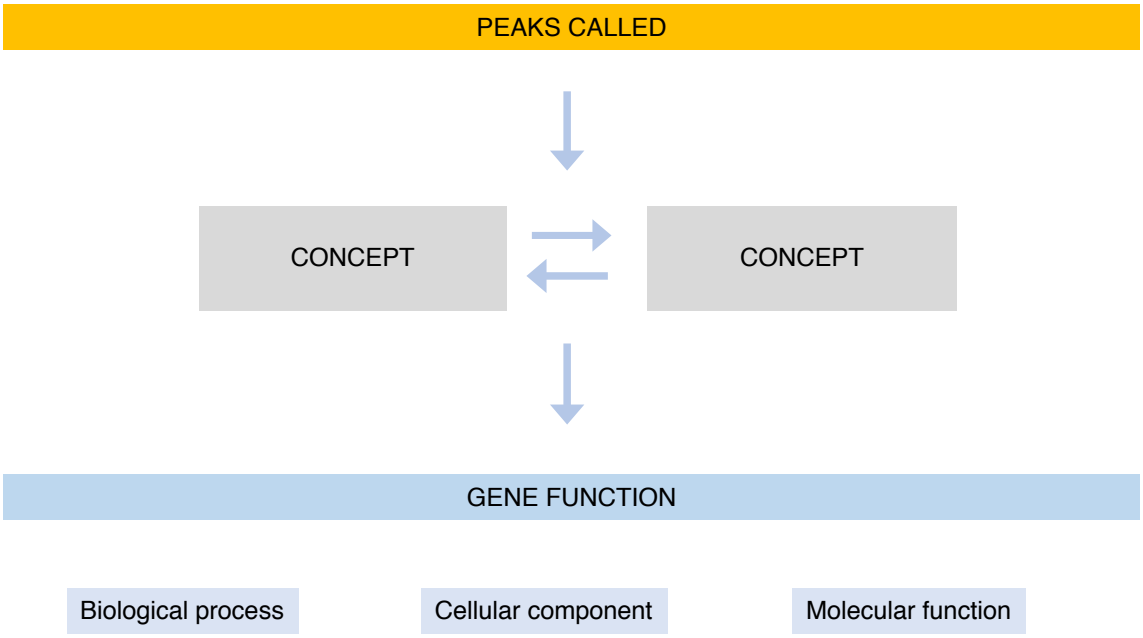


Figure 39 *Schematic of Gene Ontology analyses. In the downstream analysis of the TaDa profiling of DSX neurons in the Drosophila CNS, called peaks are associated to genes which are then assessed for biological function. Genes are grouped according to biological process, cellular component, and molecular function.*

In brain TaDa, 79 enriched GO terms were defined under the *biological process* ontology. As anticipated, 'nervous system development' ($p<0.01$, 77 gene matches, 17%), 'neurogenesis' ($p<0.05$, 69 gene matches, 15%) and similar terms were located in this ontology (Figure 40A). These were expected given the known function of *dsx* in the development of the nervous system in the SDH (Lee et al., 2002; Rezával et al., 2016; Rideout et al., 2010; Robinett et al., 2010). Figure 40A lists a subset of the 79 enriched terms ranked by number of gene matches with redundant terms removed. In the *cellular component* ontology, enriched terms included 'cell cortex' ($p<0.01$, 21 gene matches, 5%) and 'cytoplasmic region' ($p<0.01$, 25 gene matches, 6%). 'DNA binding transcription factor activity' ($p<0.05$, 36 gene matches, 8%) was noted in the *molecular*

function ontology, very much as expected given *dsx*'s role as a central regulator of developmental processes (Burtis et al., 1989; Nagoshi et al., 1988).

The GO enrichment analysis for Dsx^{M}-Dam head TaDa was based on approximately three times the number of starting genes as compared to brain TaDa (1511 versus 461). As such, it was not surprising that 194 enriched terms were defined in the *biological process* ontology as compared to the 79 in brain TaDa. As previously, terms such as 'nervous system development' ($p<0.01$, 178 gene matches, 12%), 'neuron projection development' ($p<0.01$, 84 gene matches, 6%) and similar were enriched as expected. Additionally, we also saw terms such as 'immune system process' ($p<0.01$, 78 gene matches, 5%) defined here. This could reflect the known function of the *dsx*-expressing fat body (Bownes and Hames, 1977; Haunerland, 1996; Lazareva et al., 2007; Meister et al., 1997; Yongmei Xi, 2015), a cell group that were also profiled in head TaDa (Figure 40B). In the *cellular component* ontology, most significantly enriched terms included 'cell periphery' ($p<0.01$, 183 gene matches, 12%) and 'plasma membrane' ($p<0.01$, 165 gene matches, 11%). Interestingly, enriched terms including 'cofactor binding' ($p<0.01$, 69 gene matches, 5%) and 'protein binding' ($p<0.05$, 281 gene matches, 19%) were enriched in the *molecular function* ontology, reflecting DSX's potential mode of action. Positively, neuronal functions and processes came up several times as expected across both DSX brain and head TaDa rounds. Indeed, we also positively observe a further delineation in brain-specific and fat-body specific enrichment terms as described here. See Appendix for the *cellular component* and *molecular function* ontologies (redundant terms removed) for Dsx^{M}-Dam brain and head TaDa respectively.

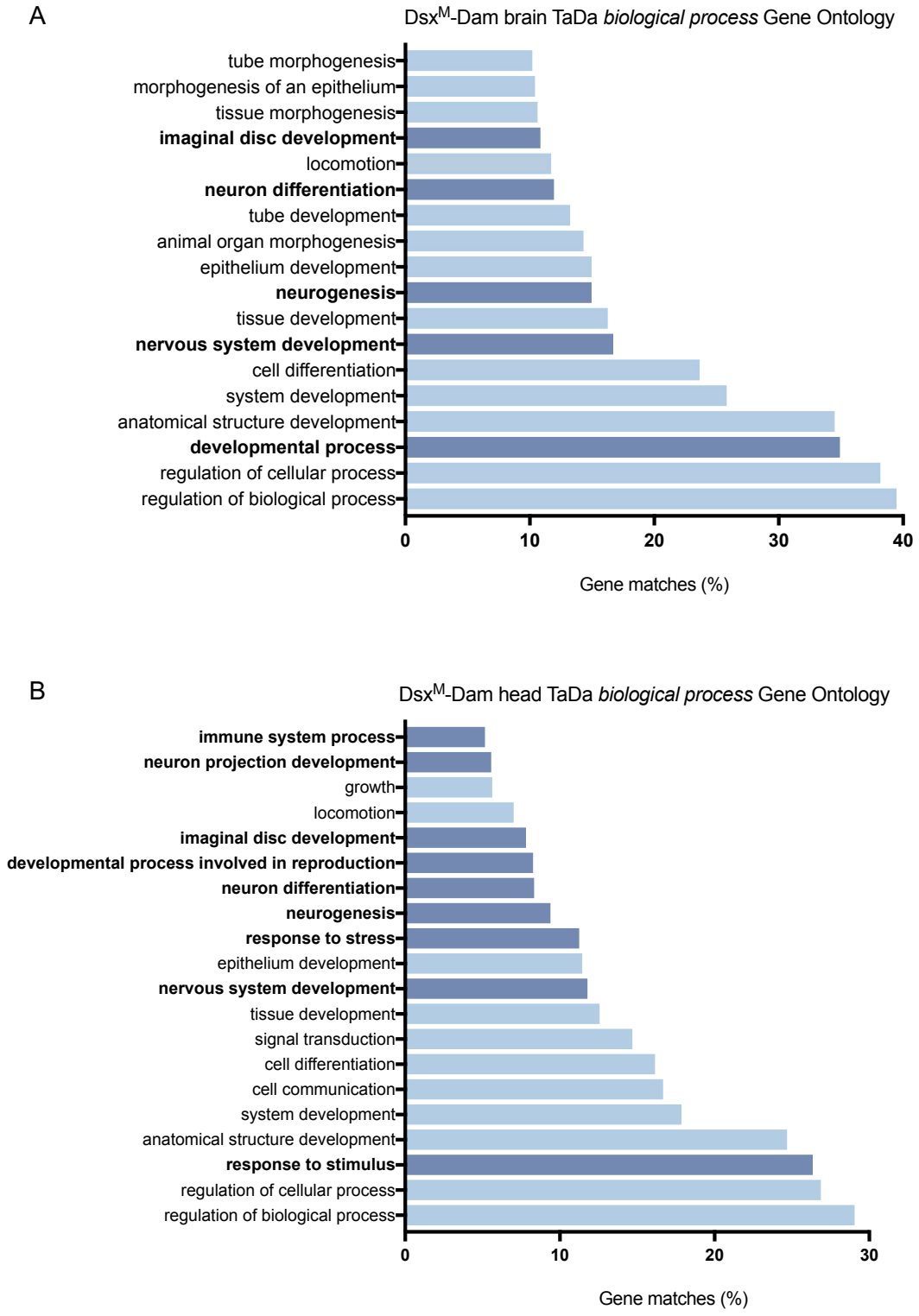


Figure 40 *Gene Ontology enrichment analysis for Dsx^{M}-Dam brain TaDa (A) and head TaDa (B) biological process ontology. For Dsx^{M}-Dam brain TaDa, 79 statistically significant enrichment terms were defined overall, and 194 for Dsx^{M}-Dam head TaDa ($p<0.05$, Holm-Bonferroni test correction). Significantly enriched terms relating to known DSX function in the nervous system and fat body are highlighted. For both, GO analysis completed on peaks/ genes common to biological replicates one and two. Chart details most significantly enriched GO terms with >45 (A) and >75 (B) gene matches respectively (redundant terms removed).*

5.3.10 SEQUENCE MOTIFS HAVE CONJECTURED BIOLOGICAL FUNCTION

In genetics, a *sequence motif* is a conserved sequence of nucleotides or amino acids that is widespread and is either conjectured or shown to have biological significance. Motifs often indicate sequence-specific binding sites for proteins such as TFs or nucleases, or are involved in important processes at the RNA level including ribosome binding or transcription termination (D'Haeseleer, 2006). Previously, information on binding sites and their determination was investigated experimentally with techniques such as DNase footprinting, and gel-shift or reporter construct assays (Fried and Crothers, 1981; Galas et al., 1978; Garner et al., 1981). SELEX, Selective Evolution of Ligands by Exponential enrichment, investigates binding affinities to artificial sequences. SELEX, or *in vitro* selection, is a method that enables the simultaneous screening of diverse pools of RNA or DNA molecules (up to 10^{15} nucleic acids) for enrichment for a specific feature. Separation or selection of functional molecules from a predominantly non-functional pool of DNA or RNA occurs, for instance, through column chromatography. Three groups developed the technique independently in 1990 (Ellington and Szostak, 1990; Joyce, 1989; Tuerk and Gold, 1990). It has since gained prevalence as an extremely useful tool in molecular biology. More recently, computational methods have garnered momentum in the identification of putative binding sites. In such methods, overrepresented and/or conserved DNA patterns upstream of functionally related genes are searched to locate regulatory sequence motifs. The richness of experimentally and computationally derived sequence motifs makes them important tools in computational biology. Indeed, motifs have become effective in deciphering the regulatory programme of individual genes and defining genetic regulatory networks.

5.3.11 PUBLISHED DSX MOTIFS VARY DEPENDING ON EXPERIMENTATION METHOD

THE REPORTED DSX MOTIF DERIVED FROM SELEX EXPERIMENTATION

Landmark studies have defined DSX DNA binding specificity biochemically (Erdman et al., 1996; Murphy et al., 2007; Yi et al., 1999). Initially, putative consensus binding sites for DSX were based on sites present in the *Yolk protein* fat body enhancer which have been shown to interact with DSX proteins *in vitro* (Burtis et al., 1991) and *in vivo* (Coschigan and Wensink, 1993). This sequence is unlikely however to accurately reflect optimal DSX binding given the small number of sites examined. Early SELEX experiments determining the consensus binding site of the DSX DNA-binding domain, isolated and characterised DSX-binding sequences from a pool of random sequence oligonucleotides. The group defined a consensus DSX-binding site as a 13 bp palindromic sequence (G/A)nnAC(A/T)A(T/A)GTnn(C/T). Structurally, the motif is two-half sites around a central (A/T) base pair (Erdman et al., 1996). Effectivity of the technique had previously been evidenced by the identification of the consensus site for the *Drosophila hairy* protein (van Doren et al., 1994). Another study used a similar selection method to identify high-affinity DSX DNA-binding domain sites by isolating oligonucleotides containing functional binding sites from a pool of random oligonucleotides using immunoprecipitation of protein/ DNA complexes (Pollock and Treisman, 1990; Yi et al., 1999). Putative oligonucleotide binding sites were tested for binding using Electrophoretic Mobility Shift Assays (EMSA). Shifted oligonucleotides were cloned and sequenced, and their DNA sequences compared to identify potential binding site preferences. All 58 sequenced oligonucleotides bound by DSX contained a close match to Erdman's defined 13 bp palindromic consensus DSX binding site (Yi and Zarkowyer, 1999).

THE REPORTED DSX MOTIF DERIVED FROM DAMID EXPERIMENTATION

The 13 bp palindromic Erdman-defined DSX binding motif is expected to appear by chance every 2 Kb in the *Drosophila melanogaster* genome. *In vivo* information is therefore important to specify biologically relevant *bona fide* DSX-binding sites. In 2011, a DamID method was employed to globally locate, putative Dsx^F-binding loci (Luo et al., 2011). The approach was one of the first to couple DamID to NGS rather than array-based methods that were more typical for the period. With array-based methods, GATC-methylated fragments were isolated using PCR and thus potentially introducing PCR bias. Recovery of fragments was dependent on methylation of GATC sites at both ends, proving problematic for large fragments generated from GATC-poor regions, which are less efficiently amplified. Their modified approach involved each GATC site being evaluated separately, by generating two sequence tags for each methylated GATC site independent of the methylation status of nearby GATCs. To reduce PCR bias, <13 PCR cycles were used in their library preparation to generate uniform PCR products (125 +/- 1 bp). As typical with DamID experiments, a Dam-only control was used as background for a measure of non-specific methylation (van Steensel and Henikoff, 2000). Given that the *Drosophila* genome contains TF colocalisation hotspots where TFs accumulate at higher frequency through protein-protein interactions (Moorman et al., 2006), the group used an additional control, in which Dam was fused to a DSX protein with a mutation specifically in its DNA-binding domain (R91Q). This control was unable to bind the *Yp1* DSX-binding site (Erdman and Burtis, 1993; Zhang et al., 2006) and *bona fide* DSX-binding sites, but theoretically would bind to TF hotspots. Their approach identified 650 DSX-binding loci globally,

extracting a new consensus 13 bp palindromic DSX-binding sequence, GCAACAATGTTGC, statistically similar to Erdman's defined DSX motif. This motif defines the outside six bases, unlike Erdman's, suggesting these bases may specifically contribute to interactions with DSX protein. The enriched motif was identified using MEME (Bailey and Elkan, 1994). As input sequences, 43 regions were selected where the distribution of GATC sites were relatively even and dense. Regions were trimmed to ~ 1.5 Kb to remove extraneous sequence without risk of removing potential DSX-binding site.

To verify the identified binding motif, *in vitro* analyses using EMSA assayed the affinity of recombinant DSX protein to a probe containing the new consensus DSX-binding site (GCAACAATGTTGC), with two competitor sequences differing in the peripheral six bases (atgACAATGTcat and cgtACAATGTacg). Theoretically, since the same amount of protein and probe are available in all three reactions, a decrease in intensity of the shifted band reflects the relative affinity of the competitor to the DSX protein. Erdman's binding site predicts that each sequence would have a similar affinity, whereas the new consensus DSX-binding sequence predicts the first sequence would have higher affinity compared to the other two competitor sequences. The EMSA results showed that the second and third sequence had three-fold or lower affinity compared to the new consensus sequence. Their results suggested that the new consensus sequence contains more information about the identity of an optimal DSX-binding site.

THE REPORTED DSX MOTIF DERIVED FROM CHIP-SEQ EXPERIMENTATION

As discussed in the introduction to this thesis, experiments involving ChIP-seq on S2 cells expressing tagged Dsx^M or Dsx^F were completed to determine where DSX binds in the *D. melanogaster* genome in the Goodwin lab (Clough et al., 2014). Using the MEME algorithm, the group identified a motif statistically similar to the DSX position weight matrix defined by Yi and Zarkower, 1999, using SELEX experimentation. The group used *de novo* motif identification under occupied ChIP-seq regions (Tomtom E value <0.01). Indeed, enrichment of sequences matching the Erdman-defined DSX consensus binding site [(G/A)nnAC(A/T)A(T/A)GTnn(C/T)] was identified under peaks ($p<0.01$, Fisher's Exact Test).

Coupling datasets generated by independent protein-DNA interaction screening methods aids in the verification of *bona fide* TF binding sites. Therefore, the statistical similarity between the putative binding sites defined by SELEX (Erdman *et al*., 1996), EMSA (Yi and Zarkowyer, 1999), DamID (Luo et al., 2011) and ChIP (Clough et al., 2014) is remarkable. Whilst each method takes a different approach, for example experimentally, with competitor sequences (Yi and Zarkower, 1999) or bioinformatically, with the consideration of TF colocalisation hotspots (Luo et al., 2011), strikingly the core defined 7-mer remained consistent. Further, the definition of the DSX binding sequence using DamID experimentation in some ways validates our searching for the DSX motif in our DSX brain and head TaDa datasets.

5.3.12 Locating the DSX motif (Yi and Zarkower, 1999) in called peaks

In line with previously discussed experiments, in this study we aim to locate the DSX binding motif in our TaDa DSX brain and head datasets. We opted to search for the Erdman-defined SELEX DSX motif (Erdman et al., 1996) corroborated with positional weight matrix EMSA data from Yi and Zarkower, 1999. This is because this motif marked the highest per base information content published for the DSX motif, the similarity of this motif compared to other defined motifs (Luo et al., 2011), as well as the use of this motif for searching in a DSX fat body protein-DNA profiling experiment (Clough et al., 2014). Using Bioconductor in R, we plotted the DSX motif sequence logo using positional PWM data identified through the SELEX/ EMSA experimentation method (Erdman et al., 1996; Yi and Zarkower, 1999; Figure 41A). We aimed to search for this DSX binding sequence in our FDR 0.05 called peaks. We conducted searches for both the entire 13 bp motif, as well as the central 7-mer excluding the 3-mer on either side of the Yi and Zarkower, 1999 defined motif. We searched for the latter because significant cogency is observed in the central 7-mer DSX motif described in various independent screening methods (Clough et al., 2014; Erdman et al., 1996; Luo et al., 2011; Yi et al., 1999). Indeed, as visualised in the sequence logo (Figure 40A), the central 7-mer represents the 7 bp with the highest likelihood to represent *bona fide* DSX binding.

We searched for either motif using FIMO (Find Individual Motif Occurrences) version 5.0.5 within the MEME-Suite (Grant et al., 2011). In both brain and head TaDa we searched for the motif in peaks common to both Dsx^{M}-Dam biological replicates one and two. In both experiments, the central 7-mer was identified more frequently

compared to the 13-mer. This was expected, given the shorter motif length and the probability of its presence in sequences of length averaging 500 bp to 3 Kb. Further, FIMO is statistically more effective at identifying motifs of shorter length. In brain TaDa, the full 13-mer DSX motif was located in 104 of 556 peaks (19%), the 7-mer in 143 peaks (26%). In head TaDa, the full 13-mer DSX motif was located in 416 of 1365 peaks (31%), the 7-mer in 588 peaks (43%) (Figure 41B). We identified the top twenty enriched genes in the Dsx^M brain and Dsx^M head datasets containing the 13-mer and 7-mer versions of the DSX motif (Figure 41C and D). Presenting the data this way provides a more global view of DSX-binding architecture across brain and head datasets. Across both datasets, the 13-mer and 7-mer motif was located in the *Spec2*, *CG14567*, and *Cyp28d1* genes within the top twenty most enriched. The Cdc42 Small Effector Protein (*Spec2*) is a protein coding gene predicted to be involved in Rho protein signal transduction. *Cdc42* regulates a plethora of cellular activities including kinase signalling (Pirone et al., 2000). The second is the uncharacterised protein, *CG14567*. Interestingly, the Probable cytochrome P450 28d1 (*Cyp28d1*), alongside twelve other P450 cytochromes are downregulated by non-sperm components of mating in female *Drosophila* (McGraw et al., 2004). P450 insect cytochromes are known to consist of a broad class of enzymes involved in detoxification and biosynthesis of ecdysteroids and juvenile hormones (Feyereisen, 1999; Wilson, 2001). It has been speculated that downregulation of *Cyp28d1* is part of a trade-off where females allocate resources away from detoxification and towards reproduction (McGraw et al., 2004).

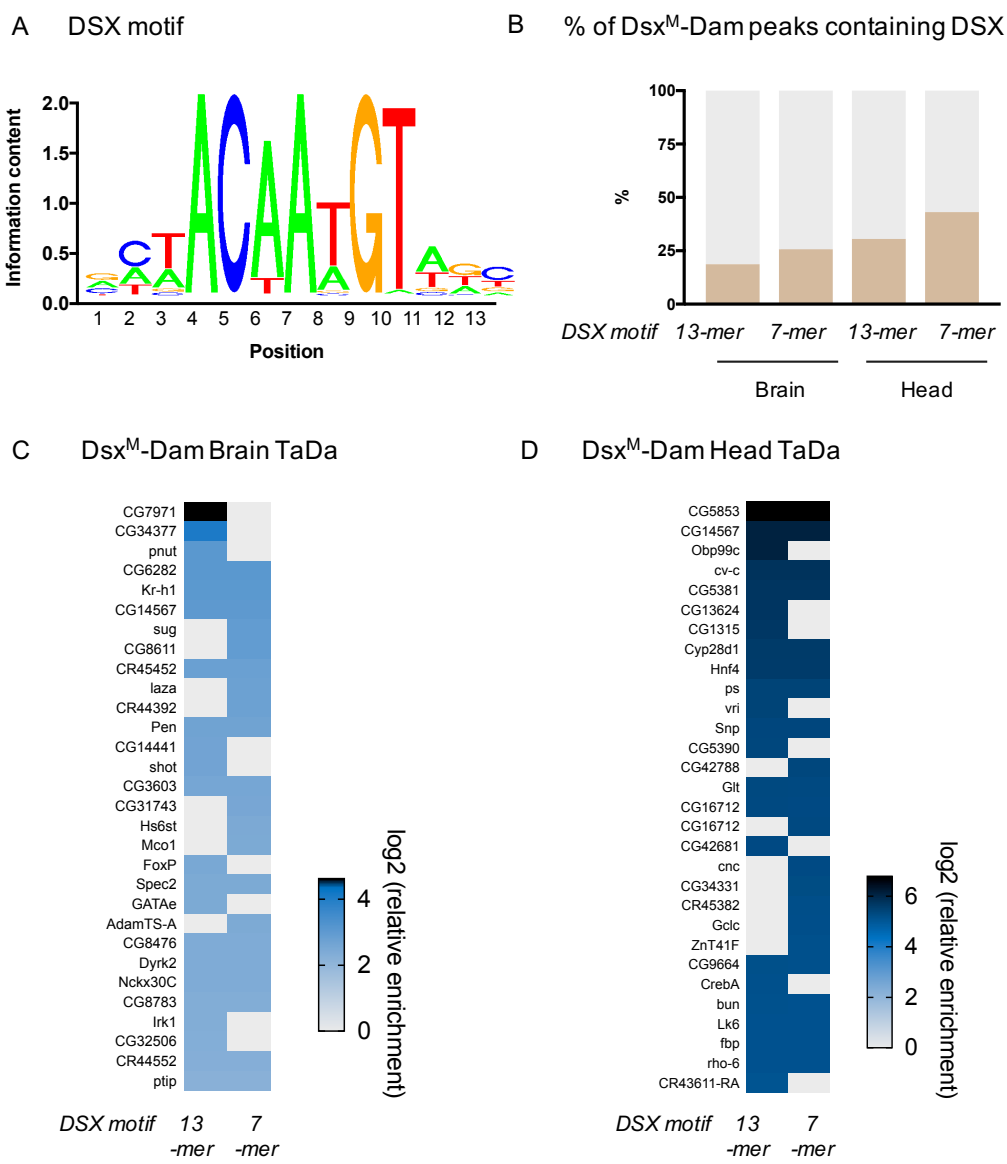


Figure 41 *13-mer DSX sequence logo as specified by Yi and Zarkower, 1999 DSX binding Position Weight Matrix (PWM) (A). Proportion of TaDa-seq peaks containing the 13-mer and 7-mer DSX motif in peaks common to both Dsx^{M}-Dam biological replicates in brain TaDa and head TaDa (B). Top twenty enriched genes in the Dsx^{M} brain (C) and head (D) datasets containing the 13-mer or 7-mer DSX motif consecutively. For C and D, peaks called to FDR 0.05 using find_peaks (Marshall and Brand, 2015). Note scale difference in C and D. Heat maps generated in PRISM v7.0d. For A, sequence logo generated using Bioconductor in R. The central 7-mer has the highest information content.*

5.3.13 Predicting regulatory features using an integrative genomics method

When searching for motifs in sequencing data, locating an already predefined binding motif alone is a biased approach. Therefore, we independently searched for binding

motifs in the TaDa Dsx^{M} brain and Dsx^{M} head called peaks using i-*cis*Target (Herrmann et al., 2012; Imrichová et al., 2015). The i-*cis*Target method locates *cis*-regulatory modules (CRMs) by ranking conserved regions in the *Drosophila* genome, and has been implemented by similar recent *Drosophila* TF DamID studies (Cheetham et al., 2018; Neville et al., 2014). We identified significant motifs and determined an optimal subset of genomic regions predicted as direct targets in the Dsx^{M} brain and head datasets. The top-ranked motifs identified in the Dsx^{M} brain and head datasets are shown in Figure 42A and B respectively. The most significantly enriched motif in the Dsx^{M} brain dataset (enrichment score of 5.2) was one previously identified as a binding site of the transcription factor *Adf1* (transfac_pro__M07860, Figure 42C). *Adf1* acts as a transcriptional activator of the alcohol dehydrogenase gene, *Adh* (Pile and Cartwright, 2000). *Adf1* is implicated in the regulation of developmental plasticity in the brain influencing crucial aspects of dendrite development (England et al., 1992; Timmerman et al., 2013). We determined the genes in the Dsx^{M} brain dataset associated with the top-ranked CRMs containing this motif. Genes included known *dsx* binding target *bab1* (Williams et al., 2008) and fellow sex-determination gene, *fru*.

The most significantly enriched motif in the Dsx^{M} head dataset (enrichment score of 9.6) was one previously identified as a binding site of the pannier (*pnr*) gene (transfac_pro_M02756, Figure 42D). *Pnr* encodes a zinc-finger transcription factor of the GATA family and is involved in several developmental processes during embryonic and imaginal development (Herranz and Morata, 2001). Genes associated with the top-ranked CRMs containing this motif included *dsx*. The highly significant transfac_pro_M02756 motif centrally contains the GATA motif. In *Drosophila*, GATA TFs are a family of zinc finger-motif DNA-binding proteins expressed in multiple

tissues in the developing embryo and involved in development (Lin et al., 1995; Winick et al., 1993). Indeed, other significantly enriched motifs in this dataset included *GATAd* transcription factors (enrichment scores of 8.7 and 8.8). Another significantly enriched motif associated with the Trithorax-like (*Trl*) transcription factor was identified (flyfactorsurvey_Trl_FlyReg, Figure 42E). *Trl* is a BTB-Zn-finger transcriptional regulator that encodes a *GAGA* transcription factor involved in chromatin modification (Granok et al., 2001).

We expected i-*cis*Target to locate the DSX motif in a proportion of called peaks, especially given our FIMO analyses identified the motif in a significant proportion of called peaks in both the Dsx^{M}-Dam brain and head datasets. We infer the results from the i-*cis*Target analysis are important in a different way however, namely the motifs identified here as predicted regulatory features could act alongside DSX as cofactors. The literature summarised here provides support for this theory as the motifs identified are largely associated with transcriptional activators or regulators involved in developmental processes. These could act alongside DSX to impinge its function downstream the SDH.

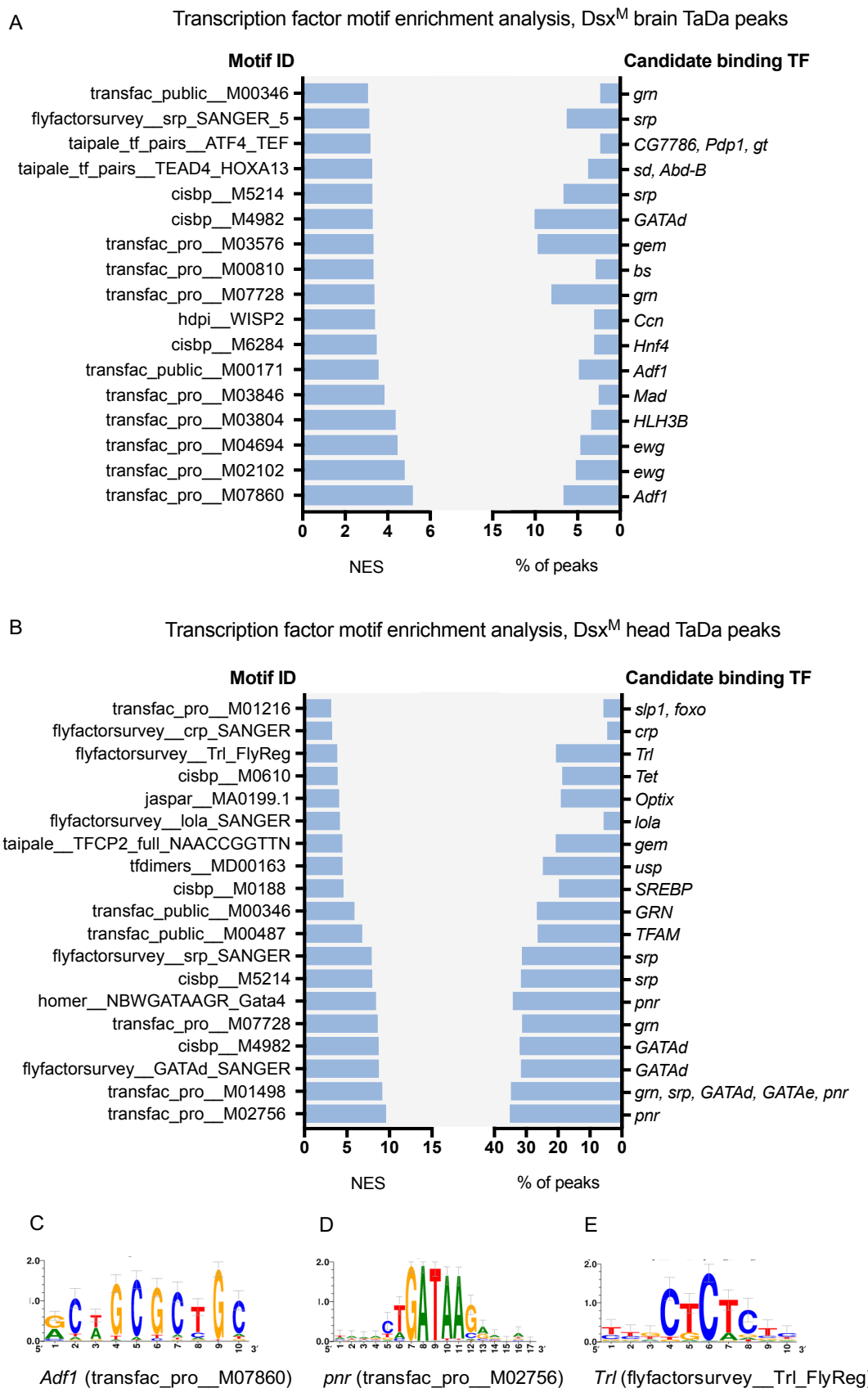


Figure 42 *The integrative genomics method, i-cisTarget (Imrichová et al., 2015; Herrmann et al., 2012), was used for the prediction of regulatory features. Enriched transcription factor motifs were identified in*

TaDa Dsx^M brain (A) and head (B) peaks. Ranked Motif IDs are presented according to the NES (normalised enrichment scores). The percentage of peaks containing the feature are indicated. Candidate binding TFs associated and corresponding to the Motif IDs are listed. Sequence logo for the Adf1 motif (C) which was most significantly enriched in the Dsx^M brain peaks is shown. Sequence logos for pnr (D) a member of the GATA-TF family, and the transcriptional regulator, Trl (E), significantly enriched in the Dsx^M head peaks are shown.

Three enriched motifs were identified across Dsx^M brain and head called peaks, grain (*grn*), GATA factor d (*GATAd*) and serpent (*srp*) (Figure 43A-D). Interestingly, these are three of five different GATA factor genes (others include pannier and GATAe) that have been reported as essential in the development and identity of multiple tissues including the brain (Martínez-Corrales et al., 2019). *Grain*, GATA-binding factor C, encodes a transcription factor involved in organ morphogenesis and regulating expression of receptors and adhesion molecules involved in axon guidance including *unc-5* and *Fas2* (Brown and Castelli-Gair Hombría, 2000; Zarin et al., 2012). GATA factor d is moderately expressed in the adult brain, and lowly expressed in the adult head more broadly (FlyAtlas.org), as well as in the adult midgut (Buchon et al., 2013). Serpent, Box A-binding factor, encodes a transcription factor involved in the development of the fat body and lymph gland. *srp* also binds a 5'-(T/A)GATAA-3' sequence element located in the larval promoters of all known *Adh* genes (Abel et al., 1993; Rehorn et al., 1996).

We implemented the i-*cis*Target method (Herrmann et al., 2012; Imrichová et al., 2015) on called peaks identified to contain the 13-mer DSX motif (based on the PWM from Yi and Zarkower, 1999) using FIMO (Figure 41B) as a means of validating the latter result (Figure 43E). Indeed, we located both the DSX motif and DMRT1 (*doublesex* and *mab-3* Related Transcription Factor 1), the mammalian homolog of *Drosophila*

DSX, in a significant proportion of called peaks (Figure 43E, G and H). The Sterol regulatory element binding protein (*SREBP*) motif was most significantly enriched in these peaks (enrichment score of 9.04, Figure 43F). *SREBPs* are transcription factors involved in the regulation of fatty acid and lipid synthesis (Dobrosotskaya et al., 2002). *SREBP* is known to be broadly expressed in *D. melanogaster* heads and nervous system (Aradska et al., 2015).

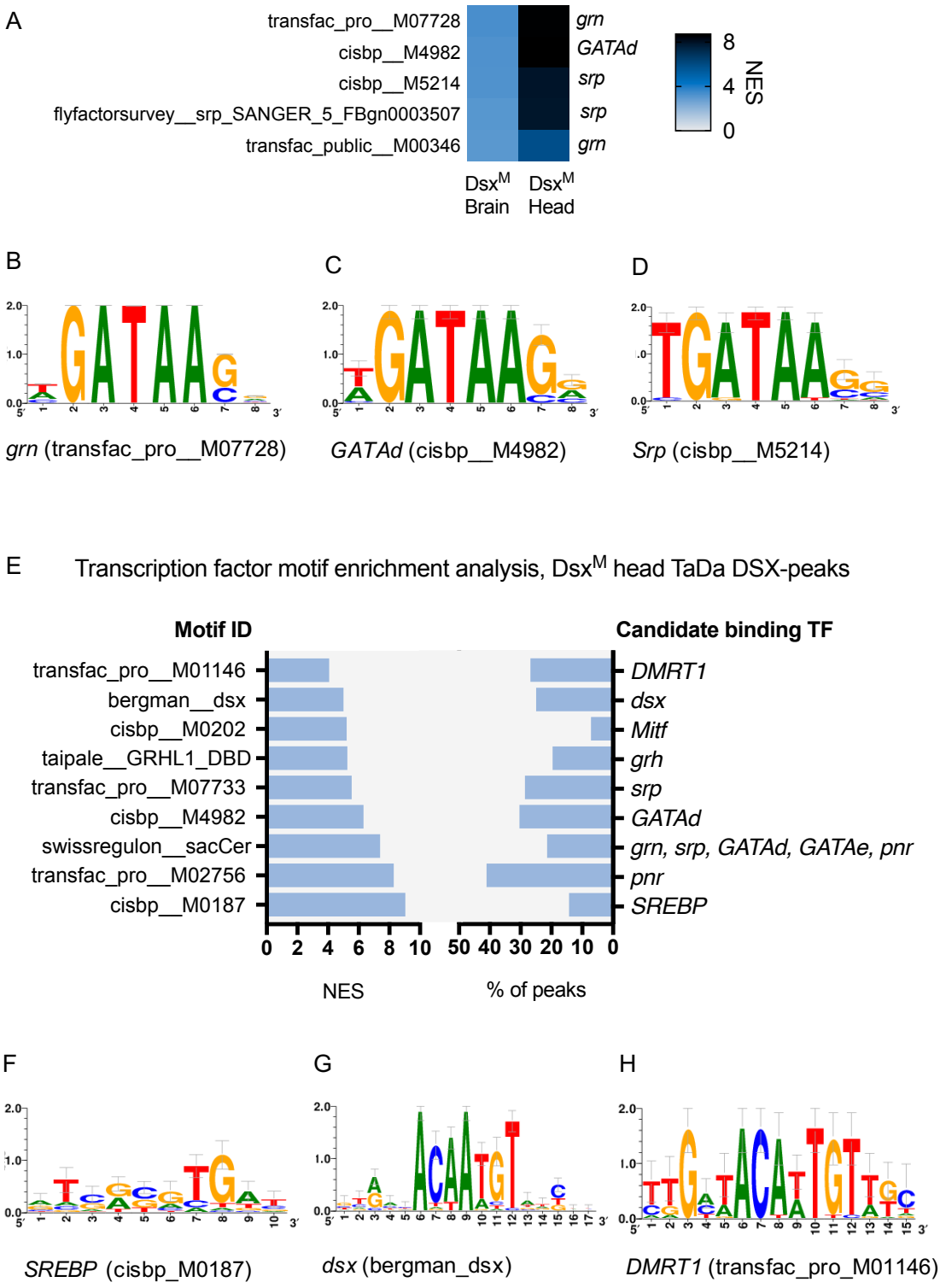


Figure 43 *The integrative genomics method, i-cisTarget (Imrichová et al., 2015; Herrmann et al., 2012), was used for the prediction of regulatory features. Heat map showing most highly enriched (NES) motif IDs and associated candidate binding TFs identified in both Dsx^{M} brain and head peaks (A). Sequence logos for three of five members of the GATA factor genes are enriched in these peaks, grn (B), GATAd (C), and Srp (D) shown. The i-cisTarget method was used to predict regulatory features in Dsx^{M} head*

peaks identified to contain the 13-mer DSX PWM (Yi and Zarkower, 1999) using FIMO (E). As previously, ranked Motif IDs are presented according to the NES. Percentage of peaks containing the feature are indicated with associated candidate binding TFs. Sequence logos for the most highly enriched motif, SREBP (F), alongside the expected dsx motif (G), and the mammalian dsx homolog, DMRT1 (H) are shown. Heat map made in PRISM v7.0d.

5.3.14 *DE NOVO* MOTIF ANALYSIS

In *de novo* analyses, computational tools can locate one or more novel signals, putative candidate motifs, from sets of multiple input biological sequences. Numerous motif discovery algorithms have been published and reviewed over the past decade, with nuanced differences in their search algorithms and parameters (Weirauch et al., 2013). HOMER (Hypergeometric Optimisation of Motif EnRichment) is a suite of tools for NGS analysis and motif discovery (Heinz et al., 2010). The *de novo* motif discovery tool, designed for regulatory element analysis, is based on a differential motif discovery algorithm. HOMER inputs two sets of sequences and searches for regulatory elements that are specifically enriched in one set relative to the other. HOMER uses ZOOPS scoring (zero or one occurrence per sequence), accounts for sequence bias, and uses hypergeometric enrichment calculations (or binomial) to determine motif enrichment. One of the most difficult tasks in motif discovery is deciding which, if any, of the discovered motifs are 'real'. Two complementary approaches aid in investigating the validity and function of the discovered motifs: is the occurrence of the motif statistically significant, and is the function of the motif already known? We employed HOMER v4.10 for *de novo* motif enrichment analysis in Dsx^{M} brain and head peaks, implemented via the Mac OSX Terminal and Perl. We further compare discovered motifs with comprehensive motif databases to ascertain whether function of discovered motifs are known.

In both the Dsx^{M} brain and head datasets, we looked for motif enrichment in peaks common to both Dsx^{M}-Dam biological replicate 1 and 2. The HOMER *de novo* analysis located 26 and 24 statistically significant ($p<0.05$) motifs respectively. The GATA motif is enriched across both datasets (Figure 44A), where it is located in 390 Dsx^{M} brain peaks (61%, p=1e-115), and 530 Dsx^{M} head peaks (39%, p=1e-157). In the Dsx^{M} head dataset, the motif for the mammalian *doublesex* and *mab-3* Related Transcription Factor 3 (DMRT3, MA0610.1/ Jaspar) protein coding gene, the mammalian homolog of *Drosophila* DSX (Figure 44B), is identified in 40 separate peaks (3%, p=1e-15). See Appendix for the top ten ranked enriched motifs identified in the Dsx^{M}-Dam brain and head datasets respectively.

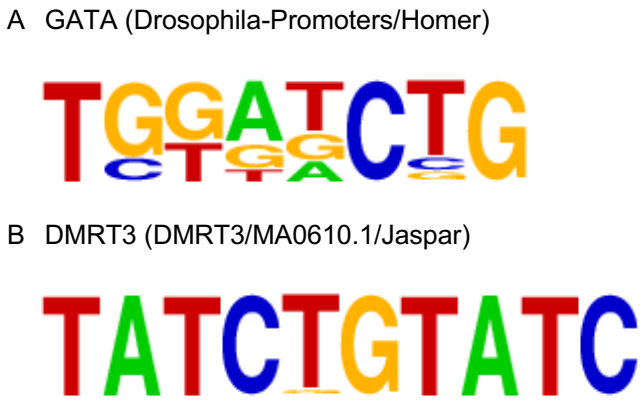


Figure 44 *Sequence logo for the GATA motif (A) and DMRT3 (B) identified by HOMER de novo analyses. The GATA feature is enriched in Dsx^{M} brain and head peaks. DMRT3, the mammalian homolog of Drosophila DSX, is located de novo in Dsx^{M} head peaks.*

5.3.15 DOWNSTREAM ANALYSES FOR DSX FEMALE-DAM DATASETS

As visualised in the area-quantitative Venn diagrams (Figure 32B, D and F) we saw significant variation in our Dsx^F-Dam biological replicates in both brain and head TaDa datasets. We hence assess the Dsx^M-Dam and Dsx^F-Dam datasets separately, as detailed in the bioinformatics downstream analysis flow chart (Figure 33). We try three computational approaches to analyse the Dsx^F-Dam TaDa biological replicates in the head dataset to attempt to understand the source of the variation. Firstly, we merge the sequencing reads from both Dsx^F-Dam replicates, and compare these to the single Dsx^M-Dam replicates. Secondly, we visually assess the unsubtracted binding enrichment profiles of the Dsx^F-Dam and Dam-only controls separately. This allows us to visually compare biological replicates and select the 'most likely' Dam-only replicate to use for background subtraction based on their similarity to published Dam-only profiles and dissimilarity to DSX-Dam profiles. Thirdly, we randomly reduce the Dam-only female control background levels to similar sequencing read levels as the Dsx^F-Dam datasets (~30%).

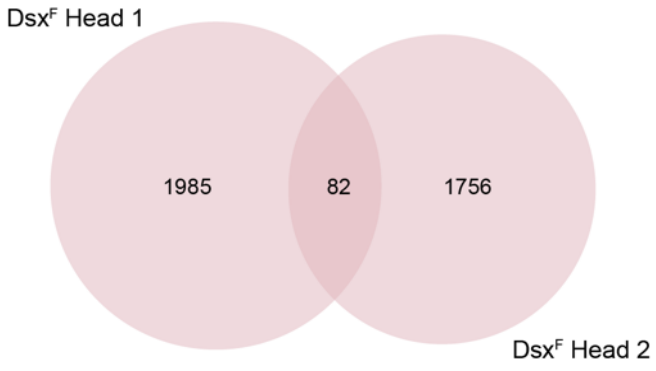


Figure 45 *Significant variation is seen in peak overlap in DSX Female-Dam replicates 1 and 2 in brain and head TaDa. Peaks called to FDR 0.05 (find_peaks, Marshall and Brand, 2015). Area-quantitative Venn diagrams generated in BioVenn (Hulsen et al., 2008).*

POTENTIAL LIBRARY PREPARATION ISSUE WITH DSX FEMALE-DAM REPLICATES

Assessing the alignment data generated in Bowtie 2, we noted the overall alignment of both Dsx^{F}-Dam (UAS-Dam-*dsx^{F}*/ *dsx^{Gal4}*) biological replicates was only ~20% (Table 14). Both Dsx^{M}-Dam (UAS-Dam-*dsx^{M}*/ *dsx^{Gal4}*) biological replicates had overall alignment rates closer to the expected ~70% as also seen in both 'Dam-only' (UAS-Dam/ *dsx^{Gal4}*) male and female controls. The mapping difference could be attributed to a problem with the generation of the initial library preparation.

Biological sample	Reads (*n*)	Mapped 1 (*n*)	Mapped >1 (*n*)	Alignment (%)
UAS-Dam-*dsx^{M}*/ *dsx^{Gal4}* 1	31 196 100	21 151 400	1 298 600	72
UAS-Dam-*dsx^{M}*/ *dsx^{Gal4}* 2	34 765 900	16 957 900	1 858 000	54
UAS-Dam-*dsx^{F}* / *dsx^{Gal4}* 1	38 114 900	5 431 800	2 094 300	20
UAS-Dam-*dsx^{F}* / *dsx^{Gal4}* 2	39 908 300	5 389 300	2 446 600	20
UAS-Dam/ *dsx^{Gal4}* ♂ 1	41 071 000	23 546 000	5 870 900	72
UAS-Dam/ *dsx^{Gal4}* ♂ 2	44 013 200	20 651 800	4 763 600	58
UAS-Dam/ *dsx^{Gal4}* ♀ 1	33 414 000	14 724 000	4 648 400	58
UAS-Dam/ *dsx^{Gal4}* ♀ 2	38 858 300	17 916 300	4 892 000	59

Table 14 *DSX head TaDa alignment rates of DamID sequencing reads to the Drosophila melanogaster genome. We note a significantly lower number of reads mapped in both Dsx^{F}-Dam biological replicates as compared to either Dsx^{M}-Dam or male and female Dam controls. Indeed, alignment scores were slightly lower than expected in all biological replicates in head TaDa. We used Bowtie 2, an ultrafast, memory-efficient tool, for the alignment of sequencing reads to the Drosophila melanogaster Ensembl BDGP6 reference genome.*

MERGING DSX FEMALE-DAM SEQUENCING READ BIOLOGICAL REPLICATES

Rationale for merging both Dsx^F-Dam replicates from head TaDa derives from the significantly smaller number of aligned reads in these replicates as compared to Dsx^M-Dam and male and female Dam controls. We combined the sequencing reads for Dsx^F-Dam biological replicate one and two and took the DamID-seq processing forward as one sample. To do this, we converted Dsx^F-Dam 1 and 2 aligned .bam files to .bed files, combined them, and converted back to sorted .bam files. We then processed this newly generated .bam file containing merged data with the damidseq_pipeline and find_peaks (Marshall and Brand, 2015) using the same parameters as previously. However, our analysis revealed no significant difference in called peak numbers between the original Dsx^F-Dam biological replicates 1 and 2, and the merged replicates (Table 15). Peaks were called to FDR 0.05 in both samples as previously.

Trial	Experimental sample	Called peaks (*n*) FDR 0.05
Original data	Dsx^F-Dam 1 x Dam ♀ v1	2067
Original data	Dsx^F-Dam 2 x Dam ♀ v2	1838
Merged Dsx^F-Dam biological replicate 1 and 2 Vs Dam ♀ 1	Dsx^F-Dam 1 and 2 x Dam ♀ v1	1944

Table 15 *Comparing called peak numbers in Dsx^F-Dam biological replicates 1 and 2, with a trial sample containing peaks called from the reads of biological replicates 1 and 2 merged. Peaks called to FDR 0.05 in all experimental sample trials. There is no statistical significant difference between called peak numbers between either biological replicate and the merged dataset (unpaired t-test, $p>0.05$).*

We viewed the processed TaDa-seq data using the Integrated Genome Browser (IGB 9.0.2), an open-source genome browser that enables visualisation of genomic data sets such as alignments, sequence data, and gene models. Viewing and visually comparing the Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates, we saw a level of variation dependent upon which Dam control replicate was used as the background sequence for subtraction. We visualised the .bedgraph enrichment profile of each Dam male and female biological control replicate, and selected the 'most likely' Dam controls for analysis. In this selection process, used solely for the bioinformatics experimental trial described here, we analysed IGB enrichment profiles for similarity within sex-specific biological replicates and between sexes, choosing 'most likely' Dam controls based on their overall broad similarity to each other and dissimilarity to the enrichment profiles of the Dsx^{M}-Dam and Dsx^{F}-Dam (Dam unsubtracted) biological replicates. We also based this selection process on 'Dam-only' enrichment profiles of published DamID-seq datasets including an experiment profiling DSX in the fat body tissue (Clough et al., 2014).

We compared the original datasets with called peak numbers in the Dsx^{F}-Dam 1 and 2 merged replicates assessed with either Dam-only female control sample. Merging both Dsx^{F}-Dam biological replicates and comparing to either Dam female control generated similar results to the original datasets, where each Dsx^{F}-Dam biological replicate is compared to either Dam control (Table 16A and B).

A

	Trial	Experimental sample
1	Original data	Dsx^F-Dam 1 x Dam ♀ v1
2	Dsx^F-Dam 1 Vs alternative Dam ♀ 2	Dsx^F-Dam 1 x Dam ♀ v2
3	Original data	Dsx^F-Dam 2 x Dam ♀ v2
4	Dsx^F-Dam 2 Vs alternative Dam ♀ 1	Dsx^F-Dam 2 x Dam ♀ v1
5	Merged Dsx^F-Dam biological replicate 1 and 2 Vs Dam ♀ 1	Dsx^F-Dam 1 and 2 x Dam ♀ v1
6	Merged Dsx^F-Dam biological replicate 1 and 2 Vs alternative Dam ♀ 2	Dsx^F-Dam 1 and 2 x Dam ♀ v2

B

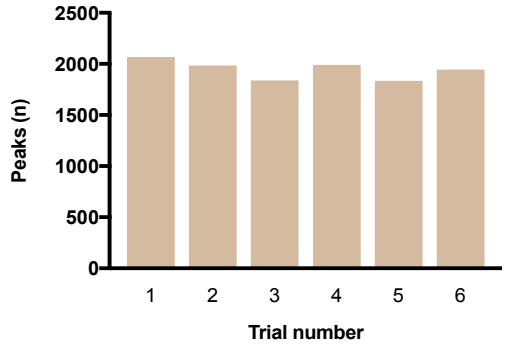


Table 16 *Summary of called peak numbers in Dsx^F-Dam biological replicates in head TaDa, with samples assessed with alternative Dam controls, as well as to merged Dsx^F-Dam samples compared to alternative Dam control (A). Peak numbers do not statistically differ in any of the six experimental trials (B). Peaks called to FDR 0.05, one-way ANOVA, $p>0.05$.*

RANDOMLY REDUCING DAM-ONLY SEQUENCE BACKGROUND

Visual analysis of the Dsx^F-Dam replicates IGB enrichment profiles suggested overall enrichment in these samples appeared lower compared to the Dsx^M-Dam replicates. We therefore aimed to randomly reduce the Dam-only female control read count of the samples used as background for the Dsx^F-Dam biological replicates. Whilst obscuring peaks in this way is not recommended through the potential introduction of bias (Marshall et al., 2015), it could allow us to assess the Dsx^F-Dam biological replicates without being 'washed out' by the Dam-only background signal. We randomly downscaled the Dam-only female one control to the same read depth (~30% of the reads) as the Dsx^F-Dam biological replicate one. Statistically, called peak numbers were not significantly different to the original dataset (unpaired t-test, $p>0.05$; Table 17).

Trial	Experimental sample	Called peaks (*n*, FDR 0.05)
Original data	Dsx^F-Dam v1 x Dam ♀ v1	2067
Original data Vs downscaled Dam ♀ 1	Dsx^F-Dam v1 x downscaled Dam ♀ v1	1948

Table 17 *Summary of called peak numbers in downscaled Dsx^F-Dam biological replicate one compared to original data. Head TaDa biological replicates, peaks called to FDR 0.05.*

We believe the variability in the Dsx^F-Dam biological replicates in both the brain and head datasets stemmed from problems in the library preparation stage, as revealed by low mapping of sequenced reads in the Bowtie 2 analysis. Indeed, the low mapping

could be explained by the over-amplification of starting genomic material in the library preparation stage. However, as described in a previous chapter of this thesis, we meticulously controlled for consistent DSX-Dam methylation smears in PCR-amplified samples before they were sent for library preparation and NGS. Indeed, this meant we used a larger amount of starting animal material in the Dsx^{F}-Dam samples compared to the Dsx^{M}-Dam samples in the DamID protocol, in line with fewer induced Dsx^{F} cells. The three bioinformatic trials described here aimed to assess the binding architecture of Dsx^{F}-Dam by identifying called peak locations across biological replicates accounting for low mapping of Dsx^{F}-Dam sequenced reads, and Dam-only binding enrichment pattern reproducibility, and its downscaling to reduce loss of Dsx^{F}-Dam bound signal. Each of the three analyses identified statistically similar (each unpaired t-test, $p<0.05$) numbers of called peaks. However, locations of called peaks differ and hence their overlap remains small. Given this variation, it is impossible to draw biological conclusions from the Dsx^{F}-Dam datasets.

5.4 DISCUSSION

To help elucidate the role of DSX in the nervous system, we undertook a targeted DamID approach. We aimed to identify the association of Dsx^M and Dsx^F proteins with specific regions in the genome to determine DSX occupancy in the brain and head of male and female adult animals, and locate DSX occupancy in different cell types (neural cells versus fat body). The series of bioinformatic analyses presented here primarily assess the biological integrity of the NGS data generated, and draw biological conclusions of the conducted analyses. We carried out two TaDa-seq screens profiling DSX cells in dissected brain tissue ('brain TaDa') and whole heads ('head TaDa') in male and female adult animals. This approach theoretically allowed us to profile both DSX neural cells in the brain as well as the DSX-expressing fat body cells encapsulating the brain by subtracting the datasets. Further, the dual screen approach enabled the winnowing of putative target genes for further analysis; this will be described in the following chapter.

NGS profiling data was processed using the damid_seq pipeline (Marshall and Brand, 2015) to generate DamID enrichment profiles, as utilised in several recently published studies (Cheetham et al., 2018; Ho et al., 2019; Marshall et al., 2017). Each biological replicate was subjected to a number of analyses to find significantly bound regions (peaks, protein-DNA interactions) according to a False Discovery Rate model, using find_peaks (Marshall and Brand, 2015). Our assessments of called peak numbers across two FDRs, and visualisation of enrichment profiles in IGB, allowed us to evaluate the integrity of the NGS biological replicates between and across screens. These are typical

methods for the assessment of biological reproducibility in NGS datasets (for example, Cheetham et al., 2018; Clough et al., 2014).

Comparing peak numbers across both brain and head TaDa, we saw considerably higher numbers called in head TaDa. This is in line with expectations, as head TaDa incorporates profiling a higher number of cells, specifically both DSX-expressing neural cells in the brain and fat body tissue. Our statistical analyses revealed similar numbers of called peaks within both Dsx^{M}-Dam biological replicates in brain and head TaDa, respectively (both unpaired t-test, $p>0.05$; Table 13). Indeed, we noted similar numbers of peaks common to both Dsx^{M}-Dam biological replicates in brain and head TaDa (both unpaired t-test, $p>0.05$; Figure 32A and C). These findings, alongside visual inspection of enrichment binding profiles in Dsx^{M}-Dam biological replicates in brain and head TaDa, suggested a good level of biological reproducibility across both Dsx^{M}-Dam 1 and 2. Observing the location of called peaks, we saw a relatively even distribution of peaks across chromosomes in brain and head TaDa, including across the X and Y sex chromosomes in the Dsx^{M}-Dam brain dataset (Figure 31A, B, E and F). This is in line with published work, which suggests this spread of binding targets is expected (Clough et al., 2014; Luo et al., 2011, 2015).

Analyses of the Dsx^{F}-Dam datasets across both brain and head TaDa, however, were considerably different. For both the number of peaks called, as well as looking at peaks common to both biological replicates within either brain or head TaDa, we saw significant levels of variation (unpaired t-tests for each, $p>0.05$; Table 13; Figure 32B and D). This led us to separate our handling of the Dsx^{M}-Dam and Dsx^{F}-Dam datasets,

tailoring the downstream bioinformatic analyses accordingly. There are comparatively smaller numbers of Dsx^F neurons in the female CNS (~300-400) than Dsx^M neurons in the male CNS (~400-700; Lee et al., 2002; Rezával et al., 2016; Rideout et al., 2010), which are profiled with our TaDa-seq analyses. Perhaps then, the biological variation we observe across replicates stems from profiling smaller numbers of induced neural populations. This is unlikely however given a recent published study which aimed to map TF occupancy using minimal numbers of mouse neural stem cells with a modified TaDa approach, started with 10^6 and 10^4 cells and found reproducible TF signature peaks were detected with as few as ~1000 cells (Tosti et al., 2018). In both head and brain TaDa, considerably higher numbers of Dsx^F-Dam cells were profiled. Perhaps, the variation could be due to a problem with the Dsx^F-Dam transgenic line – indeed, whilst verified by sequencing, PCR genotyping, and overexpression using dsx^{Gal4} (Rideout et al., 2010) – we observed significant variation in the sex-ratios of offspring. The biological variation we observed across replicates mean it is impossible to draw conclusions about the sex-specificity of target genes.

Given the biological variation we observe in both Dsx^F-Dam biological replicates in brain and head TaDa, we took a different bioinformatics approach to handle the datasets as compared to Dsx^M-Dam (Figure 33). Firstly, we noted significantly fewer Dsx^F-Dam reads were mapped to the genome than anticipated in head TaDa (Table 14). This could be attributed to a problem with the initial library preparation method, where gDNA material was over-amplified. Merging Dsx^F-Dam reads from both biological replicates, and calling peaks as previously, generated peak numbers not significantly different to either biological replicate alone (Table 15, both unpaired t-test, $p>0.05$). Secondly, from visual inspection of the Dsx^F-Dam enrichment plots, we noticed higher than expected

Dam-only enrichment in the female samples compared to either the Dsx^F-only or Dam-only male enrichment plots, as well as biological variation across both Dam-only biological samples. Perhaps this 'over enrichment' caused real Dsx^F-only signal to be lost through normalisation. Reducing Dam-only read counts to Dsx^F levels did not result in significantly different numbers of called peaks (Table 17, unpaired t-tests, $p>0.05$) and nor did assessing Dsx^F biological replicates with alternative controls (Table 16, unpaired t-tests, $p>0.05$). Importantly however, whilst each iteration of peak calling identified similar numbers of peaks, assessments of peak locations revealed vast variability. This variability meant it was impossible to draw biological conclusions from the Dsx^F-Dam datasets.

Apart from investigating peak overlaps, another method to assess the integrity of the dataset is to look for known binding targets. Here, we looked for binding of known *dsx* target genes including *bab1* (Figure 38A; Williams et al., 2008) and the *Yolk protein* genes (Burtis et al., 1991; Coschigano et al., 1993; Hutson et al., 2003). We further investigated the enrichment of *fru* given its pivotal role alongside *dsx* in the SDH (Figure 38B), and the Fatty acid desaturase (*Desat1*) gene known to have a role in courtship behaviour in *Drosophila* (Figure 38C; Bousquet et al., 2012; Grillet et al., 2006; Ueyama et al., 2005). We observed the majority of these described targets located in one or all of the Dsx^M-Dam or Dsx^F-Dam biological replicates in brain and head TaDa, albeit with variation. In brain TaDa, we identified *bab1* binding in both Dsx^M-Dam biological replicates, *fru* in all Dsx^M-Dam and Dsx^F-Dam biological replicates, and *Desat1* in one Dsx^M-Dam biological replicate. In head TaDa, whereas *bab1* did not appear in any biological replicate, *fru* was identified in both Dsx^M-Dam biological replicates and one Dsx^F-Dam biological replicate, and *Desat1* in all Dsx^M-Dam and

Dsx^{F}-Dam biological replicates. Interestingly, we did not see binding of either *Yp 1* and *Yp 2* in any dataset. This variation in binding of known targets in biological replicates is in line with previous DSX-fat body DamID-seq experiments (Clough et al., 2014; Li et al., 2015). It could be the result of the inherent nature of the DamID protein-DNA interaction detection method. Given TFBS are not ensured to be symmetrically distributed between methylated GATC sites, there can be significant variation introduced at the peak calling stage dependent on the caller implemented. This is confounded further by various methods for peak to gene assignment. Nonetheless, despite variation induced through the experimental method, the successful identification of the majority of known targets in the majority of the Dsx^{M}-Dam and Dsx^{F}-Dam replicates indicates the robustness of the datasets for downstream bioinformatic analyses.

Our analyses of the top forty enriched genes in the Dsx^{M}-Dam brain and head datasets show just ~20% were common (Figure 37C), suggesting the top forty most highly enriched genes in the Dsx^{M}-Dam head dataset were mostly fat body specific. Perhaps this tendency towards tissue-specificity could have been inferred from Dsx^{M}-Dam occupancy characteristics, which differed across brain and head datasets: brain Dsx^{M}-Dam bound largely an equal share of intronic and exonic regions, whereas over half of all gene annotations were in intronic regions in head Dsx^{M}-Dam. (Figure 35A). Indeed, the latter correlates with differences in peak lengths across Dsx^{M}-Dam brain and head datasets (Figure 35B).

To investigate the functional relationship between putative Dsx^M target genes, we performed a global gene ontological (GO) enrichment analysis of biological functions. Our analyses highlighted significant variation between brain and head TaDa as demarked by tissue function. Namely, in head TaDa specifically, enriched terms such as 'immune system process' ($p<0.01$, 78 gene matches) were defined reflecting the function of the fat body as a multifunctional tissue (Bownes et al., 1977; Haunerland, 1996; Lazareva et al., 2007; Meister et al., 1997; Yongmei Xi, 2015). Both brain and head TaDa defined enriched terms attributable to the role of DSX in development in the SDH, such as 'nervous system development' and 'neurogenesis' (Coschigano et al., 1993; Lee et al., 2002; Rezával et al., 2016; Rideout et al., 2010; Zarkower, 2013). The GO analyses confirm the robustness of the dataset, defining enriched terms specific to known *dsx* function in the SDH, the known function of fat body, and defining these terms specifically in the brain and head datasets, respectively.

Following this, we used the i-*cis*Target method (Herrmann et al., 2012; Imrichová et al., 2015) to predict Dsx^M regulatory features. A motif associated to the transcription factor *Adf1* (transfac_pro__M07860, Figure 42B) was most significantly enriched in the Dsx^M brain dataset. Pannier (*pnr*) was the most significantly enriched motif in the Dsx^M head dataset (transfac_pro_M02756, Figure 42D). Both *Adf1* and *pnr* play fundamental roles in the regulation of key developmental processes. Indeed, *Adf1* is implicated in the regulation of developmental plasticity in the brain, influencing crucial aspects of dendrite development (England et al., 1992; Timmerman et al., 2013). *Pnr* encodes a transcription factor of the GATA family (Herranz et al., 2001). Given the defined function of *Adf1* and *pnr*, we muse these transcription factors function as co-factors alongside *dsx* downstream in the SDH. It is indeed striking to locate motifs associated

with two genes involved heavily with development, given the crucial role of *dsx* in development. Unlike ChIP experiments that provide a snapshot of protein-DNA interactions at the exact moment assayed, DamID experiments identify all protein-DNA interactions that have occurred globally up to the time point assayed. Identifying these genes involved in development at the adult animal time-point could thefore be expected given the role of *dsx* in the SDH. Notably, looking at binding motifs common to both Dsx^{M}-brain and Dsx^{M}-head peaks, we identified three of five GATA factor genes, grain (*grn*), GATA factor d (*GATAd*) and serpent (*srp*) (Figure 43B-D) as most significantly enriched (Figure 43A). In *Drosophila*, the GATA TFs are essential in the development and identity of multiple tissues including the brain (Martínez-Corrales et al., 2019). They are commonly recruited transcriptional co-activators for many TFs, including Forkhead box proteins (Lam et al., 2013). Alongside GATA-motif containing *pnr*, this finding provides support for our theory that GATA factors function alongside *dsx* as co-factors to impinge its function.

Independent of the i-*cis*Target method, we searched for *de novo* significant sequence motifs enriched in our Dsx^{M}-Dam genomic occupancy data to delineate DNA-binding specificities of Dsx^{M} in the brain and head datasets. Our HOMER *de novo* discovered motifs corroborated those discovered by the i-*cis*Target method, with the GATA motif enriched across both datasets in a significant proportion of peaks (Figure 44A), as well as the DMRT3 motif in the Dsx^{M} head dataset (Figure 44B). One reason the DSX motif may have been discovered *de novo* in called peaks in the Dsx^{M}-head but not the Dsx^{M}-brain dataset could be due to differences in peak shape. However, given that the peak lengths in the brain dataset were significantly lower than in the head dataset (IQR ~700-1,500 bp versus ~1,100-2,700 bp), it would be more likely to locate the motif in Dsx^{M}-

brain peaks. This may be offset however by the smaller number of peaks in the brain dataset compared to the head dataset, which diminishes the statistical chance to identify the DSX motif. Manual searches for the Erdman et al., 1996-defined DSX motif revealed its presence at seemingly random locations within called peaks, across the start, mid-point and end of each peak, as well as sometimes appearing up to twice in each peak. This is anticipated because induced methylation signals are not expected to be symmetrical at TFBS (Li et al., 2015). This makes the process of peak to gene annotation difficult. Here, we assigned DSX peaks to genes occurring in the gene body + 1 Kb upstream of TSS and + 1 Kb downstream of TTS. Whilst the method biases towards longer genes, it takes into consideration the variation in gene body lengths and captures intronic enhancers.

Our analyses directly searching for the DSX motif in called peaks across both datasets revealed the 7-mer DSX motif was located in 43% of called peaks in the Dsx^{M}-head dataset and in 26% of called peaks in the Dsx^{M}-brain dataset (Figure 41B). Additionally, when we assessed these peaks known to contain the DSX motif using i-*cis*Target, both *Drosophila dsx* and the mammalian homolog DMRT1, were identified in 25% and 27% of called peaks respectively. Whilst *dsx* could bind DNA directly to regulate transcription, it could also regulate transcription indirectly forming a complex with other TFs to activate transcription where it is not directly involved in DNA binding (Burtis et al., 1991; Coschigano et al., 1993; Ghosh et al., 2019; Hutson et al., 2003; Shirangi et al., 2009; Williams et al., 2008). For the latter, identification of a *dsx* binding motif in these TaDa datasets would be difficult, and thus could provide an explanation for why we do not identify the *dsx* motif *de novo* in one dataset.

One limitation of DamID-seq data is that induced adenine methylation signals are not symmetrically distributed at transcription factor binding sites. In the future, we could implement a non-parametric peak calling algorithm for DamID-seq (Li et al., 2015) that has been developed to combat this bias. As proof of concept, this method has been applied to DamID-seq experiments profiling Dsx^M and Dsx^F-expressing fat body in adult animals (Clough et al., 2014; Li et al., 2015). Analysis by this method revealed significant reproducibility across biological replicates, and further when compared to homologous ChIP-seq peaks.

Our Doublesex-DNA interaction profiling of Dsx^M neuronal populations in adult *D. melanogaster* brains and heads across two TaDa-seq rounds, presented here, has generated a rich set of target genes for further analysis. This is particularly important as there are just a few defined direct targets of *dsx* in the CNS, and *dsx* is known to play a pivotal role in the SDH. Our analyses suggest that *dsx* binds thousands of genes, and comparing DSX brain and head datasets, we observe a tendency towards tissue-specificity. DamID profiling of DSX neuronal populations in adult animals provides a history of DSX-DNA interactions through development. It is therefore suitable that enriched GO terms such as 'nervous system development' and GATA TF motifs were significantly enriched in both Dsx^M brain and head datasets. Further, our i-*cis*Target (Herrmann et al., 2012; Imrichová et al., 2015) analyses, as well as *de novo* motif (HOMER, Heinz et al., 2010) and known motif searches, have aimed to winnow genes lists to 'higher confidence' targets. The broad range of *dsx* target genes identified here corroborate and exemplify the broad mode of action of *dsx* as a TF and its major role in the SDH.

6 CHASING TARGETS

CHARACTERISING DSX PUTATIVE TARGET GENES

6.1 INTRODUCTION

The TaDa-seq experiments profiling Dsx^M populations in the adult *D. melanogaster* brain and head presented in this thesis have generated a rich list of putative DSX targets which have a broad range of functions. Our bioinformatic analyses suggest these targets have a tendency towards tissue-specificity. To delineate the latter further, we compare these screens with a published DamID-seq experiment profiling DSX in the fat body (Clough et al., 2014).

6.1.1 DSX-FAT BODY DAMID-SEQ IN *DROSOPHILA* (CLOUGH ET AL., 2014)

Despite extensive research into DSX function over the last half century, there are still just a handful of defined DSX target genes (Burtis et al., 1991; Coschigano et al., 1993; Luo et al., 2011; Shirangi et al., 2009; Williams et al., 2008). Indeed, those target genes that have been defined cannot fully explain the plethora of behaviours regulated by DSX. In 2014, a study combined a series of genome-wide DSX occupancy analyses including ChIP-seq and DamID-seq on both Dsx^M and Dsx^F isoforms with comparative genomic analyses (Clough et al., 2014). ChIP-seq was conducted on S2 cells expressing tagged Dsx^M or Dsx^F, DamID in Dsx^F-expressing adult ovary, and Dsx^M and Dsx^F adult male and female fat body in transgenic flies, followed by either sequencing (DamID-seq) or hybridisation to microarrays (DamID-chip). DSX is known to function in maintaining sexually dimorphic gene expression in both organs, hence their investigation (Burtis et al., 1989; Kimura et al., 2005; Taylor and Truman, 1992; Waterbury et al., 1999).

The *Drosophila* fat body (FB) is a multifunctional tissue involved in energy storage, immune response, and nutritional sensing, where it is able to release energy according to demands (Bownes et al., 1977; Haunerland, 1996; Meister et al., 1997; Yongmei Xi, 2015). The FB has also been shown to secrete vital factors related to brain development and body size (Sousa-Nunes et al., 2011), and has been proposed to have a role in male courtship behaviour (Lazareva et al., 2007). The FB is composed of large, lipid-filled cells, and is considered equivalent to the vertebrate adipose tissue and liver in its storage and major metabolic functions (Yongmei Xi, 2015). In *Drosophila* larvae, individual cells of the FB exist through metamorphosis as fat cells distributed through the body cavity of the pupa (Butterworth, 1972; Hoshizaki et al., 1995; Nelliot et al., 2006). In newly eclosed adults, the fat cells persist, and later undergo cell death and get replaced by sheets of fat cells, otherwise known as the adult FB. In the adult, the FB surrounds the brain in the fly head, as well as in the abdomen. Fully differentiated fat cells are not easily recognised until 3-4 days post-eclosion (Figure 46) (Aguila et al., 2007).

The yolk proteins in *Drosophila*, produced by three X-linked genes, *Yolk Protein 1, 2, and 3* (*Yp1*, *Yp1*, and *Yp3*), are synthesised in the FB (Bownes et al., 1977, 1978; Hutson et al., 2003). *Yp1* and *Yp2* are two of a small handful of confirmed direct DSX target genes (Burtis et al., 1991; Coschigano et al., 1993; Luo et al., 2015; Shirangi et al., 2009; Williams et al., 2008). They provide a nutritional supply and bind conjugated hormones necessary for embryonic development (Bownes et al., 1988; Butterworth et al., 1992). At a molecular level, the regulation of yolk protein gene expression by DSX is well defined: Dsx^M and Dsx^F directly bind *in vitro* (Burtis et al., 1991) and *in vivo* (Coschigano et al., 1993) to a number of specific sites within the fat body enhancer

(FBE) sequence. Binding of Dsx^M to the FBE significantly reduces expression of the yolk protein genes. Dsx^F, however, cooperates with additional factors, such that binding of the same FBE sequences activates transcription (An et al., 1995a, 1995b). In this way, DSX can act as either a transcriptional activator or repressor (Suzuki et al., 2013).

In addition to the sex-specific role of the FB in the production of yolk proteins in females (Bownes, 1994), the FB is further involved in male courtship behaviour (Dauwalder, 2008; Lazareva et al., 2007). The *takeou*t (*to*) gene has been identified to be expressed male-specifically in the FB that surrounds the brain. Mutating *to* affects male courtship behaviour, and given it is known to interact with *fru*, it has been suggested the two may function together in the regulation of courtship behaviour. Male *to* mutant flies, whilst able to perform all steps of the courtship ritual to *wild type* females, initiated and maintained these behaviours at significantly lower rates (Dauwalder et al., 2002). Given expression of Fru^M in the CNS is needed to establish the ability to court, perhaps fat body factors, such as *to*, function with the CNS to regulate its function. Given the FB's endocrine function, it has been proposed that the FB secretes factors into the haemolymph which in turn interact with the brain. Indeed, due to the discovery of the Takeout protein in the haemolymph, this hypothesis has garnered prominence (Dauwalder, 2008; Lazareva et al., 2007). The model suggests that soluble circulating factors could play a role in the control of *Drosophila* sexual behaviour, as is the case in vertebrates where sexual behaviour is under hormonal control. Importantly, however, how these proteins cross or signal through the blood-brain barriers to interact with *fru* is unknown (Dauwalder, 2008).

A B

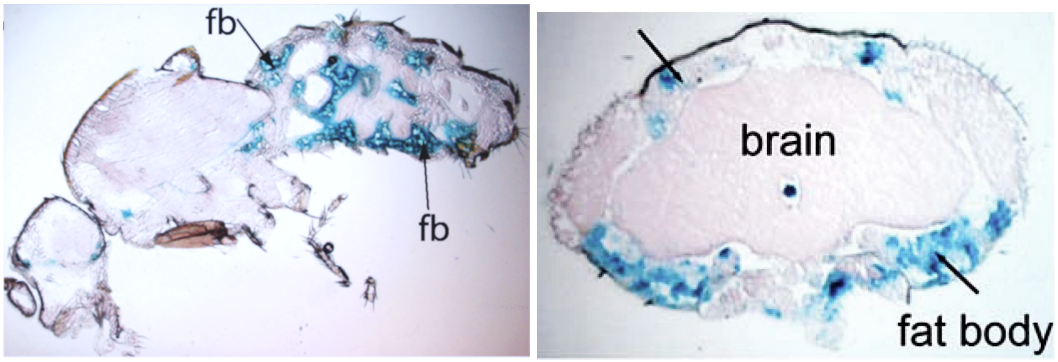


Figure 46 *Fat body expression in adult Drosophila melanogaster in the abdomen (A) and frontal section of fly head (B). For A, fat body specific Gal4 activity driven by a 3.1 Kb Lsp2 promoter fragment with UAS-lacZ reporter gene, staining in one-week old flies. The Larval serum protein (Lsp2) gene has been shown to be specifically expressed in the larval and adult fat body (Benes et al., 1990a). Image taken from Lazareva et al., 2007. For B, takeout-Gal4/ UAS-lacZ flies were stained with X-gal. Fat body cells are marked in blue, arrows. The fat body surrounds the brain. Image modified from Dauwalder et al., 2002.*

Overarchingly, the Clough et al., 2014 DSX-fat body DamID-seq study found that Dsx^{M} and Dsx^{F} bound thousands of the same targets in multiple tissues in males and females, although the same targets had varying sex- and tissue- specific functions. DSX occupancy was observed close to known DSX targets. Indeed, *Yp1* and *Yp2* showed strong DSX occupancy in the fat body and ovary, where these genes are highly expressed, but weak occupancy in S2 cells. The *bab1* locus showed strong DSX occupancy in all samples (Figure 47; Clough et al., 2014). The study found DSX may directly or indirectly regulate a broad set of transcription factor genes, including a number that regulate *dsx* expression (Clough et al., 2014). The types of target genes predicted by the analyses illustrated potentially why DSX influences such a diverse set of developmental processes. Target genes involved in paracrine (e.g. WNT and DPP) and endocrine (e.g. insulin and ecdysone) signalling were predicted, suggesting DSX expression can have far-reaching effects on the development of surrounding cells and

beyond. DSX is already known to modulate paracrine signalling pathways in the genital disc (Ahmad and Baker, 2002; Gorfinkiel et al., 2003; Keisman et al., 2001) and gonad (DeFalco et al., 2008; Oliver et al., 1993; Wawersik et al., 2005). Alongside signalling molecules such as neuropeptides, transcriptional regulators, a majority of which have sex-specific expression (Barmina et al., 2005; Chatterjee et al., 2011; Williams et al., 2008), were another major class of potential DSX targets. One could speculate that by activating or repressing transcription factors, DSX could delegate regulation to activate pathways that proceed largely without further input by DSX.

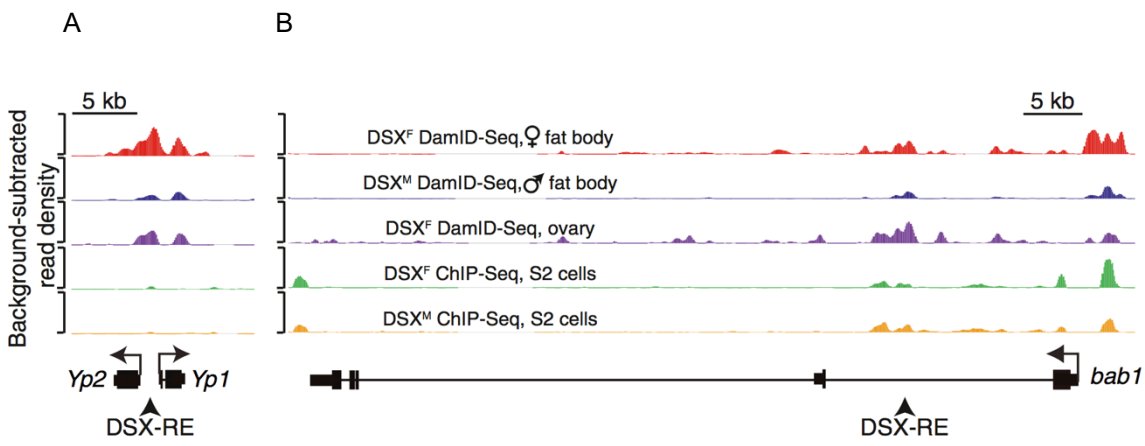


Figure 47 *DSX Occupancy and binding sites. Scaled read-density plots (background subtracted, arbitrary scale) for five replicated genome-wide analyses (as labelled). Binding shown to known DSX binding target genes: Yp1, Yp2 (A) and bab1 (B) loci. FlyBase gene models show transcription start site (directional arrows), coding exons (thick rectangles), noncoding regions (thin rectangles), introns (lines), and known DSX response elements (DSX-RE). Modified from Clough et al., 2014.*

6.1.2 NEUROPEPTIDES FUNCTION IN THE REGULATION OF PHYSIOLOGY AND BEHAVIOUR

Neuropeptides form a diverse range of signalling molecules, short chain-length peptides, synthesised through the enzymatic cleavage of larger polypeptide chain

precursors (Elphick et al., 2018; Jékely et al., 2018). Peptidergic communication was first discovered at the turn of the twentieth century (Bayliss and Starling, 1902) – in the context of peptide hormones secreted from endocrine glands. It was later recognised peptides may be synthesised in and secreted from neurons (Johnson, 1962; Knowles and Bern, 1966; Knowles, 1951; Olivecrona, 1954; Worthington, 1966), alongside 'classical' small-molecule neurotransmitters (Hökfelt et al., 1980). More recently, it has become clear that neuropeptides nuance neuronal communication, a facet paramount in behavioural plasticity (Jékely et al., 2018; Koh et al., 2003; Stein et al., 2007; Taghert and Nitabach, 2012; van den Pol, 2012).

As described in the introduction to this thesis, sex-specific splicing of *dsx* and *fru* in the SDH determines the formation of male- or female-specific neuronal circuitry. Fru^M and Dsx^M function together to specify male-specific circuitry (Billeter et al., 2006; Demir et al., 2005; Kimura et al., 2008; Rideout et al., 2007, 2010; Sanders and Arbeitman, 2008; Yamamoto, 2008; Yamamoto et al., 2013b), and Dsx^F to specify female-specific circuitry (Rideout et al., 2010; Rezával et al., 2012; Rezával et al., 2016). It is highly probable that both neuropeptides and their receptors are expressed in these circuits involved in the regulation of sexual behaviour (Anderson, 2016; Aranha et al., 2018; Ávila et al., 2011; Billeter et al., 2018; Carmel et al., 2016; Castellanos et al., 2013; Dickson, 2008; Jang et al., 2017; Kim et al., 2016; Kim et al., 2017; Li et al., 2011; Manoli et al., 2013; Neville et al., 2012; Pavlou et al., 2013; Sellami et al., 2015; Yamamoto et al., 2013). For instance, the neuropeptide Natalisin has been shown to reduce courtship initiation latency (Jiang et al., 2013). Indeed, other neuropeptides have further been implicated in the regulation of courtship behaviour: signalling by Neuropeptide F (NPF) (functionally related to the vertebrate neuropeptide Y, NPY, and

distinct from short neuropeptide F, sNPF), and Allatostatin B (AstB), result in increased courtship (Jang et al., 2017; Lee et al., 2006). Inversely, it has been shown that direct genetic ablation of NPF neurons resulted in decreased male courtship activity (Lee et al., 2006), possibly because NPF neurons are involved in detecting the female sex pheromone, whose production is regulated by insulin/ insulin-like growth factor signalling (IIS) (Gendron et al., 2014; Kuo et al., 2012). Indeed, females with increased IIS appear more attractive to males and vice versa (Kuo et al., 2012).

6.1.3 Background on key neuropeptides of interest

Taking a candidate gene approach is an effective method to select suitable target genes based on known biological, physiological, or functional relevance to the experimental question. Whilst this approach is limited by existing knowledge, it nonetheless provides an important starting point for further examination of putative targets. Alongside the rich list of Dsx^M targets identified in the TaDa brain and head analyses, we also independently employ a candidate gene approach to select a small number of putative DSX target genes for further analysis. To do this, we combine the results from a published ChIP-seq experiment on S2 cells expressing tagged Dsx^M or Dsx^F (Clough et al., 2014), and a yeast-one hybrid screen of *Drosophila* neuropeptide regulatory element DNA baits versus a *Drosophila* TF library (Hens, 2017 unpublished). This method allowed for the identification of 'high confidence' DSX targets for which, here, we present the relevant background literature. Comparisons of these datasets, and these targets specifically, with the DSX TaDa brain and head datasets is further carried out here.

DIURETIC HORMONE 31 (*DH31*)

Dh31, 117 amino acids (AAs) in length, belongs to the diuretic hormone class 2 family in *Drosophila* and is implicated in a broad range of functions. *Dh31* has been implicated in sexual behaviour in that signalling by *hector* (the receptor to *Dh31*) results in increased courtship. The *hector* receptor is broadly expressed in the adult brain and is thought to overlap with *fru* neuronal populations (Li et al., 2011). The receptor to *Dh31* has also been shown to mediate temperature preference rhythm, a function essential for maintaining homeostasis (Goda et al., 2018).

Dh31 is the fly homolog of the vertebrate neuropeptide calcitonin, and acts as a circadian wake-promoting signal that alerts the fly in anticipation of dawn. It is expressed in a subset of dorsal circadian clock neurons that also express the receptor for the circadian neuropeptide pigment-dispersing factor (PDF). Analysing loss-of-function and gain-of-function mutants shows Dh31 mediates sleep-specific circadian output in *Drosophila* by suppressing sleep late at night (Kunst et al., 2014). *Dh31* is also expressed in one subtype of enteroendocrine cells: evolutionarily conserved gastrointestinal cells in the posterior midgut. RNAi knockdown of *Dh31* gene expression in the midgut has been shown to increase lifespan (Takeda et al., 2018).

NEUROPEPTIDE LIKE PRECURSOR 1 (*NPLP1*)

Nplp1 (or CG3441), 309 AAs in length, is a neuropeptide precursor gene that encodes eleven putative peptides. Four of these peptides – NAP, MTYamide, IPNamide, and VQQ – were isolated and identified in pioneering peptidomics experiments using the

Drosophila CNS (Baggerman et al., 2005; Overend et al., 2012). *Nplp1* is conserved across several insect orders (Hummon et al., 2006; Li et al., 2008; Riehle et al., 2002). Characterising *Nplp1* in *Sarcophaga bullata* (Verleyen et al., 2009), the grey flesh fly, showed localisation of *Nplp1* to 28 neurons in the larval CNS. A similar neuronal pattern was conserved for *Nplp1-3* in the *Drosophila* larval CNS, and noted 18 h after egg lating (AEL) (Figure 48A) (Baumgardt et al., 2007; Verleyen et al., 2004). In adult *Drosophila*, *Nplp1* expression is exclusively restricted to the brain and thoracicoabdominal ganglion (flyatlas.org), and in larvae, *Nplp1* expression is exclusively restricted to the CNS. NAP and MTYamide are both expressed in the larval CNS and IPNamide in the ventral ganglion of the third-instar larval and adult brain (Figure 48B) (Verleyen et al., 2004).

Through screening peptide libraries in *Drosophila* S2 cells, it has been shown the *Nplp1*-VQQ peptide is a ligand for *Gyc76c*. In the adult fly, *Gyc76c* expression is highest in immune and stress-sensing epithelial tissues including Malphigian tubules and the midgut. It has been stipulated therefore that *Nplp1*-VQQ acts in the immune/ stress response (Overend et al., 2012). Further, *Nplp1* may have a role in regulating hindgut activity (Fontana and Crews, 2012), given its localisation in MP1 neurons: CNS midline neurons that use the neurotransmitter PDF in larval segments A8/9, and that likely innervate the hindgut.

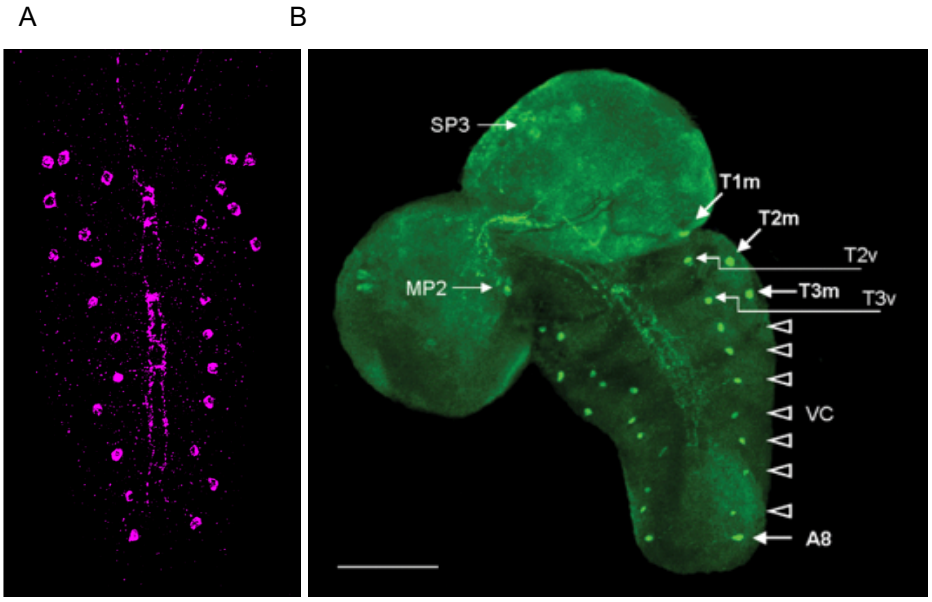


Figure 48 *Wild type Nplp1 VNC expression, 18 h AEL. w^{1118} strain, proNplp1 antibody staining (image from Baumgardt et al., 2007) (A). IPNamide immunoreactivity in Drosophila third-instar larval CNS (B). Most expression noted in the ventral ganglion, abdominal A8, ventral cells (VC, open arrows), and in the thoracic T1-3m and T2-3v cells. Scale bar = 100µm (image from Verleyen et al., 2004).*

TACHYKININ 1 (*TK1*)

Insect tachykinins have been extensively studied and identified to have roles in a number of physiological and behavioural processes. In *Drosophila*, the neuropeptide Tachykinin (*Tk*), 289 AAs in length, is expressed in a male-specific small cluster of $FruM^{+}$ neurons (Figure 49). Most notably, activation of this neuronal cluster increases inter-male aggression, whilst not affecting male-female typical courtship behaviours. Further, removing selected sensory or contextual cues as means to reduce aggressiveness was overcome by *Tk* neuronal activation, suggesting its function is to promote aggressive arousal or motivation (Asahina et al., 2014). *Tk* has further been proposed to be essential for pheromone detection in a gustatory neural circuit. *Drosophila tachykinin* (DTK) signalling is thought to inhibit courtship behaviours

through communicating anti-aphrodisiac pheromone signals from the gustatory neurons on the foreleg to central brain neurons. The pheromone remains on the female to prevent further approaches by other males. In this way, sex specific DTK signalling increases the reproductive success of the first male (Shankar et al., 2015).

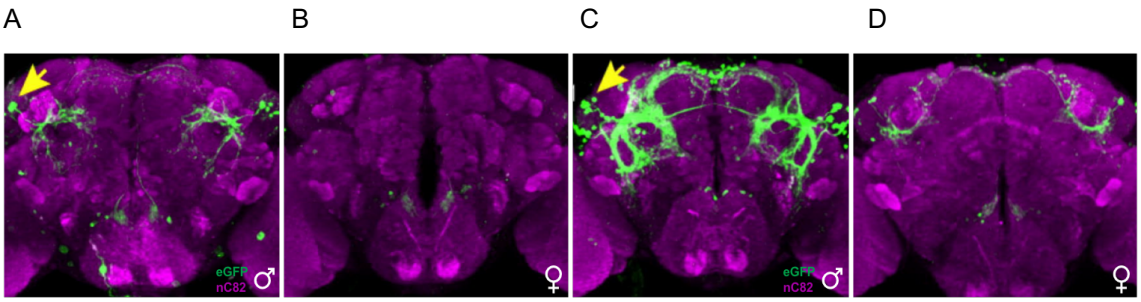


Figure 49 *Asahina et al., 2014 identified two Drosophila Tachykinin-Gal4 lines that strongly increased inter-male aggression. Expression in brains of Tk-Gal4-1; UAS-mCD8; GFP male (A) and female (B) adult animals, and Tk-Gal4-2; UAS-mCD8; GFP male (C) and female (D) adult animals. Immunostaining with anti-GFP antibody shown in green and neuropil marker counterstained with antibody to nC82 shown in magenta (Image from Asahina et al., 2014).*

Dh31, *Tk1* and *Nplp1* peptides are each expressed in large populations of neurons, pointing towards pleiotropic roles in physiology and behaviour. Both *Dh31* and *Tk1* have been shown to be involved in some way with sexual behaviour (Li et al., 2011; Shankar et al., 2015). Investigating their association with *dsx* could help to delineate their role in courtship.

6.2 AIMS

1. Analyse published DSX-fat body DamID-seq data (Clough et al., 2014) using the same computational pipelines and parameters used to assess DSX brain and head TaDa datasets.

2. Functionally compare DSX-fat body DamID-seq datasets (Clough et al., 2014) and DSX brain and head TaDa-seq datasets (this study) for tissue-specificity.

3. Take a candidate gene approach to follow up a small number of putative DSX target genes through exploratory expression analyses.

6.3 METHODS

6.3.1 PROCESSING DSX-FAT BODY DAMID-SEQ DATASETS (CLOUGH ET AL., 2014)

The DamID-seq DSX-fat body occupancy datasets were available via the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2012) series accession number GSE49480. fastq files were downloaded from the online server and henceforth assessed in the same manner in which datasets from DSX brain and head TaDa were treated in this study. We used .fastq files as input to damidseq_pipeline (Marshall and Brand, 2015). The final log2 ratio files were output in bedGraph format. Samples were aligned to *Drosophila melanogaster* Ensembl BDGP6 to allow direct comparison with brain and head TaDa datasets. We used PAVIS version 02-05-2018 (Huang et al., 2013) for annotating and visualising the DamID-seq datasets, using the called peaks .bed output from find_peaks (Marshall and Brand, 2015) at FDR 0.05 as input files, annotating genes with *D. melanogaster* flybase version R6.01. Occupancy analyses were also completed in PAVIS (Huang et al., 2013). We implemented i-*cis*Target (Herrmann et al., 2012; Imrichová et al., 2015) for the prediction of regulatory features. Peaks were searched using the 'full analyses' containing the entire motif collection (version 5.0): PWMs, modERN, TF binding sites, non-TF binding sites, histone modifications and RNA polymerase. Genomic coordinate .bed peaks were supplied, and the *Drosophila melanogaster* Ensembl BDGP6 reference genome specified.

6.3.2 YEAST-ONE HYBRID (Y1H) SCREENING OF PROTEIN-DNA INTERACTIONS

The y1H system is a technique used to study interactions between transcription factors and DNA (Deplancke et al., 2004). y1H relies on the general principle of the yeast-two

hybrid (y2H) assay, where proteins are exogenously expressed in yeast, and *in vivo* interactions are measured by downstream activation of reporter gene constructs. In y2H, the physical interaction between two proteins (the 'bait' and 'prey', which are fused to the AD and DBD domain of the Gal4 transcriptional activator, resectively) allows for the combination of the AD and DBD domains, generating a functional transcriptional unit that turns on expression of a reporter gene. In y1H, detection is based on the physical interaction between a single prey protein and a bait DNA sequence upstream of a reporter gene. The prey is fused to a transcriptional AD and tested against a library of bait DNA sequences as putative TF response elements. Positive protein-DNA interactions bring the prey-AD into close contact with the promoter element, activating downstream transcription of the reporter gene.

6.3.3 DSX ChIP-seq in S2 cells (Clough et al., 2014)

Clough et al., 2014 performed ChIP-seq on S2 cells transfected with either v5-tagged Dsx^M (pMT5.1-DSXM-V5-HisB) or Dsx^F (pMT5.1-DSXF-V5-HisB) constructs (Garrett-Engele et al., 2002) and pCoBlast (Invitrogen, Carlsbad, CA, USA) as the selection plasmid using Effectene (Qiagen, Valencia, CA, USA). ChIP was done using anti-V5 tag monoclonal antibody (Inivitrogen, Carlsbad, CA, USA) on Protein G coupled Dynabeads (Invitrogen, Carlsbad, CA, USA). Three biological replicates for each ChIP sample were performed. ChIP-seq libraries were constructed with the genomic DNA sample preparation kit and were sequenced on a GA1 (Illumina, San Diego, CA, USA). Reads from all biological replicates were pooled prior to mapping and alignment to the genome, FlyBase annotation version 5.46 using Bowtie 0.12.7 (Langmead et al., 2009). The Window Tag Density (WTD) method of peak calling from the ChIP-seq analysis program SPP (version 1.11) called peaks with an FDR of 0.01

(Kharchenko et al., 2008). The WTD method uses a sliding window calculating the geometric average of binding positions on the positive and negative strand. The window size is estimated by SPP based on binding characteristics. These datasets were available to download via the NCBI Gene Expression Omnibus (GEO) (Barrett et al., 2012) series accession number GSE49480. We downloaded two .bed.gz files of called peaks for Dsx^{M} and Dsx^{F} biological replicates as well as the two associated .bedgraph enrichment profiles for these samples.

6.3.4 Gateway cloning of Split-Gal4 neuropeptide^{AD} DNA constructs

The Gateway Cloning system (Invitrogen) is a versatile cloning technique for the study of gene expression (Chee and Chin, 2015). The system is an *in vitro* version of the integration and excision recombination reactions that take place when lambda phage infects bacteria. When recombination of attachment sites from the phage (*attP*) and the bacteria (*attB*) occurs, the phage integrates into the bacterial genome flanked by two new recombination sites (*attL* left and *attR* right). The technology involves two important reactions: BP and LR. In the BP reaction, the Gateway "Entry Clones" are generated in two steps. First, the *attB1* and *attB2* sequences are added to the 5' and 3' end of the gene fragment using gene-specific PCR primers and PCR amplification. Secondly, the PCR amplification products are mixed with the Gateway Donor vector, pDONRP4-P1R, and the proprietary 'BP Clonase' enzyme mix. BP Clonase catalyses the recombination and insertion of the *attB*-PCR product into the attP recombination sites in the pDONRP4-P1R vector. Once the fragment is incorporated into the donor vector, it is termed Entry Clone, with *attL* sites flanking the gene fragment. Entry Clones are then cloned into the *attR*-site containing destination vector, pDSPp65AD, generating pDSPp65AD-gene fragment DNA constructs, with the Gateway LR

reaction. The proprietary 'LR Clonase' is used, generating an "Expression Clone" containing the neuropeptide RE (Figure 50).

We requested and obtained the neuropeptide REs of interest from K.Hens (2017) as Gateway-compatible Entry clones which contained the neuropeptide RE flanked with *attL* sites in the pDONRP4-P1R vector (Invitrogen) (Hens et al., 2011). We thus employed the Gateway cloning technique for the generation of Split-Gal4 neuropeptide^{AD} genetic constructs. Entry clones were cloned into the *attR*-site containing destination vector, pDSPp65AD, using the proprietary LR Clonase in the Gateway LR reaction, generating pDSPp65AD-neuropeptide^{AD} DNA construct Expression clones.

6.3.5 Assessing Split-Gal4 neuropeptide^{AD} DNA constructs

DNA Sanger sequencing was employed to assess the integrity of the neuropeptide^{AD} Split-Gal4 constructs. 20 bp forward sequencing primers within the neuropeptide gene RE for each construct, and a common 20 bp reverse sequencing primer within the p65AD vector backbone were designed to do this. Namely, *Dh31*-F1 (CCC ATC GAG CAT CGT CTC TC), *Nplp1*-F1 (GAT GGT TCG ATG TGT GCT GC) and *Tk1*-F1 (CGA TCG CGA TGG TTG CAA TT) forward primers, and the common reverse primer p65AD-Rev1 (TTT ATA CCG CTG CGC TCG AT). All primers were designed, mapped and cloning planned in Geneious 10.2.3 (Build 2017.07.10).

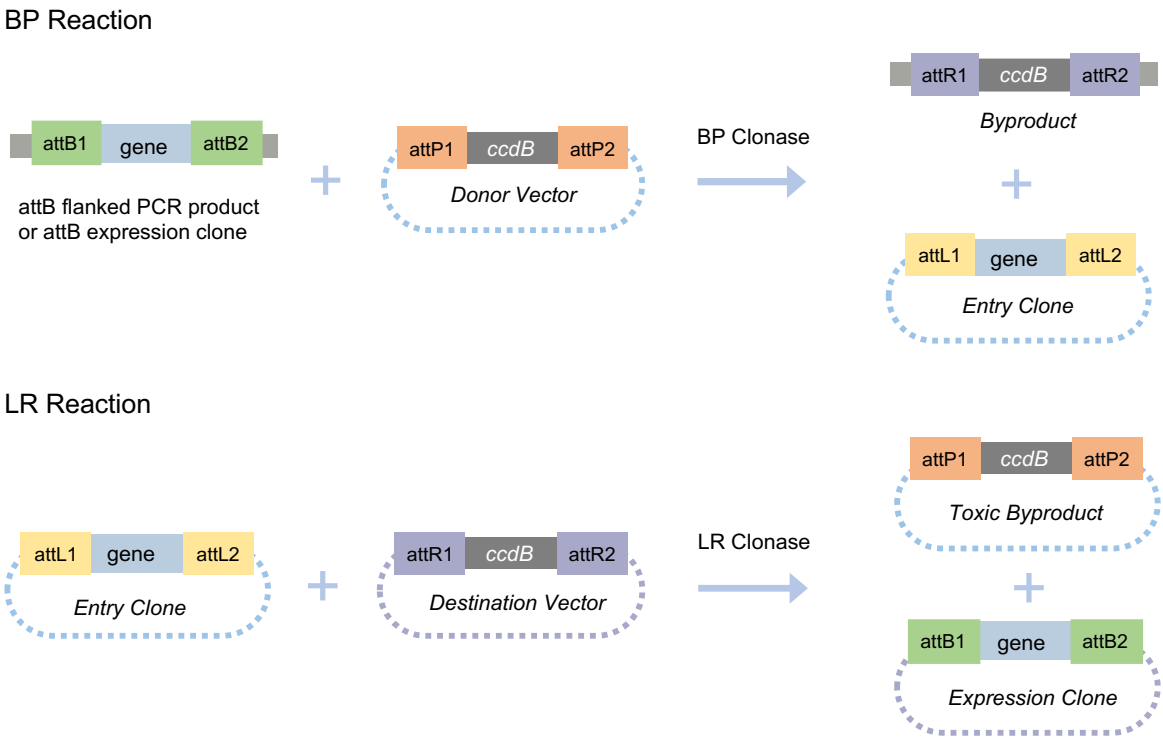


Figure 50 *Schematic of Gateway Cloning method. The BP reaction generates an attL-flanked Entry Clone, and the LR reaction generates an Expression Clone containing all the components required for gene expression.*

6.4 RESULTS

6.4.1 DSX BRAIN AND HEAD TADA (THIS STUDY) VERSUS DSX-FAT BODY DAMID (CLOUGH ET AL., 2014)

In this study we conducted two TaDa genome-wide experimental screens as a means to understand how DSX impinges function in the nervous system. Our TaDa profiling of DSX-expressing cells in the brain and whole head of adult *D. melanogaster* flies revealed a tendency towards tissue-specific target genes (neural versus fat body cells). We compare the results of these screens with previously published DamID-seq datasets profiling the fat body tissue where *dsx* is known to be widely expressed (Bownes et al., 1978; Clough et al., 2014) to delineate this tissue-specificity further. The fat body organ from male and female adults was chosen specifically for DamID profiling since *dsx* is known to play a role in maintaining sexually dimorphic gene expression there. The Clough et al., DamID-seq experiments were conducted using low basal expression in the absence of a Gal4 driver to segway problems associated with known Dam expression toxicity and experimental artefacts due to DSX overexpression (2014). Our comparisons aim to functionally assess the relationships between DSX gene enrichment across tissue types allowing us to potentially delineate neural- and fat body- specific genes in the *D. melanogaster* adult.

6.4.2 PROCESSING DSX-FAT BODY DAMID-SEQ DATA (CLOUGH ET AL., 2014)

The DSX-fat body DamID-seq datasets (Clough et al., 2014) were analysed bioinformatically in the same manner as the DSX brain and head TaDa-seq datasets

(this study) to allow for direct comparison. The 'per base sequencing quality' was assessed with FastQC version 0.11.5 (Andrews, 2010) and visualised using MultiQC version 0.9 (Ewels et al., 2016) and had reliable base calls across all reads in each biological replicate assessed. Further, the 'sequence length distribution' confirmed the majority of sequencing reads were of similar length, and the average sequence quality per read was high (Phred scores: 30-40). NGS datasets were processed using damidseq_pipeline (Marshall and Brand, 2015) at the single-end sequencing resolution the experiment was completed at.

PEAK CALLING USING FIND_PEAKS (MARSHALL AND BRAND, 2015)

Peaks were called as previously (find_peaks pipeline, Marshall and Brand, 2015) at two defined False Discovery Rates: 0.01 and 0.05. Table 18 summarises peak numbers for the fat body Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates at FDR 0.01 and FDR 0.05. Peak numbers for Dsx^{M}-Dam and Dsx^{F}-Dam replicates in brain and head TaDa at the same FDRs are included below for comparison. We saw sound reproducibility in terms of similar numbers of peaks called amongst both the fat body Dsx^{M}-Dam biological replicates (1 and 2) as well as both fat body Dsx^{F}-Dam biological replicates (1 and 2), at both FDRs (unpaired t-test, $p<0.05$). Overall, we saw a similar number of called peaks in the DSX-fat body experiment (Clough et al., 2014) and our DSX head TaDa experiment which profiled both DSX brain neural cells and DSX-expressing fat body surrounding the brain.

DSX fat body DamID-seq (Clough et al., 2014)				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	1888	1615	1749	1716
FDR 0.05	2199	2525	2575	2495
DSX brain TaDa-seq (*this study*)				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	449	538	473	113
FDR 0.05	821	1255	1276	306
DSX head TaDa-seq (*this study*)				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	784	890	1203	1720
FDR 0.05	2104	2551	2067	1838

Table 18 *Summary of called peak numbers in fat body Dsx^{M}-Dam and Dsx^{F}-Dam biological replicates (top, Clough et al., 2014). For comparison, summary of called peak numbers in Dsx^{M}-Dam and Dsx^{F}-Dam replicates in brain TaDa (middle) and head TaDa (bottom). Peak calling completed using find_peaks (Marshall and Brand, 2015), and peaks called at FDR 0.01 and 0.05.*

Area-quantitative Venn diagrams visually summarise the overlap between called peak numbers in biological replicates. Assessing the datasets in this way reveals not only whether numbers of called peaks were similar, but also whether the same peaks were called across biological replicates. We saw statistically significant reproducibility between both biological replicates in both the Dsx^{M}-Dam samples (Figure 51A) and Dsx^{F}-Dam samples (Figure 51; unpaired t-test, $p<0.05$).

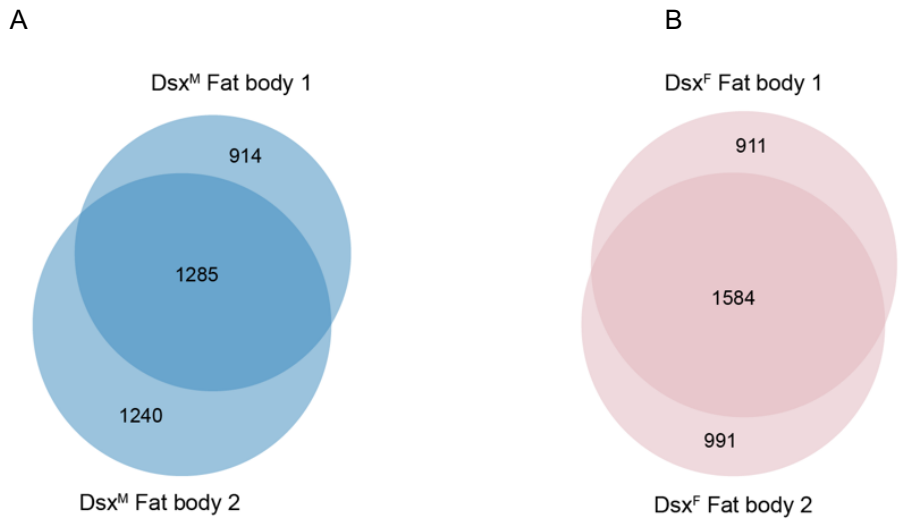


Figure 51 *Area-quantitative Venn diagram of overlapping peaks in Dsx^{M}-Dam fat body replicate one and two (A), and Dsx^{F}-Dam fat body replicate one and two (B), Clough et al., 2014. Peaks called to FDR 0.05 (find_peaks, Marshall and Brand, 2015). Venn diagrams generated in BioVenn (Hulsen et al., 2008).*

GENERATING GENOME-SCALE CANDIDATE GENE LISTS

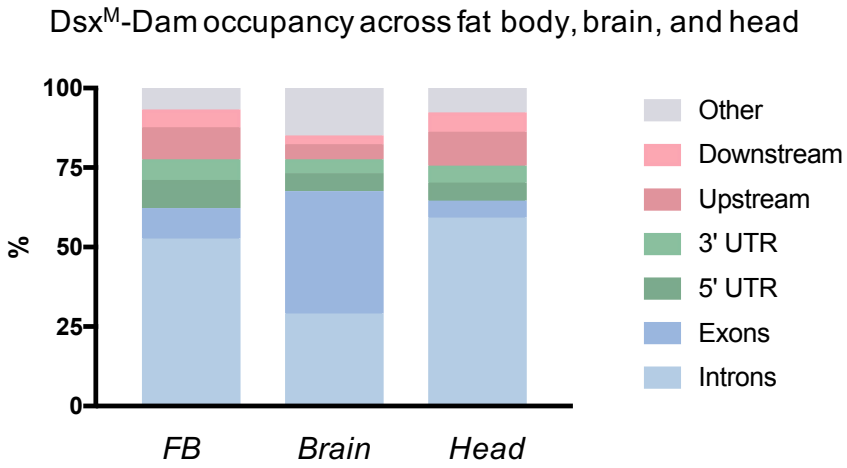
We employed the gene body + 1 Kb upstream of TSS + 1 Kb downstream of TTS peak to gene annotation method, as with brain and head TaDa, for the same reasons as discussed in the previous chapter. This method further allowed us to directly compare the gene lists for the DSX-fat body DamID-seq experiments at FDR 0.05 (Clough et al., 2014) with our DSX brain and head TaDa-seq experiments.

Across both fat body Dsx^{M}-Dam replicates (1 and 2) and Dsx^{F}-Dam replicates (1 and 2) we saw statistically significant peak overlap (unpaired t-test, $p<0.05$). We therefore assigned genes to peaks common to both the biological replicates. For fat body Dsx^{M}-Dam, we found that 93% of the loci were associated with genes, with occupancy significantly overrepresented in intronic regions (Figure 52A). We saw similar results

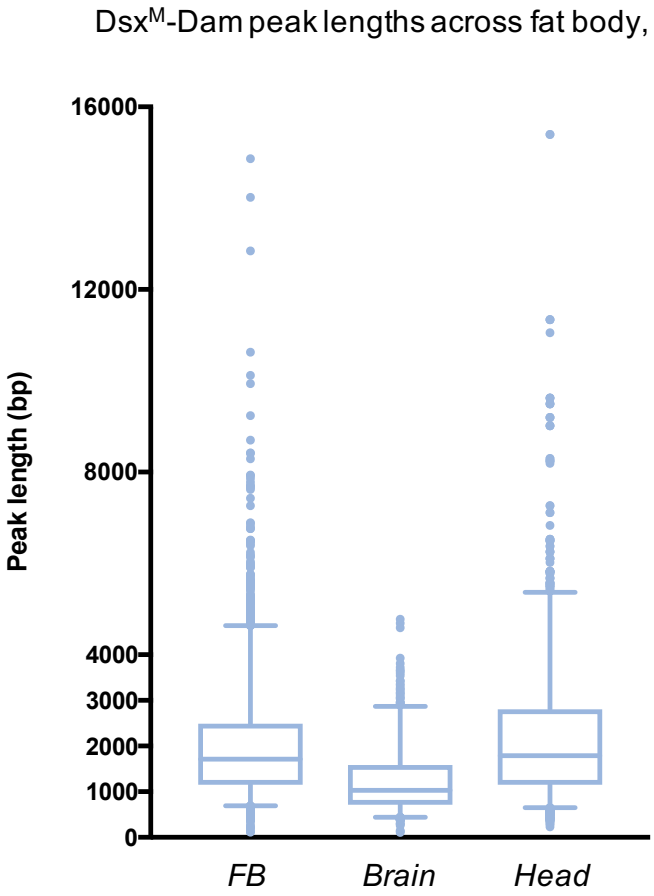
for the fat body Dsx^F-Dam samples: 91% of the loci were associated with genes, and occupancy was also significantly overrepresented in intronic regions (Figure 52C). These results differ significantly from our DSX brain TaDa profiling experiment where we saw a bias towards exonic regions in both Dsx^M-Dam and Dsx^F-Dam replicates. They are however similar to our Dsx^M-Dam samples in our head TaDa profiling experiment (and one of two Dsx^F-Dam biological replicates), this makes sense because head TaDa also profiles DSX-expressing fat body tissue encapsulating the brain. In Figure 52, the 'Other' category denotes the percentage of peaks not associated with genes.

Comparing peak lengths across Dsx^M-Dam datasets, interestingly the fat body and head share similar binding characteristics, median for both ~1,600 bp and IQR ~1,100-2,400 and ~1,200-2,700 respectively. This could explain the coherence we observe in intronic versus exonic occupancy in the samples (Figure 52A and B). Whilst the spread of peak lengths appear larger for both Dsx^M-Dam and Dsx^F-Dam fat body DamID-seq biological replicates compared to their sex-specific brain and head TaDa-seq datasets, the difference is particularly pronounced for the Dsx^F-Dam biological replicates where a number of peaks were called between 6,000-9,000 bp. We infer the differences in occupancy across Dsx^F-Dam biological replicates is in line with the varying peak length IQRs observed in biological replicates, and hence the biological variation we observe in these replicates.

A



B



(figure continued on next page)

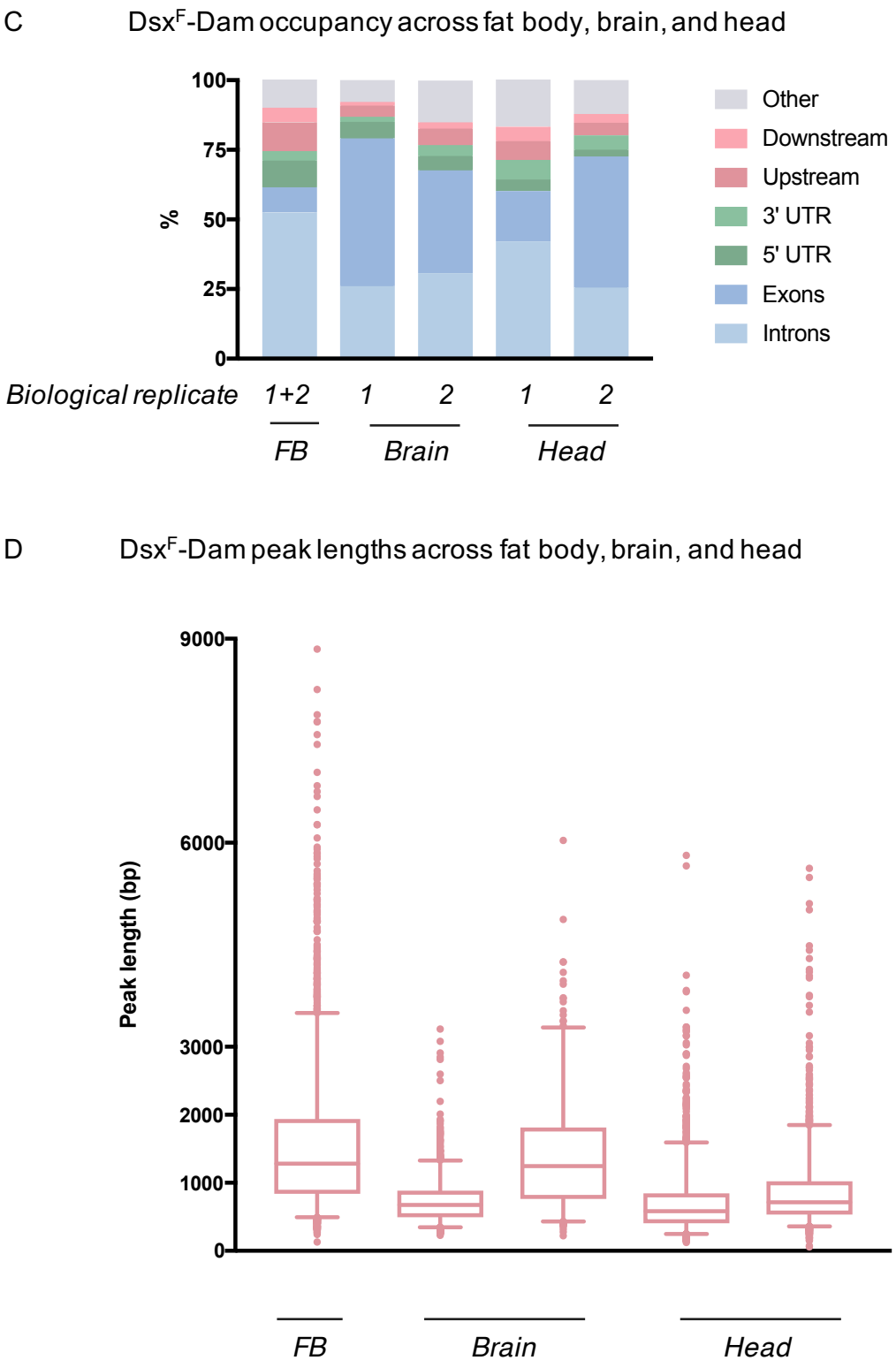


Figure 52 *Distribution of peaks in relation to genes, occupancy analyses, across the fat body 'FB' (Clough et al., 2014), brain TaDa and head TaDa Dsx^{M}-Dam replicates (A) and Dsx^{F}-Dam replicates (C). Biological replicates one and two combined for Dsx^{M}-Dam given their statistical reproducibility in both brain and head TaDa; Dsx^{F}-Dam biological replicate one and two combined in fat body dataset.*

For fat body, both Dsx^M and Dsx^F isoform occupancy was significantly overrepresented in intronic regions.

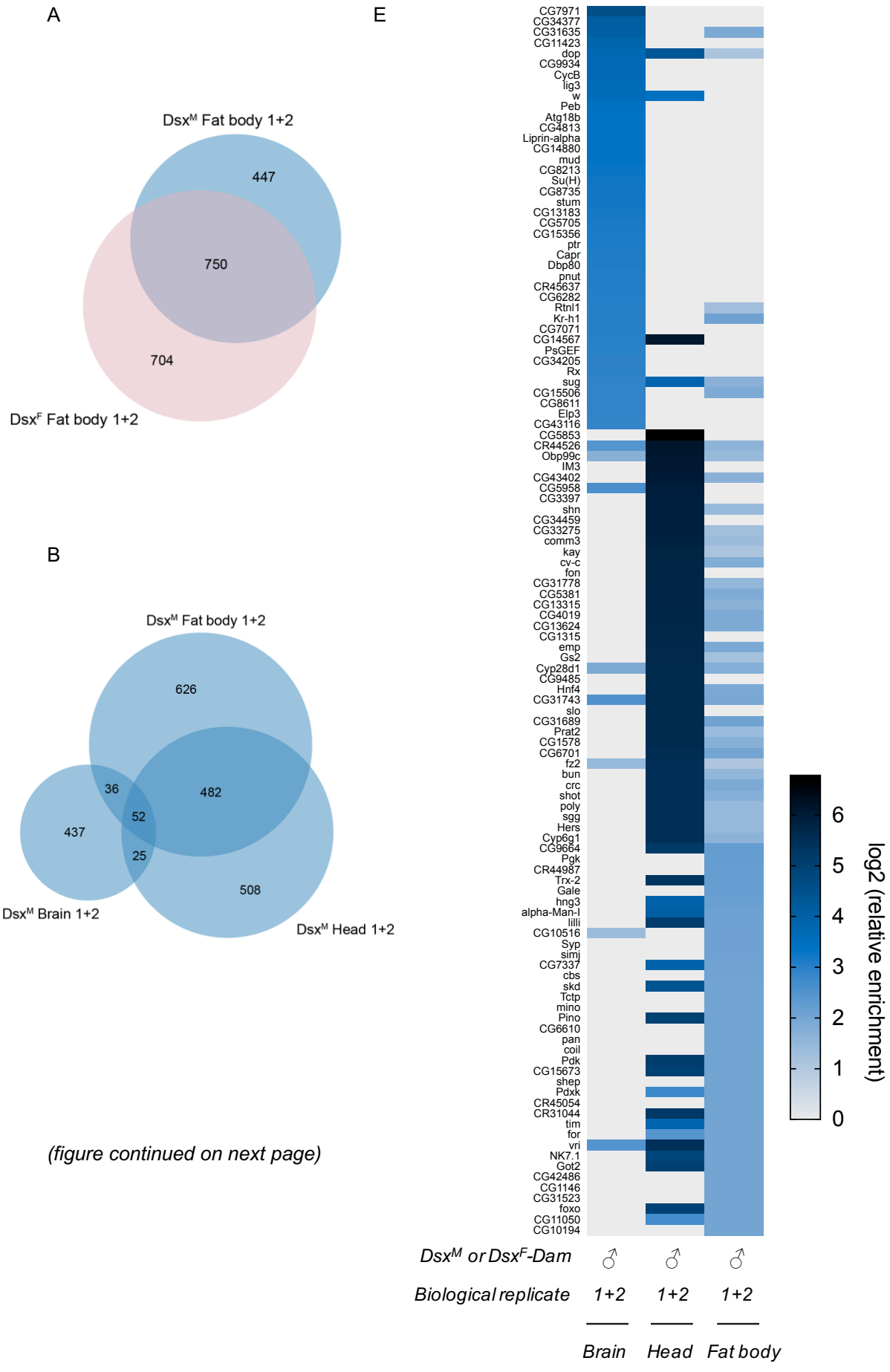
Box plots compare called peak lengths (bp) in fat body, brain, and head Dsx^M-Dam biological replicates (B), and Dsx^F-Dam biological replicates (D). Box represents the Interquartile Range, middle line is the median value, and the whiskers represent the 5 to 95 percentiles of the data. Dsx^F-Dam biological replicate one and two combined in fat body dataset given their statistical reproducibility.

6.4.3 DSX CANDIDATE TARGET GENES HAVE A TENDENCY TOWARDS TISSUE-SPECIFICITY

Analysis of called peaks across fat body Dsx^M-Dam biological replicates showed a high degree of overlap, as did the fat body Dsx^F-Dam biological replicates (Clough et al., 2014; unpaired t-test, $p<0.05$). We thus annotated genes to peaks common to both Dsx^M-Dam biological replicates, and the same for Dsx^F-Dam biological replicates. Firstly, comparing generated Dsx^M-Dam and Dsx^F-Dam gene lists, we saw a significant proportion (40%) of the genes common to both isoforms (Figure 53A). This result was in line with the original paper where both DSX isoforms were reported to have similar gene occupancy patterns (Clough et al., 2014). We completed a series of further comparisons between the DamID-seq experiments profiling the DSX-fat body isoforms (Clough et al., 2014), and our brain and head TaDa experiments profiling DSX neural tissue. Firstly, looking at the Dsx^M-Dam isoform specifically, and comparing the fat body dataset (Clough et al., 2014) with our brain and head TaDa datasets, we see significant overlap (534 genes, 26%) between the fat body dataset and head TaDa dataset (unpaired t-test, $p<0.05$). This could have been predicted as head TaDa also profiles DSX-expressing fat body tissue encapsulating the brain in the *Drosophila* head. Significantly less overlap is noted between the fat body and brain TaDa profiling DSX neural cells (88 genes, 4%) (Figure 53B).

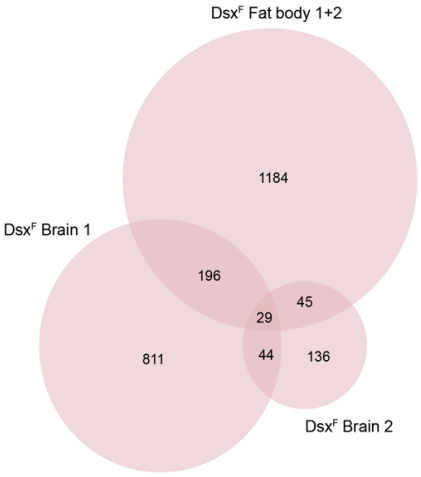
Given the variation we observed within the Dsx^F-Dam biological replicates in brain and head TaDa, it is impossible to draw meaningful biological conclusions about this isoform. However, to gauge how different the datasets were we compared the common peaks/ genes from the Dsx^F-Dam fat body sample (Clough et al., 2014), with both the Dsx^F-Dam individual biological replicates in brain (Figure 53C) and head TaDa (Figure 53D). In the former, we saw a small overlap between the fat body Dsx^F-Dam and brain TaDa Dsx^F-Dam biological replicate one (225 genes, 9%), and two (73 genes, 3%). Indeed, there is a slightly higher degree of overlap, on average, between the fat body Dsx^F-Dam (Clough et al., 2014) and the Dsx^F-Dam biological replicates from head TaDa. 305 genes (8%) with biological replicate one, and 171 (5%) with biological replicate two.

Comparing the top forty genes enriched across the Dsx^M-Dam brain, head and fat body datasets (Figure 53E) affirmed our observations that DSX putative target genes have a tendency towards tissue-specificity. We observe a significant gene overlap between Dsx^M-Dam head and fat body datasets, and far fewer common genes between these two datasets and the brain dataset. This is in line with expectations given the Dsx^M-Dam head dataset profiles fat body cells encapsulating the brain. Genes that are enriched in the head and fat body datasets seem to have a broad mode of function. For example, genes include schnurri (*shn*) that encodes a zinc finger C_2H_2 TF involved in Dpp signalling (Torres-Vazquez et al., 2000), thioredoxin-2 (*Trx-2*), involved in lifespan and oxidative stress (Svensson and Larsson, 2007), timeless (*tim*), involved in circadian rhythm feedback (Glossop et al., 1999), and the *Drosophila* forkhead transcription factor O (FOXO) involved in insulin signalling (Jünger et al., 2003).



(figure continued on next page)

C



D

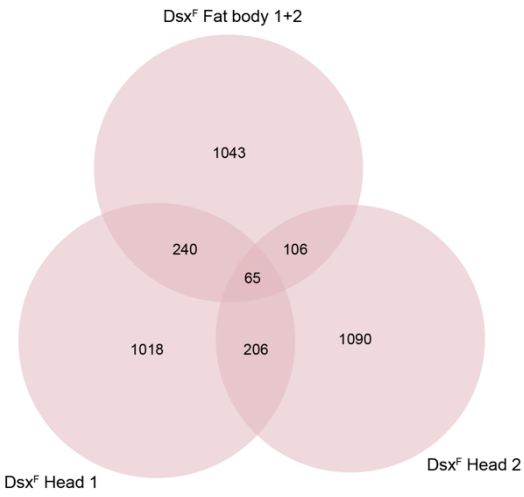


Figure 53 *Area-quantitative Venn diagrams overlapping generated gene lists between DSX-fat body DamID-seq Dsx^M-Dam and Dsx^F-Dam, Clough et al., 2014 (A), between fat body Dsx^M-Dam (Clough et al., 2014), brain TaDa Dsx^M-Dam, and head TaDa Dsx^M-Dam (B), between fat body Dsx^F-Dam (Clough et al., 2014) and brain TaDa Dsx^F-Dam biological replicate one and two (C), and between fat body Dsx^F-Dam (Clough et al., 2014), and head TaDa Dsx^F-Dam biological replicate one and two (D). For fat body samples (Clough et al., 2014), genes annotated to peaks common to both biological replicates in Dsx^M-Dam and Dsx^F-Dam samples. For Dsx^M-Dam samples in both brain and head TaDa, genes annotated to peaks common to both biological replicates. Area-quantitative Venn diagram generated in BioVenn (Hulsen et al., 2008).*

Heat map directly comparing the top forty enriched genes in Dsx^M-Dam brain, head, and fat body samples (E). Genes are ranked according to log2 relative enrichment. Peaks/ genes common to both Dsx^M-Dam biological replicates were assessed in the brain, head and fat body datasets. Heat map made in PRISM v7.0d.

6.4.4 Comparing candidate gene function in DSX- fat body, brain and head using GO

Gene Ontology enrichment analyses were completed on gene lists derived from fat body Dsx^M-Dam and Dsx^F-Dam samples to firstly gauge, and secondly compare, their functional classification. For both fat body Dsx^M-Dam and Dsx^F-Dam, we conducted the enrichment analysis on gene lists derived from peaks common to both biological replicates. For fat body Dsx^M-Dam, 256 enriched GO terms were defined under the *biological process* ontology. As reported in the original paper (Clough et al., 2014),

enriched terms were identified for a plethora of different coherent groups of genes in ontologies, supporting the idea DSX controls a large range of pathways and functions. Similar to both the Dsx^{M}-brain and Dsx^{M}-head enrichment analyses, enriched terms 'neurogenesis' ($p<0.01$, 137 gene matches, 11%) and 'nervous system development' ($p<0.01$, 176 gene matches, 15%) were identified. Alongside this, however, terms such as 'regulation of metabolic process' ($p<0.01$, 249 gene matches, 21%) and 'response to stress' ($p<0.01$, 166 gene matches, 14%) specific to fat body function were defined here and in Dsx^{M}-head analysis. Figure 54A lists a subset of 17 of these enriched terms, with key functional terms highlighted, ranked by number of gene matches with redundant terms removed. Ten enriched terms were defined in the *cellular component* ontology, where 'cytoplasm' ($p<0.01$, 458 gene matches, 38%) was the most significantly enriched term. 24 enriched terms were identified in the *molecular function* ontology including 'DNA-binding transcription factor binding' ($p<0.01$, 17 gene matches, 2%).

For fat body Dsx^{F}-Dam, 279 enriched GO terms were defined under the *biological process* ontology. As with the fat body Dsx^{M}-Dam dataset, and Dsx^{M}-brain and Dsx^{M}-head enrichment analyses, enriched terms 'neurogenesis' ($p<0.01$, 170 gene matches, 12%) and 'nervous system development' ($p<0.01$, 205 gene matches, 14%) were identified. Similarly, 'regulation of metabolic process' ($p<0.01$, 272 gene matches, 19%) and 'response to stress' ($p<0.01$, 185 gene matches, 13%) specific to fat body function were defined here. Figure 54B lists a subset of 16 of these enriched terms ranked by number of gene matches with redundant terms removed. 12 significantly enriched terms were defined under the *cellular component* ontology, and ten in the *molecular function* ontology. Interestingly in the latter, terms such as 'cofactor binding' ($p<0.01$, 69 gene matches, 5%) and 'coenzyme binding' ($p<0.05$, 45 gene matches, 3%)

were enriched. Given fat body Dsx^{M}-Dam and Dsx^{F}-Dam bind a largely overlapping subset of genes, it is not surprising that functional enrichment terms do not differ significantly. See Appendix for *cellular component* and *molecular function* ontologies (redundant terms removed) for fat body Dsx^{M}-Dam and Dsx^{F}-Dam respectively. GO analyses were completed on specific gene lists uploaded to FlyMine v45.1, the Holm-Bonferroni correction was applied, maximum p-value <0.05.

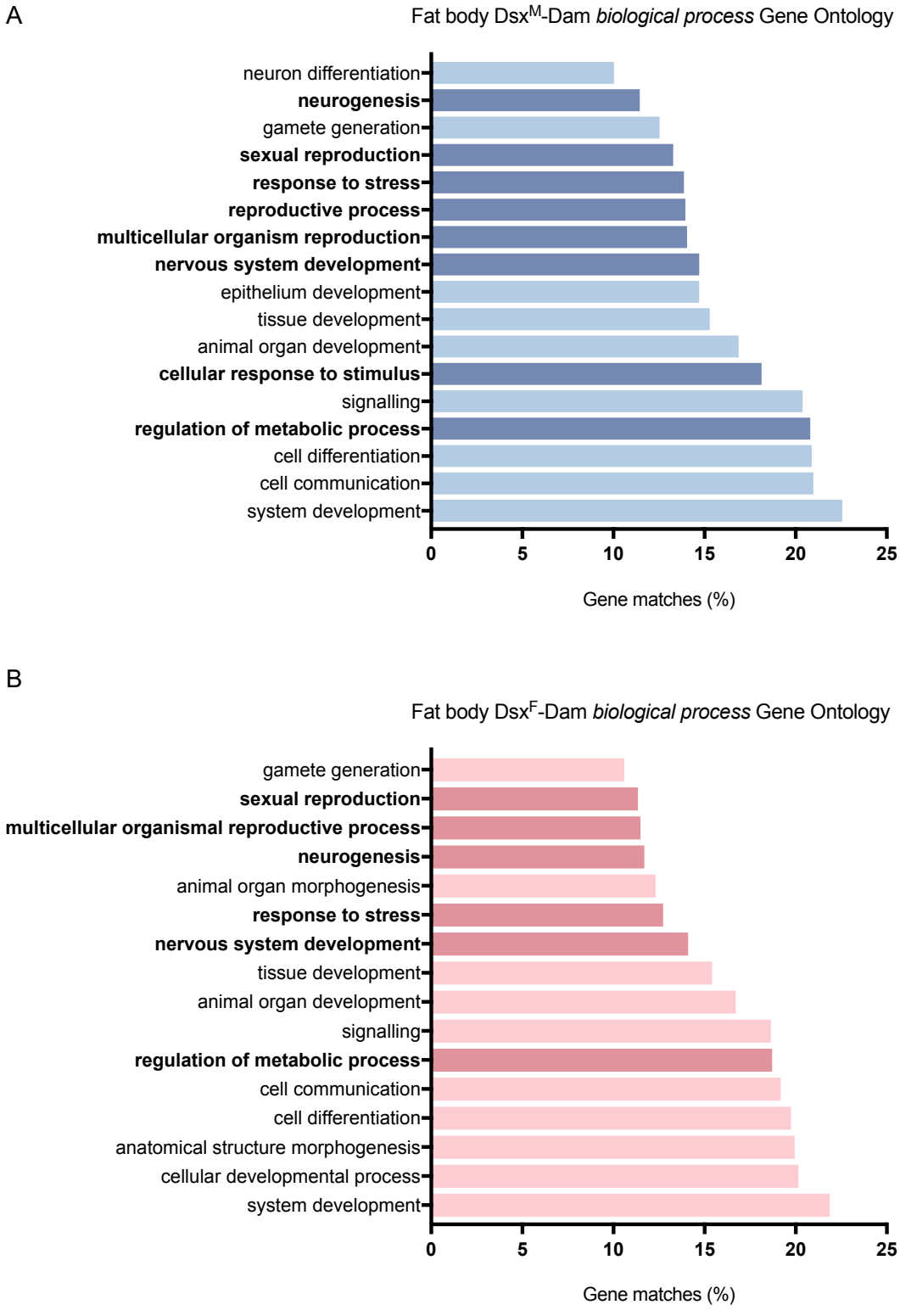


Figure 54 *Gene Ontology enrichment analysis for Clough et al., 2014 DamID-seq fat body Dsx^{M}-Dam (A) and fat body Dsx^{F}-Dam (B) 'biological process' ontology. For fat body Dsx^{M}-Dam, 256 statistically significant enrichment terms were defined overall, and 279 for fat body Dsx^{F}-Dam ($p<0.05$, Holm-Bonferroni test correction). Significantly enriched terms relating to expected DSX function in the fat body are highlighted. For both, GO analysis completed on peaks/ genes common to biological replicates one*

and two. Chart details most significantly enriched GO terms with >150 gene matches (redundant terms removed). Particularly relevant GO terms are presented in darker colours.

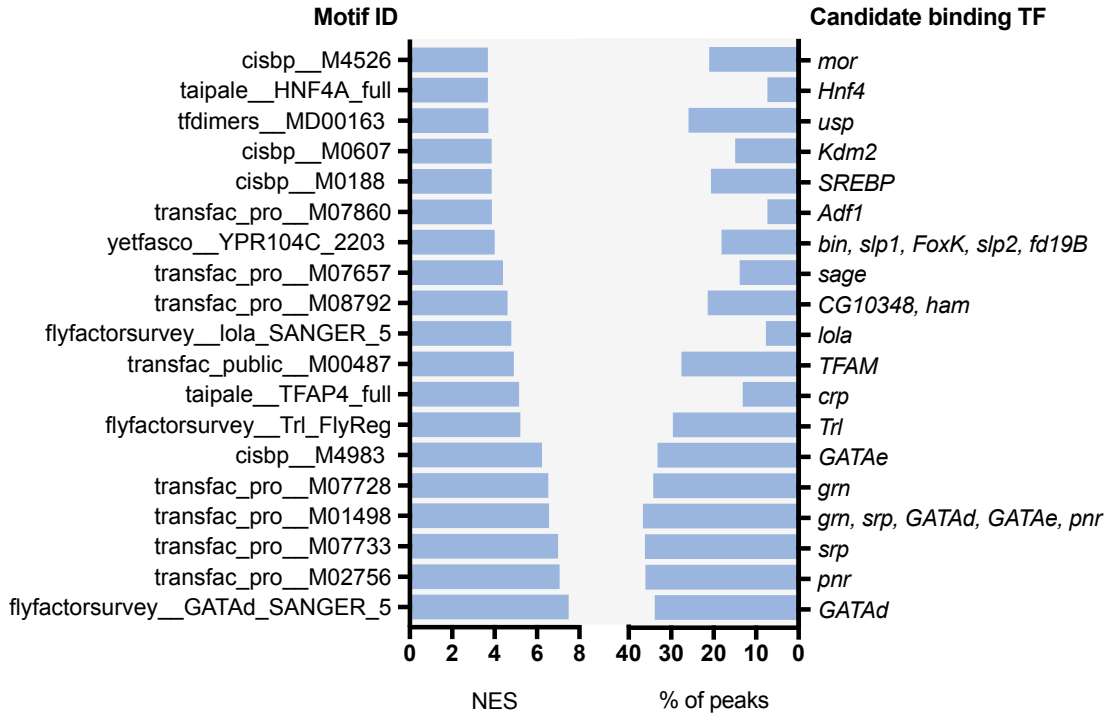
6.4.5 COMPARING I-*CIS*TARGET REGULATORY FEATURE PREDICTIONS IN DSX- FAT BODY, BRAIN AND HEAD DATASETS

We implemented the i-*cis*Target method to search for binding motifs in the Dsx^{M} fat body called peaks common to both biological replicates 1 and 2 (FDR 0.05) (Herrmann et al., 2012; Imrichová et al., 2015). The i-*cis*Target method identifies *cis*-regulatory modules (CRMs) by ranking conserved regions in the *Drosophila* genome. We identified significant motifs and determined an optimal subset of genomic regions predicted as direct targets in the Dsx^{M} fat body dataset (Clough et al., 2014). We took the same approach here as for the TaDa Dsx^{M} brain and head datasets described in the previous chapter. The top-ranked motifs identified in the Dsx^{M} fat body dataset for which a position weight matrix was available are shown in Figure 55A. The top five enriched motifs identified were all five members of the *Drosophila* GATA transcription factor family. In *Drosophila*, the GATA transcription factors are essential in the development and physiology of multiple tissues including the brain (Martínez-Corrales et al., 2019). Sequence logos for the GATA factors are listed according to their enrichment levels in Figure 55B. We determined the genes in the Dsx^{M} fat body dataset associated with the top-ranked CRMs containing the most significantly enriched motif, GATAd. These included *dsx* and fellow sex-determination gene, *fru*. Positively, we identified similar binding motifs in the Dsx^{M} head dataset where DSX-expressing neural and fat body cells were profiled. In that dataset, these same five GATA factors were identified as the most significantly enriched motifs. In the Dsx^{M} brain dataset, three of the five GATA factors similarly appeared (GATAd, *srp* and *grn*), albeit at significantly

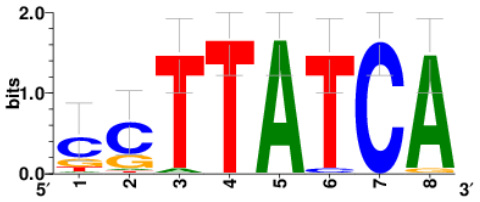
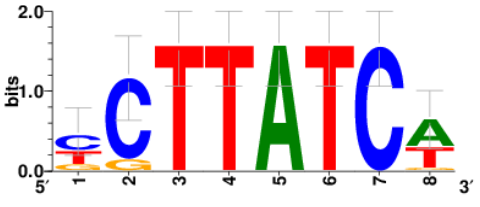
lower enrichment levels. There, the *Adf1* motif (transfac_pro__M07860) was most significantly enriched. *Adf1* is implicated in the regulation of developmental plasticity in the brain influencing crucial aspects of dendrite development (England et al., 1992; Timmerman et al., 2013).

A

Transcription factor motif enrichment analysis, Dsx^{M} fat body DamID peaks



B

Rank	Gene	Motif ID	Sequence logo
1	GATAd	flyfactorsurvey__GATAd	
2	Pannier	transfac_pro__M02756	
3	Serpent	transfac_pro__M07733	

4	Grain	transfac_pro__M07728	
5	GATAe	cisbp__M4983	

Figure 55 *i-cisTarget (Herrmann et al., 2012; Imrichová et al., 2015) was used to predict regulatory features in Dsx^{M} fat body called peaks. Enriched transcription factor motifs were identified and ranked (A). Motif IDs are presented according to the NES (normalised enrichment scores). The percentage of peaks containing the feature are indicated. Candidate binding TFs associated and corresponding to the Motif IDs are listed. Sequence logos for each member of the Drosophila GATA factor genes enriched in called peaks are ranked according to NES and listed (B).*

6.4.6 SYSTEMATIC APPROACH TO SELECTING TARGET GENES FOR FURTHER ANALYSES

Given the large number of putative target genes generated from our DSX TaDa-seq experiments, it was necessary to take a number of potential approaches to winnow gene lists. For example, we could have narrowed lists according to genes which possess the Erdman-defined DSX motif (Erdman et al., 1996), or one of our located *de novo* motifs. Our comparisons of generated gene lists from DSX-brain and head TaDa-seq experiments with the DamID-seq DSX-fat body experiments (Clough et al., 2014) allowed us to ascertain neuronal-specific or fat body-specific genes. Indeed, this candidate gene approach allowed the narrowing of gene lists. Pragmatically speaking however, the number was still far too high to sensibly attempt to follow up. Ultimately therefore, we compared our gene lists generated from our TaDa screens with additional data from independent protein-DNA screens (Clough et al., 2014; Hens, 2017

unpublished) as a means to select a far smaller number of 'high confidence' target genes.

YEAST-ONE HYBRID (Y1H) SCREENING OF *DROSOPHILA* REGULATORY ELEMENTS (HENS, 2017 UNPUBLISHED)

In 2011, Hens et al., developed and validated a high-throughput yeast-one hybrid platform to screen DNA elements of interest versus an 'almost complete' *Drosophila* TF repertoire containing 85% of predicted *Drosophila* TFs. To generate the *Drosophila* TF open reading frame (ORF) library, the group determined by bioinformatic analyses (Adryan and Teichmann, 2006) and manual curation that the *Drosophila* genome contains 755 sequence-specific TF-coding genes. Existing cDNA collections and *de novo* cloning methods were merged to create 722 (96%) Gateway-compatible Entry clones (Invitrogen) containing the ORF of each TF. The group sequence-verified a number of Entry clones for each TF using a high-throughput sequencing-based method (Massouras et al., 2010) confirming 692 TFs (92%) of which the majority were fully sequence-verified (588 or 78%).

The Gateway-compatible *Drosophila* TF library contains both DSX and FRU. In 2017, Hens utilised the platform to conduct a y1H screen of neuropeptide regulatory element DNA baits versus the *Drosophila* TF library (Hens, 2017 unpublished). Positive hits for FRU across two experimental replicated screens included a section of the RE of the *eve-stripe*, *Hug*, and *upd3* genes. For DSX, again across two experimental replicated screens, positive hits included a segment of the RE of *Dh31*, *Nplp1* and *Tk1* genes.

DSX ChIP-seq in S2 cells (Clough *et al.*, 2014)

In 2014 the Goodwin lab, completed a series of occupancy experiments to determine where DSX binds in the *D. melanogaster* genome. One of such experiments was ChIP-seq on S2 cells expressing tagged Dsx^M or Dsx^F (Clough et al., 2014). We viewed called peak files in Integrated Genome Browser (IGB 9.0.2) and conducted a targeted search for the y1H positive hits (Hens et al., 2011; Hens, 2017 unpublished). Figure 56 combines the positive hits from the y1H screen of *Drosophila* transcription factors (Hens et al., 2011; Hens, 2017 unpublished) and associated called peaks in the S2 cell DSX ChIP-seq datasets for *Dh31*, *Tk1* and *Nplp1*. Regulatory elements were split into 1 to 1.5 Kb overlapping sections to be tested in the y1H (top grey horizontal bars), positive hits indicated as yellow bars. Peaks were called in both ChIP-seq S2 Dsx^M and Dsx^F datasets for *Nplp1*. For *Tk1*, peaks were similarly called in both DSX ChIP-seq datasets albeit protein-DNA enrichment visually seems at the tail end of the called peak. For the *Dh31* RE y1H hit, we observe a called peak in the ChIP-seq S2 Dsx^F dataset, but not in the Dsx^M dataset although a peak is called nearby (<500 bp).

In our DSX TaDa head datasets, we observe a called peak (FDR 0.05 find_peaks, Marshall and Brand, 2015) for *Tk* in both Dsx^F-Dam replicate one and Dsx^M-Dam replicate one. Peaks were also called for *Dh31* in Dsx^F-Dam replicate two in head TaDa. In our DSX TaDa brain datasets, the receptor to *Dh31*, *Dh31-R*, is a positive hit in both Dsx^M-Dam biological replicates, and the Dsx^F-Dam biological replicate two.

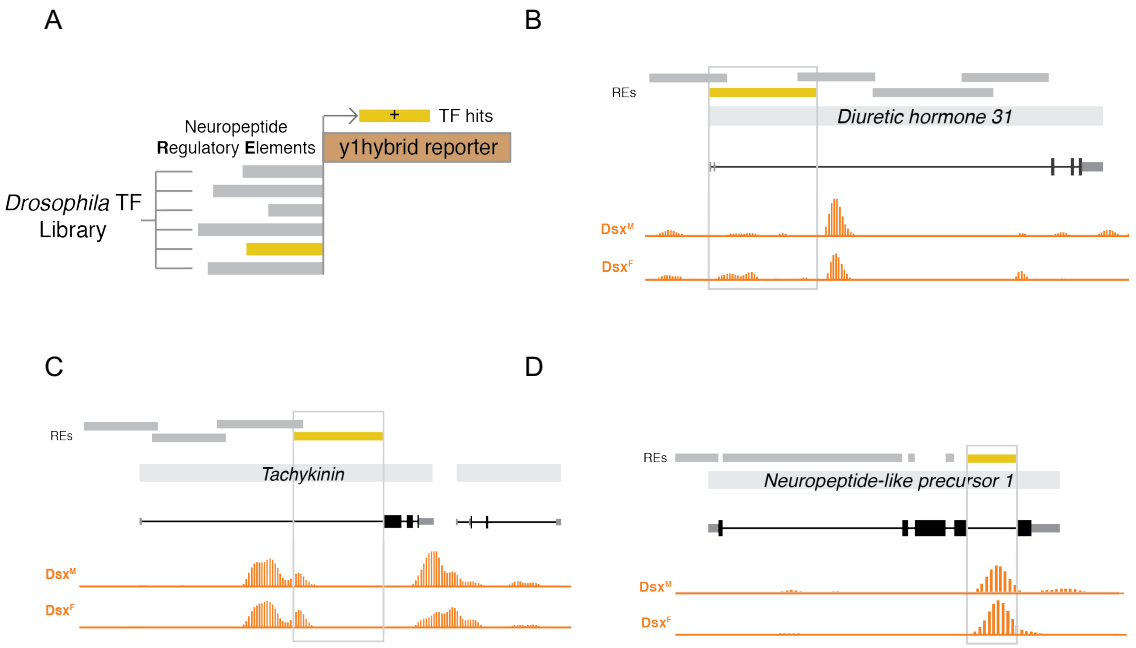


Figure 56 *DSX Occupancy and binding sites. The yeast-one hybrid platform enables high-throughput screening of DNA elements of interest versus a library containing 85% of predicted Drosophila TFs. The Drosophila TF library, containing DSX and FRU, was tested with a library containing neuropeptide Regulatory Elements (A). Scaled read density plots (background subtracted, arbitrary scale) from three replicated occupancy experiments. Dsx^M and Dsx^F ChIP-Seq in S2 cells, mapped to Diuretic hormone 31 (B), Neuropeptide-like precursor 1 (C), and Tachykinin 1 (D) gene maps. FlyBase gene models showing coding exons (thick rectangles), noncoding regions (thin rectangles), and introns (lines). y1H neuropeptide RE's highlighted yellow.*

6.4.7 Assessing Split-Gal4 neuropeptide^{AD} putative candidate gene constructs

Transgenic constructs were generated with the *Dh31*, *Tk1* and *Nplp1* REs identified in the Hens et al., 2011 y1H screen using the Gateway cloning technique. We took a similar approach as previously in the characterisation of DSX TaDa transgenic constructs to assess our Split-Gal4 neuropeptide^{AD} constructs. Namely, we used DNA Sanger sequencing to confirm that the neuropeptide Regulatory Elements (RE) of interest were inserted *in frame* and in the right orientation. Sequencing primers spanned regions from within the neuropeptide gene RE for each construct, to within the p65AD

vector backbone: *Dh31*-F1, *Nplp1*-F1 and *Tk1*-F1 forward primers, and the common reverse primer p65AD-Rev1 (Figure 57).

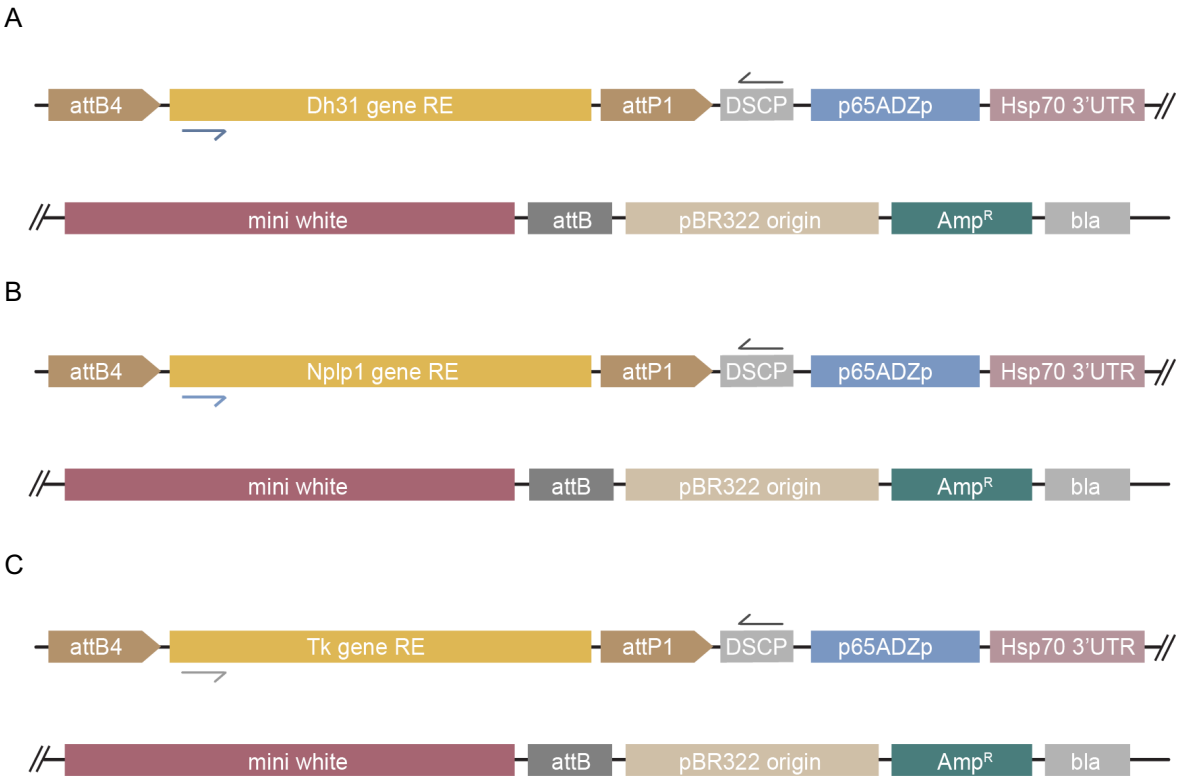


Figure 57 *Linear plasmid map of Split-Gal4 neuropeptide^{AD} DNA constructs: Dh31^{AD} (A), Nplp1^{AD} (B), and Tk1^{AD} (C). Blue and grey arrows specify forward sequencing primer locations, black arrows specify reverse sequencing primer locations.*

MICROINJECTION OF SPLIT-GAL4 NEUROPEPTIDE^{AD} CONSTRUCTS INTO *DROSOPHILA* EMBRYOS

Microinjection of genetic constructs into *Drosophila* embryos was completed at the BestGene Inc. facility in Chino Hills CA, U.S.A. Expression Clones containing the neuropeptide RE were injected into an attP40 strain (BDSC 25709) using PhiC31 integrase for the same reasons as described in the generation of UAS-Dam-*dsx^{M}* and

UAS-Dam-*dsx*^{F}. See Transgenics chapter, sub-section 3.4.1 for full details of methodology to prepare samples for microinjection. Approximately 200 *Drosophila* embryos were initially injected with *attB* DNA sample, with a survival rate of ~50% to larvae (Table 19).

n	*Dh31*$^{Gal4\text{-}AD}$	*Nplp1*$^{Gal4\text{-}AD}$	*Tk1*$^{Gal4\text{-}AD}$
Injected embryos	~200	~200	~200
Surviving larvae	~100	~100	~100
Crossing G_0 adult to *yw*	~55	~55	~55
Red eye G_1 adult	5	5	5
1st round balancing cross	5	5	5
Balanced line	5	5	5

Table 19 *Microinjection of Split-Gal4 neuropeptide^{AD} constructs in Drosophila melanogaster embryos. Survival rates from embryos to larvae, and subsequent crossing scheme for balancing.*

BALANCING SPLIT-GAL4 NEUROPEPTIDE^{AD} TRANSGENIC FLIES

In addition to the initial microinjection of the TaDa constructs into *Drosophila* embryos, flies were balanced at the facility in BestGene Inc. Selected survival G_0 adults were first back-crossed to *yw* (*yellow white* genetic background) to generate stable transformants. Screening was conducted for the loss of w^+ transformants. All transformants were picked from individually injected G_0. The G_1 adult transformants were expanded (red eye) by crossing to *yw* again. G_2 transformant flies were balanced with FM7i for X, CyO for the 2nd, and TM3 (stubble, Sb) on the 3rd. The final genotype of the balanced stocks received from BestGene Inc was:

w- / FM7i; CyO; neuropeptide^{AD*}/ TM3 (Sb)

where neuropeptide^{AD*} is *Dh31*-RE$^{Gal4\text{-}AD}$, *Nplp1*-RE $^{Gal4\text{-}AD}$ or *Tk1*-RE$^{Gal4\text{-}AD}$.

6.4.8 Imaging expression of Split-Gal4 neuropeptide^{AD} transgenic flies

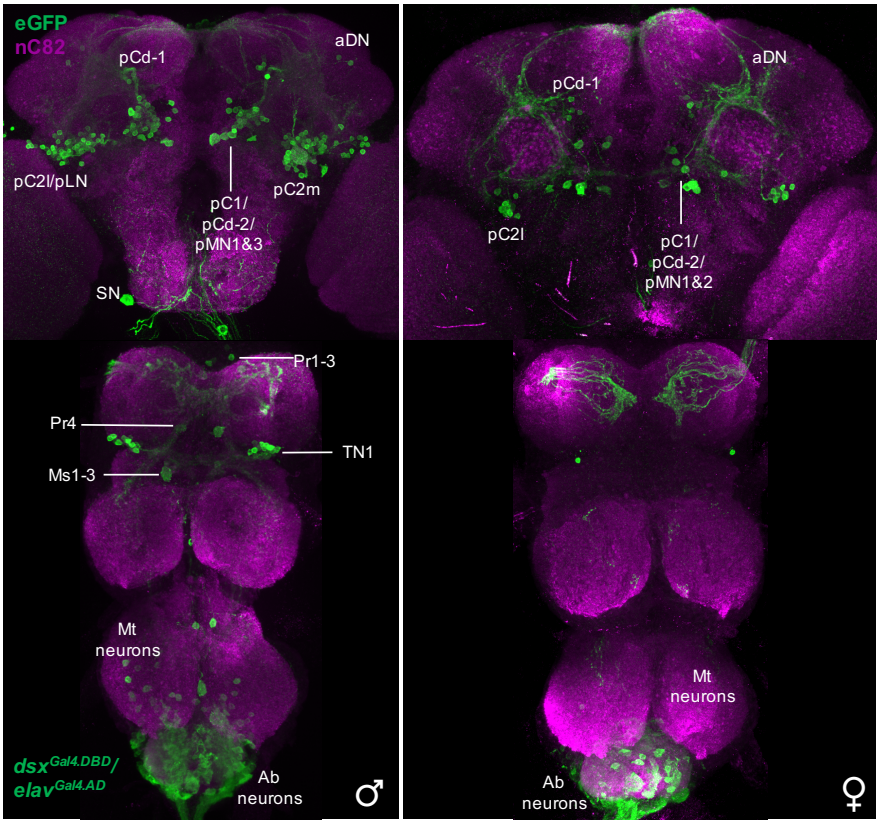
To elucidate whether the neuropeptides *Dh31*, *Nplp1* and *Tk1* are true *dsx* target genes, we conducted a series of expression analyses looking for neuronal overlap between our Split-Gal4 neuropeptide^{AD} generated transgenic lines and *dsx*. To do this, we utilised the Split-Gal4 system combining our neuropeptide^{AD} transgenic lines with the *dsx*$^{Gal4\text{-}DBD}$ allele and enhanced Green Fluorescent Protein (eGFP), theoretically driving eGFP expression in neurons co-expressing neuropeptide^{AD}/ *dsx*$^{Gal4\text{-}DBD}$. We also combined the neuropeptide^{AD} transgenic lines with the pan-neuronal reporter, embryonic lethal abnormal vision (*elav*$^{Gal4\text{-}DBD}$) and eGFP. This tags eGFP expression to all neuropeptide^{AD}-expressing neurons. Prior to dissection, flies were genotyped to ensure each contained the appropriate transgenes. We used the sequencing primers discussed above for each neuropeptide^{AD} (*Dh31*-F1, *Nplp1*-F1 and *Tk1*-F1 forward primers, p65AD-Rev1 reverse primer). For *dsx*$^{Gal4\text{-}DBD}$, we used the primers discussed in the Transgenics chapter, *dsx*^{DBD}.F1 and *dsx*^{DBD}.R1. Adult animal (5-7 day-old) brain and Ventral Nerve Cord's (VNC) were dissected in PBS.

dsx-expressing neurons in the adult CNS are sexually dimorphic: there are 400-700 *dsx*-expressing neurons in the male CNS (Lee et al., 2002; Pavlou et al., 2016; Rideout et al., 2010) and 300-400 *dsx*-expressing neurons in the female CNS (Pavlou et al., 2016; Rideout et al., 2010). Neuronal expression analyses using the *dsx*^{Gal4} allele (Rideout et

al., 2010) demonstrates these significant differences in *dsx*-expressing neuronal circuitry in the adult CNS (Figure 58A). Quantification of neuronal numbers in males showed dsx^{Gal4} to be expressed in ~280 neurons in the brain (Kimura et al., 2015; Rideout et al., 2010; Robinett et al., 2010), ~75 neurons in the thoracic ganglia (Shirangi et al., 2016), and ~275 in the abdominal ganglion (Abg). In females, reported numbers of dsx^{Gal4} neurons in the brain vary between ~50 and ~140 (Kimura et al., 2015; Lee et al., 2002; Rezával et al., 2016; Rideout et al., 2010; Robinett et al., 2010; Sanders et al., 2008), and ~310 neurons in the Abg (Kimura et al., 2015; Pavlou et al., 2016; Rideout et al., 2010).

In the male brain, dsx^{Gal4} neurons are spread across anterior dorsal neurons (aDN), posterior clusters (pC1 and pC2), and the suboesophageal neurons (SN). pC2 is divided into lateral and medial clusters (pC2l and pC2m, respectively) based on their projections. The posterior dorsal cluster (pCd-1 and pCd-2), and two distinctive posterior Medial Neurons (pMN1 and pMN2) are located close to cluster pC1. In the VNC, there are prothoracic TN1 neurons and several single and paired TN2 thoracic neurons. TN2 neurons are distinguished by their segmental identity as prothoracic (Pr), mesothoracic (Ms), or metathoracic (Mt). The abdominal cluster (Ab) is present in males and females. dsx^{Gal4} neurons in the female brain are spread across pC1, pC2, pCd, aDN, and pMN1 and 2 discrete clusters (Rezával et al., 2016; Rideout et al., 2010). Relative to males, females have 85% fewer cells in the medial pC1 cluster, 80% fewer cells in the pC2 cluster, and 50% fewer neurons in the pCd cluster. Further, dsx^{Gal4} was not expressed in the thoracic ganglia of females (Figure 58B; Rideout et al., 2010; Robinett et al., 2010).

A



B

Neuronal clusters		Male	Female
Brain	pC1	57 +/- 5.0	9 +/- 2.0
	pC2	77 +/- 3.1	11 +/- 1.9
	pC3	14 +/- 1.0	6 +/- 1.4
	aDN	2 +/- 0	2 +/- 0
	SN	1 +/- 0	0 +/- 0
Ventral Nerve Cord	TN1	22 +/- 1.7	0 +/- 0
	TN2	7 +/- 3.0	0 +/- 0
	MtAbg	22 +/- 4.9	12 +/- 3.8
	Abg	254 +/- 18.4	300 +/- 15.3

Figure 58 *Imaging sexually dimorphic expression of dsx^{Gal4} in the CNS of male and female adult Drosophila melanogaster (A). The dsx^{Gal4} allele (Rideout et al., 2010) revealed higher numbers of dsx-expresing neurons in the adult male brain and thoracic ganglia compared to females, whilst females had*

higher numbers in the abdominal ganglion. 5-7 day-old male and female brain and VNC, $dsx^{Gal4} \cap elav^{AD}$>mCD8::eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta) (Nojima unpublished, 2017). (B) Rideout et al., 2010 dsx^{Gal4} CNS neuronal cluster counts of male and female adults. Counts represent mean +/- standard deviation. Single-labelling with GFP applied to CNSs from dsx^{Gal4} flies expressing nGFP, n≥5.

We visualised no neuronal co-expression between *dsx* and *Dh31*-RE neuronal populations (Figure 59A and B). However, we saw *Dh31*-RE neuronal expression in the ventral nerve cord (VNC) of males and females in a sexually monomorphic pattern when visualising all *Dh31*-RE-expression in neuronal cells (Figure 59C and D). We observed sex-specific neuronal co-expression with *Tk1*-RE and *dsx*. In males, five neuronal clusters including pC1, pC2l, pC2m, pMN1 and pMN3 were observed in the brain, and two further clusters, Pr1 and TN1, in the VNC (Figure 60A; nomenclature of the dsx^{+} neuronal clusters is based on Kimura et al., 2015 and Nojima unpublished, 2017). No neuronal co-expression was observed in females (Figure 60B). With ELAV, substantial numbers of *Tk1*-RE neurons were visualised in the brain and VNC in both males and females (Figure 60C and D). We also observed sex-specific neuronal co-expression between *dsx* and *Nplp1*-RE. Precisely, we note a single cluster of neurons in the female brain, labelled pMN1 (Figure 60B; Figure 61C and D). In males, further neuronal clusters are observed in the brain including pMN3, pC1 and pC2l and further expression in the VNC (Figure 60A; Figure 61A and B). With ELAV, significant numbers of *Nplp1*-RE neurons were observed in the brain and VNC in both males and females (Figure 61A and D).

The *Dh31*-RE, *Tk1*-RE and *Nplp1*-RE immunohistochemistry was completed in colloboration with a senior postdoc in the lab, T. Nojima. All fly crosses were set up personally prior to dissection and imaging by Nojima (2017, unpublished).

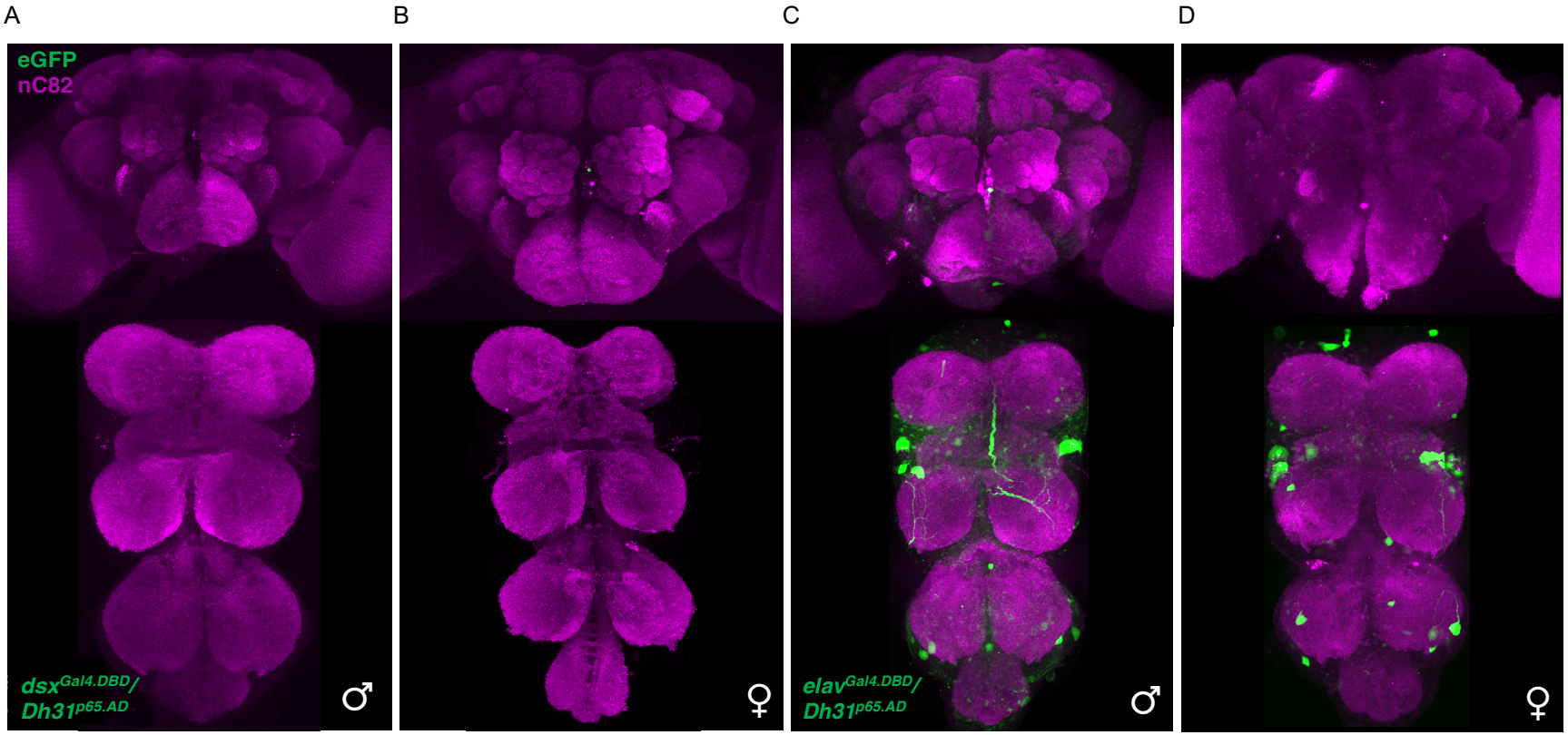


Figure 59 *Imaging dsx$^{Gal4.DBD}$/Dh31$^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (A) and female (B) animals. No dsx/ Dh31-RE neuronal overlap observed. Imaging elav$^{Gal4.DBD}$/Dh31$^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (B) and female (C) animals. Expression of a small number of Dh31-RE neurons in both male and female VNC observed. mCD8 and eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta).*

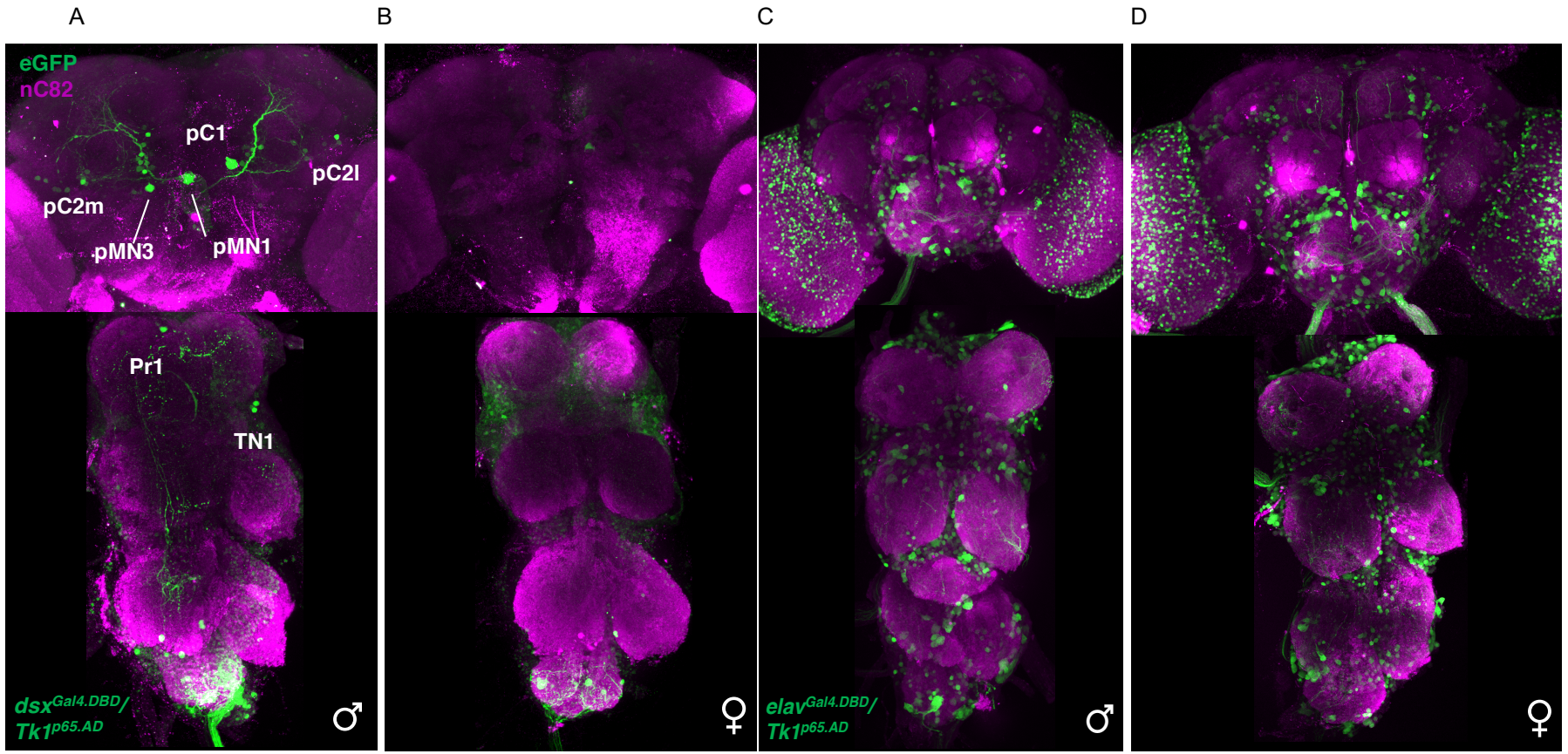


Figure 60 *Imaging dsx$^{Gal4.DBD}$/Tk1$^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (A) and female (B) animals. Sex-specific neuronal expression observed. Male animals show dsx$^{Gal4.DBD}$/Tk1$^{p65.AD}$ expression in five neuronal clusters in the male brain and two further clusters in the VNC. Imaging elav$^{Gal4.DBD}$/Tk1 $^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (C) and female (D) animals. mCD8 and eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta).*

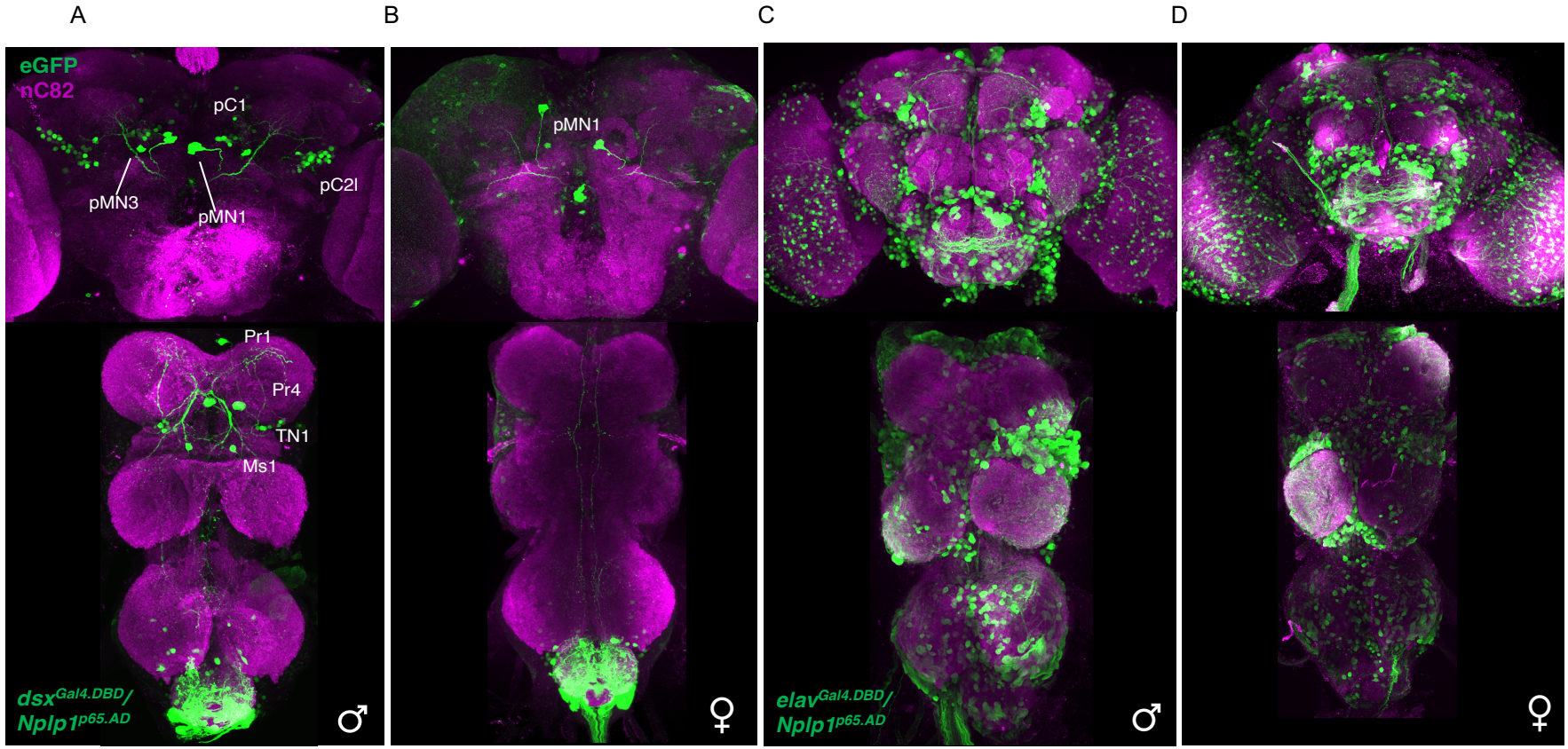


Figure 61 *Imaging dsx$^{Gal4.DBD}$/Nplp1$^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (A) and female (B) animals. A single cluster of neurons was observed sex-specifically in the female brain, labelled pMN1. In males, further neuronal clusters are observed in the brain including pMN3, pC1 and pC2l and further expression in the VNC. Imaging elav$^{Gal4.DBD}$/Nplp1 $^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain and VNC in male (C) and female (D) animals. Broad Nplp1-RE expression is observed in male and female brain and VNCs. mCD8 and eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta).*

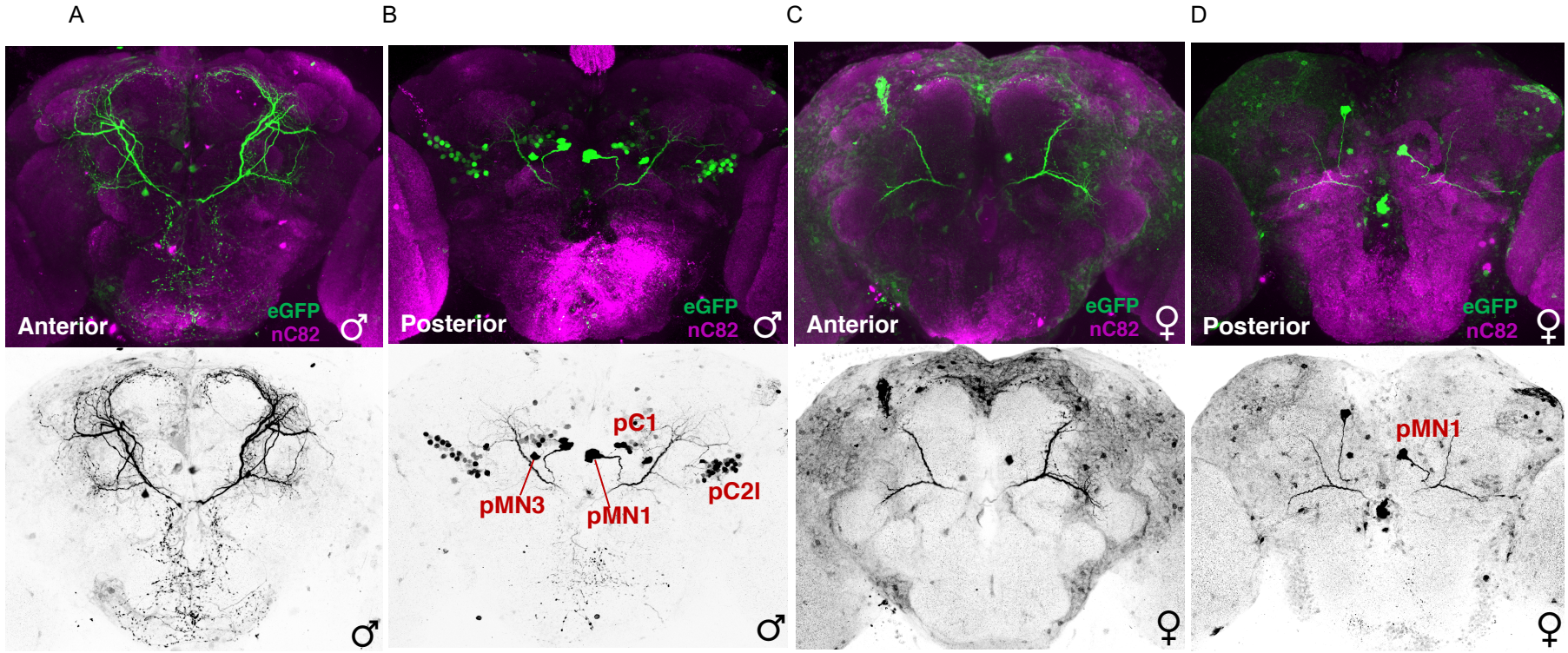


Figure 62 *Anterior and posterior view of dsx$^{Gal4.DBD}$/Nplp1$^{p65.AD}$ neuronal co-expression in the adult (5-7 day-old) brain in male (A and B) and female (C and D) animals. Projection pattern of sex-specific pMN1 single neuronal cluster in females highlighted. mCD8 and eGFP staining shown in green. Neuropil counterstained with antibody to nC82 (magenta).*

6.5 DISCUSSION

Previous attempts to unravel the DSX transcriptional network include Dsx^M or Dsx^F DamID-seq on adult fat body in transgenic flies, ChIP-seq on S2 cells expressing tagged Dsx^M or Dsx^F (Clough et al., 2014), or a yeast one-hybrid screen of *Drosophila* TFs including DSX (Hens et al., 2011; Hens, 2017 unpublished). Indeed, increasing the types of alternative protein-DNA interaction methods increases the likelihood of identifying *bona fide* binding sites. Hence, here we incorporate our TaDa-seq DSX brain and head datasets with previously published DSX occupancy analyses. More specifically, the functional comparisons of our DSX head and brain TaDa datasets with the published DSX fat body DamID-seq dataset (Clough et al., 2014) allowed for a delineation of tissue specific DSX binding targets across male adult animals.

The DSX-fat body DamID screen (Clough et al., 2014) and our DSX brain and head TaDa screens conducted here differ experimentally. Nevertheless, our direct comparisons should be biologically meaningful, because the homologous protein-DNA interaction method is identical across both approaches. Further, the DSX-fat body DamID-seq dataset (Clough et al., 2014) and the DSX brain and head TaDa-seq datasets (this study) were processed in the same way and peaks called using find_peaks (Marshall and Brand, 2015) to the same FDRs for direct comparison (Table 18). Interestingly, similar numbers of peaks were called across the fat body DamID and head TaDa datasets. Perhaps this is in line with the higher number of cells profiled in these datasets as compared to brain TaDa. Indeed, occupancy of Dsx^M-Dam in both datasets was also similar, occupying largely intronic regions (Figure 52A). Assessment of peak lengths revealed similar coherence for the Dsx^M-Dam replicates in both datasets:

~1,600 bp median for both, and IQR ~1,100-2,400 bp and ~1,200-2,700 bp respectively (Figure 52B). Comparisons of brain and head TaDa datasets with the fat body DamID dataset highlight a tendency towards tissue-specificity in DSX binding targets. For example, we note a 26% overlap (534 genes) between Dsx^{M}-Dam fat body and Dsx^{M}-Dam head datasets. This overlap is significantly less between the fat body DamID and brain TaDa dataset (88 genes, 4%) (Figure 53B). The head TaDa datasets also profile *dsx*-expressing fat body cells that surround the brain. Assessing the top forty genes enriched across the Dsx^{M}-Dam brain, head and fat body datasets (Figure 53E) affirm this theory, and highlight genes common to specifically fat body function in the DSX fat body and head datasets. Enriched terms such as 'regulation of metabolic process' were defined in the GO analyses in Dsx^{M}-Dam and Dsx^{F}-Dam fat body datasets (Figure 54A and B) and Dsx^{M}-Dam head dataset.

The sex-specific developmental programs of both the nervous system and fat body are dependent on DSX, yet likely to be highly divergent. We postulated this is achieved through Dsx^{M} and Dsx^{F} being recruited to different loci. As discussed in the previous bioinformatics chapter however, given the variation we observed in Dsx^{F}-Dam biological replicates in both brain and head TaDa, it was impossible to draw conclusions about the sex-specificity of target genes. In the DSX fat body DamID dataset (Clough et al., 2014), both DSX isoforms were found to have similar occupancy patterns binding thousands of the same targets in both sexes (Figure 53A). The DSX fat body DamID experimental data most strongly supports the idea that whilst DSX could bind certain targets, regulation of the gene would nonetheless be dependent on the combinatorial activity of other gene-specific factors. Indeed, a major class of putative DSX binding targets predicted in the DSX fat body analyses encode transcriptional

regulators, a proportion of which have sex-specific expression patterns (Barmina et al., 2005; Chatterjee et al., 2011; Williams et al., 2008). This 'combinatorial model' is supported in the literature. For instance, the known direct *dsx* binding target *bab1* locus is regulated by an enhancer region containing both DSX and homeobox protein binding sites that regulate expression, sex-specifically, along the anterior/ posterior axis (Williams et al., 2008).

Predictions of regulatory features across the called peaks in the fat body DamID dataset and brain and head TaDa datasets revealed strikingly coherent results. Namely, in both the DSX fat body peaks and the DSX head peaks (where DSX-expressing neural and fat body cells were profiled) the most highly enriched motifs were significantly similar. These motifs each belonged to the *Drosophila* GATA factor family. GATA transcription factors are known to be fundamental in the development and identity of multiple tissues including the brain (Martínez-Corrales et al., 2019). The most significantly enriched motif, dGATA, was located in *dsx* and *fru* CRMs. Another of these TFs, Serpent, has been implicated in fat cell differentiation in *Drosophila* (Sam et al., 1996). These findings strongly suggest that GATA factors function as co-factors alongside DSX to impinge its function downstream the SDH. Identifying homology across these datasets provides support for the cogency of DSX DamID and TaDa profiling experiments. Indeed, this is heightened as the most significantly enriched motifs identified in the brain TaDa experiment differed. Furthermore, it is interesting to note that across the three experiments, the i-*cis*Target method does not locate the *dsx* binding motif, perhaps because the stringency of the parameters albeit standardised are too restrictive. Applying the i-*cis*Target method to the peaks confirmed to contain the

Erdman et al., 1996-defined *dsx* motif significantly locates the *dsx* motif in all three datasets, adding support to this idea.

In addition to the DSX brain and head TaDa-seq screens conducted in this study as well as the DSX fat body DamID-seq screen (Clough et al., 2014), we further functionally compare these target genes with those found by analogous protein-DNA interaction screening methods. This approach allowed us to winnow gene lists, increasing the likelihood of identifying true *bona fide* DSX binding sites. We utilised datasets from a yeast-one hybrid screen of neuropeptide regulatory element DNA baits versus an 'almost complete' *Drosophila* TF repertoire containing 85% of predicted *Drosophila* TFs (Hens, 2017 unpublished), as well as published ChIP-seq datasets assaying S2 cells expressing tagged Dsx^{M} or Dsx^{F} (Clough et al., 2014). We chose the y1H screen of *Drosophila* TFs versus *Drosophila* neuropeptide Regulatory Elements because of the pivotal known role of neuropeptides as regulators in a range of animal physiological processes and behaviour, including a role in the control of sexual behaviour in *Drosophila* (Terhzaz et al., 2007; Vanden Broeck, 2001). The y1H screen revealed three putative *dsx* neuropeptide targets: *Nplp1*, *Tk1*, and *Dh31* (Hens, 2017 unpublished). Positively, the ChIP-seq Dsx^{M} and Dsx^{F}-expressing S2 cell datasets revealed significant peaks in Dsx^{M} and Dsx^{F} isoforms for both *Nplp1* and *Tk1*, as well as a Dsx^{F} peak close to (<500bp) the *Dh31* RE (Clough et al., 2014). For *Dh31*, the proximity of the binding site to the RE could be explained by an artefact of ChIP – namely, a result of cross-linking. Further, when we compared these findings with our TaDa-seq screens, we identified called peaks for *Tk1* in Dsx^{M} and Dsx^{F} replicates, and *Dh31* in Dsx^{M} replicates in the DSX head datasets. These findings suggested these hits could be *bona fide* DSX binding sites, and thus warranted further assessment.

Following generation of transgenic constructs containing the y1H REs from our neuropeptides of interest, and their subsequent microinjection into *Drosophila* embryos, our series of expression analyses aimed to delineate their potential neuronal co-expression with *dsx*. Our rationale here is that neuronal co-expression between *dsx* and the REs of our neuropeptides of interest could point towards function. Combining the neuropeptide^{AD} transgenic lines with the pan-neuronal reporter, *elav*$^{Gal4\text{-}DBD}$, and eGFP enabled us to visualise all the AD-expressing neurons in the whole animal. We observed a sex-difference between DSX isoforms and binding of the *Dh31*-RE in the y1H system. This is in line with the *Dh31* gene literature which suggests a sex-specific role of *Dh31*. For example, sex differences in the *Dh31* loss-of function mutation have been observed in hypersensitivity to stress conditions, locomotor activity (Bretz, 2009) and increases in levels of sleep (Kunst et al., 2014). We visualised no co-expression between *dsx* and the *Dh31*-RE neuronal sub-populations in male or female animals (Figure 59A and B). This could suggest *Dh31* is not a *bona fide* DSX neural target, or perhaps simply the RE does not function out of context. The finding is not wholly unexpected given that *Dh31* only appears in head TaDa Dsx^{F}-Dam replicate two, and the known primary function of *Dh31* is in the regulation of fluid secretion (Coast et al., 2001).

Literature has both implicated *Tk1* in the inhibition of the display of male courtship behaviours (Shankar et al., 2015), and pointed towards sex-specific expression. For instance, *D. melanogaster* males but not females are reported to have a small cluster of sexually dimorphic Fru^{M}+ neurons that express Tachykinin. Activation of these neurons invoked inter-male aggressive behaviours without an effect on courtship (Asahina et al.,

2014). Our expression analyses are in line with these findings, in that we observe sex-specific neuronal co-expression between the *Tk1* RE and *dsx*. In males, five neuronal clusters including pC1 and pMN1 were observed in the brain, and two further clusters including TN1 in the VNC (Figure 60A). pC1 has previously been identified as a cluster of *dsx*-expressing neurons (Lee et al., 2000, 2002; Ren et al., 2016), and alongside P1, the P1/ pC1 neuronal cluster is regarded as a fundamental higher order command centre that functions to initiate male-type courtship behaviour (Kimura et al., 2008). Indeed, activation of pC1 and pC2l neurons is sufficient to induce courtship behaviour in males (Kimura et al., 2015). Further, pMN1, pMN3 and TN1 have been reported to express *dsx* in the brain and VNC respectively (Kimura et al., 2015; Lee et al., 2002; Rideout et al., 2010; Robinett et al., 2010; Sanders et al., 2008; Zhou et al., 2014). In females, no neuronal co-expression was identified (Figure 60B). In line with previous literature, anatomically, these expression analyses strongly suggest *dsx*/ *Tk1*+ neurons may be involved in male typical courtship behaviours.

Nplp1 expression has been confirmed in the adult brain via mass spectroscopy (Verleyen et al., 2004), and indeed in our expression analyses, we observed sex-specific neuronal co-expression between *dsx* and the *Nplp1* RE in the adult brain. Precisely, we note a single cluster of neurons in the female brain, pMN1 (Figure 60B; Figure 61C and D). Whilst pMN2 is reported to be involved in egg laying behaviours (Kimura et al., 2015), behavioural functions of pMN1 are still unknown. In males, further neuronal clusters are observed in the brain including pMN3, pC1 and pC2l and further expression in the VNC, neuronal clusters Pr1, Pr4, TN1 and Ms1, as well as in the abdominal ganglion (Figure 60A, Figure 61A and B). Given the findings from these expression analyses, we propose the *dsx*/ *Nplp1*-RE neuronal populations are likely to be involved

in the female post mating response. We coupled results from various analogous protein-DNA interaction screens with the expression analyses described here, to unite *in vivo* and *in vitro* binding information respectively as a means of identifying *bona fide* DSX binding genes, as well as delineate the function and sex-specificity of these targets.

The RE-fusion driver neuropeptide^{AD} lines generated here may not necessarily faithfully recapitulate expression patterns of the endogenous genes. This is because single RE-segments may not account for the complex spatial and temporal aspects of gene expression (Levine and Tjian, 2003). Hence confirmation of these findings using the published *dsx* antibody (for example, Sanders and Arbeitman, 2008) as a next step could validate these results. Further, in the future, testing our behavioural hypotheses would be crucial to elucidate the functional role of these neuropeptides in sexual physiology and behaviour. For *Nplp1*, by combining our split-Gal4 neuropeptide^{AD} transgenic lines with the novel *dsx*-specific split-Gal4 hemidriver (*dsx*$^{Gal4\text{-}DBD}$) (Pavlou et al., 2016) and a brain-specifically expressed flippase recombinase (*Otd-FLP*) (Asahina et al., 2014), we could spatially restrict expression to intersect the *dsx*/ *Nplp1*-RE neuronal populations in the brain only. In combination with a Gal4/FLP-responsive reporter or effector, we could confirm intersection of *dsx*/ *Nplp1*-RE neuronal populations in the brain only using UAS>stop>eGFP, and complete a series of necessity versus sufficiency experiments by activation of *dsx*/ *Nplp1*-RE using UAS>stop>TrpA1 or their inactivation using UAS>stop>TNT. For *dsx*/ *Tk1*-RE, behaviourally a similar intersectional approach could be employed to genetically intersect and activate *dsx*/ *Tk1*-RE neuronal populations in the brain with the aim of assaying effects on courtship behaviour in male adults. This could be particularly interesting as to date, FruM+ *Tk* neurons have been shown to affect adult behaviour, but not *dsx*/ *Tk*. For *Dh31*-RE, first

experimental steps could be to employ alternative techniques to ascertain neuronal co-expression such as assaying endogenous expression using antibody staining. The guinea pig and rabbit anti-*Dh31* antibody have been shown to accurately capture *Dh31* expression (Goda et al., 2016; Kunst et al., 2014). Further, alternative genetic strategies such as LexA-lexAop (Lai and Lee, 2006) could be introduced through the generation of new transgenic lines containing the *Dh31*-RE to assay neuronal co-expression with *dsx*.

7 CONCLUSIONS

Our attempts to elucidate the DSX transcriptional network using the novel targeted DamID method (Southall et al., 2013) mark a significant step forward from previous DamID and ChIP approaches (Clough et al., 2014; Luo et al., 2011). Targeted DamID employs the Gal4/UAS system to enable conditional, targeted expression of the Dam-fusion whilst avoiding problems with toxicity or potential artefacts from overexpression, as seen with the original DamID approach (Marshall et al., 2016; Southall et al., 2013; van Steensel et al., 2000). We employed this strategy to profile Dsx^{M} and Dsx^{F} neuronal populations in the brain and head. There are relatively small numbers of *dsx* neurons, ~280 in the male brain (Kimura et al., 2015; Rideout et al., 2010; Robinett et al., 2010) and ~140 in the female brain (Kimura et al., 2015; Lee et al., 2002; Rezával et al., 2016; Rideout et al., 2010; Robinett et al., 2010; Sanders et al., 2008). Yet Dsx^{M} and Dsx^{F} isoforms are crucial for proper sexual development in *Drosophila* (Coschigano and Wensink, 1993). Given there are just a few defined direct *dsx* target genes in the CNS (Burtis et al., 1991; Coschigano et al., 1993; Hutson et al., 2003; Shirangi et al., 2009; Williams et al., 2008), our study provides a rich set of putative targets for further investigation.

7.1.1 GENERATING DSX TADA TRANSGENIC FLIES FOR TADA-SEQ

Targeted DamID harnesses the phenomenon of low frequency ribosome reinitiation (Child et al., 1999; Luukkonen et al., 1995; Southall et al., 2013; van Blokland et al., 2011), with the Dam-fusion encoded as a secondary ORF on a bicistronic expression construct downstream of a UAS enhancer. Coupled with a Gal4 driver, Dam-fusion translation occurs at low levels, allowing tissue-specific methylation of target sequences without associated toxicity. We employed a PCR-based cloning method to insert Dsx^{M} and Dsx^{F} coding peptides into the secondary ORF of the pUAST-attB-LT3-NDam TaDa vector, with mCherry encoded on the primary ORF. DSX TaDa constructs were assessed using DNA Sanger sequencing, as well as a colony PCR/ gel electrophoresis-based approach. Constructs were injected into fly embryos by means of a site-specific integration approach using PhiC31 integrase. This minimised expression differences across UAS-Dam-dsx^{M} and UAS-Dam-dsx^{F} and Dam control flies, and enabled their direct comparison. We completed a series of DSX overexpression analyses using the dsx^{Gal4} allele (Rideout et al., 2010), to ascertain whether overexpression of Dsx^{M} and Dsx^{F} encoded on the secondary ORF was able to direct some aspects of development. Positively, our assays revealed respective masculinisation and feminisation of certain secondary sexual characteristics, suggesting DSX-Dam binds appropriate endogenous *dsx* target genes. Interestingly, sex-bias experiments using the X-linked character, *Bar*, revealed each of UAS-Dam-dsx^{M}, UAS-Dam-dsx^{F}, and UAS-Dam control crosses with dsx^{Gal4} produced more female progeny, although only in UAS-Dam-dsx^{M} crosses, was this difference statistically different from numbers of males. We propose this sex-bias is attributed to an experimental caveat where whilst the age of flies involved in crosses was controlled for, this age difference (potentially up to four days) could bias the sex of progeny. It is known females mated to older males produce more daughters (Mange,

1970). Our experiments suggested that our novel DSX TaDa transgenic constructs were functional and can be taken forward for TaDa profiling of *dsx*-expressing cells.

7.1.2 Optimising TaDa to profile DSX-DNA interactions in the *Drosophila* Brain and Head

TaDa protein-DNA interaction screens require very small numbers of induced cells for profiling compared to existing techniques: less than 10,000 cells compared to 4-6 million needed for ChIP-seq (Southall et al., 2013). We aimed to profile Dsx^{M} and Dsx^{F} neuronal populations in the brain and head of adult *Drosophila* flies. We used the published dsx^{Gal4} allele (Rideout et al., 2010) to drive the DSX TaDa transgenic flies encoding the DSX coding peptide on the secondary ORF, in all *dsx*-expressing cells. Verification that the appropriate *dsx* cells were induced was carried out by DNA Sanger sequencing and PCR genotyping using the gDNA from whole adult animals. Preliminary experimental TaDa rounds revealed large inconsistencies in the final 'Dam methylation smears' within biological replicates despite meticulous attention to experimental conditions. Our delineation of the five stages of the TaDa protocol allowed a logical method to optimise the selective extraction and amplification of DSX-Dam-methylated DNA from *D. melanogaster* adult brains and whole heads. We found most significant variation stemmed from the initial gDNA extraction kit used (QIAamp versus DNeasy) and downstream of this digestion times with the methylation sensitive restriction enzymes DpnI and DpnII. For the brain TaDa dataset we noticed a sex-specific difference in our experimental assays, whereby Dsx^{F}-Dam samples showed inconsistent methylation smears despite the experimental optimisations described. Increasing initial gDNA starting amounts for these samples (fifty to seventy brains for brain TaDa), brought experimental consistency at the final 'Dam methylation smears'

step. Here, we thus successfully adapt the pioneering Marshall et al., 2016 experimental protocol for the TaDa-seq profiling of Dsx^{M} and Dsx^{F} -expressing cells in *D. melanogaster* brain and whole head transgenic tissue.

7.1.3 DSX BRAIN AND HEAD TADA-SEQ BIOINFORMATIC ANALYSIS

Drawing biological interpretation from the generation of sequencing data is arguably the most important step in any genome-scale experiment. The TaDa method presents novel computational challenges for peak calling and the identification of TFBS. We implement the damidseq_pipeline and find_peaks pipeline (Marshall and Brand, 2015), developed and released with the TaDa experimental protocol (Marshall et al., 2016), for processing raw NGS datasets. Assessing biological reproducibility at the level of called peaks, we observed good overlap in Dsx^{M}-Dam replicates in both the brain and head dataset, but found significant variation in all four Dsx^{F}-Dam replicates. The latter variation, we speculate, could arise from over-amplification of Dsx^{F}-Dam induced cells in the library preparation stage of TaDa. Or indeed, perhaps a problem with the original UAS-Dam-dsx^{F} transgenic line – whilst confirmed by DNA Sanger sequencing, PCR genotyping, and dsx^{Gal4} overexpression assays – we did note large variation in numbers of male and female progeny across individual crosses in the sex-bias assays. Given the biological variation we observed in Dsx^{F}-Dam biological replicates, it was impossible to directly compare Dsx^{M}-Dam and Dsx^{F}-Dam datasets. Binding of known *dsx* target genes *bab1*, *fru*, and *Desat1* were identified across the Dsx^{M}-Dam brain and head datasets, confirming their integrity. Enriched terms such as “neurogenesis” and “nervous system development” were identified in the Dsx^{M}-Dam replicates in both datasets in our gene ontology analyses as expected. Further terms such as “response to stimulus” and “immune system process” were defined in the head dataset in line with

fat body function, cells that were also profiled in the head dataset. The i-*cis*Target method (Herrmann et al., 2012; Imrichová et al., 2015) was implemented for the prediction of regulatory features, identifying members of the *Drosophila* GATA TF family to be significantly enriched in both brain and head datasets. This is striking because the GATA family are essential in the development and identity of multiple tissues including the brain (Martínez-Corrales et al., 2019). It seems likely therefore that these factors function alongside *dsx* as co-factors. The wealth of putative Dsx^{M} target genes generated across our two TaDa screens could allow for the more thorough characterisation of the *dsx* machinery and delineation of its role in sexual physiology and behaviour.

7.1.4 CHARACTERISING DSX PUTATIVE TARGET GENES

Comparing our DSX brain and head TaDa-seq datasets with published DSX-fat body DamID-seq datasets (Clough et al., 2014) allowed us to ascertain the tissue specificity of our target genes. Indeed, our analyses revealed coherence between the Dsx^{M}-Dam head dataset and the Dsx^{M}-Dam fat body dataset, expected given the presence of the DSX-expressing fat body cells encapsulating the brain profiled in the head dataset. Further, subtraction of the DSX brain and head datasets theoretically identified putative target genes specific to fat body cells in the brain. Interestingly, similar to both the Dsx^{M}-Dam brain and head datasets, our i-*cis*Target analyses defined members of the GATA family most significantly as predicted regulatory features.

Parallel to our TaDa-seq profiling of DSX populations in the adult brain and head, we independently analysed analogous DSX protein-DNA interaction screens to identify

'high confidence' DSX target genes. These include the published Dsx^M and Dsx^F S2-cell ChIP-seq datasets (Clough et al., 2014), and a yeast-one hybrid screen of *Drosophila* TFs (Hens, 2017 unpublished). A candidate gene approach was employed, and we selected the neuropeptides *Nplp1*, *Tk1* and *Dh31* for further analysis. These were positively enriched in both protein-DNA interaction screens. We generated transgenic lines of these genes of interest and conducted expression analyses looking for their neuronal overlap with *dsx*. We observed sex-specific neuropeptide/ *dsx* neuronal overlap with *Tk1*-RE and *Nplp1*-RE.

In *Drosophila*, the neuropeptide Tachykinin has previously been defined to co-express $FruM^+$ but not *dsx* neurons and has been shown to be involved in aggressive but not sexual behaviour (Asahina et al., 2014). Our expression studies however suggest the *Tk1*-RE may overlap with *dsx* neuronal populations in the brain and VNC. We hence speculate whether *Tk1* could be involved in the regulation of sexual physiology and behaviour. In adult *Drosophila*, *Nplp1* expression is exclusively restricted to the brain and thoracicoabdominal ganglion (flyatlas.org). Its involvement in sexual behaviour is unknown. Our expression analyses reveal sex-specificity, with a single cluster of *Nplp1*-RE/ *dsx* neurons labelled in the female brain, pMN1. The anatomically distinct but morphologically similar *dsx*-expressing neuronal cluster, pMN2, is known to be involved in the post mating response (Kimura et al., 2015). The behavioural function of pMN1 is unknown, and until now, a Gal4-driver targeting pMN1 neurons precisely did not exist. Given the function of the morphologically similar, *dsx*-expressing, pMN2 neurons in the female postmating response, we speculate a similar role for pMN1 neurons involving *dsx* and *Nplp1*-RE.

7.2 FUTURE DIRECTION

Here, we aimed to lay the groundworks for the further delineation of the role of *dsx* in sexual development and behaviour. The study is heading towards building a DSX developmental *connectome*, whereby similar TaDa profiling experiments to those described here could be conducted on Dsx^M and Dsx^F neuronal populations in pupal and larval animals. These studies could build a DSX developmental gene profile, where, for example, one could compare stage-specific and sex-specific genes across developmental time. A similar approach has previously been taken to map the developmental profile of Fru^M proteins across development using DamID (Neville et al., 2014). For the experiments proposed, the generation of a dsx^{DBD}, UAS-Dam-X* recombinant line coupled with a neuron-specific reporter to genetically target *dsx* neurons solely would be imperative. This would negate the need for the technically difficult and time-consuming process of the manual dissection of animal material. Nonetheless, thorough optimisation of the experimental TaDa protocol would have to be done for profiling pupal and larval animals, as had been completed for adult animals here. An additional challenge would be the even smaller number of cells involved for TaDa profiling at these developmental time points. The generation of this Split-Gal4 recombinant line would potentially allow the profiling of smaller subsets of *dsx* neurons. For example, one could analyse *dsx* neuronal populations known to be involved in the control of coordinating *Drosophila* male copulation. Indeed, *dsx* glutamatergic motor neurons innervate muscles of the genitalia and enable genital attachment, and *dsx* GABAergic inhibitory neurons mediate genital uncoupling by inhibiting key motor neurons (Pavlou et al., 2016). Published Split-Gal4 lines targeting these neuronal populations are readily available. These could be combined with the

TaDa Split-Gal4 line to specifically allow profiling of these subsets of Dsx^M and Dsx^F neuronal populations.

Across both TaDa brain and head screens, we have noted significant variation across Dsx^F-Dam biological replicates. This has made the process of defining *bona fide* Dsx^F binding sites, and drawing biological conclusions from these, impossible. Indeed, while comparisons with alternative protein-DNA interaction detection methods such as Dsx^F tagged S2-cell ChIP-seq experiments could aid in this process, the delineation of experimental noise from true binding targets is compounded with a separate experimental procedure. Repeating a third TaDa experimental round profiling Dsx^F neurons in the brain, seems unlikely to bring cogency to the previous datasets. Perhaps then, a modification of the experimental approach could be employed with additional controls, such as controlling for TF colocalisation hotspots where TFs accumulate at higher frequency through protein-protein interactions (Moorman et al., 2006). In this case, one could fuse Dam to the Dsx^F protein with a mutation specifically in the DNA-binding domain, rendering it unable to bind *bona fide* Dsx^F binding sites, but theoretically binding to TF colocalisation hotspots. This approach has been successfully employed in published DamID studies (Luo et al., 2011).

The two TaDa Dsx^M and Dsx^F brain and head screens described in this study have provided a wealth of putative target genes for further analysis. Prior to this study there have been few defined direct *dsx* target genes in the CNS, and those that have been described cannot explain the plethora of developmental pathways and behaviours regulated by *dsx*. Our candidate gene approach brings together DSX binding data from

analogous protein-DNA interaction methods, alongside TaDa, to chase three neuropeptides. Our generation of transgenic lines and subsequent expression analyses revealed sex-specific expression and *dsx* co-expression with the Regulatory Element of *Tk1* as well as *Nplp1*. The expression pattern of *dsx*/ *Nplp1*-RE in females in particular, as described above, signifies in our mind 'low-hanging fruit' in its likelihood to be attached to a female post mating sexual behaviour under the control of *dsx*. In the immediate future, then, a series of necessity versus sufficiency behavioural experiments could be carried out to ascertain the behavioural phenotype. This would be done by intersecting the *dsx*/ *Nplp1*-RE neuronal cluster in the female brain with Otd-FLP (Asahina et al., 2014), and coupling with expression of the thermosensitive TrpA1 (thermosensitive transient receptor potential) ion channel for the activation (Berni et al., 2010), or TNT (tetanus toxin light chain) for the inactivation of neurons (Sweeney et al., 1995). For the latter, alternative strategies such as RNA interference (RNAi) could be used for the knockdown of *dsx*/ *Nplp1*-RE neurons (Fire et al., 1998).

In the past century, only few genes of sex determination and differentiation have been identified. Indeed, whilst DSX has been studied for over fifty years, the targets defined to date are sparse and cannot explain the plethora of sexually dimorphic morphologies and behaviour regulated by *dsx*. Here, we combined an extensive TaDa study of Dsx^{M} and Dsx^{F} isoforms in adult *Drosophila* brain and head, generating gene lists which highlight *dsx*'s broad mode of action. Dsx^{M}-Dam brain and head binding targets revealed a tendency towards tissue specificity, which we confirmed through a comparison with a published DSX-fat body DamID-seq screen. We further propose GATA TFs act as co-factors with *dsx* to impinge its function. We compared our TaDa screens with independent protein-DNA interaction screens, increasing the likelihood of

identifying *bona fide* DSX binding targets. Expression analyses of one such target, the neuropeptide *Nplp1,* with *dsx,* reveal expression in a single cluster in the female brain (pMN1), which we postulate is involved in the female post mating response. The results presented here reaffirm *dsx*'s pleiotropic role in the development and regulation of sexual physiology and behaviour, and point towards a novel mode of action for doing so. The wealth of targets identified pave the way for better delineation of the *dsx* machinery and an understanding of how it functions to control sex differentiation and behaviour.

8 REFERENCES

Abel, T., Michelson, A. M., and Maniatis, T. (1993). A Drosophila GATA family member that binds to Adh regulatory sequences is expressed in the developing fat body. Development, *119*, 623–633.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., et al. (2000). The Genome Sequence of Drosophila melanogaster. Science, *287*, 2185–2195.

Adryan, B., and Teichmann, S. A. (2006). FlyTF: a systematic review of site-specific transcription factors in the fruit fly Drosophila melanogaster. Bioinformatics, *22*, 1532–1533.

Aguila, J. R., Suszko, J., Gibbs, A. G., and Hoshizaki, D. K. (2007). The role of larval fat cells in adult Drosophila melanogaster. The Journal of Experimental Biology, *210*, 956–963.

Ahmad, S. M., and Baker, B. S. (2002). Sex-specific deployment of FGF signaling in Drosophila recruits mesodermal cells into the male genital imaginal disc. Cell, *109*, 651–661.

Albert, E. A., Puretskaia, O. A., Terekhanova, N. V., Labudina, A., and Bökel, C. (2018). Direct control of somatic stem cell proliferation factors by the Drosophila testis stem cell niche. Development, *145*, dev156315.

Amrein, H., and Thorne, N. (2005). Gustatory perception and behavior in Drosophila melanogaster. Current Biology, *15*, R673–R684.

An, W., and Wensink, P. C. (1995a). Integrating sex- and tissue-specific regulation within a single Drosophila enhancer. Genes and Development, *9*, 256–266.

An, W., and Wensink, P. C. (1995b). Three protein binding sites form an enhancer that regulates sex- and fat body-specific transcription of Drosophila yolk protein genes. The EMBO Journal, *14*, 1221–1230.

Anand, A., Villella, A., Ryner, L. C., Carlo, T., Goodwin, S. F., Song, H. J., Gailey, D. A., Morales, A., Hall, J. C., Baker, B. S., et al. (2001). Molecular genetic dissection of the sex-specific and vital functions of the Drosophila melanogaster sex determination gene fruitless. Genetics, *158*, 1569–1595.

Anderson, D. J. (2016). Circuit modules linking internal states and social behaviour in flies and mice. Nature Reviews Neuroscience, *17*, 692.

Andrews, S. (2010). *FastQC: a quality control tool for high throughput sequence data*.

Aradska, J., Bulat, T., Sialana, F. J., Birner-Gruenberger, R., Erich, B., and Lubec, G. (2015). Gel-free mass spectrometry analysis of Drosophila melanogaster heads. Proteomics, *15*, 3356–3360.

Aranha, M. M., and Vasconcelos, M. L. (2018). Deciphering Drosophila female innate behaviors. Current Opinion in Neurobiology, *52*, 139–148.

Arbeitman, M. N., Fleming, A. A., Siegal, M. L., Null, B. H., and Baker, B. S. (2004). A genomic analysis of Drosophila somatic sexual differentiation and its regulation. Development, *131*, 2007–2021.

Asahina, K., Watanabe, K., Duistermars, B. J., Hoopfer, E., González, C. R., Eyjólfsdóttir, E. A., Perona, P., and Anderson, D. J. (2014). Tachykinin-expressing neurons control male-specific aggressive arousal in Drosophila. Cell, *156*, 221–235.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, *25*, 25–29.

Auerbach, R. K., Euskirchen, G., Rozowsky, J., Lamarre-Vincent, N., Moqtaderi, Z., Lefrancois, P., Struhl, K., Gerstein, M., and Snyder, M. (2009). Mapping accessible chromatin regions using Sono-Seq. Proceedings of the National Academy of Sciences, *106*, 14926–14931.

Aughey, G. N., and Southall, T. D. (2016). Dam it's good! DamID profiling of protein-DNA interactions. Wiley Interdisciplinary Reviews: Developmental Biology, *5*, 25–37.

Ávila, V., Fernández, J., Quesada, H., and Caballero, A. (2011). An experimental evaluation with Drosophila melanogaster of a novel dynamic system for the management of subdivided populations in conservation programs. Heredity, *106*, 765–774.

Baggerman, G., Boonen, K., Verleyen, P., De Loof, A., and Schoofs, L. (2005). Peptidomic analysis of the larval Drosophila melanogaster central nervous system by two-dimensional capillary liquid chromatography quadrupole time-of-flight mass spectrometry. Journal of Mass Spectrometry, *40*, 250–260.

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. International Conference on Intelligent Systems for Molecular Biology, *2*, 28–36.

Baker, B. S., and Ridge, K. A. (1980). Sex and the single cell. I. On the action of major loci affecting sex determination in Drosophila melanogaster. Genetics, *94*, 383-423.

Baker, B. S., Taylor, B. J., and Hall, J. C. (2001). Are complex behaviors specified by dedicated regulatory genes? Reasoning from Drosophila. Cell, *105*, 13–24.

Baltimore, D. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in virions of RNA tumour viruses. Nature, *226*, 1209–1211.

Barmina, O., Gonzalo, M., McIntyre, L. M., and Kopp, A. (2005). Sex- and segment-specific modulation of gene expression profiles in Drosophila. Developmental Biology, *288*, 528–544.

Barras, F., and Marinus, M. G. (1989). The great GATC: DNA methylation in E. coli. Trends in Genetics, *5*, 139–143.

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Research, *41*, D991–D995.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D. E., Wang, Z., Wei, G.,

Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell, *129*, 823–837.

Bashaw, G. J., and Baker, B. S. (1997). The Regulation of the Drosophila msl-2 Gene Reveals a Function for Sex-lethal in Translational Control. Cell, *89*, 789–798.

Bassett, A., and Liu, J.-L. (2014). CRISPR/Cas9 mediated genome engineering in Drosophila. Methods, *69*, 128–136.

Bateman, J. R., Lee, A. M., and Wu, C. (2006). Site-specific transformation of Drosophila via phiC31 integrase-mediated cassette exchange. Genetics, *173*, 769–777.

Baumgardt, M., Miguel-Aliaga, I., Karlsson, D., Ekman, H., and Thor, S. (2007). Specification of neuronal identities by feedforward combinatorial coding. PLoS Biology, *5*, e37.

Bayliss, W. M., and Starling, E. H. (1902). The mechanism of pancreatic secretion. The Journal of Physiology, *28*, 325–353.

Bayrer, J. R., Zhang, W., and Weiss, M. A. (2005). Dimerization of doublesex is mediated by a cryptic ubiquitin-associated domain fold: Implications for sex-specific gene regulation. Journal of Biological Chemistry, *280*, 32989–32996.

Beall, E L, and Rio, D. C. (1997). Drosophila P-element transposase is a novel site-specific endonuclease. Genes and Development, *11*, 2137–2151.

Beall, Eileen L, Mahoney, M. B., and Rio, D. C. (2002). Identification and analysis of a hyperactive mutant form of Drosophila P-element transposase. Genetics, *162*, 217–227.

Bellen, H. J., Levis, R. W., Liao, G., He, Y., Carlson, J. W., Tsang, G., Evans-Holm, M., Hiesinger, P. R., Schulze, K. L., Rubin, G. M., et al. (2004). The BDGP gene disruption project. Genetics, *167*, 761–781.

Bellen, H. J., Tong, C., and Tsuda, H. (2010). 100 years of Drosophila research and its impact on vertebrate neuroscience: a history lesson for the future. Nature Reviews Neuroscience, *11*, 514-522.

Belote, J. M., McKeown, M. B., Andrew, D. J., Scott, T. N., Wolfner, M. F., and Baker, B. S. (1985). Control of sexual differentiation in Drosophila melanogaster. Cold Spring Harbor Symposia on Quantitative Biology, *50*, 605–614.

Bender, W., Akam, M., Karch, F., Beachy, P. A., Peifer, M., Spierer, P., Lewis, E. B., and Hogness, D. S. (1983). Molecular genetics of the Bithorax complex in Drosophila melanogaster. Science, *221*, 23–29.

Benes, H., Spivey, D. W., Miles, J., Neal, K., and Edmondson, R. G. (1990). Fat-body-specific expression of the Drosophila Lsp-2 gene. SAAS Bulletin, Biochemistry and Biotechnology, *3*, 129–133.

Bennet-Clark, H. C., and Ewing, A. W. (1969). Pulse interval as a critical parameter in the courtship song of Drosophila melanogaster. Animal Behaviour, *17*, 755–759.

Bentley, D. R. (2006). Whole-genome re-sequencing. Current Opinion in Genetics and Development, *16*, 545–552.

Berni, J., Muldal, A. M., and Pulver, S. R. (2010). Using Neurogenetics and the warmth-gated ion channel TRPA1 to study the neural basis of behavior in Drosophila. Journal of Undergraduate Neuroscience Education, *9*, A5–A14.

Bernstein, E. (2005). RNA meets chromatin. Genes and Development, *19*, 1635–1655.

Bianchi-Frias, D., Orian, A., Delrow, J. J., Vazquez, J., Rosales-Nieves, A. E., and Parkhurst, S. M. (2004). Hairy transcriptional repression targets and cofactor recruitment in Drosophila. PLoS Biology, *2*, E178.

Bier, E. (2005). Drosophila, the golden bug, emerges as a tool for human genetics. Nature Reviews Genetics, *6*, 9–23.

Billeter, J.-C., Rideout, E. J., Dornan, A. J., and Goodwin, S. F. (2006). Control of male sexual behavior in Drosophila by the sex determination pathway. Current Biology, *16*, R766–R776.

Billeter, J.-C., and Wolfner, M. F. (2018). Chemical cues that guide female reproduction in Drosophila melanogaster. Journal of Chemical Ecology, *44*, 750–769.

Bischof, J., Maeda, R. K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific C31 integrases. Proceedings of the National Academy of Sciences, *104*, 3312–3317.

Blat, Y., and Kleckner, N. (1999). Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell, *98*, 249–259.

Boivin, A., and Dura, J. M. (1998). In vivo chromatin accessibility correlates with gene silencing in Drosophila. Genetics, *150*, 1539–1549.

Bousquet, F., Nojima, T., Houot, B., Chauvel, I., Chaudy, S., Dupas, S., Yamamoto, D., and Ferveur, J.-F. (2012). Expression of a desaturase gene, desat1, in neural and nonneural tissues separately affects perception and emission of sex pheromones in Drosophila. Proceedings of the National Academy of Sciences, *109*, 249–254.

Bownes, M. (1994). The regulation of the yolk protein genes, a family of sex differentiation genes in Drosophila melanogaster. BioEssays, *16*, 745–752.

Bownes, M., and Hames, B. D. (1977). Accumulation and degradation of three major yolk proteins in Drosophila melanogaster. The Journal of Experimental Zoology, *200*, 149–156.

Bownes, M., and Hames, B. D. (1978). Genetic analysis of vitellogenesis in Drosophila melanogaster: the identification of a temperature-sensitive mutation affecting one of the yolk proteins. Journal of Embryology and Experimental Morphology, *47*, 111–120.

Bownes, M., Shirras, A., Blair, M., Collins, J., and Coulson, A. (1988). Evidence that insect embryogenesis is regulated by ecdysteroids released from yolk proteins. Proceedings of the National Academy of Sciences, *85*, 1554–1557.

Brand, A. H., and Perrimon, N. (1993). Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. Development, *118*, 401–415.

Bretz, C. (2009). *Genetic Dissection of DH31 Signaling in Drosophila* (Doctoral dissertation, Wake Forest University).

Bridges, C. B. (1916). Non-disjunction as proof of the chromosome theory of heredity (concluded). Genetics, *1*, 107–163.

Bridges, C. B. (1917). Deficiency. Genetics, *2*, 445–465.

Bridges, C. B. (1921). Triploid intersexes in Drosophila melanogaster. Science, *54*, 252–254.

Bridges, C. B. (1922). The origin of variations in sexual and sex-limited characters. The American Naturalist, *56*, 51–63.

Bridges, C. B. (1925). Elimination of chromosomes due to a mutant (Minute-N) in Drosophila melanogaster. Proceedings of the National Academy of Sciences, *11*,

701–706.

Bridges, C. B. (1932). Recombination and crossing-over. The American Naturalist, *66*, 571–574.

Bridges, C. B. (1935). Salivary chromosome maps: with a key to the banding of the chromosomes of Drosophila melanogaster. Journal of Heredity, *26*, 60–64.

Bridges, C. B. (1939). Cytological and genetic basis of sex. Sex and Internal Secretions, 15–63.

Brinster, R. L., and Palmiter, R. D. (1986). Introduction of genes into the germ line of animals. Harvey lectures, *80*, 1-38.

Brody, T., and Cravchik, A. (2000). Drosophila melanogaster G protein-coupled receptors. The Journal of Cell Biology, *150*, F83-8.

Brooks, J. E., Blumenthal, R. M., and Gingeras, T. R. (1983). The isolation and characterization of the Esherichia Coli DNA adenine methylase (dam) gene. Nucleic Acids Research, *11*, 837–851.

Brown, S., and Hombría, J. C. G. (2000). Drosophila grain encodes a GATA transcription factor required for cell rearrangement during morphogenesis. Development, *127*, 4867–4876.

Buchon, N., Osman, D., David, F. P., Fang, H. Y., Boquete, J. P., Deplancke, B., and Lemaitre, B. (2013). Morphological and molecular characterization of adult midgut compartmentalization in Drosophila. Cell Reports, *3*, 1725-1738.

Bulyk, M. L. (2006). Analysis of sequence specificities of DNA-Binding proteins with protein binding microarrays. Methods in enzymology, *410*, 279-299.

Bulyk, M. L., Gentalen, E., Lockhart, D. J., and Church, G. M. (1999). Quantifying DNA-protein interactions by double-stranded DNA arrays. Nature Biotechnology, *17*, 573–577.

Burtis, K. C., Coschigano, K. T., Baker, B. S., and Wensink, P. C. (1991). The doublesex proteins of Drosophila melanogaster bind directly to a sex-specific yolk protein gene enhancer. The EMBO Journal, *10*, 2577–2582.

Burtis, K. C. (1993). The regulation of sex determination and sexually dimorphic differentiation in Drosophila. Current Opinion in Cell Biology, *5*, 1006–1014.

Burtis, K. C., and Baker, B. S. (1989). Drosophila doublesex gene controls somatic sexual differentiation by producing alternatively spliced mRNAs encoding related sex-specific polypeptides. Cell, *56*, 997–1010.

Butterworth, F. M. (1972). Adipose tissue of Drosophila melanogaster. V. Genetic and experimental studies of an extrinsic influence on the rate of cell death in the larval fat body. Developmental Biology, *28*, 311–325.

Butterworth, F. M., Burde, V. S., and Bownes, M. (1992). Mutant yolk proteins lead to female sterility in Drosophila. Developmental Biology, *154*, 182–194.

Camara, N., Whitworth, C., Dove, A., and Van Doren, M. (2019). Doublesex controls specification and maintenance of the gonad stem cell niches in Drosophila. Development, *146*, dev170001.

Camara, N., Whitworth, C., and Van Doren, M. (2008). The creation of sexual dimorphism in the Drosophila soma. Current topics in developmental biology, *83*, 65-107.

Carl, S. H., and Russell, S. (2015). Common binding by redundant group B Sox proteins is evolutionarily conserved in Drosophila. BMC Genomics, *16*, 292.

Carmel, I., Tram, U., and Heifetz, Y. (2016). Mating induces developmental changes in the insect female reproductive tract. Current Opinion in Insect Science, *13*, 106–113.

Castellanos, M. C., Tang, J. C. Y., and Allan, D. W. (2013). Female-biased dimorphism underlies a female-specific role for post-embryonic Ilp7 neurons in Drosophila fertility. Development, *140*, 3915–3926.

Chapman, K. B., and Wolfner, M. F. (1988). Determination of male-specific gene expression in Drosophila accessory glands. Developmental Biology, *126*, 195–202.

Chapman, T., Bangham, J., Vinti, G., Seifried, B., Lung, O., Wolfner, M. F., Smith, H. K., and Partridge, L. (2003). The sex peptide of Drosophila melanogaster: Female post-mating responses analyzed by using RNA interference. Proceedings of the National Academy of Sciences of the United States of America, *100*, 9923–9928.

Charlesworth, B. (1996). The evolution of chromosomal sex determination and dosage compensation. Current Biology, *6*, 149–162.

Chatterjee, S. S., Uppendahl, L. D., Chowdhury, M. A., Ip, P.-L., and Siegal, M. L.

(2011). The female-specific Doublesex isoform regulates pleiotropic transcription factors to pattern genital development in Drosophila. Development, *138*, 1099–1109.

Cheetham, S. W., Gruhn, W. H., van den Ameele, J., Krautz, R., Southall, T. D., Kobayashi, T., Surani, M. A., and Brand, A. H. (2018). Targeted DamID reveals differential binding of mammalian pluripotency factors. Development, *145*, dev170209.

Chen, W., Zollman, S., Couderc, J. L., and Laski, F. A. (1995). The BTB domain of bric à brac mediates dimerization in vitro. Molecular and Cellular Biology, *15*, 3424-3429.

Chen, Y., Negre, N., Li, Q., Mieczkowska, J. O., Slattery, M., Liu, T., Zhang, Y., Kim, T.-K., He, H. H., Zieba, J., et al. (2012). Systematic evaluation of factors influencing ChIP-seq fidelity. Nature Methods, *9*, 609–614.

Chee, J. Y., and Chin, C.F. (2015). Gateway cloning technology: advantages and drawbacks. Cloning Transgenes, *4*, 138.

Child, S. J., Miller, M. K., and Geballe, A. P. (1999). Translational control by an upstream open reading frame in the HER-2/neu transcript. The Journal of Biological Chemistry, *274*, 24335–24341.

Choksi, S. P., Southall, T. D., Bossing, T., Edoff, K., de Wit, E., Fischer, B. E., van Steensel, B., Micklem, G., and Brand, A. H. (2006). Prospero acts as a binary switch between self-renewal and differentiation in Drosophila neural stem cells. Developmental Cell, *11*, 775–789.

Christiansen, A. E., Keisman, E. L., Ahmad, S. M., and Baker, B. S. (2002). Sex comes in from the cold: the integration of sex and pattern. Trends in genetics, *18*, 510-516.

Chung, D., Kuan, P. F., Li, B., Sanalkumar, R., Liang, K., Bresnick, E. H., Dewey, C., and Keleş, S. (2011). Discovering Transcription Factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. PLoS Computational Biology, *7*, e1002111.

Cline, T. W., and Meyer, B. J. (1996). Vive la différence: males vs females in flies vs worms. Annual Review of Genetics, *30*, 637–702.

Clough, E., Jimenez, E., Kim, Y.-A., Whitworth, C., Neville, M. C., Hempel, L. U., Pavlou, H. J., Chen, Z.-X., Sturgill, D., Dale, R. K., et al. (2014). Sex- and tissue-specific functions of Drosophila Doublesex transcription factor target genes. Developmental Cell, *31*, 761–773.

Clyne, J. D., and Miesenböck, G. (2008). Sex-specific control and tuning of the pattern generator for courtship song in Drosophila. Cell, *133*, 354–363.

Clynen, E., Reumer, A., Baggerman, G., Mertens, I., and Schoofs, L. (2010). Neuropeptide biology in Drosophila. Advances in Experimental Medicine and Biology, *692*, 192–210.

Coast, G. M., Webster, S. G., Schegg, K. M., Tobe, S. S., and Schooley, D. A. (2001). The Drosophila melanogaster homologue of an insect calcitonin-like diuretic peptide stimulates V-ATPase activity in fruit fly Malpighian tubules. The Journal of Experimental Biology, *204*, 1795–1804.

Cook, R. (1979). The courtship tracking of Drosophila melanogaster. Biological Cybernetics, *34*, 91–106.

Cook, R., and Connolly, K. (1973). Rejection responses by female Drosophila melanogaster : their ontogeny, causality and effects upon the behaviour of the courting male. Behaviour, *44*, 142–165.

Cooley, L., Kelley, R., and Spradling, A. (1988). Insertional mutagenesis of the Drosophila genome with single P elements. Science, *239*, 1121–1128.

Coschigano, K. T., and Wensink, P. C. (1993). Sex-specific transcriptional regulation by the male and female doublesex proteins of Drosophila. Genes and Development, *7*, 42–54.

Crickmore, M. A., and Vosshall, L. B. (2013). Opposing dopaminergic and gabaergic neurons control the duration and persistence of copulation in Drosophila. Cell, *155*, 881.

Crocker, A., and Sehgal, A. (2010). Genetic analysis of sleep. Genes and development, *24*(12), 1220-1235.

Cronmiller, C., and Salz, H. K. (1994). The feminine mystique: the initiation of sex determination in Drosophila. In *Molecular Genetics of Sex Determination* (pp. 171-203). Academic Press.

D'Haeseleer, P. (2006). What are DNA sequence motifs? Nature Biotechnology, *24*, 423–425.

Daniels, S. B., Chovnick, A., and Boussy, I. A. (1990). Distribution of hobo transposable elements in the genus Drosophila. Molecular Biology and Evolution, *7*, 589–606.

Dauwalder, B. (2008). Systems behavior: of male courtship, the nervous system and beyond in Drosophila. Current Genomics, *9*, 517–524.

Dauwalder, B. (2011). The roles of fruitless and doublesex in the control of male courtship. In *International review of neurobiology* (Vol. 99, pp. 87-105). Academic Press.

Dauwalder, B., Tsujimoto, S., Moss, J., and Mattox, W. (2002). The Drosophila takeout gene is regulated by the somatic sex-determination pathway and affects male courtship behavior. Genes and Development, *16*, 2879–2892.

Day, D. S., Luquette, L. J., Park, P. J., and Kharchenko, P. V. (2010). Estimating enrichment of repetitive elements from high-throughput sequence data. Genome Biology, *11*, R69.

de Castro, J. P., and Carareto, C. M. A. (2004). Canonical P elements are transcriptionally active in the saltans group of Drosophila. Journal of Molecular Evolution, *59*, 31–40.

Deardorff, M. A., Wilde, J. J., Albrecht, M., Dickinson, E., Tennstedt, S., Braunholz, D., Mönnich, M., Yan, Y., Xu, W., Gil-Rodríguez, M. C., et al. (2012). RAD21 mutations cause a human cohesinopathy. American Journal of Human Genetics, *90*, 1014–1027.

DeFalco, T., Camara, N., Le Bras, S., and van Doren, M. (2008). Nonautonomous sex determination controls sexually dimorphic development of the Drosophila gonad. Developmental Cell, *14*, 275–286.

Demir, E., and Dickson, B. J. (2005). fruitless splicing specifies male courtship behavior in Drosophila. Cell, *121*, 785–794.

Deplancke, B., Dupuy, D., Vidal, M., and Walhout, A. J. M. (2004). A gateway-compatible yeast one-hybrid system. Genome Research, *14*, 2093–2101.

Devi, T. R., Amruthavalli, C., and Shyamala, B. V. (2013). Evolution of sex comb from

the primitive bristle pattern in Drosophila is associated with modification in the developmental regulatory protein dachshund. Genesis, *51*, 97–109.

Dickson, B. J. (2008). Wired for Sex: the neurobiology of Drosophila mating decisions. Science, *322*, 904–909.

Dobrosotskaya, I. Y., Seegmiller, A. C., Brown, M. S., Goldstein, J. L., and Rawson, R. B. (2002). Regulation of SREBP processing and membrane lipid production by phospholipids in Drosophila. Science, *296*, 879–883.

Doe, C. Q. (2008). Neural stem cells: Balancing self-renewal with differentiation. Development, *135*, 1575–1587.

Donlea, J. M., Pimentel, D., and Miesenböck, G. (2014). Neuronal machinery of sleep homeostasis in Drosophila. Neuron, *81*, 860–872.

Doudna, J. A., and Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. Science, *346*, 1258096.

Doupé, D. P., Marshall, O. J., Dayton, H., Brand, A. H., and Perrimon, N. (2018). Drosophila intestinal stem and progenitor cells are major sources and regulators of homeostatic niche signals. Proceedings of the National Academy of Sciences of the United States of America, *115*, 12218–12223.

Dudai, Y., Jan, Y. N., Byers, D., Quinn, W. G., and Benzer, S. (1976). dunce, a mutant of Drosophila deficient in learning. Proceedings of the National Academy of Sciences, *73*, 1684–1688.

Ejima, A., Nakayama, S., and Aigaki, T. (2001). Phenotypic association of spontaneous ovulation and sexual receptivity in virgin females of Drosophila melanogaster mutants. Behavior Genetics, *31*, 437–444.

Ellington, A. D., and Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. Nature, *346*, 818–822.

Elphick, M. R., Mirabeau, O., and Larhammar, D. (2018). Evolution of neuropeptide signalling systems. The Journal of Experimental Biology, *221*, jeb151092.

England, B. P., Admon, A., and Tjian, R. (1992). Cloning of Drosophila transcription factor Adf-1 reveals homology to Myb oncoproteins. Proceedings of the National Academy of Sciences, *89*, 683–687.

Erdman, S. E., and Burtis, K. C. (1993). The Drosophila doublesex proteins share a

novel zinc finger related DNA binding domain. The EMBO Journal, *12*, 527–535.

Erdman, S. E., Chen, H. J., and Burtis, K. C. (1996). Functional and genetic characterization of the oligomerization and DNA binding properties of the Drosophila doublesex proteins. Genetics, *144*, 1639–1652.

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, *32*, 3047–3048.

Ewen-Campen, B., Yang-Zhou, D., Fernandes, V. R., González, D. P., Liu, L.-P., Tao, R., Ren, X., Sun, J., Hu, Y., Zirin, J., et al. (2017). Optimized strategy for in vivo Cas9-activation in Drosophila. Proceedings of the National Academy of Sciences, *114*, 9409–9414.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature, *421*, 448–453.

Feschotte, C., and Pritham, E. J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annual Review of Genetics, *41*, 331–368.

Feyereisen, R. (1999). Insect P450 Enzymes. Annual Review of Entomology, *44*, 507–533.

Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, A. C., and Perrimon, N. (1990). The orthodenticle gene encodes a novel homeo domain protein involved in the development of the Drosophila nervous system and ocellar visual structures. Genes and Development, *4*, 1516–1527.

Finnegan, D. J. (1992). Transposable elements. Current Opinion in Genetics and Development, *2*, 861–867.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature, *391*, 806–811.

Fischer, J. A., Giniger, E., Maniatis, T., and Ptashne, M. (1988). GAL4 activates transcription in Drosophila. Nature, *332*, 853–856.

Fiston-Lavier, A.-S., Singh, N. D., Lipatov, M., and Petrov, D. A. (2010). Drosophila melanogaster recombination rate calculator. Gene, *463*, 18–20.

Fontana, J. R., and Crews, S. T. (2012). Transcriptome analysis of Drosophila CNS

midline cells reveals diverse peptidergic properties and a role for castor in neuronal differentiation. Developmental Biology, *372*, 131–142.

Fried, M., and Crothers, D. M. (1981). Equilibria and kinetics of lac repressor-operator interactions by polyacrylamide gel electrophoresis. Nucleic Acids Research, *9*, 6505–6525.

Furey, T. S. (2012). ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. Nature Reviews Genetics, *13*, 840–852.

Gailey, D. A., Billeter, J.-C., Liu, J. H., Bauzon, F., Allendorfer, J. B., and Goodwin, S. F. (2006). Functional conservation of the fruitless male sex-determination gene across 250 Myr of insect evolution. Molecular Biology and Evolution, *23*, 633–643.

Galas, D. J., and Schmitz, A. (1978). DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Research, *5*, 3157–3170.

Galton, F. (1874). *English Men of Science Their Nature and Nurture*. London: Macmillan and Co.

Garner, M. M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Research, *9*, 3047–3060.

Garrett-Engele, C. M., Siegal, M. L., Manoli, D. S., Williams, B. C., Li, H., and Baker, B. S. (2002). intersex, a gene required for female sexual development in Drosophila, is expressed in both sexes and functions together with doublesex to regulate terminal differentiation. Development, *129*, 4661–4675.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proceedings of the National Academy of Sciences, *109*, E2579–E2586.

Gaspar, J. M. (2018). Improved peak-calling with MACS2. BioRxiv, 496521.

Gehring, W. J. (1996). The master control gene for morphogenesis and evolution of the eye. Genes to Cells, *1*, 11–15.

Gendron, C. M., Kuo, T.-H., Harvanek, Z. M., Chung, B. Y., Yew, J. Y., Dierick, H. A., and Pletcher, S. D. (2014). Drosophila life span and physiology are modulated by sexual perception and reward. Science, *343*, 544–548.

Germann, S., and Gaudin, V. (2011). Mapping in vivo protein-DNA interactions in plants by DamID, a DNA adenine methylation-based method. Methods in Molecular Biology, *754*, 307–321.

Germann, S., Juul-Jensen, T., Letarnec, B., and Gaudin, V. (2006). DamID, a new tool for studying plant chromatin profiling in vivo , and its use to identify putative LHP1 target loci. The Plant Journal, *48*, 153–163.

Ghosh, N., Bakshi, A., Khandelwal, R., Rajan, S. G., and Joshi, R. (2019). The Hox gene Abdominal-B uses Doublesex ^{F} as a cofactor to promote neuroblast apoptosis in the Drosophila central nervous system. Development, *146*, dev175158.

Glossop, N. R., Lyons, L. C., and Hardin, P. E. (1999). Interlocked Feedback Loops Within the Drosophila Circadian Oscillator. Science, *286*, 766–768.

Goda, T., Doi, M., Umezaki, Y., Murai, I., Shimatani, H., Chu, M. L., Nguyen, V. H., Okamura, H., and Hamada, F. N. (2018). Calcitonin receptors are ancient modulators for rhythms of preferential temperature in insects and body temperature in mammals. Genes and Development, *32*, 140–155.

Goda, T., Tang, X., Umezaki, Y., Chu, M. L., Kunst, M., Nitabach, M. N., and Hamada, F. N. (2016). Drosophila DH31 Neuropeptide and PDF Receptor Regulate Night-Onset Temperature Preference. The Journal of Neuroscience, *36*, 11739–11754.

Godt, D., Couderc, J. L., Cramton, S. E., and Laski, F. A. (1993). Pattern formation in the limbs of Drosophila: bric a brac is expressed in both a gradient and a wave-like pattern and is required for specification and proper segmentation of the tarsus. Development, *119*, 799–812.

Goens, G., Rusu, D., Bultot, L., Goval, J. J., and Magdalena, J. (2009). Characterization and quality control of antibodies used in ChIP assays. Methods in Molecular Biology, *567*, 27–43.

Goldstein, A. (1964). *Biostatistics. An introductory text*. MacMillan Co., New York.

González-Aguilera, C., Ikegami, K., Ayuso, C., de Luis, A., Íñiguez, M., Cabello, J., Lieb, J. D., and Askjaer, P. (2014). Genome-wide analysis links emerin to

neuromuscular junction activity in Caenorhabditis elegans. Genome Biology, *15*.

Goodwin, S. F., Taylor, B. J., Villella, A., Foss, M., Ryner, L. C., Baker, B. S., and Hall, J. C. (2000). Aberrant splicing and altered spatial expression patterns in fruitless mutants of Drosophila melanogaster. Genetics, *154*, 725–745.

Gorfinkiel, N., Sánchez, L., and Guerrero, I. (2003). Development of the Drosophila genital disc requires interactions between its segmental primordia. Development, *130*, 295–305.

Gottschling, D. E. (1992). Telomere-proximal DNA in Saccharomyces cerevisiae is refractory to methyltransferase activity in vivo. Proceedings of the National Academy of Sciences, *89*, 4062–4065.

Granok, H., Leibovitch, B. A., and Elgin, S. C. R. (2001). A heat-shock-activated cDNA encoding GAGA factor rescues some lethal mutations in the Drosophila melanogaster Trithorax-like gene. Genetics Research, *78*, 13–21.

Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. Bioinformatics, *27*, 1017–1018.

Gratz, S. J., Cummings, A. M., Nguyen, J. N., Hamm, D. C., Donohue, L. K., Harrison, M. M., Wildonger, J., and O'Connor-Giles, K. M. (2013). Genome engineering of Drosophila with the CRISPR RNA-Guided Cas9 nuclease. Genetics, *194*, 1029–1035.

Greenspan, R. J. (2004). E pluribus unum, ex uno plura: quantitative and single-gene perspectives on the study of behavior. Annual Review of Neuroscience, *27*, 79–105.

Greenspan, R. J., and Ferveur, J.-F. (2000). Courtship in Drosophila. Annual Review of Genetics, *34*, 205–232.

Grillet, M., Dartevelle, L., and Ferveur, J. F. (2006). A Drosophila male pheromone affects female sexual receptivity. Proceedings of the Royal Society of London, *273*, 315–323.

Groth, A C, Olivares, E. C., Thyagarajan, B., and Calos, M. P. (2000). A phage integrase directs efficient site-specific integration in human cells. Proceedings of the National Academy of Sciences of the United States of America, *97*, 5995–6000.

Groth, Amy C, Fish, M., Nusse, R., and Calos, M. P. (2004). Construction of transgenic Drosophila by using the site-specific integrase from phage phiC31. Genetics, *166*, 1775–1782.

Guo, Y., Mahony, S., and Gifford, D. K. (2012). High resolution genome wide binding event finding and motif discovery reveals transcription factor spatial binding constraints. PLoS Computational Biology, *8*, e1002638.

Gutierrez-Triana, J. A., Mateo, J. L., Ibberson, D., Ryu, S., and Wittbrodt, J. (2016). iDamIDseq and iDEAR: An improved method and computational pipeline to profile chromatin-binding proteins. Development, *143*, 4272–4278.

Hall, J. C. (1979). Control of male reproductive behavior by the central nervous system of Drosophila: dissection of a courtship pathway by genetic mosaics. Genetics, *92*, 437–457.

Hanahan, D. (1983). Studies on transformation of Escherichia coli with plasmids. Journal of Molecular Biology, *166*, 557–580.

Handler, A. M., and James, A. A. (2000). *Insect transgenesis: methods and applications*. CRC Press.

Hansen, M. C., Nederby, L., Roug, A., Villesen, P., Kjeldsen, E., Nyvold, C. G., and Hokland, P. (2015). Novel scripts for improved annotation and selection of variants from whole exome sequencing in cancer research. MethodsX, *2*, 145–153.

Harmanci, A., Rozowsky, J., and Gerstein, M. (2014). MUSIC: identification of enriched regions in ChIP-Seq experiments using a mappability-corrected multiscale signal processing framework. Genome Biology, *15*, 474.

Hartenstein, V., Spindler, S., Pereanu, W., and Fung, S. (2008). The development of the Drosophila larval brain. In *Brain development in Drosophila melanogaster* (Vol. 628, pp. 1-31). Springer, New York.

Hass, M. R., Liow, H. haw, Chen, X., Sharma, A., Inoue, Y. U., Inoue, T., Reeb, A., Martens, A., Fulbright, M., Raju, S., et al. (2015). SpDamID: marking DNA bound by protein complexes identifies Notch-dimer responsive enhancers. Molecular Cell, *59*, 685–697.

Haunerland, N. H. (1996). Insect storage proteins: gene families and receptors. Insect Biochemistry and Molecular Biology, *26*, 755–765.

Hauser, F., Cazzamali, G., Williamson, M., Blenau, W., and Grimmelikhuijzen, C. J. P. (2006). A review of neurohormone GPCRs present in the fruitfly Drosophila melanogaster and the honey bee Apis mellifera. Progress in Neurobiology, *80*, 1–19.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y. C., Laslo, P., Cheng, J. X., Murre, C., Singh, H., and Glass, C. K. (2010). Simple combinations of lineage-determining transcription factors prime cis-Regulatory Elements required for macrophage and B cell identities. Molecular Cell, *38*, 576–589.

Hellman, L. M., and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. Nature Protocols, *2*, 1849–1861.

Hens, K., Feuz, J. D., Isakova, A., Iagovitina, A., Massouras, A., Bryois, J., Callaerts, P., Celniker, S. E., and Deplancke, B. (2011). Automated protein-DNA interaction screening of Drosophila regulatory elements. Nature Methods, *8*, 1065–1073.

Herranz, H., and Morata, G. (2001). The functions of pannier during Drosophila embryogenesis. Development, *128*, 4837–4846.

Herrmann, C., Van de Sande, B., Potier, D., and Aerts, S. (2012). i-cisTarget: an integrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic Acids Research, *40*, e114–e114.

Hewes, R. S., and Taghert, P. H. (2001). Neuropeptides and neuropeptide receptors in the Drosophila melanogaster genome. Genome Research, *11*, 1126–1142.

Ho, T.-Y., Wu, W.-H., Hung, S.-J., Liu, T., Lee, Y.-M., and Liu, Y.-H. (2019). Expressional profiling of carpet glia in the developing Drosophila eye reveals its molecular signature of morphology regulators. Frontiers in Neuroscience, *13*, 244.

Hökfelt, T., Johansson, O., Ljungdahl, A., Lundberg, J. M., and Schultzberg, M. (1980). Peptidergic neurones. Nature, *284*, 515–521.

Hong, C. S., Park, B.-Y., and Saint-Jeannet, J.-P. (2007). The function of Dmrt genes in vertebrate development: It is not just about sex. Developmental Biology, *310*, 1–9.

Horn, C., and Handler, A. M. (2005). Site-specific genomic targeting in Drosophila. Proceedings of the National Academy of Sciences of the United States of America, *102*, 12483.

Hoshizaki, D. K., Lunz, R., Johnson, W., and Ghosh, M. (1995). Identification of fat-

cell enhancer activity in Drosophila melanogaster using P-element enhancer traps. Genome, *38*, 497–506.

Huang, W., Loganantharaj, R., Schroeder, B., Fargo, D., and Li, L. (2013). PAVIS: a tool for Peak Annotation and Visualization. Bioinformatics, *29*, 3097–3099.

Hulsen, T., de Vlieg, J., and Alkema, W. (2008). BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. BMC Genomics, *9*, 488.

Hummon, A. B., Richmond, T. A., Verleyen, P., Baggerman, G., Huybrechts, J., Ewing, M. A., Vierstraete, E., Rodriguez-Zas, S. L., Schoofs, L., Robinson, G. E., et al. (2006). From the Genome to the Proteome: uncovering peptides in the Apis brain. Science, *314*, 647–649.

Hutson, S. F., and Bownes, M. (2003). The regulation of yp3 expression in the Drosophila melanogaster fat body. Development Genes and Evolution, *213*, 1–8.

Imrichová, H., Hulselmans, G., Kalender Atak, Z., Potier, D., and Aerts, S. (2015). i-cisTarget 2015 update: generalized cis-regulatory enrichment analysis in human, mouse and fly. Nucleic Acids Research, *43*, W57–W64.

Ito, H., Fujitani, K., Usui, K., Shimizu-Nishikawa, K., Tanaka, S., and Yamamoto, D. (1996). Sexual orientation in Drosophila is altered by the satori mutation in the sex-determination gene fruitless that encodes a zinc finger protein with a BTB domain. Proceedings of the National Academy of Sciences of the United States of America, *93*, 9687–9692.

Ito, H., Sato, K., Koganezawa, M., Ote, M., Matsumoto, K., Hama, C., and Yamamoto, D. (2012). Fruitless recruits two antagonistic chromatin factors to establish single-neuron sexual dimorphism. Cell, *149*, 1327–1338.

Ivics, Z., Mátés, L., Yau, T. Y., Landa, V., Zidek, V., Bashir, S., Hoffmann, O. I., Hiripi, L., Garrels, W., Kues, W. A., et al. (2014). Germline transgenesis in rodents by pronuclear microinjection of Sleeping Beauty transposons. Nature Protocols, *9*, 773–793.

Jaenisch, R., Jähner, D., Nobis, P., Simon, I., Löhler, J., Harbers, K., and Grotkopp, D. (1981). Chromosomal position and activation of retroviral genomes inserted into the germ line of mice. Cell, *24*, 519–529.

Jain, D., Baldi, S., Zabel, A., Straub, T., and Becker, P. B. (2015). Active promoters give rise to false positive "Phantom Peaks" in ChIP-seq experiments. Nucleic Acids Research, *43*, 6959–6968.

Jang, Y.-H., Chae, H.-S., and Kim, Y.-J. (2017). Female-specific myoinhibitory peptide neurons regulate mating receptivity in Drosophila melanogaster. Nature Communications, *8*, 1630.

Jékely, G., Melzer, S., Beets, I., Kadow, I. C. G., Koene, J., Haddad, S., and Holden-Dye, L. (2018). The long and the short of it – a perspective on peptidergic regulation of circuits and behaviour. The Journal of Experimental Biology, *221*, jeb166710.

Jiang, H., Lkhagva, A., Daubnerová, I., Chae, H.-S., Šimo, L., Jung, S.-H., Yoon, Y.-K., Lee, N.-R., Seong, J. Y., Žitňan, D., et al. (2013). Natalisin, a tachykinin-like signaling system, regulates sexual activity and fecundity in insects. Proceedings of the National Academy of Sciences of the United States of America, *110*, E3526-34.

Jiang, Z., Wu, X.-L., Michal, J. J., and McNamara, J. P. (2005). Pattern profiling and mapping of the fat body transcriptome in Drosophila melanogaster. Obesity Research, *13*, 1898–1904.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science, *337*, 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. ELife, *2*, 471.

Johnson, B. (1962). Neurosecretion and the transport of secretory material from the corpora cardiaca in aphids. Nature, *196*, 1338–1339.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-Wide Mapping of in Vivo Protein-DNA Interactions. Science, *316*, 1497–1502.

Joyce, G. F. (1989). Amplification, mutation and selection of catalytic RNA. Gene, *82*, 83–87.

Jünger, M. A., Rintelen, F., Stocker, H., Wasserman, J. D., Végh, M., Radimerski, T., Greenberg, M. E., and Hafen, E. (2003). The Drosophila forkhead transcription

factor FOXO mediates the reduction in cell number associated with reduced insulin signaling. Journal of Biology, *2*, 20.

Kaminker, J. S., Bergman, C. M., Kronmiller, B., Carlson, J., Svirskas, R., Patel, S., Frise, E., Wheeler, D. A., Lewis, S. E., Rubin, G. M., et al. (2002). The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biology, *3*, res0084-1.

Karess, R. E., and Rubin, G. M. (1984). Analysis of P transposable element functions in Drosophila. Cell, *38*, 135–146.

Kaufman, P. D., and Rio, D. C. (1991). Germline transformation of Drosophila melanogaster by purified P element transposase. Nucleic Acids Research, *19*, 6336–6336.

Keisman, E. L., Christiansen, A. E., and Baker, B. S. (2001). The sex determination gene doublesex regulates the A/P organizer to direct sex-specific patterns of growth in the Drosophila genital imaginal disc. Developmental Cell, *1*, 215–225.

Kelley, R. L., Wang, J., Bell, L., and Kuroda, M. I. (1997). Sex lethal controls dosage compensation in Drosophila by a non-splicing mechanism. Nature, *387*, 195–199.

Kharchenko, P. V., Tolstorukov, M. Y., and Park, P. J. (2008). Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nature Biotechnology, *26*, 1351–1359.

Kidwell, M. G., Kidwell, J. F., and Sved, J. A. (1977). Hybrid dysgenesis in Drosophila melanogaster: a syndrome of aberrant traits including mutation, sterility and male recombination. Genetics, *86*, 813–833.

Kim, M., Park, S., Cho, W., Gu, B., and Kim, C.-H. (2016). Neuropeptide Substance-P-conjugated Chitosan Nanofibers as an active modulator of stem cell recruiting. International Journal of Molecular Sciences, *17*, 68.

Kim, S. M., Su, C.-Y., and Wang, J. W. (2017). Neuromodulation of innate behaviors in Drosophila. Annual Review of Neuroscience, *40*, 327–348.

Kimura, K. I., Hachiya, T., Koganezawa, M., Tazawa, T., and Yamamoto, D. (2008). Fruitless and Doublesex coordinate to generate male-specific neurons that can initiate courtship. Neuron, *59*, 759–769.

Kimura, K. I., Ote, M., Tazawa, T., and Yamamoto, D. (2005). Fruitless specifies

sexually dimorphic neural circuitry in the Drosophila brain. Nature, *438*, 229–233.

Kimura, K. I., Sato, C., Koganezawa, M., and Yamamoto, D. (2015). Drosophila ovipositor extension in mating behavior and egg deposition involves distinct sets of brain interneurons. PLOS ONE, *10*, e0126445.

Kind, J., Pagie, L., De Vries, S. S., Nahidiazar, L., Dey, S. S., Bienko, M., Zhan, Y., Lajoie, B., De Graaf, C. A., Amendola, M., et al. (2015). Genome-wide maps of nuclear lamina interactions in single human cells. Cell, *163*, 134–147.

Kladde, M. P., and Simpson, R. T. (1994). Positioned nucleosomes inhibit Dam methylation in vivo. Proceedings of the National Academy of Sciences of the United States of America, *91*, 1361–1365.

Knowles, F., and Bern, H. A. (1966). Function of neurosecretion in endocrine regulation. Nature, *210*, 271–272.

Knowles, F. G. W. (1951). Hormone production within the nervous system of a crustacean. Nature, *167*, 564–565.

Koganezawa, M., Kimura, K., and Yamamoto, D. (2016). The neural circuitry that functions as a switch for courtship versus aggression in Drosophila males. Current Biology, *26*, 1395–1403.

Koh, H.-Y., Vilim, F. S., Jing, J., and Weiss, K. R. (2003). Two neuropeptides colocalized in a command-like neuron use distinct mechanisms to enhance its fast synaptic connection. Journal of Neurophysiology, *90*, 2074–2079.

Konopka, R. J., and Benzer, S. (1971). Clock mutants of Drosophila melanogaster. Proceedings of the National Academy of Sciences, *68*, 2112-2116.

Kopp, A., and True, J. R. (2002). Phylogeny of the oriental Drosophila melanogaster species group: a multilocus reconstruction. Systematic Biology, *51*, 786–805.

Korzelius, J., Naumann, S. K., Loza-Coll, M. A., Chan, J. S., Dutta, D., Oberheim, J., Gläßer, C., Southall, T. D., Brand, A. H., Jones, D. L., et al. (2014). Escargot maintains stemness and suppresses differentiation in Drosophila intestinal stem cells . The EMBO Journal, *33*, 2967–2982.

Kozhevnikova, E. N., Leshchenko, A. E., and Pindyurin, A. V. (2018). An inducible DamID system for profiling interactions of nuclear lamina protein component lamin B1 with chromosomes in mouse cells. Biochemistry, *83*, 586–594.

Koziol, M. J., Bradshaw, C. R., Allen, G. E., Costa, A. S. H., Frezza, C., and Gurdon, J. B. (2016). Identification of methylated deoxyadenosines in vertebrates reveals diversity in DNA modifications. Nature Structural and Molecular Biology, *23*, 24–30.

Krashes, M. J., Keene, A. C., Leung, B., Armstrong, J. D., and Waddell, S. (2007). Sequential use of mushroom body neuron subsets during Drosophila odor Memory processing. Neuron, *53*, 103–115.

Kravitz, E. A., and Huber, R. (2003). Aggression in invertebrates. Current Opinion in Neurobiology, *13*, 736-743.

Kroll, K. L., and Kirschner, M. W. (1999). Easy passage: germline transgenesis in frogs. Proceedings of the National Academy of Sciences of the United States of America, *96*, 14189–14190.

Kubli, E. (2003). Sex-peptides: seminal peptides of the Drosophila male. Cellular and Molecular Life Sciences, *60*, 1689–1704.

Kunst, M., Hughes, M. E., Raccuglia, D., Felix, M., Li, M., Barnett, G., Duah, J., and Nitabach, M. N. (2014). Calcitonin gene-related peptide neurons mediate sleep-specific circadian output in Drosophila. Current Biology, *24*, 2652–2664.

Kuo, T.-H., Fedina, T. Y., Hansen, I., Dreisewerd, K., Dierick, H. A., Yew, J. Y., and Pletcher, S. D. (2012). Insulin signaling mediates sexual attractiveness in Drosophila. PLoS Genetics, *8*, e1002684.

Lai, S.-L., and Lee, T. (2006). Genetic mosaic with dual binary transcriptional systems in Drosophila. Nature Neuroscience, *9*, 703–709.

Lakovaara, S., and Saura, A. (1982). Evolution and speciation in the Drosophila obscura group. *Genetics and biology of Drosophila* (pp. 1–59). London Academic Press.

Lam, E. W.-F., Brosens, J. J., Gomes, A. R., and Koo, C.-Y. (2013). Forkhead box proteins: tuning forks for transcriptional harmony. Nature Reviews Cancer, *13*, 482–495.

Langmead, B., and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-

efficient alignment of short DNA sequences to the human genome. Genome Biology, *10*, R25.

Lazareva, A. A., Roman, G., Mattox, W., Hardin, P. E., and Dauwalder, B. (2007). A role for the adult fat body in Drosophila male courtship behavior. PLoS Genetics, *3*, e16.

Lebo, M. S., Sanders, L. E., Sun, F., and Arbeitman, M. N. (2009). Somatic, germline and sex hierarchy regulated gene expression during Drosophila metamorphosis. BMC Genomics, *10*, 80.

Lebrun, E., Fourel, G., Defossez, P.-A., and Gilson, E. (2003). A methyltransferase targeting assay reveals silencer-telomere interactions in budding yeast. Molecular and Cellular Biology, *23*, 1498–1508.

Lee, G., Bahn, J. H., and Park, J. H. (2006). Sex- and clock-controlled expression of the neuropeptide F gene in Drosophila. Proceedings of the National Academy of Sciences, *103*, 12580–12585.

Lee, G., Foss, M., Goodwin, S. F., Carlo, T., Taylor, B. J., and Hall, J. C. (2000). Spatial, temporal, and sexually dimorphic expression patterns of the fruitless gene in theDrosophila central nervous system. Journal of Neurobiology, *43*, 404–426.

Lee, G., Hall, J. C., and Park, J. H. (2002). Doublesex gene expression in the central nervous system of Drosophila melanogaster. Journal of Neurogenetics, *16*, 229–248.

Lee, P. Y., Costumbrado, J., Hsu, C. Y., and Kim, Y. H. (2012). Agarose gel electrophoresis for the separation of DNA fragments. JoVE (Journal of Visualized Experiments), 62, e3923.

Lemeunier, F. D. J., Tsacas, L., and Ashburner, M. (1986). The melanogaster species group. In *The Genetics and Biology of Drosophila* (pp. 148–239). New York: Academic Press.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature, *424*, 147–151.

Lewis, E. B. (1978). A gene complex controlling segmentation in Drosophila. In *Genes, development and cancer* (pp. 205-217). Springer, Boston.

Lewis, E. B., and Bacher, F. (1968). Method of feeding ethylmethane sulfonate (EMS)

to Drosophila males. Drosophila Information Service. *43*,193.

Li, B., Predel, R., Neupert, S., Hauser, F., Tanaka, Y., Cazzamali, G., Williamson, M., Arakane, Y., Verleyen, P., Schoofs, L., et al. (2008). Genomics, transcriptomics, and peptidomics of neuropeptides and protein hormones in the red flour beetle Tribolium castaneum. Genome Research, *18*, 113–122.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, *25*, 1754–1760.

Li, R., Hempel, L. U., and Jiang, T. (2015). A non-parametric peak calling algorithm for DamID-Seq. PLoS ONE, *10*, 1–12.

Li, Y., Hoxha, V., Lama, C., Dinh, B. H., Vo, C. N., and Dauwalder, B. (2011). The hector G-Protein coupled receptor is required in a subset of fruitless neurons for male courtship behavior. PLoS ONE, *6*, e28269.

Lie-A-Ling, M., Marinopoulou, E., Li, Y., Patel, R., Stefanska, M., Bonifer, C., Miller, C., Kouskoff, V., and Lacaud, G. (2014). RUNX1 positively regulates a cell adhesion and migration program in murine hemogenic endothelium prior to blood emergence. Blood, *124*, e11–e20.

Lie, Y. S., and Petropoulos, C. J. (1998). Advances in quantitative PCR technology: 5′ nuclease assays. Current Opinion in Biotechnology, *9*, 43–48.

Lin, C., Garrett, A. S., De Kumar, B., Smith, E. R., Gogol, M., Seidel, C., Krumlauf, R., and Shilatifard, A. (2011). Dynamic transcriptional events in embryonic stem cells mediated by the super elongation complex (SEC). Genes and Development, *25*, 1486–1498.

Lin, W. H., Huang, L. H., Yeh, J. Y., Hoheisel, J., Lehrach, H., Sun, Y. H., and Tsai, S. F. (1995). Expression of a Drosophila GATA transcription factor in multiple tissues in the developing embryos: Identification of homozygous lethal mutants with P-element insertion at the promoter region. Journal of Biological Chemistry, *270*, 25150–25158.

Liu, H., and Kubli, E. (2003). Sex-peptide is the molecular basis of the sperm effect in Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America, *100*, 9929–9933.

Loza-Coll, M. A., Southall, T. D., Sandall, S. L., Brand, A. H., and Jones, D. L. (2014).

Regulation of Drosophila intestinal stem cell maintenance and differentiation by the transcription factor Escargot . The EMBO Journal, *33*, 2983–2996.

Luan, H., Peabody, N. C., Vinson, C. R., and White, B. H. (2006). Refined spatial manipulation of neuronal function by combinatorial restriction of transgene expression. Neuron, *52*, 425–436.

Luo, S. D., and Baker, B. S. (2015). Constraints on the evolution of a *doublesex* target gene arising from doublesex's pleiotropic deployment. Proceedings of the National Academy of Sciences, *112*, E852–E861.

Luo, S. D., Shi, G. W., and Baker, B. S. (2011). Direct targets of the D. melanogaster DSXF protein and the evolution of sexual development. Development, *138*, 2761–2771.

Luukkonen, B. G., Tan, W., and Schwartz, S. (1995). Efficiency of reinitiation of translation on human immunodeficiency virus type 1 mRNAs is determined by the length of the upstream open reading frame and by intercistronic distance. Journal of Virology, *69*, 4086–4094.

Lyne, R., Smith, R., Rutherford, K., Wakeling, M., Varley, A., Guillier, F., Janssens, H., Ji, W., Mclaren, P., North, P., et al. (2007). FlyMine: an integrated database for Drosophila and Anopheles genomics. Genome Biology, *8*, R129.

Mange, A. P. (1970). Possible nonrandom utilization of X- and Y-bearing sperm in Drosophila melanogaster. Genetics, *65*, 95–106.

Manning, A. (1967). The control of sexual receptivity in female Drosophila. Animal Behaviour, *15*, 239–250.

Manoli, D. S., Fan, P., Fraser, E. J., and Shah, N. M. (2013). Neural control of sexually dimorphic behaviors. Current Opinion in Neurobiology, *23*, 330–338.

Mardis, E. R. (2008). Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics, *9*, 387–402.

Markow, T. A. (1987). Behavioral and sensory basis of courtship success in Drosophila melanogaster. Proceedings of the National Academy of Sciences, *84*, 6200–6204.

Markow, T. A., Bustoz, D., and Pitnick, S. (1996). Sexual selection and a secondary sexual character in two Drosophila species. Animal Behaviour, *52*, 759–766.

Marshall, O. J., and Brand, A. H. (2015). damidseq-pipeline: An automated pipeline for

processing DamID sequencing datasets. Bioinformatics, *31*, 3371–3373.

Marshall, O. J., and Brand, A. H. (2017). Chromatin state changes during neural development revealed by in vivo cell-type specific profiling. Nature Communications, *8*, 2271.

Marshall, O. J., Southall, T. D., Cheetham, S. W., and 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*, 1586–1598.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, *17*, 10.

Martínez-Corrales, G., Cabrero, P., Dow, J. A. T., Terhzaz, S., and Davies, S.-A. (2019). Novel roles for GATAe in growth, maintenance and proliferation of cell populations in the Drosophila renal tubule. Development, *146*, dev178087.

Massouras, A., Decouttere, F., Hens, K., and Deplancke, B. (2010). WebPrInSeS: automated full-length clone sequence identification and verification using high-throughput sequencing data. Nucleic Acids Research, *38*, W378–W384.

Matthews, K. A., Kaufman, T. C., and Gelbart, W. M. (2005). Research resources for Drosophila: the expanding universe. Nature Reviews Genetics, *6*, 179–193.

McClintock, B. (1950). The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences, *36*, 344–355.

McGraw, L. A., Gibson, G., Clark, A. G., and Wolfner, M. F. (2004). Genes regulated by mating, sperm, or seminal proteins in mated female Drosophila melanogaster. Current Biology, *14*, 1509–1514.

McGuire, S. E., Deshazer, M., and Davis, R. L. (2005). Thirty years of olfactory learning and memory research in Drosophila melanogaster. In *Progress in Neurobiology* (Vol. 76, Issue 5, pp. 328–347). Elsevier.

Meister, M., Lemaitre, B., and Hoffmann, J. A. (1997). Antimicrobial peptide defense in Drosophila. BioEssays, *19*, 1019–1026.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., and Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research, *47*, D419–D426.

Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G.,

Alvarez, P., Brockman, W., Kim, T.-K., Koche, R. P., et al. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature, *448*, 553–560.

Modrich, P., and Lahue, R. (1996). Mismatch repair in replication fidelity, genetic recombination, and cancer biology. Annual Review of Biochemistry, *65*, 101–133.

Moorman, C., Sun, L. V., Wang, J., De Wit, E., Talhout, W., Ward, L. D., Greil, F., Lu, X. J., White, K. P., Bussemaker, H. J., et al. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proceedings of the National Academy of Sciences of the United States of America, *103*, 12027–12032.

Morgan, T. H. (1910). Sex limited inheritance in Drosophila. Science, *32*, 120–122.

Muller, H. J. (1927). Artificial transmutation of the gene. Science, *66*, 84–87.

Muller, H. J. (1932). Some Genetic Aspects of Sex. The American Naturalist, *66*, 118–138.

Murphy, D., and Carter, D. A. (1993). An Overview of Transgenic Mouse Production. In *Transgenesis Techniques* (Vol. 18, pp. 111–114). Humana Press.

Murphy, M. W., Sarver, A. L., Rice, D., Hatzi, K., Ye, K., Melnick, A., Heckert, L. L., Zarkower, D., and Bardwell, V. J. (2010). Genome-wide analysis of DNA binding and transcriptional regulation by the mammalian Doublesex homolog DMRT1 in the juvenile testis. Proceedings of the National Academy of Sciences of the United States of America, *107*, 13360–13365.

Murphy, M. W., Zarkower, D., and Bardwell, V. J. (2007). Vertebrate DM domain proteins bind similar DNA sequences and can heterodimerize on DNA. BMC Molecular Biology, *8*, 58.

Nagoshi, R. N., and Baker, B. S. (1990). Regulation of sex-specific RNA splicing at the Drosophila doublesex gene: Cis-acting mutations in exon sequences alter sex-specific RNA splicing patterns. Genes and Development, *4*, 89–97.

Nagoshi, R. N., McKeown, M., Burtis, K. C., Belote, J. M., and Baker, B. S. (1988). The control of alternative splicing at genes regulating sexual differentiation in D. melanogaster. Cell, *53*, 229–236.

Nakato, R., and Shirahige, K. (2017). Recent advances in ChIP-seq analysis: from

quality management to whole-genome annotation. Briefings in Bioinformatics, *18*, 279–290.

Nègre, N., Hennetin, J., Sun, L. V., Lavrov, S., Bellis, M., White, K. P., and Cavalli, G. (2006). Chromosomal distribution of PcG proteins during Drosophila development. PLoS Biology, *4*, 0917–0932.

Nelliot, A., Bond, N., and Hoshizaki, D. K. (2006). Fat-body remodeling in Drosophila melanogaster. Genesis, *44*, 396–400.

Nelson, J. D., Denisenko, O., and Bomsztyk, K. (2006). Protocol for the fast chromatin immunoprecipitation (ChIP) method. Nature Protocols, *1*, 179–185.

Neville, M. C., Nojima, T., Ashley, E., Parker, D. J., Walker, J., Southall, T., van de Sande, B., Marques, A. C., Fischer, B., Brand, A. H., et al. (2014). Male-specific fruitless isoforms target neurodevelopmental genes to specify a sexually dimorphic nervous system. Current Biology, *24*, 229–241.

Neville, M., and Goodwin, S. F. (2012). Genome-wide approaches to understanding behaviour in Drosophila melanogaster. Briefings in Functional Genomics, *11*, 395–404.

Nimmo, D. D., Alphey, L., Meredith, J. M., and Eggleston, P. (2006). High efficiency site-specific genetic engineering of the mosquito genome. Insect Molecular Biology, *15*, 129–136.

Nüsslein-Volhard, C., and Wieschaus, E. (1980). Mutations affecting segment number and polarity in Drosophila. Nature, *287*, 795–801.

O'Neill, L. P., and Turner, B. M. (1996). Immunoprecipitation of chromatin. Methods in Enzymology, *274*, 189–197.

Oberstein, A., Pare, A., Kaplan, L., and Small, S. (2005). Site-specific transgenesis by Cre-mediated recombination in Drosophila. Nature Methods, *2*, 583–585.

Oliphant, A. R., Brandl, C. J., and Struhl, K. (1989). Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Molecular and Cellular Biology, *9*, 2944–2949.

Olivecrona, H. (1954). Relation of the paraventricular nucleus to the pituitary gland. Nature, *173*, 1001–1001.

Oliver, B., Kim, Y. J., and Baker, B. S. (1993). Sex-lethal, master and slave: a hierarchy of germ-line sex determination in Drosophila. Development, *119*, 897–908.

Orlando, V., and Paro, R. (1993). Mapping Polycomb-repressed domains in the bithorax complex using in vivo formaldehyde cross-linked chromatin. Cell, *75*, 1187–1198.

Ostrowski, J., Kawata, Y., Schullery, D. S., Denisenko, O. N., and Bomsztyk, K. (2003). Transient recruitment of the hnRNP K protein to inducibly transcribed gene loci. Nucleic Acids Research, *31*, 3954–3962.

Otsuki, L., and Brand, A. H. (2018). Cell cycle heterogeneity directs the timing of neural stem cell activation from quiescence. Science, *360*, 99–102.

Overend, G., Cabrero, P., Guo, A. X., Sebastian, S., Cundall, M., Armstrong, H., Mertens, I., Schoofs, L., Dow, J. A. T., and Davies, S.-A. (2012). The receptor guanylate cyclase Gyc76C and a peptide ligand, NPLP1-VQQ, modulate the innate immune IMD pathway in response to salt stress. Peptides, *34*, 209–218.

Painter, T. S. (1933). A new method for the study of chromosome rearrangements and the plotting of chromosome maps. Science, *78*, 585–586.

Pan, Y., and Baker, B. S. (2014). Genetic identification and separation of innate and experience-dependent courtship behaviors in Drosophila. Cell, *156*, 236–248.

Pan, Y., Robinett, C. C., and Baker, B. S. (2011). Turning males on: activation of male courtship behavior in Drosophila melanogaster. PLoS ONE, *6*, e21144.

Parashar, N. C., Parashar, G., Nayyar, H., and Sandhir, R. (2018). N6-adenine DNA methylation demystified in eukaryotic genome: From biology to pathology. Biochimie, *144*, 56–62.

Park, D., Lee, Y., Bhupindersingh, G., and Iyer, V. R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. PLoS ONE, *8*, e83506.

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics, *10*, 669–680.

Pavlou, H. J., and Goodwin, S. F. (2013). Courtship behavior in Drosophila melanogaster: Towards a "courtship connectome." Current Opinion in Neurobiology, *23*, 76–83.

Pavlou, H. J., Lin, A. C., Neville, M. C., Nojima, T., Diao, F., Chen, B. E., White, B.

H., and Goodwin, S. F. (2016). Neural circuitry coordinating male copulation. eLife, *5*.

Pile, L. A., and Cartwright, I. L. (2000). GAGA factor-dependent transcription and establishment of DNase hypersensitivity are independent and unrelated events in vivo. Journal of Biological Chemistry, *275*, 1398–1404.

Pindyurin, A. V, Pagie, L., Kozhevnikova, E. N., van Arensbergen, J., and van Steensel, B. (2016). Inducible DamID systems for genomic mapping of chromatin proteins in Drosophila. Nucleic Acids Research, *44*, 5646–5657.

Pirone, D. M., Fukuhara, S., Gutkind, J. S., and Burbelo, P. D. (2000). SPECs, small binding proteins for Cdc42. Journal of Biological Chemistry, *275*, 22650–22656.

Polak, M., Starmer, W. T., and Wolf, L. L. (2004). Sexual selection for size and symmetry in a diversifying secondary sexual character in Drosophila bipectinata Duda (Diptera: Drosophilidae). Evolution; International Journal of Organic Evolution, *58*, 597–607.

Pollock, R., and Treisman, R. (1990). A sensitive method for the determination of protein-DNA binding specificities. Nucleic Acids Research, *18*, 6197–6204.

Raymond, C. S., Kettlewell, J. R., Hirsch, B., Bardwell, V. J., and Zarkower, D. (1999). Expression of Dmrt1 in the genital ridge of mouse and chicken embryos suggests a role in vertebrate sexual development. Developmental Biology, *215*, 208–220.

Rehorn, K. P., Thelen, H., Michelson, A. M., and Reuter, R. (1996). A molecular aspect of hematopoiesis and endoderm development common to vertebrates and Drosophila. Development, *122*, 4023–4031.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. (2000). Genome-wide location and function of DNA binding proteins. Science, *290*, 2306–2309.

Ren, Q., Awasaki, T., Huang, Y. F., Liu, Z., and Lee, T. (2016). Cell class-lineage analysis reveals sexually dimorphic lineage compositions in the Drosophila brain. Current Biology, *26*, 2583–2593.

Rezával, C., Pattnaik, S., Pavlou, H. J., Nojima, T., Brüggemeier, B., D'Souza, L. A. D., Dweck, H. K. M., and Goodwin, S. F. (2016). Activation of latent courtship circuitry in the brain of Drosophila females induces male-like behaviors. Current

Biology, *26*, 2508–2515.

Rezával, C., Pavlou, H. J., Dornan, A. J., Chan, Y.-B., Kravitz, E. A., and Goodwin, S. F. (2012). Neural circuitry underlying Drosophila female postmating behavioral responses. Current Biology, *22*, 1155–1165.

Rhee, H. S., and Pugh, B. F. (2011). Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. Cell, *147*, 1408–1419.

Rhee, H. S., and Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. Nature, *483*, 295–301.

Richmond, J. E., and Broadie, K. S. (2002). The synaptic vesicle cycle: Exocytosis and endocytosis in Drosophila and C. elegans. In *Current Opinion in Neurobiology* (Vol. 12, Issue 5, pp. 499–507). Elsevier Ltd.

Rideout, E. J., Billeter, J.-C., and Goodwin, S. F. (2007). The sex-determination genes fruitless and doublesex specify a neural substrate required for courtship song. Current Biology, *17*, 1473–1478.

Rideout, E. J., Dornan, A. J., Neville, M. C., Eadie, S., and Goodwin, S. F. (2010). Control of sexual differentiation and behavior by the doublesex gene in Drosophila melanogaster. Nature Neuroscience, *13*, 458–466.

Riehle, M. A., Garczynski, S. F., Crim, J. W., Hill, C. A., and Brown, M. R. (2002). Neuropeptides and peptide hormones in *Anopheles gambiae*. Science, *298*, 172–175.

Ringrose, L. (2009). Transgenesis in Drosophila melanogaster. In *Methods in Molecular Biology* (Vol. 561, pp. 3–19).

Robertson, G., Hirst, M., Bainbridge, M., Bilenky, M., Zhao, Y., Zeng, T., Euskirchen, G., Bernier, B., Varhol, R., Delaney, A., et al. (2007). Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods, *4*, 651–657.

Robinett, C. C., Vaughan, A. G., Knapp, J.-M., and Baker, B. S. (2010). Sex and the Single Cell. II. There is a time and place for sex. PLoS Biology, *8*, e1000365.

Rong, Y. S., and Golic, K. G. (2000). Gene targeting by homologous recombination in Drosophila. Science, *288*, 2013–2018.

Rong, Y. S., and Golic, K. G. (2001). A targeted gene knockout in Drosophila.

Genetics, *157*, 1307–1312.

Rørth, P. (1996). A modular misexpression screen in Drosophila detecting tissue-specific phenotypes. Proceedings of the National Academy of Sciences of the United States of America, *93*, 12418–12422.

Rossant, J., Nutter, L. M. J., and Gertsenstein, M. (2011). Engineering the embryo. Proceedings of the National Academy of Sciences of the United States of America, *108*, 7659–7660.

Rubin, G., and Spradling, A. (1982). Genetic transformation of Drosophila with transposable element vectors. Science, *218*, 348–353.

Ryner, L. C., Goodwin, S. F., Castrillon, D. H., Anand, A., Villella, A., Baker, B. S., Hall, J. C., Taylor, B. J., and Wasserman, S. A. (1996). Control of male sexual behavior and sexual orientation in Drosophila by the fruitless gene. Cell, *87*, 1079–1089.

Salz, H. K., and Erickson, J. W. (2010). Sex determination in Drosophila: the view from the top. Fly, *4*, 60–70.

Sam, S., Leise, W., and Hoshizaki, D. K. (1996). The serpent gene is necessary for progression through the early stages of fat-body development. Mechanisms of Development, *60*, 197–205.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (2001). *Molecular cloning : a laboratory manual*.

Sanders, L. E., and Arbeitman, M. N. (2008). Doublesex establishes sexual dimorphism in the Drosophila central nervous system in an isoform-dependent manner by directing cell number. Developmental Biology, *320*, 378–390.

Schübeler, D., and Elgin, S. C. R. (2005). Defining epigenetic states through chromatin and RNA. Nature Genetics, *37*, 917–918.

Schuldt, A. J., Adams, J. H. J., Davidson, C. M., Micklem, D. R., Haseloff, J., Johnston, D. S., and Brand, A. H. (1998). Miranda mediates asymmetric protein and RNA localization in the developing nervous system. Genes and Development, *12*, 1847–1857.

Schuster, E., McElwee, J. J., Tullet, J. M. A., Doonan, R., Matthijssens, F., Reece-Hoyes, J. S., Hope, I. A., Vanfleteren, J. R., Thornton, J. M., and Gems, D. (2010).

DamID in C. elegans reveals longevity-associated targets of DAF-16/FoxO. Molecular Systems Biology, *6*.

Sellami, A., and Veenstra, J. A. (2015). SIFamide acts on fruitless neurons to modulate sexual behavior in Drosophila melanogaster. Peptides, *74*, 50–56.

Sen, S. Q., Chanchani, S., Southall, T. D., and Doe, C. Q. (2019). Neuroblast-specific open chromatin allows the temporal transcription factor, Hunchback, to bind neuroblast-specific loci. eLife, *8*, e44036.

Shaner, N. C., Campbell, R. E., Steinbach, P. A., Giepmans, B. N. G., Palmer, A. E., and Tsien, R. Y. (2004). Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein. Nature Biotechnology, *22*, 1567–1572.

Shankar, S., Chua, J. Y., Tan, K. J., Calvert, M. E., Weng, R., Ng, W. C., Mori, K., and Yew, J. Y. (2015). The neuropeptide tachykinin is essential for pheromone detection in a gustatory neural circuit. eLife, *4*, e06914.

Shendure, J., and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechnology, *26*, 1135–1145.

Shimbo, T., Du, Y., Grimm, S. A., Dhasarathy, A., Mav, D., Shah, R. R., Shi, H., and Wade, P. A. (2013). MBD3 localizes at promoters, gene bodies and enhancers of active genes. PLoS Genetics, *9*, e1004028.

Shirangi, T. R., Dufour, H. D., Williams, T. M., and Carroll, S. B. (2009). Rapid evolution of sex pheromone-producing enzyme expression in Drosophila. PLoS Biology, *7*, e1000168.

Shirangi, T. R., and McKeown, M. (2007). Sex in flies: What "body-mind" dichotomy? Developmental Biology, *306*, 10–19.

Shirangi, T. R., Wong, A. M., Truman, J. W., and Stern, D. L. (2016). Doublesex regulates the connectivity of a neural circuit controlling Drosophila male courtship song. Developmental Cell, *37*, 533–544.

Siggs, O., and Beutler, B. (2012). The BTB-ZF transcription factors. Cell Cycle, *11*, 3358–3369.

Simpson, J. L. (1999). Disorders of sexual differentiation. American Journal of Medical Genetics, *89*, 175–175.

Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014). Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics, *15*, 121–132.

Sims, R. J. (2004). Elongation by RNA polymerase II: the short and long of it. Genes and Development, *18*, 2437–2468.

Singh, J., and Klar, A. J. (1992). Active genes in budding yeast display enhanced in vivo accessibility to foreign DNA methylases: a novel in vivo probe for chromatin structure of yeast. Genes and Development, *6*, 186–196.

Snyder, L. A., Loman, N. J., Linton, J. D., Langdon, R. R., Weinstock, G. M., Wren, B. W., and Pallen, M. J. (2010). Simple sequence repeats in Helicobacter canadensis and their role in phase variable expression and C-terminal sequence switching. BMC Genomics, *11*, 67.

Solomon, M. J., Larsen, P. L., and Varshavsky, A. (1988). Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene. Cell, *53*, 937–947.

Sousa-Nunes, R., Yee, L. L., and Gould, A. P. (2011). Fat cells reactivate quiescent neuroblasts via TOR and glial insulin relays in Drosophila. Nature, *471*, 508–512.

Southall, T. D., Gold, K. S., Egger, B., Davidson, C. M., Caygill, E. E., Marshall, O. J., and 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*, 101–112.

Spéder, P., and Brand, A. H. (2018). Systemic and local cues drive neural stem cell niche remodelling during neurogenesis in Drosophila. eLife, *7*, e30413.

Spieth, H., and Ringo, J. (1983). Mating behaviour and sexual isolation in Drosophila. In *The genetics and biology of Drosophila*. (Vol. 3c, pp. 224–270).

Spieth, H. T. (1952). Mating behavior within the genus Drosophila (Diptera). Bulletin of the American Museum of Natural History, *99*, 7.

Spradling, A. C., Stern, D. M., Kiss, I., Roote, J., Laverty, T., and Rubin, G. M. (1995). Gene disruptions using P transposable elements: an integral component of the Drosophila genome project. Proceedings of the National Academy of Sciences of the United States of America, *92*, 10824–10830.

Steglich, B., Filion, G. J., van Steensel, B., and Ekwall, K. (2012). The inner nuclear membrane proteins Man1 and Ima1 link to two different types of chromatin at the nuclear periphery in S. pombe. Nucleus, *3*, 77–87.

Stein, W., DeLong, N. D., Wood, D. E., and Nusbaum, M. P. (2007). Divergent co-transmitter actions underlie motor pattern activation by a modulatory projection neuron. European Journal of Neuroscience, *26*, 1148–1165.

Steinmann-Zwicky, M., and Nöthiger, R. (1985). A small region on the X chromosome of Drosophila regulates a key gene that controls sex determination and dosage compensation. Cell, *42*, 877–887.

Stockinger, P., Kvitsiani, D., Rotkopf, S., Tirián, L., and Dickson, B. J. (2005). Neural circuitry that governs Drosophila male courtship behavior. Cell, *121*, 795–807.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Journal of Experimental Zoology, *14*, 43–59.

Sturtevant, A. H. (1915). Experiments on sex recognition and the problem of sexual selection in Drosoophilia. Journal of Animal Behavior, *5*, 351–366.

Sturtevant, A. H. (1925). The effects of unequal crossing over at the Bar locus in Drosophila. Genetics, *10*, 117–147.

Südhof, T. C. (2004). The synaptic vesicle cycle. Annual Review of Neuroscience, *27*, 509–547.

Suzuki, M. G., and Shimada., T. (2013). Transgenic analysis of the biological functions of a Doublesex homologue in Bombyx mori. In *Madame Curie Bioscience Database*.

Sweeney, S. T., Broadie, K., Keane, J., Niemann, H., and O'Kane, C. J. (1995). Targeted expression of tetanus toxin light chain in Drosophila specifically eliminates synaptic transmission and causes behavioral defects. Neuron, *14*, 341–351.

Taghert, P. H., and Nitabach, M. N. (2012). Peptide neuromodulation in invertebrate model systems. Neuron, *76*, 82–97.

Takeda, K., Okumura, T., Terahata, M., Yamaguchi, M., Taniguchi, K., and Adachi-Yamada, T. (2018). Drosophila peptide hormones Allatostatin A and Diuretic

Hormone 31 exhibiting complementary gradient distribution in posterior midgut antagonistically regulate midgut senescence and adult Lifespan. Zoological Science, *35*, 75.

Tamirisa, S., Papagiannouli, F., Rempel, E., Ermakova, O., Trost, N., Zhou, J., Mundorf, J., Brunel, S., Ruhland, N., Boutros, M., et al. (2018). Decoding the regulatory logic of the Drosophila male stem cell system. Cell Reports, *24*, 3072–3086.

Taylor, B. J., and Truman, J. W. (1992). Commitment of abdominal neuroblasts in Drosophila to a male or female fate is dependent on genes of the sex-determining hierarchy. Development, *114*, 625–642.

Temin, H. M., and Mizutani, S. (1970). Viral RNA-dependent DNA Polymerase: RNA-dependent DNA Polymerase in Virions of Rous Sarcoma Virus. Nature, *226*, 1211–1213.

Terhzaz, S., Rosay, P., Goodwin, S. F., and Veenstra, J. A. (2007). The neuropeptide SIFamide modulates sexual behavior in Drosophila. Biochemical and Biophysical Research Communications, *352*, 305–310.

Teytelman, L., Thurtle, D. M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proceedings of the National Academy of Sciences of the United States of America, *110*, 18602–18607.

Thiriet, C., and Hayes, J. J. (2005). Replication-independent core histone dynamics at transcriptionally active loci in vivo. Genes and Development, *19*, 677–682.

Thomas, R., Thomas, S., Holloway, A. K., and Pollard, K. S. (2017). Features that define the best ChIP-seq peak calling algorithms. Briefings in Bioinformatics, *18*, 441–450.

Thorpe, H. M., and Smith, M. C. M. (1998). In vitro site-specific integration of bacteriophage DNA catalyzed by a recombinase of the resolvase/invertase family. Proceedings of the National Academy of Sciences, *95*, 5505–5510.

Thyagarajan, B., Olivares, E. C., Hollis, R. P., Ginsburg, D. S., and Calos, M. P. (2001). Site-specific genomic integration in mammalian cells mediated by Phage C31 integrase. Molecular and Cellular Biology, *21*, 3926–3934.

Tice, S. C. (1914). A new Sex-linked character in Drosophila. The Biological Bulletin, *26*, 221–230.

Timmerman, C., Suppiah, S., Gurudatta, B. V., Yang, J., Banerjee, C., Sandstrom, D. J., Corces, V. G., and Sanyal, S. (2013). The Drosophila transcription factor Adf-1 (nalyot) regulates dendrite growth by controlling FasII and Staufen expression downstream of CaMKII and neural activity. Journal of Neuroscience, *33*, 11916–11931.

Tito, A. J., Cheema, S., Jiang, M., and Zhang, S. (2016). A simple one-step dissection protocol for whole-mount preparation of adult Drosophila brains. Journal of Visualized Experiments, *118*.

Tompkins, L., Siegel, R. W., Gailey, D. A., and Hall, J. C. (1983). Conditioned courtship in Drosophila and its mediation by association of chemical cues. Behavior Genetics, *13*, 565–578.

Torres-Vazquez, J., Warrior, R., and Arora, K. (2000). schnurri is required for dpp-dependent patterning of the Drosophila wing. Developmental Biology, *227*, 388–402.

Tosti, L., Ashmore, J., Tan, B. S. N., Carbone, B., Mistri, T. K., Wilson, V., Tomlinson, S. R., and Kaji, K. (2018). Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo. Genome Research, *28*, 592–605.

Towbin, B. D., González-Aguilera, C., Sack, R., Gaidatzis, D., Kalck, V., Meister, P., Askjaer, P., and Gasser, S. M. (2012). Step-wise methylation of histone H3K9 positions heterochromatin at the nuclear periphery. Cell, *150*, 934–947.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science, *249*, 505–510.

Turner, B. M., and O'Neill, L. P. (1995). Histone acetylation in chromatin and chromosomes. Seminars in Cell Biology, *6*, 229–236.

Ueyama, M., Chertemps, T., Labeur, C., and Wicker-Thomas, C. (2005). Mutations in the desat1 gene reduces the production of courtship stimulatory pheromones through a marked effect on fatty acids in Drosophila melanogaster. Insect Biochemistry and Molecular Biology, *35*, 911–920.

Usui-Aoki, K., Ito, H., Ui-Tei, K., Takahashi, K., Lukacsovich, T., Awano, W., Nakata, H., Piao, Z. F., Nilsson, E. E., Tomida, J., et al. (2000). Formation of the male-specific muscle in female Drosophila by ectopic fruitless expression. Nature Cell Biology, *2*, 500–506.

van den Pol, A. N. (2012). Neuropeptide transmission in brain circuits. Neuron, *76*, 98–115.

van Blokland, H. J. M., Hoeksema, F., Siep, M., Otte, A. P., and Verhees, J. A. (2011). Methods to create a stringent selection system for mammalian cell lines. Cytotechnology, *63*, 371–384.

van Doren, M., Bailey, A. M., Esnayra, J., Ede, K., and Posakony, J. W. (1994). Negative regulation of proneural gene activity: hairy is a direct transcriptional repressor of achaete. Genes and Development, *8*, 2729–2742.

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

van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered Dam methyltransferase. Nature Biotechnology, *18*, 424–428.

Vanden Broeck, J. (2001). Neuropeptides and their precursors in the fruitfly, Drosophila melanogaster. Peptides, *22*, 241–254.

Venken, K. J. T., and Bellen, H. J. (2005). Emerging technologies for gene manipulation in Drosophila melanogaster. Nature Reviews Genetics, *6*, 167–178.

Venken, K. J. T., and Bellen, H. J. (2007). Transgenesis upgrades for Drosophila melanogaster. Development, *134*, 3571–3584.

Venken, K. J. T., He, Y., Hoskins, R. A., and Bellen, H. J. (2006). P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in D. melanogaster. Science, *314*, 1747–1751.

Venken, K. J. T., Schulze, K. L., Haelterman, N. A., Pan, H., He, Y., Evans-Holm, M., Carlson, J. W., Levis, R. W., Spradling, A. C., Hoskins, R. A., et al. (2011). MiMIC: A highly versatile transposon insertion resource for engineering Drosophila melanogaster genes. Nature Methods, *8*, 737–747.

Verleyen, P., Baggerman, G., Wiehart, U., Schoeters, E., Van Lommel, A., De Loof,

A., and Schoofs, L. (2004). Expression of a novel neuropeptide, NVGTLARDFQLPIPNamide, in the larval and adult brain of Drosophila melanogaster. Journal of Neurochemistry, *88*, 311–319.

Verleyen, P., Chen, X., Baron, S., Preumont, A., Hua, Y. J., Schoofs, L., and Clynen, E. (2009). Cloning of neuropeptide-like precursor 1 in the gray flesh fly and peptide identification and expression. Peptides, *30*, 522–530.

Vidal, M., and Cagan, R. L. (2006). Drosophila models for cancer research. Current Opinion in Genetics and Development, *16*, 10–16.

Villella, A., and Hall, J. C. (2008). Neurogenetics of Courtship and Mating in Drosophila. Advances in Genetics, *62*, 67–184.

Vissers, J. H. A., Froldi, F., Schrö, J., Papenfuss, A. T., Cheng, L. Y., and Correspondence, K. F. H. (2018). The Scalloped and Nerfin-1 transcription factors cooperate to maintain neuronal cell fate. Cell Reports, *25*, 1561–1576.

Vogel, M. J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. Nature Protocols, *2*, 1467–1478.

von Schilcher, F. (1976a). The function of pulse song and sine song in the courtship of Drosophila melanogaster. Animal Behaviour, *24*, 622–625.

von Schilcher, F. (1976b). The role of auditory stimuli in the courtship of Drosophila melanogaster. Animal Behaviour, *24*, 18–26.

Vosshall, L. B., Price, J. L., Sehgal, A., Saez, L., and Young, M. W. (1994). Block in nuclear localization of period protein by a second clock mutation, timeless. Science, *263*, 1606–1609.

Waterbury, J. A., Jackson, L. L., and Schedl, P. (1999). Analysis of the doublesex female protein in Drosophila melanogaster: role on sexual differentiation and behavior and dependence on intersex. Genetics, *152*, 1653–1667.

Wawersik, M., Milutinovich, A., Casper, A. L., Matunis, E., Williams, B., and van Doren, M. (2005). Somatic control of germline sexual development is mediated by the JAK/STAT pathway. Nature, *436*, 563–567.

Weirauch, M. T., Cote, A., Norel, R., Annala, M., Zhao, Y., Riley, T. R., Saez-Rodriguez, J., Cokelaer, T., Vedenko, A., Talukder, S., et al. (2013). Evaluation of

methods for modeling transcription factor sequence specificity. Nature Biotechnology, *31*, 126–134.

Widmer, Y. F., Bilican, A., Bruggmann, R., and Sprecher, S. G. (2018). Regulators of long-term memory revealed by mushroom body-specific gene expression profiling in Drosophila melanogaster. Genetics, *209*, 1167–1181.

Wigby, S., and Chapman, T. (2005). Sex peptide causes mating costs in female Drosophila melanogaster supplental experimental procedures. Current Biology, *15*, 316–321.

Williams, T. M., Selegue, J. E., Werner, T., Gompel, N., Kopp, A., and Carroll, S. B. (2008). The regulation and evolution of a genetic switch controlling sexually dimorphic traits in Drosophila. Cell, *134*, 610–623.

Wilson, J. B., Weinberg, W., Johnson, R., Yuspa, S., and Levine, A. J. (1990). Expression of the BNLF-1 oncogene of Epstein-Barr virus in the skin of transgenic mice induces hyperplasia and aberrant expression of keratin 6. Cell, *61*, 1315–1327.

Wilson, T. G. (2001). Resistance of Drosophila to Toxins. Annual Review of Entomology, *46*, 545–571.

Wines, D. R., Talbert, P. B., Clark, D. V, and Henikoff, S. (1996). Introduction of a DNA methyltransferase into Drosophila to probe chromatin structure in vivo. Chromosoma, *104*, 332–340.

Winick, J., Abel, T., Leonard, M. W., Michelson, A. M., Chardon-Loriaux, I., Holmgren, R. A., Maniatis, T., and Engel, J. D. (1993). A GATA family transcription factor is expressed along the embryonic dorsoventral axis in Drosophila melanogaster. Development, *119*, 1055–1065.

Wolfner, M. F. (1988). Sex-specific gene expression in somatic tissues of Drosophila melanogaster. Trends in Genetics, *4*, 333–337.

Wolfram, V., Southall, T. D., Günay, C., Prinz, A. A., Brand, A. H., and Baines, R. A. (2014). The transcription factors islet and lim3 combinatorially regulate ion channel gene expression. Journal of Neuroscience, *34*, 2538–2543.

Woolcock, K. J., Gaidatzis, D., Punga, T., and Bühler, M. (2011). Dicer associates with chromatin to repress genome activity in Schizosaccharomyces pombe. Nature

Structural and Molecular Biology, *18*, 94–100.

Worthington, W. C. (1966). Blood samples from the pituitary stalk of the rat: Method of collection and factors determining volume. Nature, *210*, 710–712.

Wu, F., and Yao, J. (2013). Spatial compartmentalization at the nuclear periphery characterized by genome-wide mapping. BMC Genomics, *14*, 591.

Xu, G.-L., and Bestor, T. H. (1997). Cytosine methylation targetted to pre-determined sequences. Nature Genetics, *17*, 376–378.

Yamamoto, D. (2008). Brain sex differences and function of the fruitless gene in Drosophila. Journal of Neurogenetics, *22*, 309–332.

Yamamoto, D., and Koganezawa, M. (2013). Genes and circuits of courtship behaviour in Drosophila males. Nature Reviews Neuroscience, *14*, 681–692.

Yang, Y., Zhang, W., Bayrer, J. R., and Weiss, M. A. (2008). Doublesex and the regulation of sexual dimorphism in Drosophila melanogaster: Structure, function, and mutagenesis of a female-specific domain. Journal of Biological Chemistry, *283*, 7280–7292.

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mpl8 and pUC19 vectors. Gene, *33*, 103–119.

Yi, W., and Zarkower, D. (1999). Similarity of DNA binding and transcriptional regulation by Caenorhabditis elegans MAB-3 and Drosophila melanogaster DSX suggests conservation of sex determining mechanisms. Development, *126*, 873–881.

Yongmei Xi, Y. Z. (2015). Fat Body development and its function in energy storage and nutrient sensing in Drosophila melanogaster. Journal of Tissue Science and Engineering, *06*.

Zarin, A. A., Daly, A. C., Hulsmeier, J., Asadzadeh, J., and Labrador, J.-P. (2012). A GATA/homeodomain transcriptional code regulates axon guidance through the Unc-5 receptor. Development, *139*, 1798–1805.

Zarkower, D. (2013). DMRT Genes in vertebrate gametogenesis. Current Topics in Developmental Biology, *102*, 327–356.

Zehring, W. A., Wheeler, D. A., Reddy, P., Konopka, R. J., Kyriacou, C. P., Rosbash,

M., and Hall, J. C. (1984). P-element transformation with period locus DNA restores rhythmicity to mutant, arrhythmic Drosophila melanogaster. Cell, *39*, 369–376.

Zhang, W., Li, B., Singh, R., Narendra, U., Zhu, L., and Weiss, M. A. (2006). Regulation of sexual dimorphism: mutational and chemogenetic analysis of the Doublesex DM domain. Molecular and Cellular Biology, *26*, 535–547.

Zhang, X., Germann, S., Blus, B. J., Khorasanizadeh, S., Gaudin, V., and Jacobsen, S. E. (2007). The Arabidopsis LHP1 protein colocalizes with histone H3 Lys27 trimethylation. Nature Structural and Molecular Biology, *14*, 869–871.

Zhang, Yan, Heidrich, N., Ampattu, B. J., Gunderson, C. W., Seifert, H. S., Schoen, C., Vogel, J., and Sontheimer, E. J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Molecular Cell, *50*, 488–503.

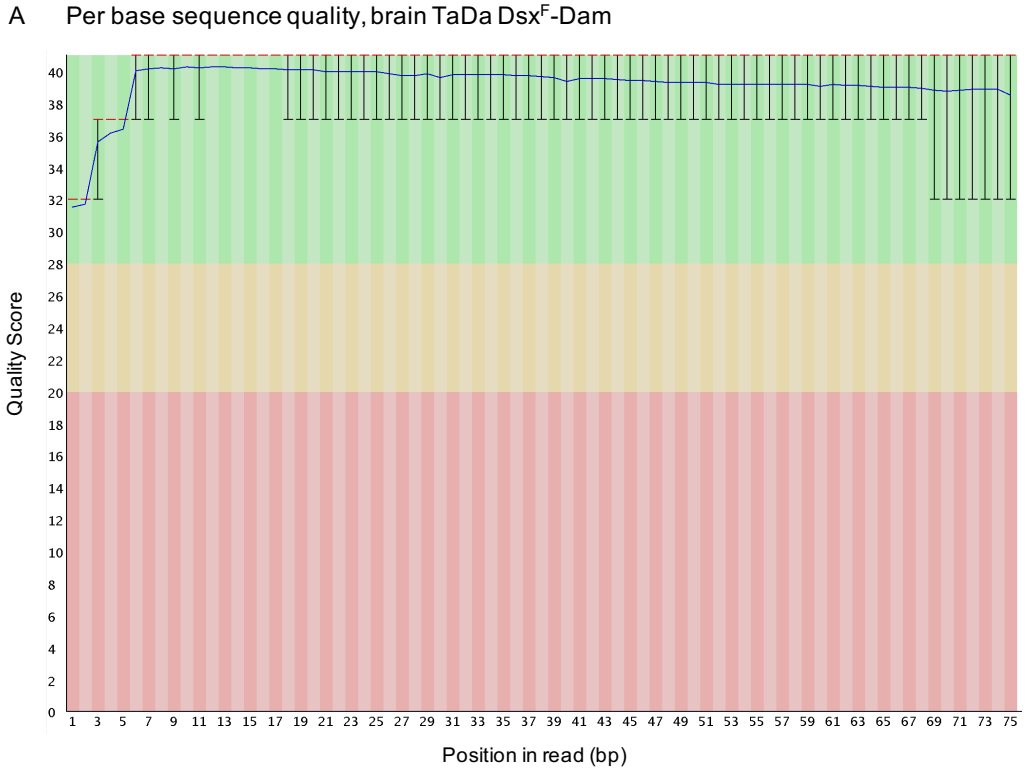
Zhang, Yong, Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biology, *9*, R137.

Zhou, C., Pan, Y., Robinett, C. C., Meissner, G. W., and Baker, B. S. (2014). Central brain neurons expressing doublesex regulate female receptivity in Drosophila. Neuron, *83*, 149–163.

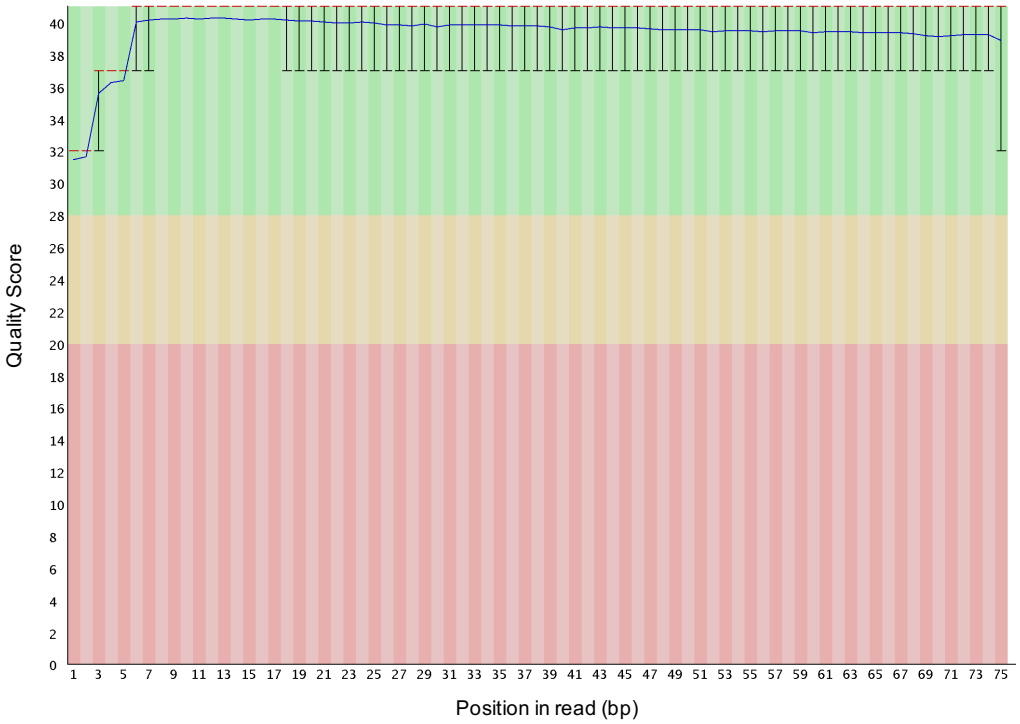
9 APPENDICES

9.1 ASSESSING READ QUALITY WITH FASTQC (ANDREWS, 2010)

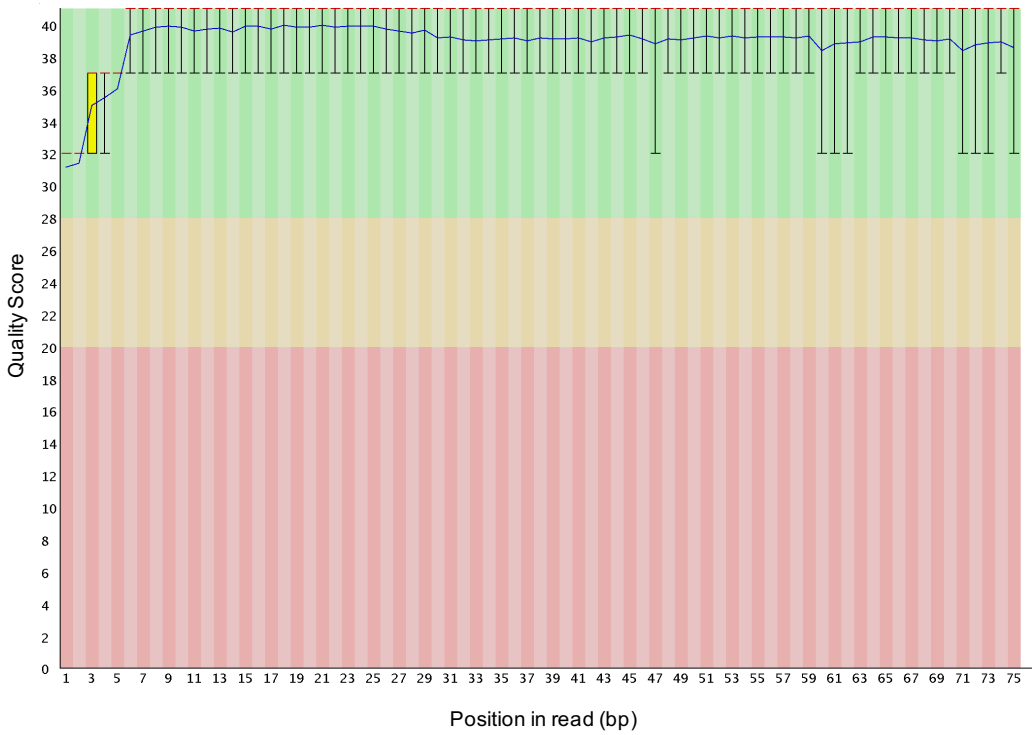
Assessing Dsx^F-Dam brain, Dsx^M-Dam and Dsx^F-Dam head TaDa read quality using FastQC (Andrews, 2010). Per base sequence quality (A-C) had reliable base calls across across every base in 75 bp reads. The graph background divides the y axis into very good quality calls (green), calls of reasonable quality (orange), and calls of poor quality (red). The yellow box represents the inter-quartile range (25-75%). The blue line represents the mean quality. The central line is the median value. The upper and lower whiskers represent the 10% and 90% points. The quality score distribution over all sequences (D-F) was high (Phred scores: 30-40). The distribution of sequence lengths over all sequences (G-I) confirmed the majority of all sequencing reads were of similar length.



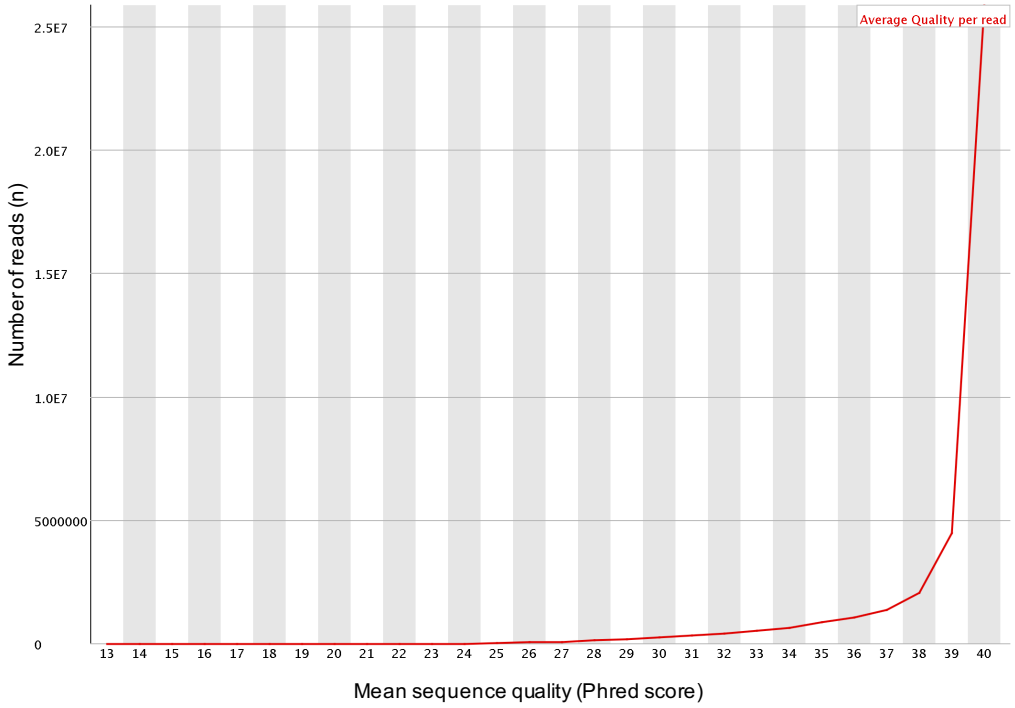
B Per base sequence quality, head TaDa Dsx^{M}-Dam



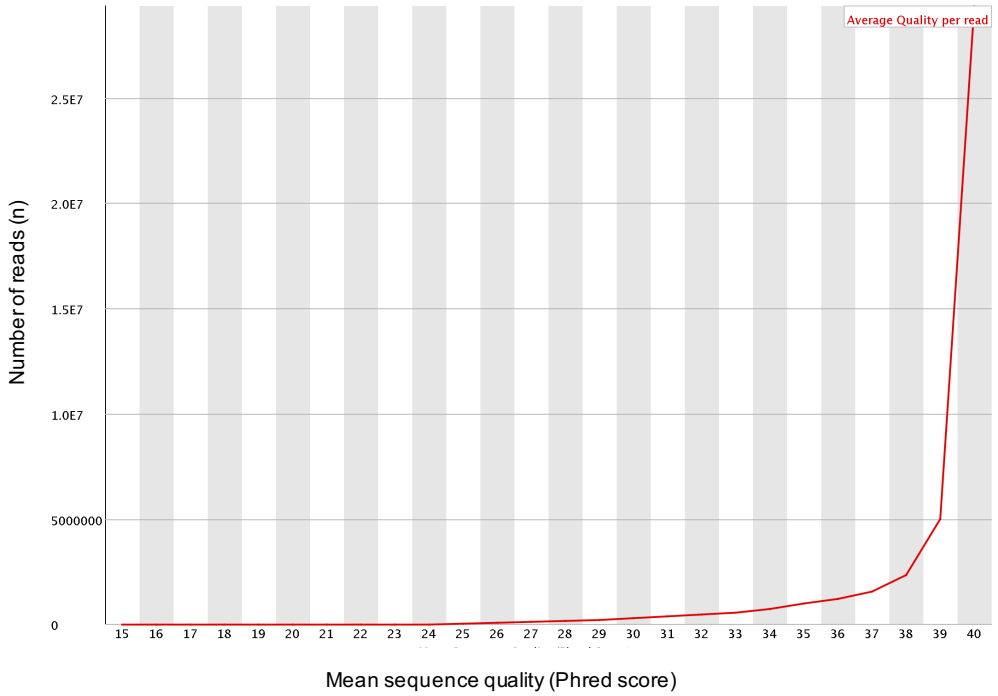
C Per base sequence quality, head TaDa Dsx^{F}-Dam



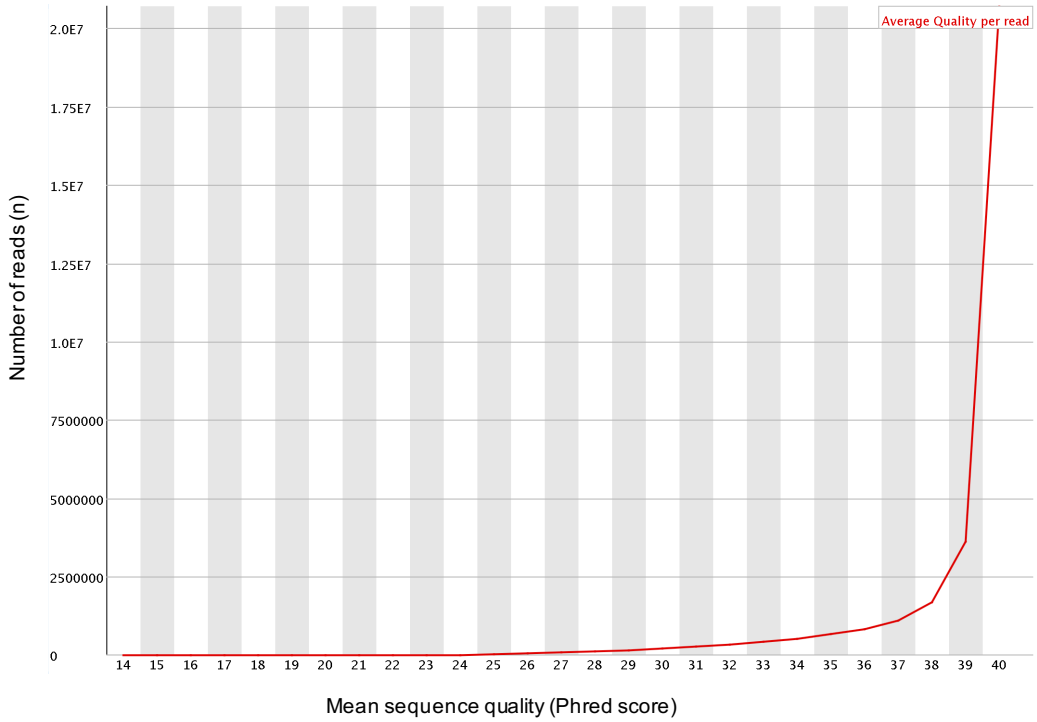
D Quality score distribution over all sequences, brain TaDa Dsx^{F}-Dam



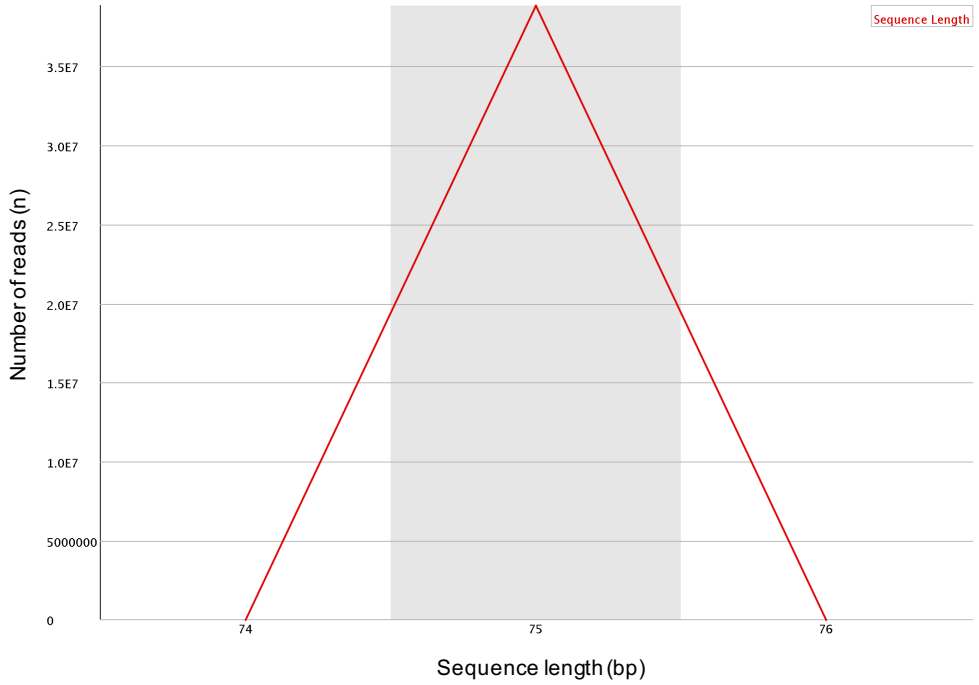
E Quality score distribution over all sequences, head TaDa Dsx^{M}-Dam



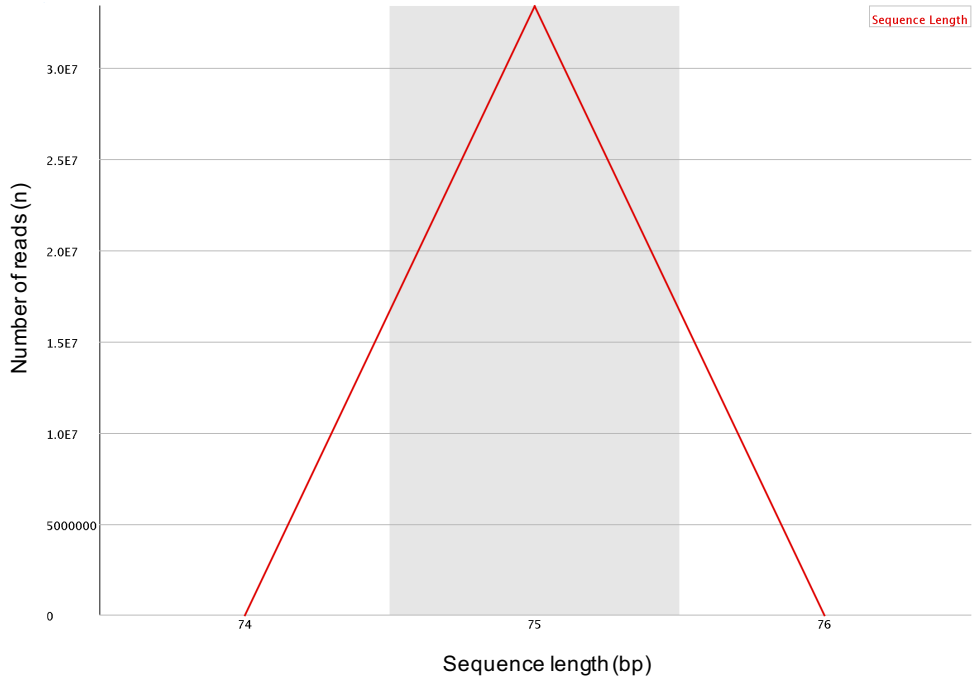
F Quality score distribution over all sequences, head TaDa Dsx^{F}-Dam



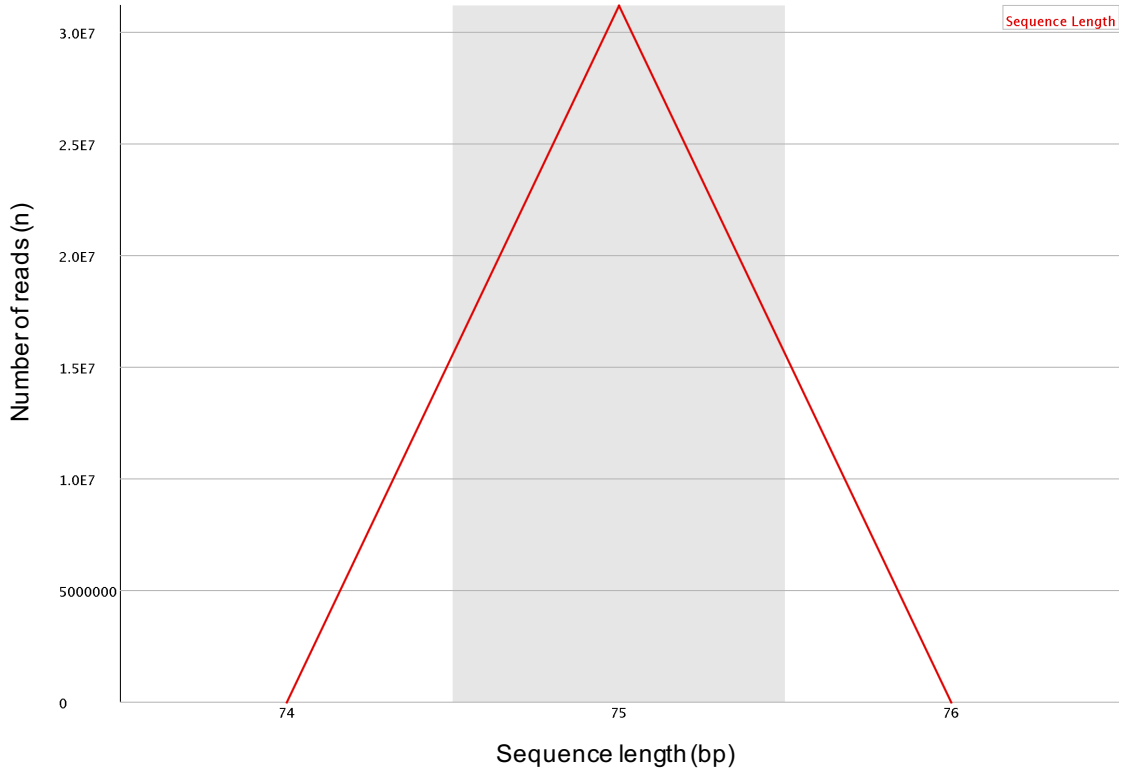
G Distribution of sequence lengths over all sequences, brain TaDa Dsx^{F}-Dam



H Distribution of sequence lengths over all sequences, head TaDa Dsx^{M}-Dam



I Distribution of sequence lengths over all sequences, head TaDa Dsx^{F}-Dam



9.2 PEAK CALLING USING FIND_PEAKS (MARSHALL AND BRAND, 2015)

Summary of called peak numbers in Dsx^{M}-Dam male and Dsx^{F}-Dam female replicates in DSX-brain TaDa-seq (top) and DSX-head TaDa-seq (bottom) for alternative sequencing orientation. Peaks called at FDR 0.01 and 0.05. Peak numbers were >95% similar in each sample.

Brain TaDa-seq				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	434	511	496	107
FDR 0.05	842	1279	1265	291
Head TaDa-seq				
	Dsx^{M}-Dam 1	Dsx^{M}-Dam 2	Dsx^{F}-Dam 1	Dsx^{F}-Dam 2
FDR 0.01	753	888	1213	1748
FDR 0.05	2001	2576	2089	1862

9.3 GO BRAIN TADA DSX^{M}-DAM

FlyMine v46.1 GO enrichment analysis for brain TaDa Dsx^{M}-Dam. Peaks/ genes common to both Dsx^{M}-Dam biological replicates were included in the enrichment analysis. For the *cellular component* ontology, 3 statistically significant enrichment terms were defined, 1 redundant term removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
cell cortex	0.001673	21
cytoplasmic region	0.002065	25

For the molecular function ontology, 4 statistically significant enrichment terms were defined, 1 redundant term removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
anion binding	0.012857	68
DNA-binding transcription factor activity	0.013083	36
small molecule binding	0.023843	66

9.4 GO HEAD TADA DSX^{M}-DAM

FlyMine v46.1 GO enrichment analysis for head TaDa Dsx^{M}-Dam. Peaks/ genes common to both Dsx^{M}-Dam biological replicates were included in the enrichment analysis. For the *cellular component* ontology, 5 statistically significant enrichment terms were defined, 2 redundant terms removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
cell periphery	1.03E-06	183
plasma membrane	6.93E-06	165
cell cortex	0.0068	34

For the *molecular function* ontology, 8 statistically significant enrichment terms were defined ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
catalytic activity	3.78E-04	433
cofactor binding	0.0037	69
ion binding	0.0112	246
anion binding	0.0227	149
lipid binding	0.0348	33
heme binding	0.0394	29
protein binding	0.0402	281
tetrapyrrole binding	0.0458	29

9.5 GO FAT BODY DSX^{M}-DAM (CLOUGH ET AL., 2014)

FlyMine v46.1 GO enrichment analysis for fat body Dsx^{M}-Dam (Clough et al., 2014). Peaks/ genes common to both Dsx^{M}-Dam biological replicates were included in the enrichment analysis. For the *cellular component* ontology, 10 statistically significant enrichment terms were defined, 4 redundant terms removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
fusome	1.87E-05	15
cytoplasm	5.31E-05	458
plasma membrane	4.63E-04	152
cell cortex	0.0162	33
supramolecular complex	0.0208	33
axon	0.0337	39

For the *molecular function* ontology, 24 statistically significant enrichment terms were defined, 11 redundant terms removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
kinase activity	8.80E-07	65
anion binding	7.34E-06	153
protein binding	4.81E-05	281
transferase activity	3.12E-04	155
catalytic activity	0.001877	408
nucleotide binding	0.0061	126
nucleoside phosphate binding	0.0061	126

actin binding	0.0067	29
DNA-binding transcription factor binding	0.0122	17
protein domain specific binding	0.0206	20
purine ribonucleotide binding	0.0252	108
adenyl ribonucleotide binding	0.0259	91
cofactor binding	0.0351	60

9.6 GO FAT BODY DSX^F-DAM (CLOUGH ET AL., 2014)

FlyMine v46.1 GO enrichment analysis for fat body Dsx^F-Dam (Clough et al., 2014). Peaks/ genes common to both Dsx^M-Dam biological replicates were included in the enrichment analysis. For the *cellular component* ontology, 12 statistically significant enrichment terms were defined, 5 redundant terms removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
cell periphery	2.23E-07	207
plasma membrane	3.80E-07	189
cell junction	3.89E-06	43
adherens junction	6.75E-04	28
anchoring junction	6.75E-04	28
cell cortex	0.0216	36
cytoplasm	0.0235	507

For the *molecular function* ontology, 10 statistically significant enrichment terms were defined, 4 redundant terms removed ($p<0.05$, Holm-Bonferroni test correction).

GO term	p-value	Matches
protein binding	1.42E-06	334
kinase activity	0.0024	63
lipid binding	0.0080	37
ion binding	0.0089	276
cofactor binding	0.0137	69
coenzyme binding	0.0491	45

9.7 *DE NOVO* MOTIF ANALYSIS: BRAIN TADA DSX^{M}-DAM

Top ten enriched motifs in brain TaDa Dsx^{M}-Dam identified *de novo* using HOMER v4.10 (Heinz et al., 2010). 26 statistically significant ($p<0.05$) motifs defined overall. We looked for motif enrichment in peaks common to both Dsx^{M}-Dam biological replicate 1 and 2. Motif best matches compared to Homer Motif Database library (May 2018). Database includes JASPAR v7, 2018 (http://jaspar.genereg.net) and DMMPMM v26-MAY-2009, 2009 (http://autosome.ru/DMMPMM/) motif collections.

Rank	Motif sequence logo	p-value	Peaks (%)	Best Match
1		1e-200	32.0	Cf2/MA0015.1/Jaspar
2		1e-133	56.0	PB0016.1_Foxj1_1/Jaspar
3		1e-115	24.2	GAT1/MA0300.1/Jaspar
4		1e-111	66.8	CES-1(Homeobox)/Homer
5		1e-103	62.3	byn/dmmpmm(Pollard)/fly
6		1e-88	40.6	GATA15/MA1016.1/Jaspar
7		1e-85	31.6	SeqBias: GCW-triplet
8		1e-74	56.0	twi/dmmpmm(Papatsenko)/fly
9		1e-66	58.7	prd/dmmpmm(Down)/fly
10		1e-63	41.9	NF1-halfsite(CTF)/Homer

9.8 *DE NOVO* MOTIF ANALYSIS: HEAD TADA DSX^{M}-DAM

Top ten enriched motifs in head TaDa Dsx^{M}-Dam identified *de novo* using HOMER v4.10 (Heinz et al., 2010). 24 statistically significant ($p<0.05$) motifs defined overall. We looked for motif enrichment in peaks common to both Dsx^{M}-Dam biological replicate 1 and 2. Motif best matches compared to Homer Motif Database library (May 2018). Database includes JASPAR v7, 2018 (http://jaspar.genereg.net) and DMMPMM v26-MAY-2009, 2009 (http://autosome.ru/DMMPMM/) motif collections.

Rank	Motif sequence logo	p-value	Peaks (%)	Best Match
1		1e-319	36.9	SeqBias: CA-repeat
2		1e-200	59.8	Cf2/MA0015.1/Jaspar
3		1e-157	32.9	PB0022.1_Gata5_1/Jaspar
4		1e-146	44.6	Aef1/dmmpmm(Pollard)/fly
5		1e-134	56.1	AIB/MA0959.1/Jaspar
6		1e-130	57.5	NtERF2(AP2/EREBP)
7		1e-129	56.8	GATA15/MA1016.1/Jaspar
8		1e-118	76.5	PB0146.1_Mafk_2/Jaspar
19		1e-41	56.0	GATA10/MA1013.1/Jaspar
24		1e-15	2.5	DMRT3/MA0610.1/Jaspar