Investigating the regulation and evolution of *tartan* during *Drosophila* male genital diversification.

Javier Figueras Jiménez

Degree awarded by
Oxford Brookes University

This thesis is submitted in partial fulfilment of the requirements of the award of Doctor of Philosophy

February 2022

Acknowledgments

In first place, I would like to thank Alistair for always supporting me, helping me through my PhD with advice and guidance while enduring my many mistakes throughout the years. Thank you Dani, for the help and suggestions in lab meetings and genitalia meetings and thank you Saad for the help with bioinformatics tools.

I also would like to thank the McGregor lab, current and former members that have helped me in many different ways. Anna, Pedro, Franzi and Alex thank you for helping me with every technique I have used or that I was not familiar with. Jo, thank you for introducing me into the *Drosophila* genitalia world and how to work with it. Thank you, Shamma, for helping me by discussing many times about our projects and sharing our troubles in the lab. Thank you, Amber R, for the help with the project and sometimes helping me when I was not able to go to the lab or we had difficulties due to covid. Thanks to all members for helping with suggestions in lab meetings that made me improve my results and the fun discussions at journal pubs I thank Sebastian, Maike, Amber H, Grace, and Emily.

A huge thanks to my partner Carolina, without her I would not be here and would have not been able to go through everything. Thank you for enduring my ups and downs and supporting me. Thanks to my little dog as well, that even though I could not play with her as much as she would have liked during the writing of this thesis, she always made me feel better with myself.

Last but not least, thank you to my family that without understanding what I do they have always supported me. To my father that always pushed me to understand nature, to my mother that always wanted me to do what I liked no matter what, to my sister

who has always helped me in any way she could and to the rest of my family for all the continuous support.

Abstract

Male genital structures have evolved rapidly in many animal groups and in some cases are the only features that can distinguish species that have recently diverged. However, we still lack knowledge about the genetic mechanisms responsible for the divergence of this trait. The *Drosophila melanogaster* species subgroup has been used to investigate this and recent studies between *D. simulans* and *D. mauritiana* have identified that the gene *tartan* (*trn*) is involved in the evolution of the surstyli (claspers) which form part of the external male genitalia. *trn* encodes a leucine-rich repeat transmembrane protein and contributes to the difference in surstyli size and the bristle number between *D. simulans* and *D. mauritiana. trn* is differentially expressed (spatially and temporarily) between the developing surstyli of these two species. This indicates that changes in the regulation of *trn* rather than the coding sequence underlie its contribution to surstyli divergence.

In my PhD project I have investigated the regulation of *trn.* First, I characterised the endogenous *trn* expression using a Trn tagged protein, and then I screened the *trn* locus to search for enhancers. I identified nine regions that are active during pupal development, three of which are active in the male genitalia, U4, U8 and D6. U4 was the most promising region because its activity overlaps with endogenous trn expression in the developing genitalia, and therefore I investigated which transcription factors were candidates to bind to this enhancer. I identified *caudal (cad)* as a likely candidate to be regulating U4. I then used RNAi to further characterise the function of *cad* during genitalia development. My results showed that the anal plates are lost upon *cad* RNAi knockdown and that this gene may regulate *trn* expression around anal plates to delimit their specification. Additionally, I investigated the evolution of the U4 enhancer sequence and activity. Analysis of the *D. mauritiana* U4 activity shows it has

expanded expression compared to the *D. melanogaster* U4. This result suggests that there has been evolution of this enhancer in *D. mauritiana* that may contribute to the larger surstyli of this species, but this remains to be directly tested. Taken together, my investigation has characterised most of the regulatory landscape of *trn* and has identified an enhancer that is active in the genitalia and that may have evolved between species. My investigation will allow us to further understand the compartmentalization and regulation of male genitalia development and the evolution of these structures.

Table of contents

Chapter 1. General introduction……14
1.1 Genital evolution……....14
1.2 The *Drosophila melanogaster* species subgroup.....17
1.3 Drosophila male genitalia development……..19
1.4 *tartan*…….....24
1.4.1 *tartan* function……..25
1.5 Regulation……..28
1.5.1 Enhancer identification……28
1.6 Aims of this thesis…….……31
Chapter 2. Investigation of the *trn* regulatory landscape.……..33
2.1 *trn* regulatory landscape……..33
2.2 Methods……..33
2.2.1 Cloning……33
2.2.2 Fly stocks and genetics……34
2.2.3 Microscope imaging……..35
2.2.4 In Situ Hybridization……..35
2.3 Results……36
2.3.1 *tartan* endogenous expression pattern……..36
2.3.2 Enhancer survey……40
2.3.2.1 Design of enhancer constructs……..40
2.3.2.2 Survey of enhancer activity of candidate regions……42

2.3.3 Enhancers driving expression in the genitalia…….49

2.4 Discussion……..54

2.4.1 Enhancer identified……54

2.4.2 U4, U8 and D6……55

Chapter 3. Investigating transcription factors that regulate *trn* in the male genital of *Drosophila*……57

3.1.1 Transcription factor binding sites……..……57

3.1.2 ATAC-seq…..……..……..…57

3.1.3 Finding binding sites using ATAC-seq…….……58

3.2 Methods…….……59

3.2.1 Motif discovery and footprinting……..59

3.2.2 RNAi Knockdown of *cad*……..60

3.2.3 Statistical analysis……61

3.3 Results……..…..62

3.3.1 Binding sites in U4……62

3.3.2 *cad* RNAi……66

3.3.2.1 NP6333>*cad*-RNAi Knockdown……67

3.3.2.2 *cad*-RNAi Knockdown using U4……70

3.4 Discussion……..74

3.4.1 Binding sites in U4……..…74

3.4.2 *cad* RNAi perturbs genital development…….75

Chapter 4. Evolution of the U4 *trn* enhancer……78

4.1.1 Chromatin accessibility……78

4.1.2 Enhancer evolution…..……79

4.2 Methods…….……81

4.2.1 Cloning……81

4.2.2 Fly stocks and genetics……81

4.2.3 Microscope imaging……..82

4.2.4 Accessible chromatin analysis……82

4.2.5 U4 sequence analysis……..82

4.3 Results……83

4.3.1 The *trn* chromatin accessibility profile……83

4.3.2 The U4 in *D. mauritiana*……84

4.3.3 U4 sequence evolution……88

4.4 Discussion……..92

4.4.1 ATAC-seq profile comparison……..92

4.4.2 U4 differential activity……93

4.4.3 U4 Evolution……...94

Chapter 5. General Discussion……..96

5.1 Enhancer screening of *trn* locus……..96

5.2 *Drosophila* male genital enhancers identified……98

5.3 Binding sites in U4……99

5.4 *cad* regulation of *trn* and genital GRN……..100

5.5 U4 chromatin profile and activity……101

5.6 U4 activity and sequence evolution……..102

5.7 Conclusions.……..……..104

References……105

Supplementary materials……125

Table of figures

Figure 1.1. Scheme of the genital periphallic and phallic structures

Figure 1.2. *D. simulans* and *D. mauritiana* male genitalia

Figure 1.3. Summary of sex-specific regulation in the genital disc of *D. melanogaster*

Figure 2.1. *trn* in situ hybridization (ISH)

Figure 2.2. Trn endogenous protein pattern

Figure 2.3. *trn* locus in *D. melanogaster* showing candidate regions

Figure 2.4. Activity of regions U1 and U2

Figure 2.5. Activity of line VT030010 and region U3

Figure 2.6. Activity of regions U4 and U5

Figure 2.7. Activity of regions U8 and Intron

Figure 2.8. Activity of VT030021 line and D2 region

Figure 2.9. Activity of regions D3 and D4

Figure 2.10. Activity of D5 and D6 regions

Figure 2.11. U2 reporter expression at 24 hAPF in the male genital

Figure 2.12. Activity of U4 in the male genital

Figure 2.13. Activity of U8 in the male genital

Figure 2.14. Schematic representation of Trn endogenous expression pattern and regions U4 and U8

Figure 3.1. Motifs found in U4 accessible chromatin

Figure 3.2. Motif comparison with TFBSs from JASPAR

Figure 3.3. Motifs found using footprints

Figure 3.4. Whole terminalia and surstyli from *cad* RNAi KD

Figure 3.5. Box and whiskers plots of the *cad* RNAi knockdown

Figure 3.6. Box and whiskers plots of the *cad* knockdown using U4-GAL4

Figure 3.7. Box and whiskers plots of the *cad* knockdown using U4-GAL4 (uncorrected)

Figure 4.1. Chromatin accessibility profile around the *trn* locus of pupal male genitalia from *D. mauritiana (D1)* and *D. simulans (w501)* generated from ATAC-seq data

Figure 4.2. Enhancer activity of region U4 from the D1 strain of *D. mauritiana* (U4D1) compared to *D. melanogaster* U4

Figure 4.3. Expression driven from *D. melanogaster* U4 and *D. mauritiana* U4.

Figure 4.4. sequence comparison among *Drosophila* species

Supplementary Figure 2.1. Orthogonal views of U4 and U8 expression patterns

Supplementary Figure 3.1. U4 alignment for *D. melanogaster, D. mauritiana* (MS17) and *D. simulans* (w^{501})

Supplementary Figure 3.2. Posterior lobes and anal plates from *cad* KD and NP6333-GAL4;UAS-Dicer

Supplementary Figure 3.3. Correlation between areas measured and tibia length

Supplementary Figure 4.1. Trn protein alignment with other *Drosophila* species

Supplementary Figure 4.2. Chromatin accessibility profile around the *trn* locus of pupal male genitalia from *D. simulans (w501)* and eye imaginal disc from *D. simulans (yvf)* generated from ATAC-seq data.

List of tables

Table 1. Summary of enhancer activity of regions tested at 24 and 46 hAPF

Table 2. Fixed differences found between *D. simulans* and *D. mauritiana*

Supplementary Table 1. Enhancer candidate primers

Supplementary Table 2. Coordinates for enhancer candidates

Supplementary Table 3. Transcription factor footprints of U4

Supplementary Table 4. Transcription factor BSs

Supplementary Table 5. Table of polymorphism among *D. simulans* and *D. mauritiana* strains.

Chapter 1. General Introduction

1.1 Genital evolution

Among animal groups with internal fertilization, the male genitals structures evolve rapidly with respect to other aspects of morphology (Eberhard, 2010; Hosken *et al.*, 2019). This divergence is seen in many different animal groups including insects (Jerry A Coyne, 1983; Hosken and Stockley, 2004; Mutanen, Kaitala and Mönkkönen, 2006; Fujisawa *et al.*, 2019), reptiles (Böhme and Ziegler, 2009) and mammals (Dixson, 1987).

Even species that have diverged recently, for example *Drosophila mauritiana* and *Drosophila simulans,* which last shared an ancestor approximately only 240,000 years ago (Garrigan *et al.*, 2012), can be distinguished by their genital structures, although their morphology is otherwise basically the same (Liu, Mercer, *et al.*, 1996).

The mechanisms underlying these rapid changes happening in a relative short period of time are still a matter of much discussion. One hypothesis to explain the evolution of genitalia is ‘Lock and Key’, whereby the male genitalia act as a key that ‘unlocks’ the female genitalia (Eberhard, 1985). This hypothesis explains the divergence of the genital structures between species as a result of the female genitalia being only available to those of the same species. Hence, the evolution of the male genitalia in directed by its ability to fit in the ‘lock’, thus leading to reproductive isolation (Eberhard, 1985). However, this hypothesis lacks support as a very general hypothesis that explains genital evolution because, for example, there is a lack of a ‘lock’ of the female in some groups and species-specific genitalia in species that have evolved while physically separated (Paulson, 1974; Eberhard, 1985; Shapiro and Porter, 1989).

There are two other main hypotheses based on sexual selection that have been proposed to explain the rapid evolution of genitalia: sexually antagonistic coevolution (SAC) and cryptic female choice (CFC) (Eberhard, 2009b, 2010).

SAC is based on the conflict that emerges between the male and female due to different interests in reproduction, thus forcing a coevolution between the male and female (Chapman *et al.*, 2003; Eberhard, 2010). In this conflictive coevolution, the male evolution is directed towards obtaining a benefit from the female or forcing the female either to mate with the male, prevent mating with other males or punishment for rejecting the male, causing a possible disadvantage for female fitness (Eberhard, 1985, 2010; Clutton-Brock and Parker, 1995; Chapman *et al.*, 2003; Masly and Kamimura, 2014). This control of the male over the copulation forces the female adaption to focus on minimizing such disadvantages by either resisting copulation with males or reducing damages from mating (Eberhard, 1985, 2010; Masly and Kamimura, 2014). Such adaptations would then lead to further male adaptions; thus, this hypothesis infers that this cyclic counteradaptation may explain the divergence of genitalia.

With CFC, the male and female have also different interests, but in this case the hypothesis is based on sexual selection from the female driven by behavioural, morphological and physiological components (Eberhard, 2009a, 2010; Eberhard and Lehmann, 2019). Using this strategy, the female may select males and allow copulation only with those that are likely to leave fitter offspring. Thus, the sexual selection is not only based on the courtship pre-copulation but also on the genital structures that provide either a better stimulation or that might fit better to the morphology of the female genitalia. This includes other aspects of the copulation that might influence on whether the female allows a series of advantages for the

copulation, such as permitting the male structures to penetrate enough for a better sperm deposition or reducing internal defences that reduce sperm viability. (Eberhard, 2009a, 2010; Eberhard and Lehmann, 2019). Therefore, with CFC the genital morphology and secondary structures that are not involved directly in the transmission of sperm might also play a role in pre-copulatory and copulatory sexual selection

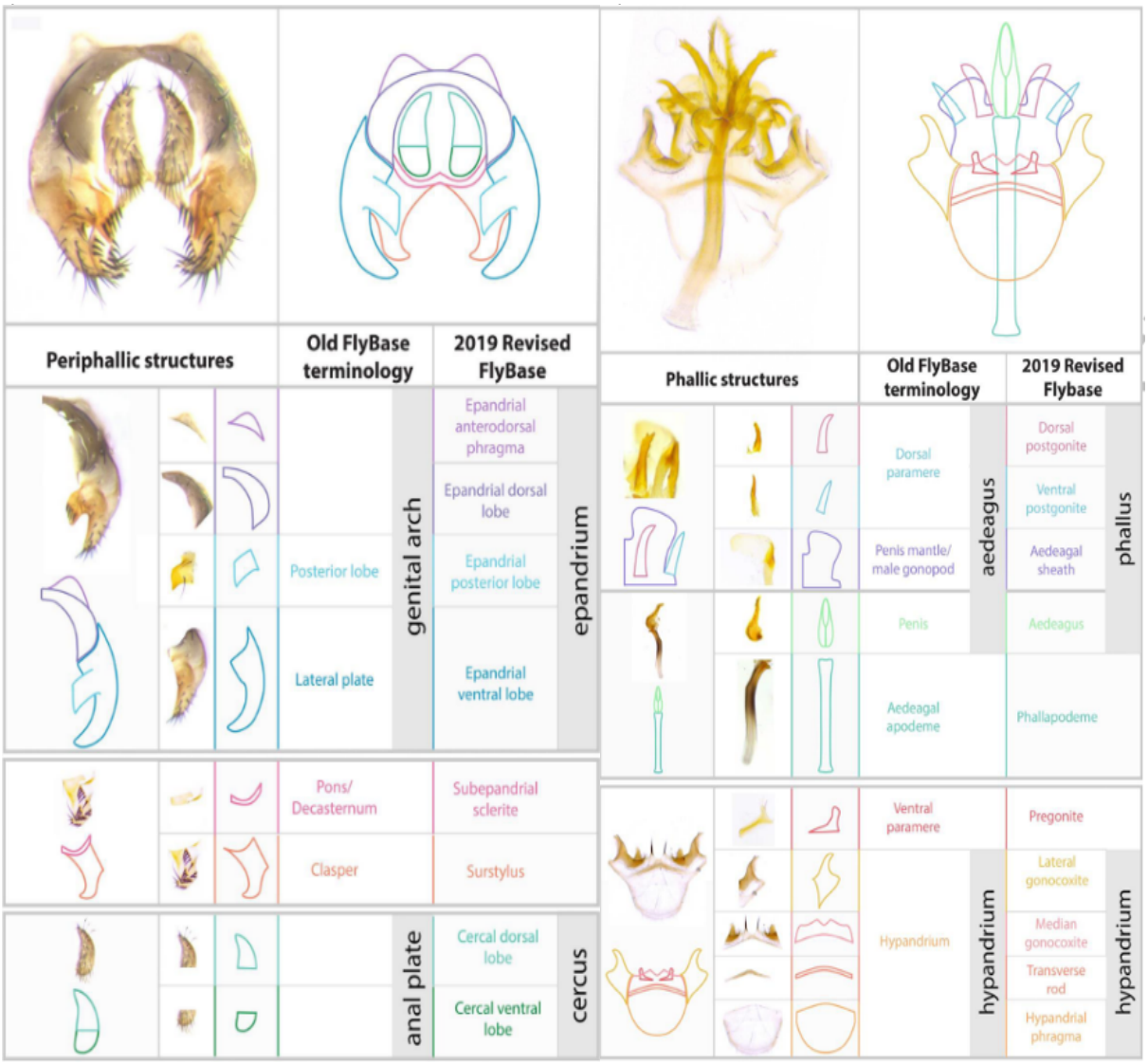


Figure 1.1. Scheme of the genital periphallic and phallic structures. Structures and subdivisions of the male genitalia of *Drosophila melanogaster* following both nomenclatures: Old FlyBase terminology and a new revised terminology. Figure from Rice *et al.*, 2019.

Both CFC and SAC correlate better with species that are polyandrous, where sexual selection as well as natural selection is a strong driving force of genital evolution, hence these species should be more divergent (Hosken and Stockley, 2004; Simmons, 2014; Eberhard and Lehmann, 2019). In support of this it has been found that both natural and sexual selection are involved in the evolution of the morphology of male genitalia in *Drosophila* species (Ritchie, 2007; House *et al.*, 2013).

1.2 The *Drosophila melanogaster* species subgroup

To study the rapid divergence of the male genital structures, the *Drosophila melanogaster* species subgroup has long been an excellent model, thanks to the differences that can be found in male genitalia of these closely related species (Coyne, 1983; Liu, Mercer, *et al.*, 1996; True *et al.*, 1997; Lang *et al.*, 2014).

Here I will first describe the male genitalia structures of these *Drosophila* species. Following a new unified nomenclature (Rice *et al.*, 2019), *Drosophila* male genitalia are formed from a set of structures that are part of the *Drosophila* male terminalia, which is formed by the male genitalia and the male analia (Fig. 1.1).

Both the genitalia and the analia develop from the genital disc (Sánchez and Guerrero, 2001; Rice *et al.*, 2019). The male genitalia can then be divided again in periphallic structures and phallic structures (Fig. 1.1)

The periphallic structures are secondary structures to the internal genitalia. Some of these structures are involved in the copulation through grasping the female but they are not directly involved in the transmission of sperm. The phallic structures are those that form the phallus that will be introduced into the female vagina, and therefore directly involved in the transfer of sperm (Rice *et al.*, 2019).

The periphallic structures are composed of the epandrium (genital arch), the cerci (anal plates), the pair of surstyli (claspers) and subepandrial sclerite (pons). The phallic structures are composed of the phallus (aedeagus), hypandrium, phallapodeme (adeagal apodeme) and pregonite (ventral paramere) (Fig. 1.1). These structures are also subdivided into smaller substructures (Fig. 1.1) (Rice *et al.*, 2019). In addition to the terminal structure, there is also the sex combs, another important secondary sexual structure that is found on the foreleg tarsi of the males and that help mediate copulation (True, 2008; Atallah *et al.*, 2009; Seeholzer *et al.*, 2018; Ahmed *et al.*, 2019).

Although the periphallic structures are not directly involved in the transmission of sperm, they strongly influence copulation. This has been demonstrated with alterations of the claspers and the posterior lobes (a subdivision of the epandrium) that resulted in changes to mating position, and consequently reducing the mating frequency and/or reproductive success (Acebes, 2003; LeVasseur-Viens, Polak and Moehring, 2015). For example, LeVasseur-Viens *et al.* (2015) assesed the role of the posterior lobes and their inffluence on intraspecific and interspecific effects of sexual selection in copulation with *D. simulans* and *D. mauritiana.* They used *D. simulans* male flies with altered posterior lobes and found that the copulation frequency of males with highly altered posterior lobes was reduced compared to those with slight alterations (LeVasseur-Viens, Polak and Moehring, 2015). Furthermore, it has been shown that ablating the sensilla on one or both claspers reduced the mating frequency in *D. melanogaster* and *D. simulans* (Acebes, 2003).

There is striking natural variation in the shape, size, and bristles of periphallic structures among species of the *D. simulans* clade: *D. mauritiana, D. sechellia* and *D. simulans* (Fig. 1.2). For example, the surstyli of *D. simulans* are smaller and has less

bristles than those of *D. mauritiana* (Fig. 1.2 B, C). The epandrial posterior lobes of *D. mauritiana* are smaller and have a straight and narrow shape, while in *D.* simulans this structure has a wide helmet-like shape (Fig. 1.2). These species last shared a common ancestor only 240,000 years ago and diverged from *D. melanogaster* about 3 million years ago (Robertson, 1988; Garrigan *et al.*, 2012).

Consequently, these closely related species represent a powerful model because of differences in their genital structures and the important roles of these structures in grasping and proprioception during copulation (Jagadeeshan and Singh, 2006; Kamimura and Mitsumoto, 2011; Mattei *et al.*, 2015).

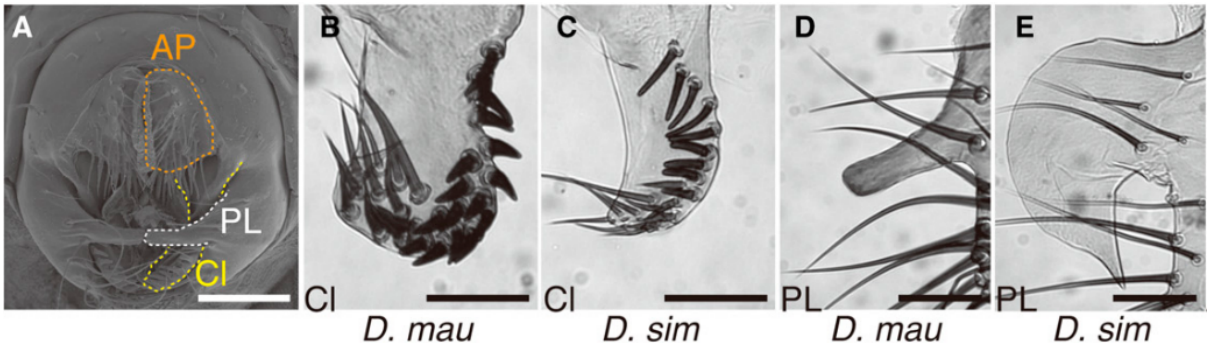


Figure 1.2. *D. simulans* and *D. mauritiana* male genitalia. (A) SEM image of *D. mauritiana* male external genitalia. Highlighted in orange the anal plate, in yellow the surstyli (clasper) and in white the epandrial posterior lobe. (B-C) *D. mauritiana* and *D. simulans* surstyli (clasper). (D-E) *D. mauritiana* and *D. simulans* epandrial posterior lobe. Figure from Tanaka *et al.*, 2015.

1.3 *Drosophila* male genitalia development

Drosophila genitalia derive from a genital disc, one of the imaginal discs set aside during embryogenesis. These discs are epithelial sac-like structures present in the larvae that form from invaginations of the embryonic ectoderm. The imaginal discs give rise to the adult structures such as the genitalia, legs, wings, eyes, etc. (Beira and Paro, 2016). Depending on the proportion of sex chromosomes, to autosomes the

genital disc will give rise to a male genital disc or a female genital disc (Fig. 1.3) (Estrada, Casares and Sánchez-Herrero, 2003).

The genital imaginal disc develops from three abdominal segments of the embryo A8, A9 and A10. The female primordium develops from the segment A8, the male primordium from A9 and the anal plate from the segment A10 (Casares *et al.*, 1997; Estrada, Casares and Sánchez-Herrero, 2003; Gorfinkiel, Sánchez and Guerrero, 2003; Chen, Christiansen and Baker, 2005). These primordia are specified initially by *Sex lethal (Sxl),* a gene that is only active in the female primordia due to the ratio of autosomes to X chromosomes (Fig. 1.3) (Keisman and Baker, 2001; Sánchez and Guerrero, 2001; Estrada, Casares and Sánchez-Herrero, 2003). After this point, there are two different scenarios depending on whether *Sxl* is active (females) or inactive (males) (Fig. 1.3). In the first case, *Sxl* activates an alternative splicing of another gene, *transformer (tra)*, that interacts with *transformer-2 (tra-2)* causing the female specific splicing of *doublesex (*Dsx^F*)* (Fig. 1.3A). In the males *Sxl* is not expressed, therefore *tra* is not spliced into the female isoform and therefore it does not interact with *tra-2,* leaving *dsx* in its default splicing to produce Dsx^M isoform (Fig. 1.3B) (Keisman and Baker, 2001; Estrada, Casares and Sánchez-Herrero, 2003).

Three segments therefore contribute to the genital disc but A8 is repressed in males and A9 repressed in females, while A10 is present in both (Fig. 1.3). These compartments are then specified by Homeotic genes (Hox) (Casares *et al.*, 1997; Sánchez and Guerrero, 2001; Estrada, Casares and Sánchez-Herrero, 2003). *abdominal-A (abd-A)* is expressed in the A8 segment and is required for the formation of female internal genitalia but not required for external genitalia (Chen, Christiansen and Baker, 2005; Foronda *et al.*, 2006; Aspiras, Smith and Angelini, 2011). *abdominal-B (abd-B)* gives rise to two different proteins Abd-B M and Abd-B R that are essential

for the morphogenesis of the imaginal disc into genitalia. These proteins repress *Distal-less (Dll)* otherwise the imaginal disc would develop into a leg or distal antennae (Fig. 1.3) (Sánchez and Guerrero, 2001; Foronda *et al.*, 2006).

Abd-B M is only present in segment A8 whereas Abd-B R is only present in A9 and these isoforms define female and male primordium respectively (Fig. 1.3) (Sánchez and Guerrero, 2001; Estrada, Casares and Sánchez-Herrero, 2003; Foronda *et al.*, 2006). Segment A10 is defined by *caudal* (*cad), cad* together with *Hedgehog (Hh)* is required for the correct development of the anal plates, where *Dll* is not repressed but induced by the Hh pathway and *cad* that form from A10 and develops into the anal structures (Moreno and Morata, 1999; Sánchez and Guerrero, 2001; Estrada, Casares and Sánchez-Herrero, 2003; Chen, Christiansen and Baker, 2005).

The fate of the genital disc is determined by whether the male or the female isoform of Dsx is expressed as these modulate the actions of different genes within the segments, for example, *hedgehog* (*Hh*), *decapentaplegic* (*dpp*) and *wingless* (*wg*) (Sánchez *et al.*, 1997; Sánchez and Guerrero, 2001). In the case of the female pathway, Abd-BR combined with Dsx^F will block the expression of *dpp* in A9 to repress this segment in females (Fig. 1.3 A). Dsx^F and Abd-BM also promote the expression of *wg* in A8 to initiate female development of this segment (Fig. 1.3 A) (Sánchez and Guerrero, 2001; Foronda *et al.*, 2006). Conversely in males, Dsx^M promotes the expression of *dpp* in A9 while together with Abd-BM it represses the expression of *wg* in A8, resulting in the repression of A8 and development of the male primordium A9 (Fig. 1.3 B) (Sánchez and Guerrero, 2001; Foronda *et al.*, 2006; Aspiras, Smith and Angelini, 2011).

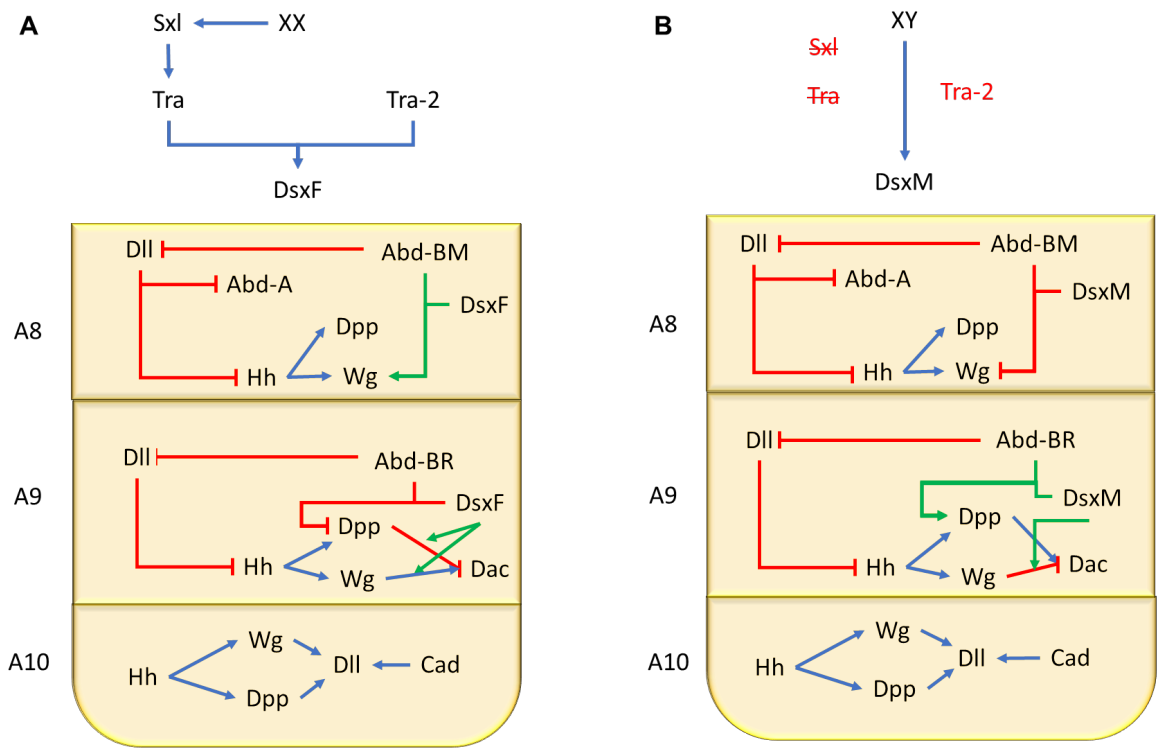


Figure 1.3. Summary of sex-specific regulation in the genital disc of *D. melanogaster*. Blue arrows indicate activation, red arrows indicate repression and green arrows indicate influencing other factors into activating or repressing a third gene (i.e., DsxF influence Wg into activating Dac and DsxM influence Wg into repressing Dac). (A) Female genital disc regulatory network, the 2X:2A chromosome ratio triggers the cascade for the development of female genital disc. (B) Male genital disc regulatory network, the ratio X:2A triggers the cascade for development of male genital disc. Model based on evidence from (Sánchez *et al.*, 1997; Moreno and Morata, 1999; Keisman and Baker, 2001; Sánchez and Guerrero, 2001; Chen, Christiansen and Baker, 2005; Foronda *et al.*, 2006; Aspiras, Smith and Angelini, 2011).

There is also a sex-specific regulation of *dachshund (dac)* mediated by Dsx. In the absence of Dsx, Dpp and Wg would both activate *dac* in A9, but the sex-specific versions of Dsx change this dynamic (Sánchez *et al.*, 1997; Keisman and Baker, 2001; Sánchez and Guerrero, 2001). Wg and Dpp have the potential to activate or repress Dac, and Dsx^M converts *wg* into a repressor of *dac* in the A9, while *dpp* will continue activating it in its domain inside of A9 (Fig. 1.3 B) (Sánchez *et al.*, 1997; Keisman and Baker, 2001; Sánchez and Guerrero, 2001). Dsx^F acts differently in A9 in females by

promoting the activation of *dac* by Wg while repressing *dpp* and possibly turning any Dpp present into a repressor of *dac* (Fig. 1.3 A), it was seen that ectopic expression of *dpp* in clones within the *dpp* domain represses *dac* within those clones (Keisman and Baker, 2001; Sánchez and Guerrero, 2001). It is important to mention that the functions of *dpp* and *wg* are restricted to their own compartments within the imaginal disc, thus they are regulating *dac* differently in these compartments (Sánchez *et al.*, 1997).

Although these studies focus on the mechanisms and genes regulating the development of male and female genital disc, there is still a lack of knowledge about the development of the genitalia (Sánchez *et al.*, 1997; Estrada, Casares and Sánchez-Herrero, 2003; Gorfinkiel, Sánchez and Guerrero, 2003). Indeed, the regulation of the differentiation of genital structures during pupal development in particular is not well understood despite studies including the discovery that a network co-opted from posterior spiracle into pupal genitalia development is required to form the posterior lobes (William J Glassford *et al.*, 2015), the generation of an atlas of TF expression for the male pupa terminalia (Vincent *et al.*, 2019) and quantitative trait locus (QTL) studies to map genes responsible of developmental and phenotypic diversification between species (Masly *et al.*, 2011; Takahara and Takahashi, 2015; Tanaka *et al.*, 2015).

1.4 *tartan*

Several studies over nearly the last 40 years have investigated genetics of genitalia variation among *Drosophila* species (Coyne, 1983; Liu, Mercer, *et al.*, 1996; True *et al.*, 1997; Masly *et al.*, 2011; Tanaka *et al.*, 2015; Fujisawa *et al.*, 2019).These studies used either quantitative trait locus (QTL) mapping, a method that allows a rough localisation of large regions in the genome responsible for morphological variation between species (Broman *et al.*, 2003), and/or introgression mapping which can provide a higher resolution (Sawamura, Davis and Wu, 2000; Frazee *et al.*, 2021).

Recently, these techniques were applied to identify *tartan* (*trn*) as one of the loci underlying differences in the morphology of the surstyli between *D. mauritiana* and *D. simulans* (Tanaka *et al.*, 2015; Hagen *et al.*, 2019). This also revealed for the first time that *trn* is involved in genitalia development.

trn was revealed as a candidate gene by introgression mapping of a previously identified QTL for clasper size variation and then found to be differentially spatially and temporally expressed in the developing genitalia of *D. mauritiana* and *D. simulans* (Hagen *et al.*, 2019). In order to confirm the divergence of *trn* between species a reciprocal hemizygosity test was performed, where males with identical genetic background carried only one functional allele of *trn* from either *D. mauritiana* or *D. simulans* (Hagen *et al.*, 2019). The surstyli of the male flies carrying a functional allele of *D. simulans* were smaller and had fewer bristles than those carrying the functional *trn* allele of *D. mauritiana* (Hagen *et al.*, 2019). Furthermore, RNAi knockdown of *trn* in *D. melanogaster* reduced the size of the surstyli (Hagen *et al.*, 2019). Note that although the related gene *capricious (caps)* works together with *trn* in other

developmental contexts, knockdown of *caps* did not affect the surstyli (Milán *et al.*, 2001; Sakurai *et al.*, 2007; Mao, Kerr and Freeman, 2008; Hagen *et al.*, 2019).

The changes in *trn* expression associated with divergence in surstylus size between species is consistent with many other studies where differences in morphology are caused by changes in gene regulation rather than coding changes, in other words, the regulation of *trn* by transcription factors (TFs) would be the responsible of the differences of *trn* expression observed between species (Koshikawa *et al.*, 2015; Glassford *et al.*, 2015; Buffry, Mendes and McGregor, 2016; Hagen *et al.*, 2019). Note that while 22 nucleotide differences were found between the coding sequence of different strains of *D. mauritiana* and *D. simulans,* only three differences were non-synonymous and these were not fixed between these species (Hagen *et al.*, 2019).

1.4.1 Trn function

trn is located on the left arm of chromosome 3 and encodes a leucine-rich repeat transmembrane protein (LRR protein) (Z Chang *et al.*, 1993).

It is known to be expressed in several tissues at different stages of the development, such as the embryo, eye-antenna imaginal disc, leg disc, optic lobe and pupal genitalia (Z Chang *et al.*, 1993; Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Sakurai *et al.*, 2007; Mao, Kerr and Freeman, 2008; Hagen *et al.*, 2019). The presence of *trn* in that many tissues and time points suggests that *trn* is involved in basic functions required for many aspects of development. One of the possible functions that *trn* has been related to is cell compartment allocation and cell to cell affinity, thus changes in *trn* expression may result in differences in cell allocation to compartments or a loss of the compartment boundaries of the tissues (Blair, 2001; Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Sakurai *et al.*, 2007; Mao, Kerr and Freeman, 2008). *trn*

encodes a LRR protein of 14 LRR extracellular domains (ECD) (Chang *et al.*, 1993; Graham *et al.*, 2019a), a family of proteins, which includes Capricious and Conectin. These proteins are thought to be involved in protein to protein interaction between cell membranes, thus mediating cell-cell interactions for cell recognition and organisation (Rothberg *et al.*, 1990; Raghavan and White, 1997; Shishido, Takeichi and Nose, 1998; MingKuan and Wei, 2012; Graham *et al.*, 2019). In other cases, Trn may act as a cell surface receptor. For instance, Trn was found to be acting as a receptor in a yeast two-hybrid screen, where it bound to another cell surface receptor, the neuronal Receptor-linked protein-tyrosine phosphatase Ptp52F, that regulates axon guidance (Bugga, Ratnaparkhi and Zinn, 2009). However in other cases it seems that the cytoplasmic domain of Trn is not required for its function, suggesting that it might act as a ligand rather than a receptor (Milán, Pérez and Cohen, 2002)

In the *Drosophila* embryo *trn* exhibits a striking pattern resembling pair rule gene expression in the early stages, which is regulated by *fushi tarazu* (*ftz*) and *empty spiracles* (*ems*). Subsequently, it might be regulated by a combination of pair rule genes and other segmentation genes, however *trn* is not required for segmentation (Chang *et al.*, 1993; Field, *et al.*, 2016).

Recent studies have provided crucial new insights into Trn function during embryogenesis (Paré *et al.*, 2019; Graham *et al.*, 2019a; Pare and Zallen, 2020). Although Trn shows similarities to proteins of the Toll like receptor family, these proteins are expressed in non-overlapping segments during embryogenesis and it has been suggested that they might be regulated by different pair-rule genes (Graham *et al.*, 2019).

However, It seems that Trn functions in combination with another cell surface protein Tenascin-major to direct planar polarity and boundary cell organisation in the embryo (Paré *et al.*, 2019). The Trn ECD was shown to interact with the ECD of Tenascin-major (Ten-m) and consequently the stripped pattern of Trn in the embryo is required for the correct localization of Ten-m, it was shown that Ten-m accumulates at the membranes of cells that are in contact with other cells where Trn is present (Paré *et al.*, 2019; Graham *et al.*, 2019a). Furthermore, while Trn interacts with Ten-m in adjacent cells, at the same time Trn represses *ten-m* expression within Trn expressing cells (Paré *et al.*, 2019). When this interaction is perturbed because of the reduction of one of these two proteins, this disrupts the cell boundaries (Paré *et al.*, 2019; Pare and Zallen, 2020).

The function of Trn in compartmentalization is not restricted to the embryonic stage, but has been observed in the imaginal discs (Bejarano *et al.*, 2008; Mao, Kerr and Freeman, 2008; Wang and Dahmann, 2020). In the eye and wing discs, Trn together with *cap* regulate boundary formation and cell adhesion (Mao, Kerr and Freeman, 2008; Wang and Dahmann, 2020). In this case, it seems that Trn and Cap are somewhat redundant because individual null mutants seem to have no effect while double mutants show cell rearrangement in these tissues (Mao, Kerr and Freeman, 2008; Wang and Dahmann, 2020). It has also be seen that Trn and Cap prevent apoptosis of wing disc cells and muscle cells, suggesting that they interact with Notch signalling for the apoptotic response (Milán, Pérez and Cohen, 2002; Artero *et al.*, 2003; Bejarano *et al.*, 2008).

In the developing genitalia, *trn* appears to be expressed at boundaries with and between structures (Hagen *et al.*, 2019). Although, the specific function of *trn* in the developing genitalia is unknown, studies of *trn* in other tissues (Milán *et al.*, 2001;

Milán, Pérez and Cohen, 2002; Bejarano *et al.*, 2008; Mao, Kerr and Freeman, 2008; Tsuboi *et al.*, 2017; Wang and Dahmann, 2020) suggest that it might be involved in boundary formation and cell survivability (Milán, Pérez and Cohen, 2002). However, it remains unclear whether Trn acts as a receptor, as a ligand or both, the knockdown of Cap did not result in any significant effect (Hagen *et al.*, 2019). Moreover, the roles of Toll receptors and Ten-M have not been studied in the genitalia.

1.5 Regulation

Clearly *trn* contributes to the rapid divergence of claspers genitalia, however the responsible enhancers and mutations therein underlying differences in *trn* expression between *D. mauritiana* and *D. simulans* have yet to be identified (Hagen *et al.*, 2019). To study how *trn* contributes to the development and divergence of the surstyli, I focused on the regulation of *trn* expression. Thus, before trying to find which are the genes that regulate *trn,* I aimed to identify the enhancers that regulate *trn*. These enhancers contain small motifs that can be bound by TFs and regulate the expression of the gene regulated by that enhancer. After the identification of enhancers that regulate *trn,* I have used them to find possible TFs that can bind to the enhancers identified, hence regulating *trn* (Alonso and Wilkins, 2005; Shlyueva, Stampfel and Stark, 2014; Buffry, Mendes and McGregor, 2016).

1.5.1 Enhancer Identification

Enhancers are DNA sequences that can facilitate the transcription of genes and are able to control the spatial and temporal expression of the genes they regulate (Banerji, Rusconi and Schaffner, 1981; Buffry, Mendes and McGregor, 2016). Enhancers are found genome-wide from short distances to long distance enhancers from the coding sequence of the genes that they regulate (Li and Arnosti, 2011; Sanyal *et al.*, 2012).

For example, Sonic hedgehog (Shh) is regulated by an enhancer that is 1 Mb from its promoter (Lettice *et al.*, 2003). Although in principle enhancers can be found at any distance from the gene they are regulating, chromatin interaction studies show that the probability of contact decreases with distance (Lieberman-Aiden *et al.*, 2009; Barbieri *et al.*, 2012) suggesting then, that most of the enhancers might be found within short distances of the coding region.

There are many different methods to identify enhancers, from predictions made using bioinformatic tools to functional analysis (Suryamohan and Halfon, 2015).

One of the most successful methods is known as “enhancer bashing”. This approach consists of placing the candidate DNA fragment that is thought to be an enhancer upstream a reporter gene that will be used to test the ability to recapitulate part of the expression pattern of a gene (Pfeiffer *et al.*, 2008; Kvon, 2015; Buffry, Mendes and McGregor, 2016). This method by itself is low throughput so it needs to be used on sequences that are likely enhancers due to their location (Kvon, 2015) or can be used to test enhancers first predicted by other methods.

One approach to predict sequences with enhancer activity is studying the conservation of these sequences, or so called phylogenetic footprinting, because it is thought that a noncoding sequence that is an enhancer would be expected to be more conserved than a noncoding sequence without any enhancer activity (Ganley and Kobayashi, 2007; Maeso *et al.*, 2013; Buffry, Mendes and McGregor, 2016). However, this method is highly dependent on the species chosen to find footprints of enhancers (Ganley and Kobayashi, 2007; Hardison and Taylor, 2012; Maeso *et al.*, 2013).

Other methods that can be used to narrow down enhancers are based on chromatin profiling and protein binding. These include immunoprecipitation (ChIP-seq), Assay

for transposable-accessible chromatin (ATAC-seq) and Chromosome Conformation Capture methods (3C, 4C, 5C and HI-C).

ChIP-seq involves crosslinking proteins to the DNA sequences they are bound to, then the DNA fragments bound by the protein of interest can be purified using an antibody against the protein and subsequently sequencing (Mundade *et al.*, 2014). This method can be used to identify the putative genome wide distribution of binding sites and enhancers bound by a specific protein (Mundade *et al.*, 2014).

Chromatin conformation capture approaches are used to map the tertiary structure of the genome (Risca and Greenleaf, 2015). 3C detects contact between sequences previously selected, 4C between a sequence and the other regions of the genome and 5C uses more than two sequences previously selected (usually multiple sequences) and detects contact between them, and finally HI-C provides information about interactions of all sequences throughout the whole genome (Risca and Greenleaf, 2015). These approaches provide information about possible enhancers may be some distance away from the gene they regulate (Bouwman and de Laat, 2011).

CRISPR/Cas9 now also allows testing of enhancers and transcription factor binding sites (Cong *et al.*, 2013; Mali *et al.*, 2013; Lopes, Korkmaz and Agami, 2016). For example, it can be used to edit specific binding sites and study the effects within the endogenous locus (López-Rivera *et al.*, 2020), or to modulate enhancers using nuclease-deactivated CAS9 to activate or repress transcription (Lopes, Korkmaz and Agami, 2016).

ATAC-seq, is a method that involves using the Tn5 transposase to perform a DNA fragmentation and at the same time tagging these fragments with sequencing adapters (Buenrostro *et al.*, 2013, 2015). With this method accessible regions will be tagged

and sequenced while the less accessible sequences are less likely to be tagged by the transposase, due to the Tn5 transposase being a hyperactive prokaryotic transposase that cuts unprotected (accessible) DNA (Buenrostro *et al.*, 2013, 2015). The sequenced fragments will reveal regions that are accessible in the tissue studied, thus facilitating the prediction of regions that are more likely to be enhancers or promoters.

1.6 Aims of this thesis

The overall aims of my thesis were to identify the enhancer(s) and TFs that regulate *trn* expression and to investigate if changes in the enhancer(s) with genital activity may have evolved to underlie the differences in the male surstyli between *Drosophila* species.

To address these aims, this thesis was divided in the following Results chapters.

Chapter 2

This chapter involved using a combination of endogenous *trn* tagged lines, ATAC-seq data and functional reporter assays.

- Using RNA in situs and an endogenous *trn* tagged line to better describe the pattern of the endogenous *trn* in developing genitalia to be able to compare it to enhancer reporter activity assays.

- Generation of reporter constructs in *D. melanogaster* to assess regions covering the *trn* locus from the closest genes upstream and downstream *trn.*

Chapter 3

This chapter used the ATAC-seq data, together with RNA-seq data from male genitalia of *D. mauritiana* and *D. simulans* to identify candidate TFs that bind to the enhancers found in Chapter 2. This also involved the use of RNAi to test candidates.

- Use of ATAC-seq to find footprints and regions with peaks to identify binding sites of TFs.

- Using RNA-seq to filter the candidate TFs to eliminate those TFs which are not expressed in the male genital of *Drosophila* species.

- Functionally testing using RNAi knockdown of predicted TFs candidate genes to determine if they are involved in male genitalia.

Chapter 4

In this chapter I will compare *Drosophila mauritiana* and *Drosophila simulans* at the sequence level and I will test enhancers to see how they differ between them.

- ATAC-seq data generated from developing genitalia of *D. mauritiana* and *D. simulans* were used to analyse the open chromatin profile of the *trn* locus to assess whether there was any difference between species.

- Comparison of *trn* sequences from different species to describe the similarities and differences in identified enhancers.

- Testing of reporters from *Drosophila mauritiana* and *Drosophila simulans* in *Drosophila melanogaster.*

Chapter 2. Investigation of the *trn* regulatory landscape

2.1 *trn* regulatory landscape

trn is located in a genomic position where the most proximal upstream gene is at approximately 24 kb (CG11251) from *trn's* TSS, while the closest downstream gene is at approximately 18 kb (CG33262) (Fig. 2.2) (Thurmond *et al.*, 2019). This genomic region contains around 42 kb of non-coding sequence, while minimal enhancers are thought to be of 200 base pairs (bp) – 1 kb (Levine, 2010; Suryamohan and Halfon, 2015). Therefore, given the complexity of trn expression in multiple different developing tissues as well as the genitalia (Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Sakurai *et al.*, 2007; Hong *et al.*, 2009; Graham *et al.*, 2019; Hagen *et al.*, 2019), this genomic region could potentially harbour several *cis*-regulatory elements of *trn.*

Therefore, in this chapter, I analysed this genomic region to investigate the regulatory landscape of *trn* as the first step towards understanding the evolution of *trn* expression in developing male genitalia between species. To do this I first further analysed *trn* expression in the developing terminalia of *D. melanogaster,* thus with the endogenous expression of *trn* characterised it can be compared to reporters. Then I carried out a reporter screen for sequences in the region with enhancer activity in this species and compared it to the endogenous expression of *trn*.

2.2 Methods

2.2.1 Cloning

Candidate enhancer regions were amplified from *D. melanogaster* (Oregon R strain) gDNA and cloned initially into pENTRY/D-TOPO vector. The primers used can be

found in Supplementary Table 1. Then, using LR gateway cloning the regions were inserted into a GAL4 vector compatible with Gateway cloning that can be integrated into specific targets in vivo genomic using the PhiC31 integrase, the pBPGUw plasmid (pBPGUw was a gift from Gerald Rubin (Addgene plasmid # 17575; http://n2t.net/addgene:17575 ; RRID:Addgene_17575)) (Pfeiffer *et al.*, 2008). This vector contains a Drosophila synthetic core promoter (DCSP) that contains TATA, Inr, MTE and DPE motifs, additionally the GAL4 CDS and yeast transcriptional terminator can easily be substituted for another driver by a directional 5' KpnI to 3' HindIII digest (Pfeiffer *et al.*, 2008). Embryos from the stock y[1] w[67c23]; P{y[+t7.7]=CaryP}attP2 (Bloomington stock 8622) were injected with plasmid pBPGUw containing the selected candidate regions and the ΦC31 integrase helper plasmid pBS130 (pBS130 was a gift from Tom Clandinin (Addgene plasmid # 26290; http://n2t.net/addgene:26290; RRID:Addgene_26290)) (Gohl *et al.*, 2011). Injected F0s were crossed to *w* ;;TM6B/TM3 and the F1s screened for transgenic individuals marked by the orange colour of their eyes. This approach allowed me to compare the activity of the enhancer fragments as reporters when they are inserted into the same place in the genome.

2.2.2 Fly stocks and genetics

All fly stocks were kept at 25 °C on standard medium yeast extract-sucrose. Fly lines with candidate regions were crossed to *y, w*;;UAS-GFP line to test their activity by assaying Green fluorescence protein (GFP) expression. *w*;;trn::NeonGreen was made and provided by Joanna F. Hagen (then at the Francis Crick Institute). This line expresses endogenous *trn* protein tagged line with fluorescent tag inserted at the N terminus, which allows us to detect the endogenous *trn* protein due to the fluorescence of the NeonGreen protein. *w*;;trn::Neongreen was crossed to *w,y*; Ecad::tdTomato to

image. *w*$^{+}$;;U4-GAL4/TM6B and *w*$^{+}$;;U8-GAL4/TM6B were crossed to *w*$^{+}$;Ecad::GFP/CyO;UAS-RedStinger/TM6B.

2.2.3 Microscope imaging

White prepupae from the crosses were collected and aged until 24 and 46 hours after puparium formation (hAPF) at last 5 pupae were examined for each line. These time points were selected because they precede important stages of the genitalia development, when the surstyli and the epandrial posterior lobe are forming and separating, thus genes involved in their regulation are thought to be active (William J Glassford *et al.*, 2015). For whole body mounting the pupal case was dissected and then the pupa was mounted on a glass slide with 80% glycerol. Images were taken using Zeiss Axiozoom stereomicroscope. For pupal genitalia dissections, the posterior part of the pupa was cut off and the genitalia was dissected in PBS 1x and mounted on a glass slide with immunostaining hydro mount liquid. Images were taken using Zeiss LSM 800 confocal using the 25x objective, 0.2% laser power of DsRed and 1% laser power for GFP.

2.2.4 In Situ Hybridization

RNA in situ hybridization (*ISH*) experiments were performed on developing genitalia in pupae collected from *D. melanogaster w*^{1118} using the protocol described in Hagen *et al.*, (2019). The probe for *trn ISH* was also described in Hagen *et al.*, (2019). *ISH* were imaged using a Zeiss Axiozoom stereomicroscope.

2.3 Results

2.3.1 *tartan* endogenous expression pattern

To better understand *trn's* regulatory landscape and enhancers that drive its expression in the developing genitalia, I first sought to better understand its expression in this tissue. Previous in situ hybridization experiment performed in *D. melanogaster* showed expression around the anal plates, phallus and base of the surstyli or joint between the surstyli and anal plates, as well as potentially in distal regions of the claspers and epandrial ventral lobes, but this was very faint (Fig. 2.1 A) (Hagen *et al.*, 2019). My *ISH* detected a stronger expression at 46 hAPF around the anal plates, in addition to a very clear and strong signal at the distal of the surstyli and epandrial ventral lobes (Fig. 2.1 B).

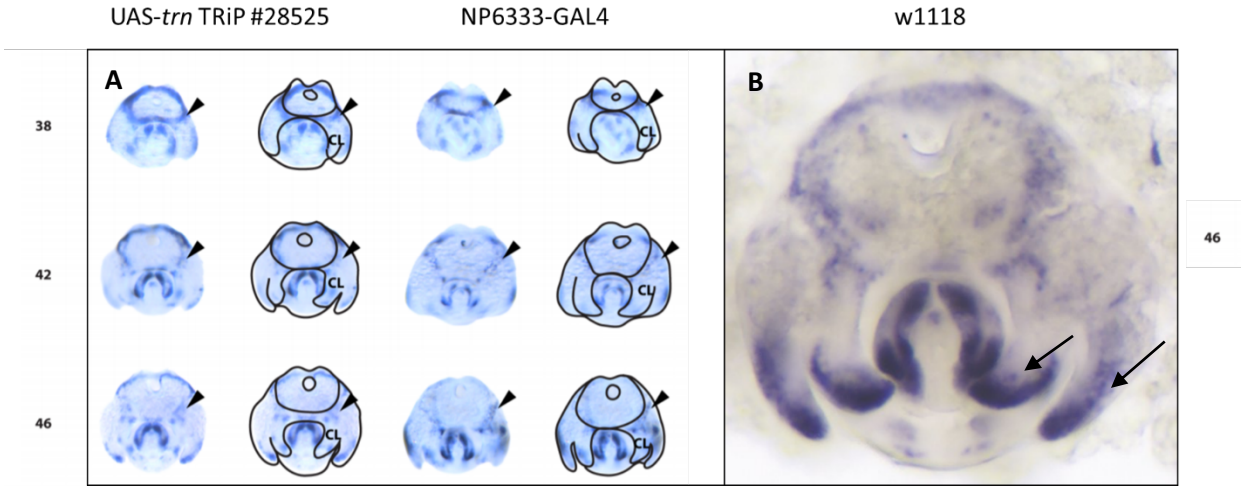


Figure 2.1. *trn in situ hybridization* (*ISH*). (A) *trn ISHs* performed in *D. melanogaster* at 38 hAPF, 42 hAPF and 46 hAPF from the lines UAS-*trn* TRIP #28525 and NP6333-GAL4 from Hagen *et al*., 2019. Surstyli are marked by CL and arrow indicates expression of *trn.* (B) *ISH* performed in *D. melanogaster* line w1118 at 46 hAPF (see methods). Arrows mark stronger expression found in comparison with previous *ISHs* at the distal of the surstyli and epandrial ventral lobes.

To help corroborate the expression of *trn* as assayed by RNA *ISH*, I then tested a line of *D. melanogaster* that expresses the endogenous Trn protein tagged with mNeonGreen (Shaner *et al.*, 2013; Hagen *et al*., unpublished)(see methods). This line

was crossed to a line in which E-Cadherin is tagged with another fluorescent protein, tdTomato (Shaner *et al.*, 2004), to mark cell membranes allowing concomitant visualisation of the general structure of the genitalia.

At 24 hAPF, Trn can be found surrounding the developing anal plate and more strongly on the lateral sides of these structures (Fig. 2.2 A, C). It is also present in the external sides of the primordium that will give rise to the epandrial ventral lobe, surstylus and posterior lobes (Fig. 2.2 A, C and J). The male genital is more developed at 35 hAPF, where we can see that the surstyli can already be distinguished from the epandrial ventral lobes and the posterior lobes are starting to show (Fig. 2.2 D, F and K). At this stage, Trn exhibits a similar pattern of that seen at 24hAPF with expression surrounding the anal plates, but at this stage it is more strongly expressed at the intersection between the surstyli and the anal plates, at the terminal part and external part of the surstyli and in the phallus (Fig. 2.2 D, F).

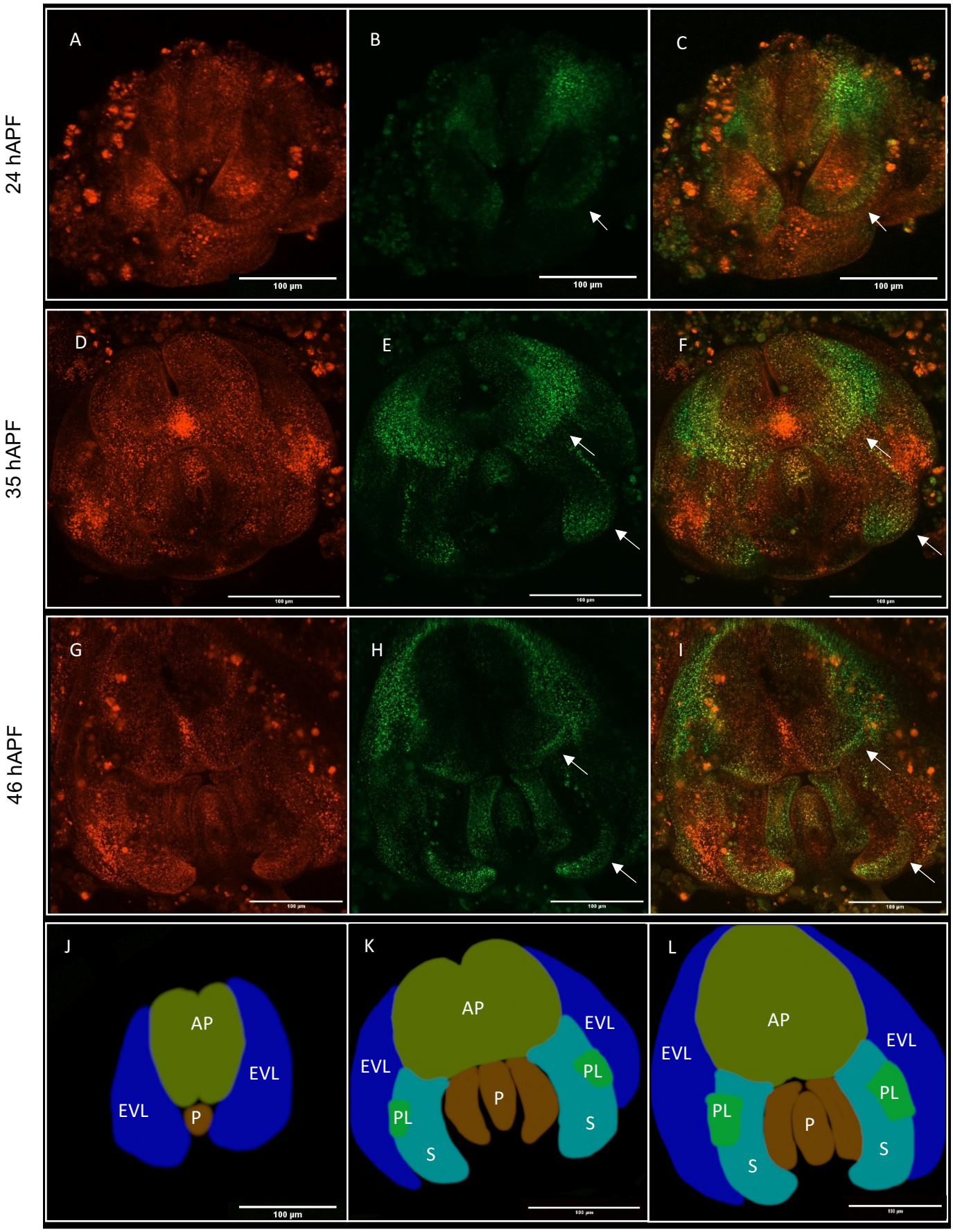


Figure 2.2. Trn endogenous protein pattern. Ecad tagged with tdTomato is shown in red and Trn tagged with Neongreen is shown in green. (A, C) Images of male

genitalia taken at 24 hAPF. (D, F) Images of male genitalia taken at 35 hAPF. (G, I) Images of male genitalia taken at 46 hAPF. (J, L) Schematic representation of the structures of the *Drosophila* male genitalia at 24 hAPF, 35hAPF and 46 hAPF consecutively. The anal plate (AP) is represented in green, surstylus (S) in light blue, posterior lobe in green (PL), epandrial ventral lobe (EVL) in blue and phallus (P) in brown. White arrows point to aspects of expression of Trn in the EVL, S and AP reported in the text.

All the genital structures are more fully resolved at 46 hAPF, where we can see that they already resemble the structures of the adult male genitalia (Fig. 2.2 G, I and L), the surstyli are fully separated and differentiated from the epandrial ventral lobe and the posterior lobes are nearly fully formed (Fig. 2.2 G, I and L). At this stage, Trn continues to be expressed in a similar pattern to previous stages but with a stronger expression at the phallus, and the surstyli, where it shows a strong domain of expression in the external region (Fig. 2.2 G, I).

These patterns are similar to those that were already identified by *ISH* previously (this work; Hagen *et al.*, 2019). The expression surrounding the anal plates is consistent with all the *ISHs* performed (Fig. 2.1 A), although the endogenous pattern is stronger than observed by Hagen *et al.,* (2019). The stronger domain seen in the surstyli in the new *ISH* I performed, is consistent with the endogenous expression found at 46 hAPF (Fig. 2.1 B and 2.2 H, I). The strong Trn expression in the distal surstylus and proximal at the base where the surstylus intersects with the anal plates may be responsible for the differences seen in surstyli size between species (Hagen *et al.*, 2019).

2.3.2 Enhancer survey

2.3.2.1 Design of enhancer constructs

In order to try to find the enhancer(s) responsible for *trn* expression in the developing genitalia that could contribute to the differences seen in *trn* expression between *Drosophila* species (Hagen *et al.*, 2019), I surveyed the *trn* locus to identify sequences with enhancer activity (Fig. 2.3). This included two existing GAL4 reporter lines from the Vienna collection (VDRC) that covered regions within *trn* locus, of which line VT030021 was already identified as a *trn* embryonic enhancer (Kvon *et al.*, 2014; Rivera *et al.*, 2019). I then divided most of the rest of the region in overlapping segments of ~3 Kb, except for the last region due to its proximity with the gene downstream and due to a coordinate mix-up U2 is ~50 bp separated from VT030010 (Fig. 6) (Supplementary Table 2).



Figure 2.3. *trn* locus in *D. melanogaster* showing candidate regions. *trn* and flanking genes are shown in blue. Regions that were taken to make constructs to test are shown in red, including the VT lines that were obtained from the Vienna collection (VDRC).

In addition to the 2 pre-existent VT lines, I cloned a total of 15 candidate regions into a vector pBPGUw (Pfeiffer *et al.*, 2008) and established for twelve of them (excluding U6, U7 and D1) at least 1 transgenic line of attP2 Bloomington stock 8622. This stock was used due to having the same docking site for the attP integration than the 2 pre-existent VT lines, thus allowing to compare all enhancers under the same genomic background. Note that unfortunately, it was not possible to generate transgenic flies

to test regions U6, U7 and D1 before submission of this thesis. I then crossed these GAL4 lines to UAS-GFP to test their enhancer activity at 24 hAPF and 46 hAPF. These timepoints were chosen because they correspond to when the male genitalia primordia have become visible but before they divide and then when the main structures are nearly fully formed (Hagen *et al.*, 2019). Therefore, I assayed for enhancers that are active early and late in the development of male genitalia in pupae. Note that these regions were tested without considering position with respect to the chromatin accessibility profile (see Chapter 4). Thus, I tested sequences that would have been overlooked if I had just focused on regions of accessible chromatin (see Chapter 4).

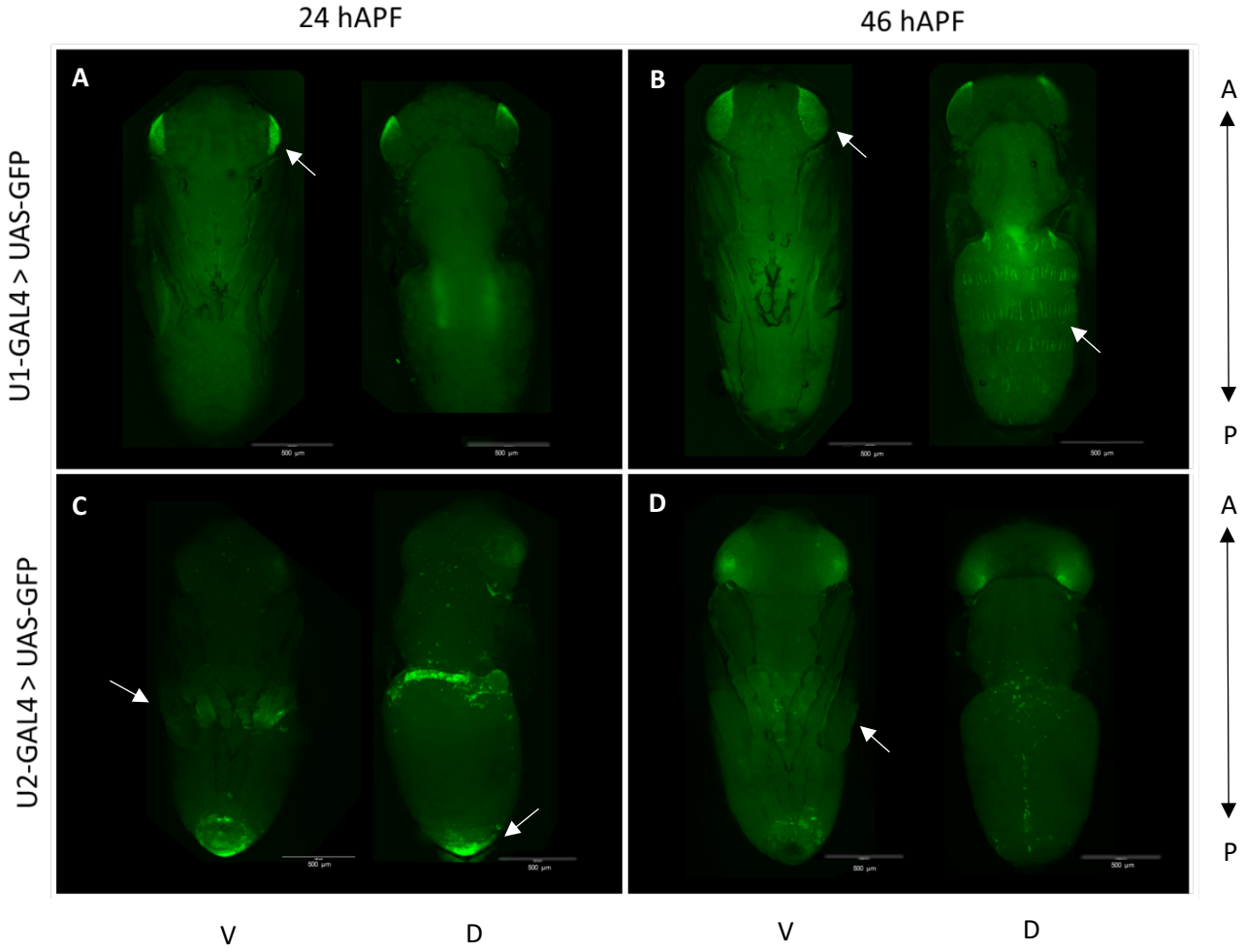


Figure 2.4. Activity of regions U1 and U2. (A, B) Expression pattern driven by U1 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. Expression is observed in the eyes at both timepoints and the abdomen at 46 hAPF. (C, D) Expression pattern of U2 at 24 hAPF (C) and 46 hAPF (D) in the whole pupae of males. U2 also drives in

the eyes at both timepoints, but more strongly at 46 hAPF. Expression is also detected at the periphery of the terminalia and in the dorsal region of the first abdominal segment. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

2.3.2.2 Survey of enhancer activity of candidate regions

Regions U1 and U2 located proximally to CG11251 (Fig. 2.3), both showed enhancer activity in pupal stages (Fig. 2.4). U1 is expressed in the eyes at both timepoints, however at 46 hAPF it also appears to drive expression in the abdominal segments (Fig. 2.4 A, B). U2 is expressed in a similar pattern at both stages with stronger expression in the eyes, but it seems to be internal instead of the expression observed to be driven by U1, that appears to be on the surface of the forming ommatidia (Fig. 2.4 C, D). U2 at 46 hAPF gives stronger expression in the eyes than at 24 hAPF (Fig. 2.4 C, D). It is also active in the periphery of the terminalia and in the first segment of the abdomen on the dorsal region and possibly ventral as well (Fig. 2.4 C, D) (Supplementary Fig. 2). This expression in the eye could potentially be associated with *trn* expression and function during eye development (Mao, Kerr and Freeman, 2008).

Line VT030010 showed no detectable expression at the stages tested (Fig. 2.5 A, B). U3 drives expression in many different tissues (Fig. 2.5 C, D). At 24 hAPF it drives GFP at the rim of the eyes, in the maxillary palps, sternum, notum, and the joints between the tarsal segments (Fig. 2.5 C, D). U3 drives a similar pattern at 46 hAPF although the eye expression is weaker (Fig. 2.5 D).

Region U4 drives expression in the genitalia and in the antennae, (Fig. 2.6 A, B). Within the genitalia the expression pattern is similar at both stages, with expression

detectable at both the most posterior part and the basal part of the genitalia (Fig. 2.6 A, B).

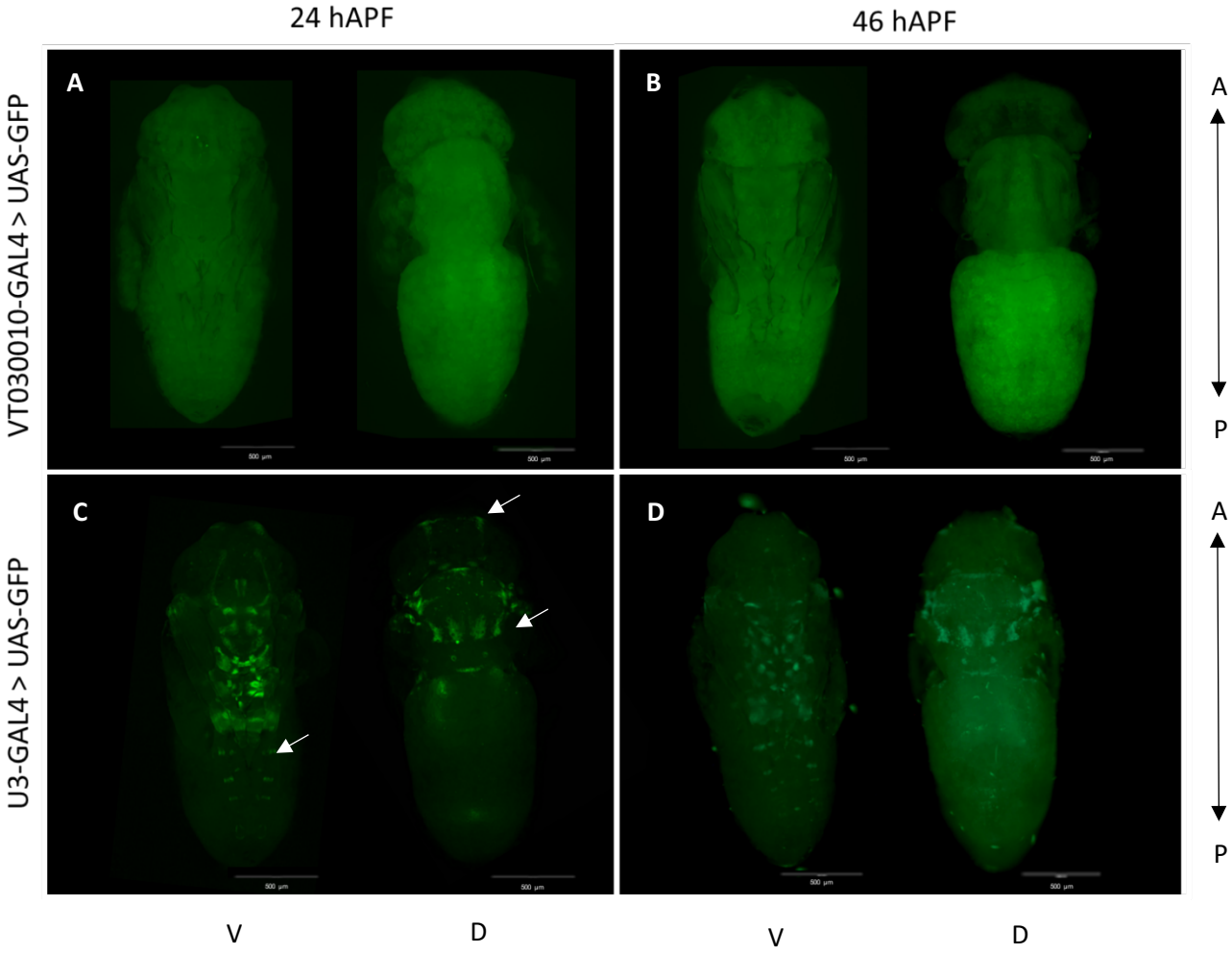


Figure 2.5. Activity of line VT030010 and region U3. (A, B) Expression pattern driven by VT030010 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. VT030010 showed no detectable expression at either stage. (C, D) Expression pattern driven by U3 at 24 hAPF (C) and 46 hAPF (D) in the whole pupae of males. U3 drives expression at both timepoints in the joints of the legs between the tarsal segments. Expression is also detected in the rim of the eyes at 24 hAPF, maxillary palp and the notum and sternum of the thorax. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

Region U5 drives expression mostly at 24 hAPF in the proximal part of the wing where it connects with the thorax (Fig. 2.6 C). It also drives expression in the maxillary palps

at 24 hAPF (Fig. 2.6 C). This expression seems to be weaker at 46 hAPF, when it may also drive expression in the internal tissues of the abdomen (Fig. 2.6 D).

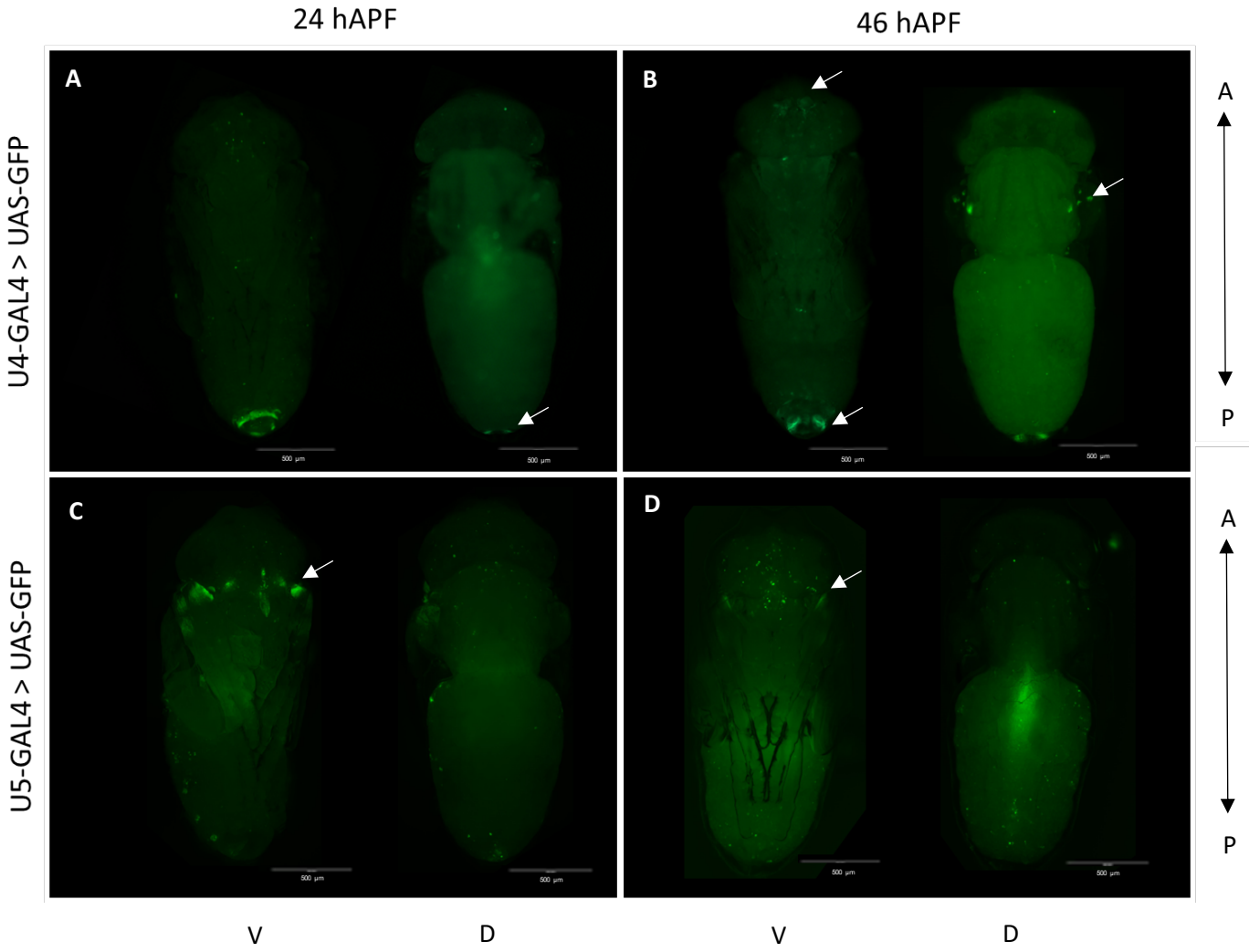


Figure 2.6. Activity of regions U4 and U5. (A, B) Expression pattern driven by U4 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. Expression driven by U4 was detected in the genitalia at both stages, at the wings at 46 hAPF (B) and a weaker expression in the antenna. (C, D) Expression pattern driven by U5 at 24 hAPF (C) and 46 hAPF (D) in the whole pupae of males. Expression of U5 was detected mainly at the proximal part of the wing with the thorax and in the maxillary palps. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

U8 is the closest candidate to the promoter of *trn*, but it does not include the TSS. This enhancer drives a very specific expression in the genitalia at both stages, where the

GFP signal is observed at two specific points in the central region of the genitalia (Fig. 2.7 A, B). It can also be seen at the eye at 24 hAPF but this is weaker at 46 hAPF (Fig. 2.7 A, B), and at 46 hAPF expression driven by U8 is also observed in the sternum.

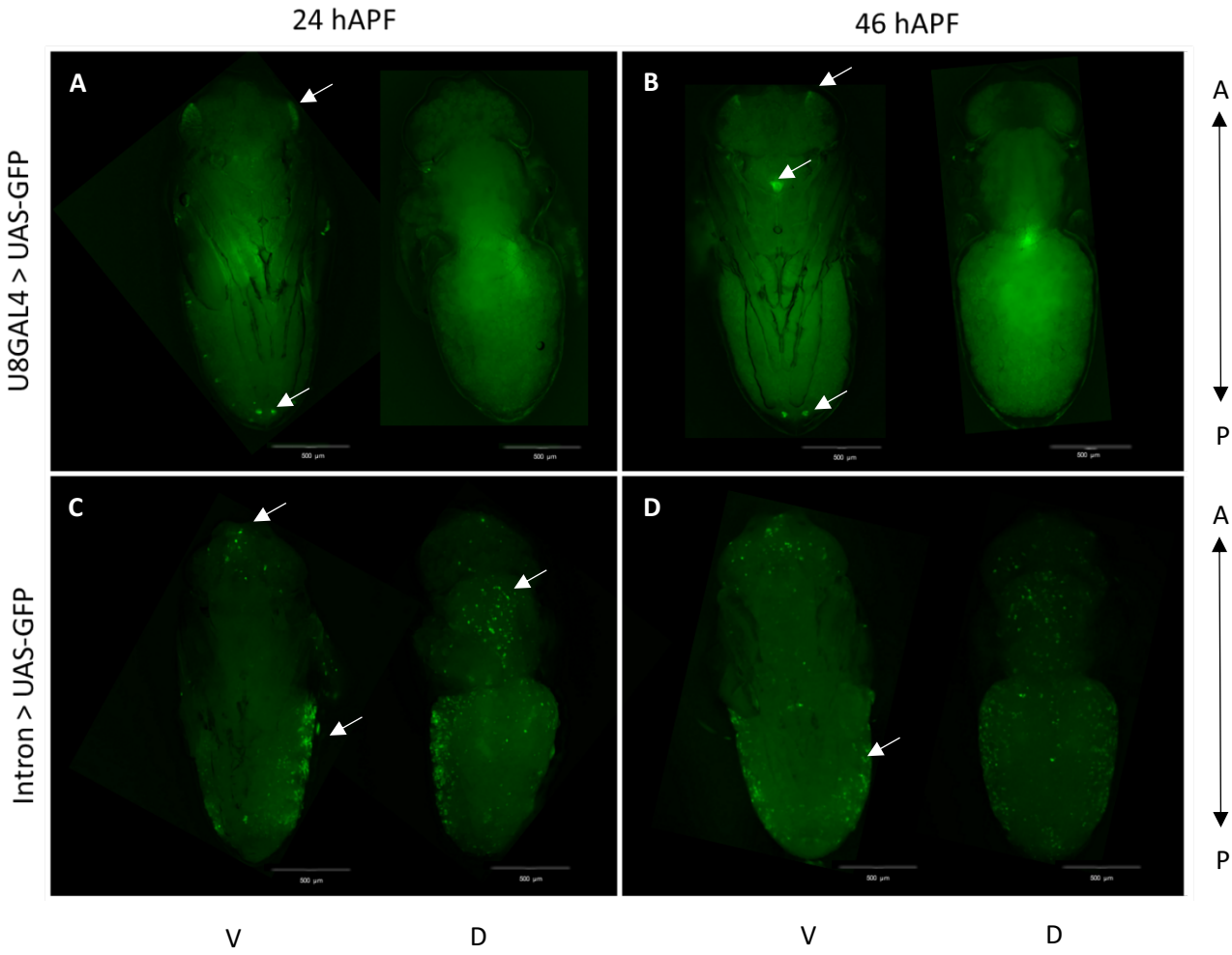


Figure 2.7. Activity of regions U8 and Intron. (A, B) Expression pattern driven by U8 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. U8 drives expression in the eyes and in the genitalia. (C, D) Expression pattern of *trn* intron at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. The intron drives expression at the notum, head and both lateral sides of the abdomen. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

The 447 bp *trn* intronic region drives expression in several different tissues including thorax and head. However, this expression lacks a pattern that can be compared to any specific body structure. However, in the abdomen the expression driven by the *trn*

intron seems to be associated with the abdominal segments (Fig. 2.7 C, D), where it is more clearly distinguishable at 24 hAPF (Fig. 2.7 C, D).

It was previously shown that region VT030021 is an embryonic enhancer of *trn* (Kvon *et al.*, 2014; Rivera *et al.*, 2019).However, it did not drive any detectable expression when I tested its activity in pupae (Fig. 2.8 A, B).

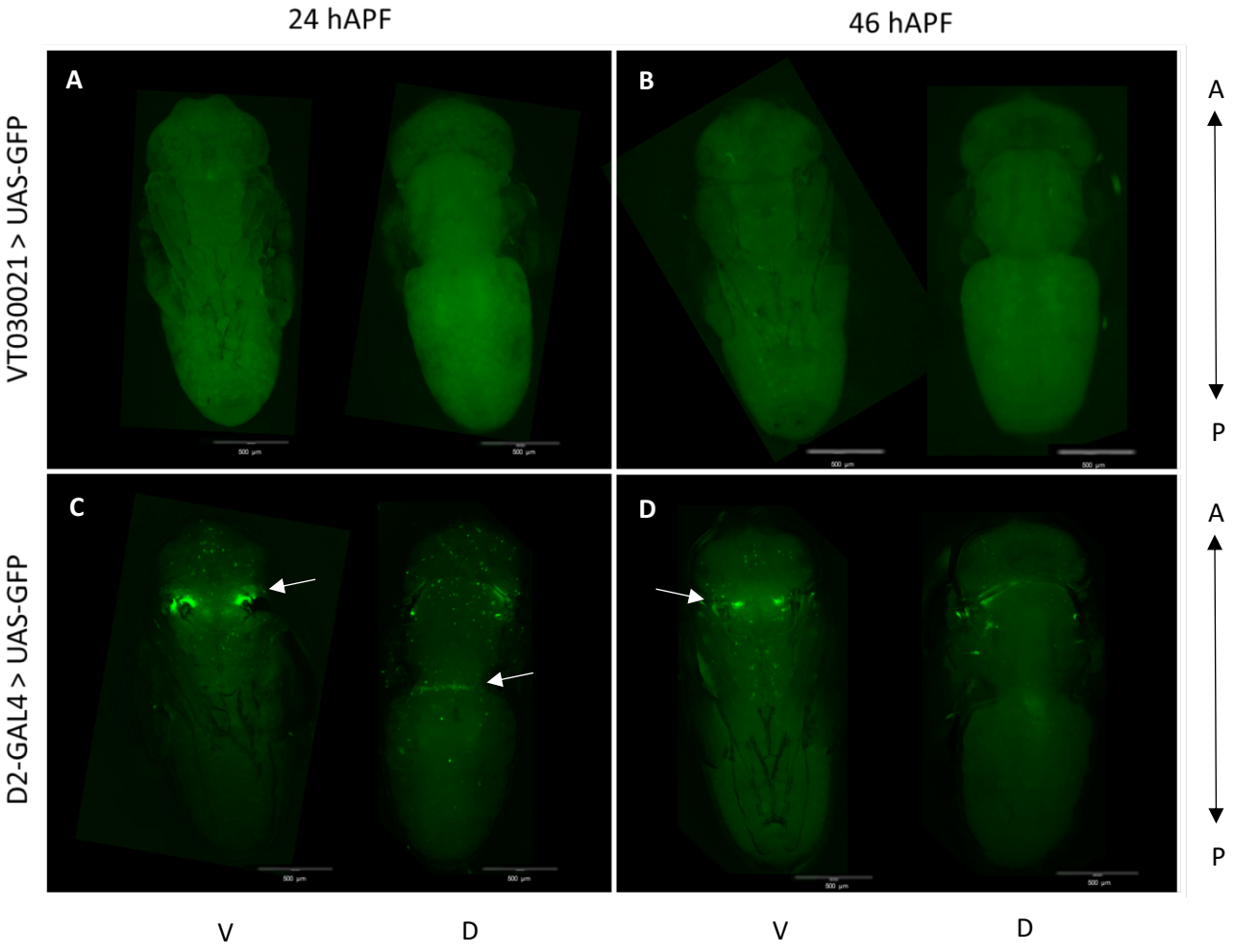


Figure 2.8. Activity of VT030021 line and D2 region. (A, B) Expression pattern driven by VT030021 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. No expression was found to be driven by VT030021 at the tested stages. (C, D) Expression pattern driven by D2 at 24 hAPF (C) and 46 hAPF (D) in the whole pupae of males. D2 drives expression at both stages between the head and the thorax, it also drives expression at 24 hAPF at the dorsal region of the first abdominal segment. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

Region D2 drives expression in a similar pattern to U5 (Figs. 2.6 C, D and Fig. 2.8 C, D). This expression seems to be located between the head and the thorax, in addition to expression found at 24 hAPF (Fig. 2.8 C) in the dorsal region of the first abdominal segment and some scattered dots in the head and thorax (Fig. 2.8 C, D).

In the case of D3, this region drives similar expression at both stages, almost everywhere in the body (Fig. 2.9 A, B). However, tissues that are formed from imaginal discs, such as legs, eyes, antenna, genitalia, and wings, appear to lack expression driven from this enhancer (Fig. 2.9 A, B).

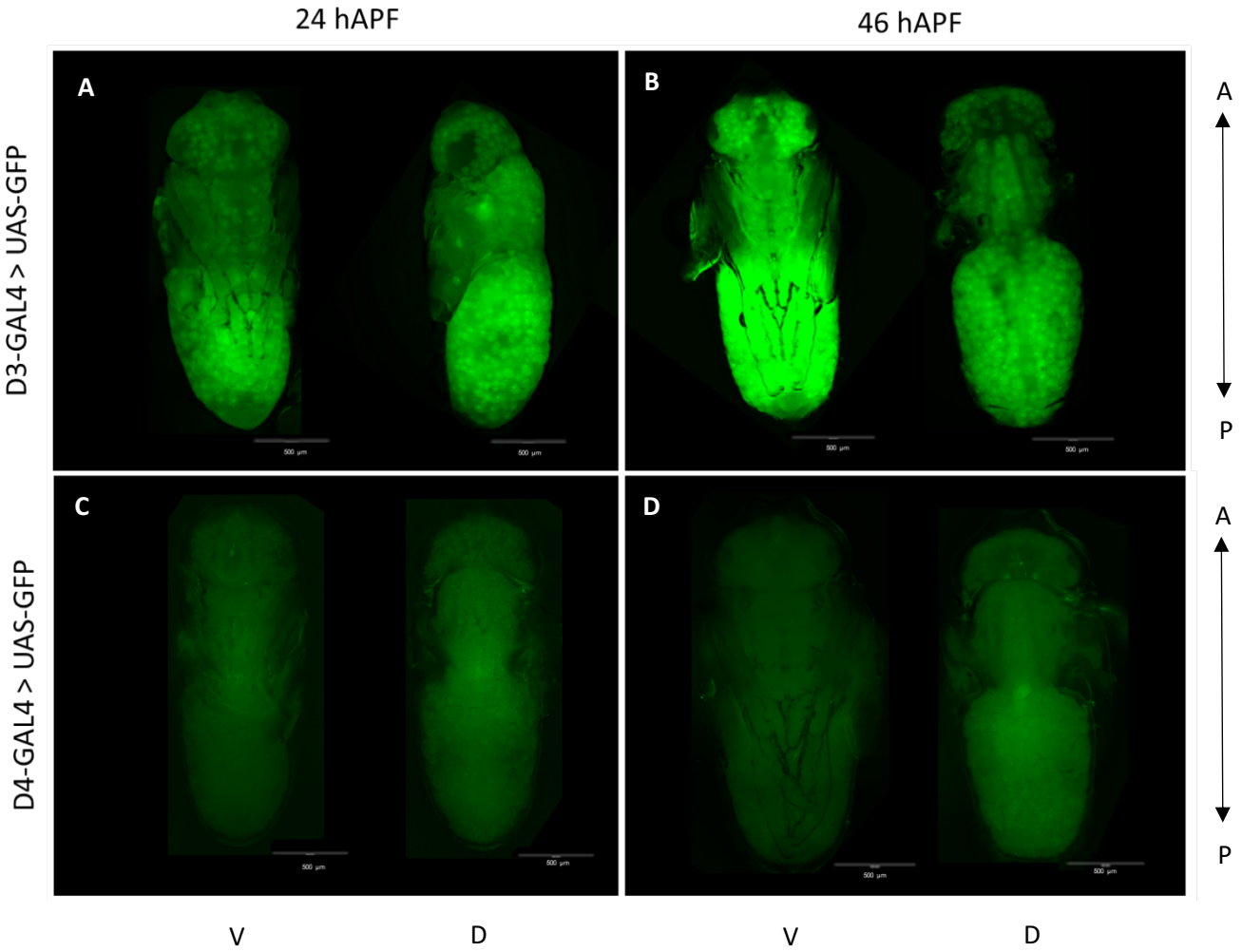


Figure 2.9. Activity of regions D3 and D4. (A, B) Expression pattern of D3 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. D3 drives expression everywhere but in the legs, genitalia, eyes, antenna, and wings. (C, D) Expression pattern driven by D4 at 24 hAPF and 46 hAPF in the whole pupae of males. No

expression was found to be driven by D4 at the tested stages. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

Regions D4 did not drive any detectable expression at the timepoint tested (Fig. 2.9 C, D). D5 which is located approximately 13 kb downstream of the *trn* coding region (Fig. 2.3), only appeared to drive expression in a few spots in the abdomen at 46 hAPF (Fig. 2.10 A, B).

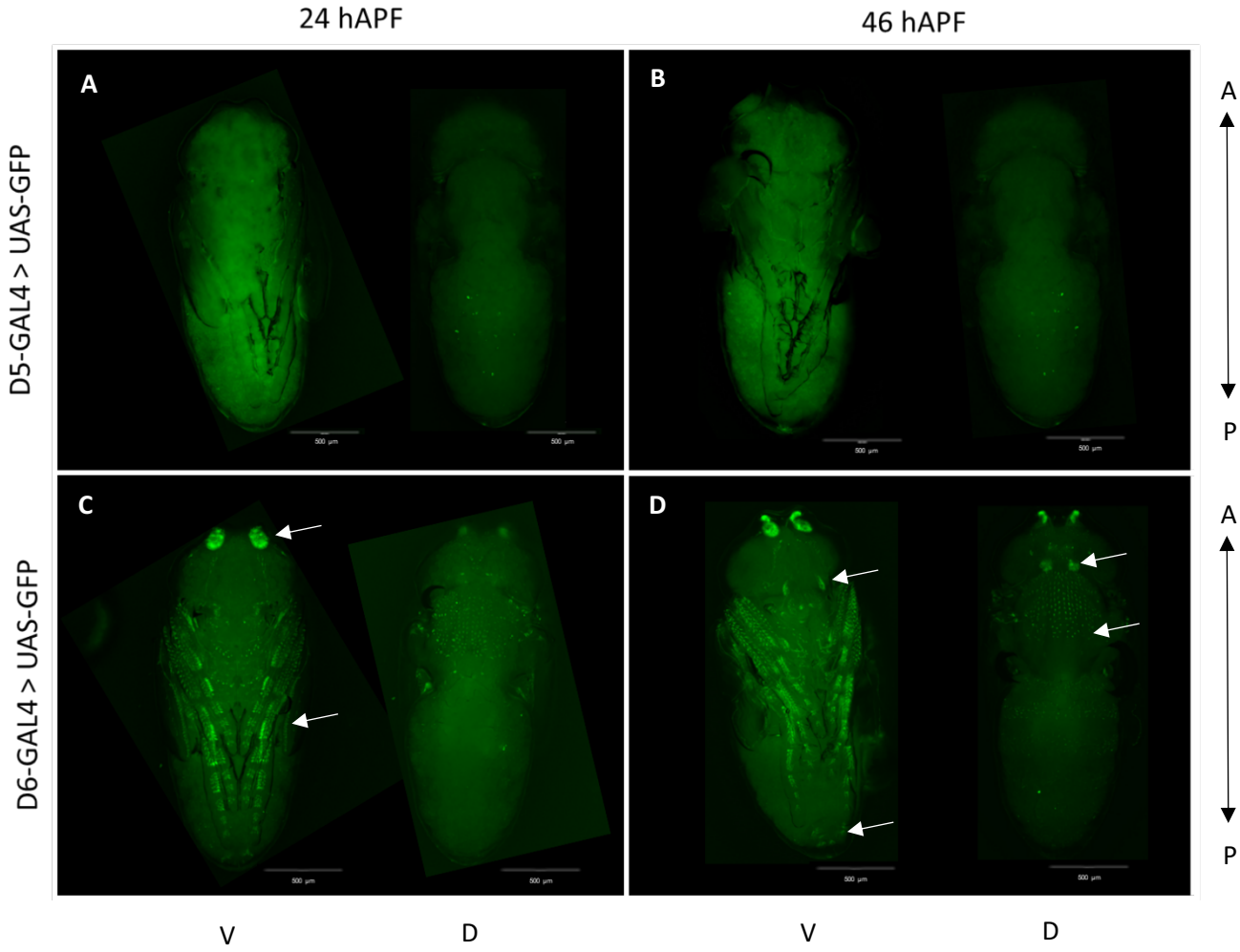


Figure 2.10. Activity of D5 and D6 regions. (A, B) Expression pattern driven by D5 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. D5 does not drive expression at these stages apart from a few spots in the abdomen at 46 hAPF. (C, D) Expression pattern driven by D6 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. D6 drives expression in many different tissues, legs, thorax, abdomen, antenna, eye and genitalia. Anterior is to the top in all images, and ventral (left) and dorsal (right) views are shown in each panel.

D6, the region closest to the downstream gene CG33262, drives expression in many different tissues (Fig. 2.10 C, D). This expression appears to be localised to the developing bristles of these tissues, including the bristles presents in the genitalia (Fig. 2.10 C, D).

In total I tested fourteen regions, and a summary of their enhancer activity is shown in Table 1.

Table 1. Summary of enhancer activity of regions tested at 24 and 46 hAPF.

Candidate	Expression driven at 24hAPF	Expression driven at 46hAPF
U1	Eyes surface	Eyes surface
U2	Eyes, first abdominal segment and periphery of the terminalia	Eyes, first abdominal segment and periphery of the terminalia
VT030010	No detected expression at this stage	No detected expression at this stage
U3	Joints of the legs, rim of the eyes, maxillary palps, notum and sternum of the thorax	Joints of the legs, maxillary palps, notum and sternum of the thorax
U4	Genitalia	Genitalia and antennae
U5	Wings and maxillary palps	Wings and maxillary palps
U8	Genitalia and eyes	Genitalia, eyes and sternum
Intron	Abdominal segments, notum of the thorax and head	Abdominal segments, notum of the thorax and head
VT030021	No detected expression at this stage	No detected expression at this stage
D2	First abdominal segment and intersection between the head and the thorax	Intersection between the head and the thorax
D3	Everywhere except tissues that form from imaginal discs	Everywhere except tissues that form from imaginal discs
D4	No detected expression at this stage	No detected expression at this stage
D5	No detected expression at this stage	Spots in the abdomen
D6	Bristles of many different tissues, including legs, thorax, abdomen, antennae, eye and genitalia	Bristles of many different tissues, including legs, thorax, abdomen, antennae, eye and genitalia

2.3.3 Enhancers driving expression in the genitalia

From the fourteen regions I tested for enhancer activity, four of them appear to drive expression in the genitalia (Table 1). At first glance U2, U4, U8 and D6 are all candidates for enhancers that drive *trn* expression in the developing genitalia (Fig.

2.2). To explore this in more detail, I imaged male genitalia specifically for the U4 and U8 enhancers driving RFP expression and Ecad tagged with GFP in order to visualise the genitalia (see Methods).

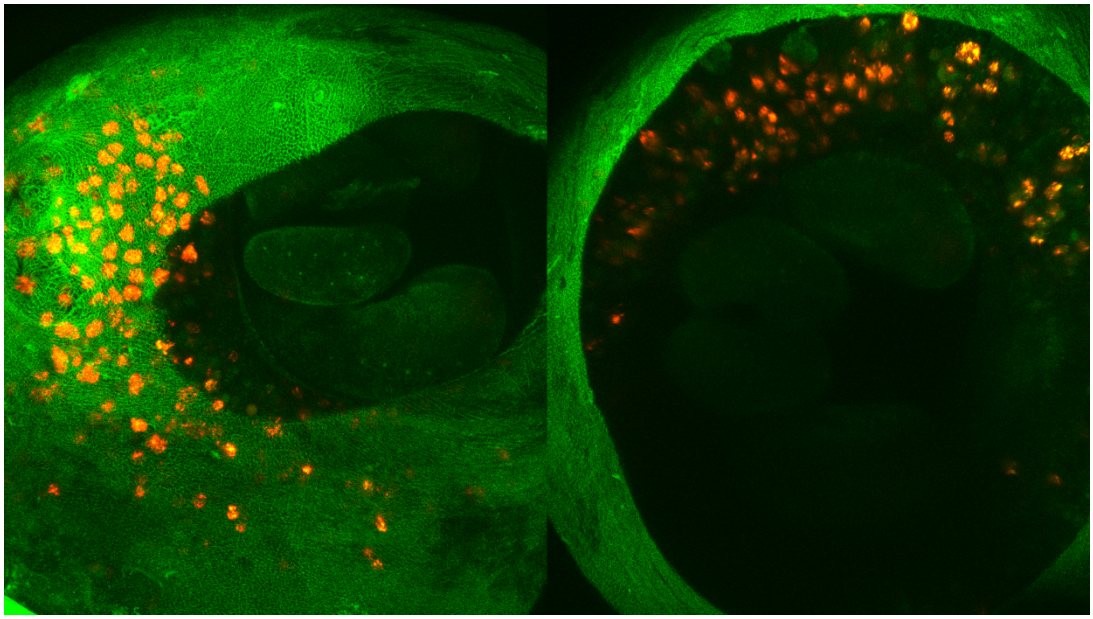


Figure 2.11. U2 reporter expression at 24 hAPF in the male genital. Captured from live imaging of the expression driven by U2 in the male genital at 24 hAPF. Ecad is shown in green, and expression driven is shown in red.

I focused on U4 and U8 because the expression of D6 appears to be generally in developing bristles rather than specific to the genitalia. Also, I observed that the expression driven by U2 was not actually located in the genitalia (Fig. 2.11), instead this expression is located in the abdominal tissue surrounding the terminalia (Fig. 2.11). Expression driven by U4 recapitulates part of the endogenous *trn* expression (Fig. 2.2 and Fig. 2.12). At 24 hAPF, U4 drives expression in lateral regions of the anal plates, where it connects to the primordium from which will form the surstyli and epandrial ventral lobes (Fig. 2.12 A, C). This expression is very similar to endogenous *trn* expression in *D. melanogaster* at the same timepoint (Fig. 2.3 A, C and Fig. 2.14). At 46 hAPF, U4 expression still observed in the anal plates, but it is also present in

the phallus, at the base of the surstyli and it can also be seen between the surstyli and the epandrial ventral lobes (Fig. 2.3 D, F). It should be noted that U4 does not recapitulate the full pattern of the endogenous Trn protein at either timepoint, but nevertheless appears to be responsible for a subset of specific expression (Fig. 2.4, Fig. 2.12 and Fig 2.14).

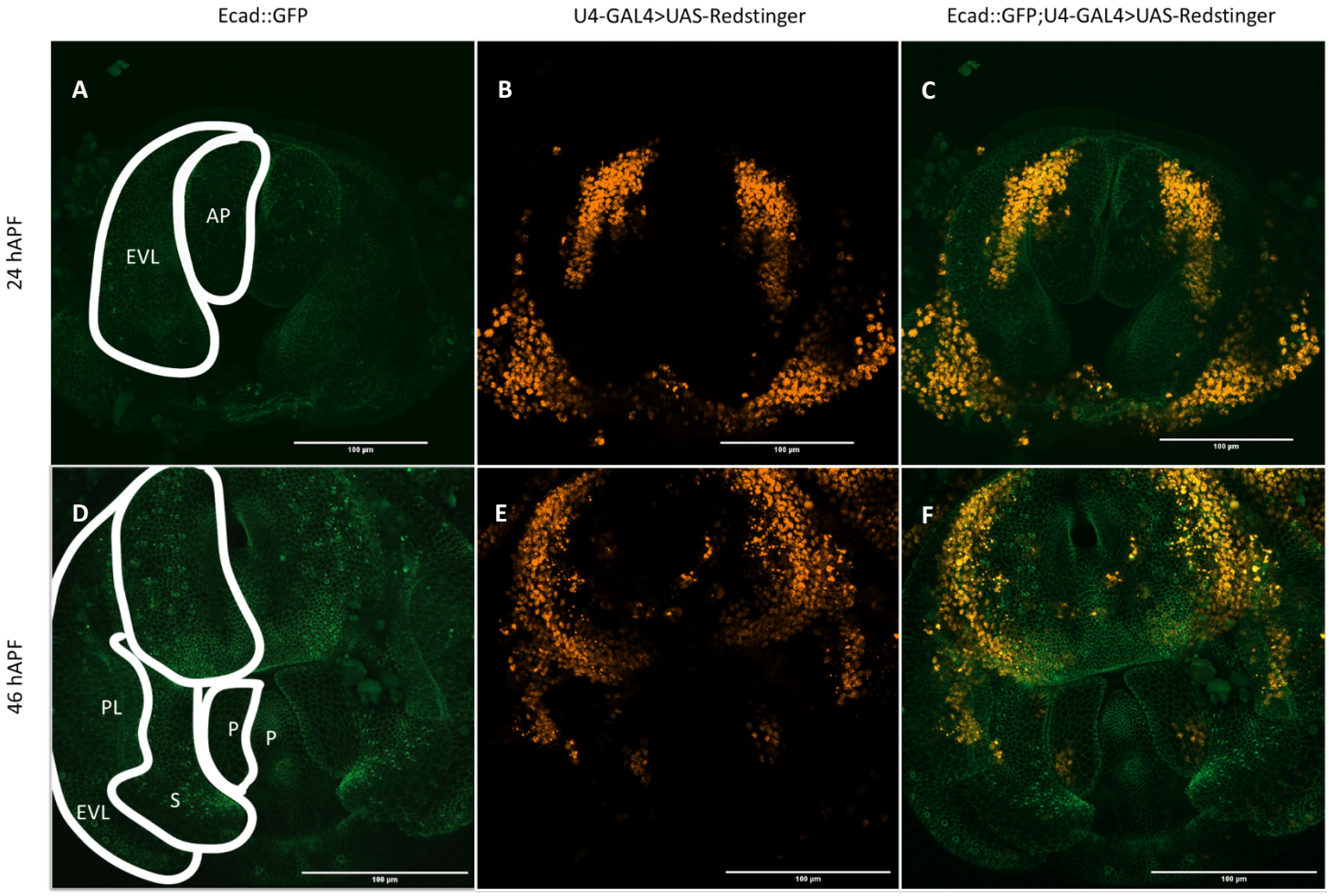


Figure 2.12. Activity of U4 in the male genital. (A-C) Expression driven by U4 at 24 hAPF mostly localised to the lateral regions of the anal plate. (D-F) U4 expression at 46 hAPF, expression localised to surrounding the anal plate, with part of the expression in the phallus and between the surstylus and the epandrial ventral lobe. Parts of the male genital are outlined in the images (A) and (D), anal plate (AP), surstylus (P), epandrial ventral lobe (EVL), posterior lobe (PL) and phallus (P).

At 24 hAPF, U8 drives a very localised expression pattern in the interior part of the primordium from which the epandrial ventral lobes and surstyli will form (Fig. 2.12 A-C). This expression appears to be specific to the distal part of developing surstyli. This is consistent with the expression driven by U8 at 46 hAPF (Fig. 2.12 D, E), which is specific to a subset of cells of the more distal part of the surstyli (Fig. 2.12 D-F). At this stage, expression can also be detected at the aedeagal sheath of the phallus (Fig. 2.12 D-F)

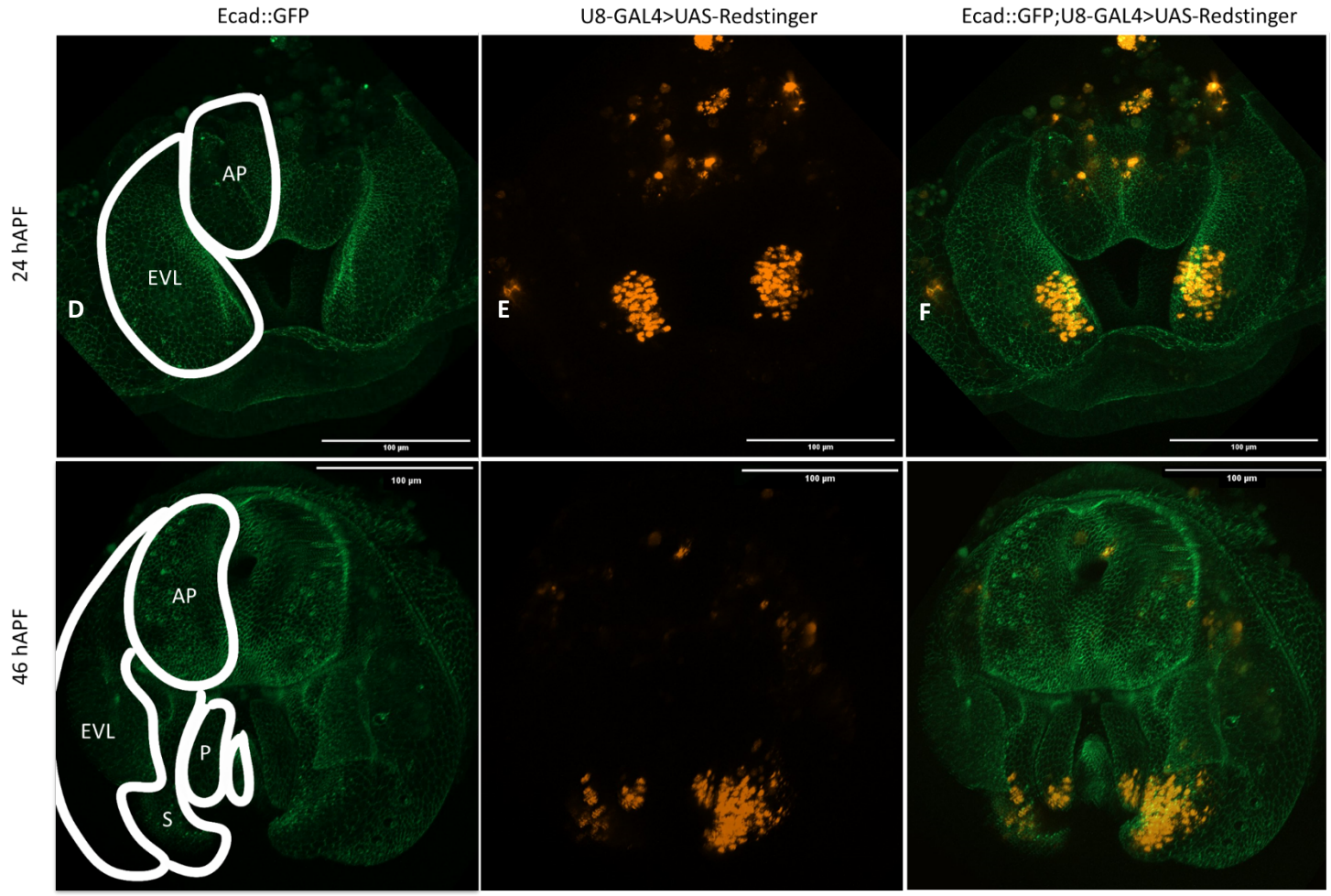


Figure 2.13. Activity of U8 in the male genital. (A-C) Expression driven by U8 at 24 hAPF localised in the epandrial ventral lobe in the internal part. (D-F) Expression driven by U8 at 46 hAPF, expression localised at the internal part of the surstyli and at the base of the aedeagal sheath of the phallus. Parts of the male genital are outlined in the images (A) and (D), Anal plate (AP), Surstylus (P), Epandrial ventral lobe (EVL), Posterior lobe (PL) and Phallus (P).

This expression driven by U8 seems to correlate with the endogenous pattern of Trn (Fig. 2.3 and Fig. 2.13), however, there are some discrepancies. At 24 hAPF, the endogenous Trn Expression seems to be present at the distal region of the primordium, but only at the external side, while U8 driven expression may overlap with this, it seems to be more localised to the interior part (Fig. 2.2 A, C, Fig. 2.13 A-C, Fig. 2.14 and Supplementary Fig. 2.1 A). At 46 hAPF there are also differences, as the Trn expression domain is reduced again to the external and more distal part of the surstyli (Fig. 2.3 G-I), while U8 driven expression is more widely spread at the distal part of the surstyli (Fig. 2.12 D-F).



Figure 2.14. Schematic representation of Trn endogenous expression pattern and regions U4 and U8. Endogenous Trn expression pattern and expression driven by regions U4 and U8 represented in red over a schematic representation of the male genital at 24 hAPF and 46 hAPF. Arrows point at expression pattern driven by regions U4 and U8 that appear to correlate with the endogenous expression of Trn. Endogenous expression represented in red, U4 expression in Green and U8 expression in blue.

2. 4 Discussion

Using a non-biased screening for cs-regulatory activity in the *trn* locus, I have identified nine regions with enhancer activity. While this approach requires more work than focussing on predictions using ATAC-seq data or other bioinformatic tools, these results suggest that a prediction biased approach can overlook certain sequences. For example, regions U8 and D6 in the ATAC-seq profile would have been overlooked due to not being specifically accessible in the data. (See Chapter 4 for further discussion of this).

2.4.1 Enhancers identified

Although the sequences with enhancer activity are localised close to the *trn* coding region, these enhancers could regulate the flanking genes or even genes that are further away (Li and Arnosti, 2011; Small and Arnosti, 2020). However, some of the enhancers identified correlate *trn* expression (Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Mao, Kerr and Freeman, 2008), which strongly suggest they are *trn* enhancers. This could be further validated, perhaps using chromatin capture technology in the future (Risca and Greenleaf, 2015).

The VT lines tested did not drive any expression in pupae at the stages examined, although VT030021 has been previously described to drive expression in the embryo (Kvon *et al.*, 2014; Rivera *et al.*, 2019), this enhancer could be silent or repressed at the pupal stage.

Regions U1, U2 and U3 drive expression in tissues, where *trn* expression has been previously described, specifically the eyes and the legs. It is interesting in the case of U3 that it is found to drive expression in the joints of the tarsal segments of the legs, which are lost when *trn* and c*ap* are mutated (Sakurai *et al.*, 2007).

For other regions with enhancer activity identified in this chapter, such as the intron or D3, it is unclear whether they regulate *trn* or not, as there is no previous evidence of *trn* expression or function in the tissues in which they are driving expression, for example, the abdominal segment expression driven by the intronic region (Fig. 2.7 C, D). Furthermore, although my focus was on identifying genital enhancers of *trn*, it would be interesting to compare the enhancer activity of the regions more systematically with *trn* expression during development.

2.4.2 U4, U8 and D6

In this chapter, I identified three enhancers that can drive expression in the pupal male genital of *D melanogaster*, which makes them candidates to be responsible for the differential regulation of *trn* in the genitalia between *D. mauritiana* and *D. simulans* (Hagen *et al.*, 2019).

U4 is the most promising enhancer in this regard because the expression it drives completely colocalises with a subset of the endogenous Trn protein expression (Fig. 2.3, Fig. 2.11 and Fig. 2.13). Further testing is required to assess whether this enhancer has evolved between species and is regulating *trn* differently between them, which I have started in Chapter 4.

On the other hand, U8 initially looked promising, but the pattern of expression does not seem to colocalise completely with the endogenous expression of Trn protein (Fig. 2.3, Fig. 2.13 and Fig. 2.14). However, it is a very specific pattern in the surstyli and overlaps somewhat with Trn specially at 46 hAPF (Fig. 2.13 and Fig. 2.14). It is possible that the U8 sequence is missing key flanking binding sites that could be needed to drive a more specific *trn* expression pattern in the surstyli. This could be

tested by making a larger version of U8 that encompasses flanking sequences in the future.

D6 is the last enhancer I found driving expression in the genitalia and this expression seems to be specifically localised to the bristles (Fig. 2.10). However, D6 is not a very specific enhancer because it is also active in many other bristles on other parts of the body. D6 is also very close to CG33262 (Fig. 2.2), but there is lack of information about this gene, and its expression has not been characterised.

Another interesting focus for these enhancers would be to test them in the female developing genitalia of *D. melanogaster*. Whether these enhancers are responsible for the differences seen between species or not, if the enhancers that control certain regulation in the male developing genitalia are the same that control the regulation in the female developing genitalia would mean that if they suffered changes in these sequences, they would evolve together, or otherwise they would do it independently.

Chapter 3. Investigating transcription factors that regulate *trn* in the male genital of *Drosophila*

3.1.1 Transcription factor binding sites

The regulation of spatial and temporal gene expression via an enhancer depends on TFs recognising combinations of binding sites to activate or repress the target gene (Crocker, Preger-Ben Noon and Stern, 2016). These binding sites are motifs of normally 6-10 nucleotides in eukaryotes (Stewart, Hannenhalli and Plotkin, 2012), which are bound by TFs with different levels of affinity (Badis *et al.*, 2009; Slattery *et al.*, 2011; Crocker, Preger-Ben Noon and Stern, 2016). High affinity binding sites seem to be less abundant than low affinity sites (Badis *et al.*, 2009; Slattery *et al.*, 2011; Crocker, Preger-Ben Noon and Stern, 2016). This suggests that the use of one or a few strong binding sites in enhancers could be selected against and it is thought that the use of multiple low affinity binding sites can confer more specific regulation because these allow the cooperation of TFs and may increase robustness (Giniger and Ptashne, 1988; Ludwig *et al.*, 2000; Crocker *et al.*, 2015; Crocker, Preger-Ben Noon and Stern, 2016). Identifying enhancers and binding sites is key to understanding gene expression and gene regulatory networks, however this is not straightforward despite recent advances.

3.1.2 ATAC-seq

ATAC-seq is a method to look for accessible chromatin (Buenrostro *et al.*, 2013) and since it was first developed it has increasingly been used more than other methods like FAIRE-seq or DNase-seq to assay for accessible chromatin (Yan *et al.*, 2020).

This is due to ATAC-seq providing useful information about chromatin accessibility but requiring less starting material and time than FAIRE-seq and DNase-seq and resulting in less background than the former (Buenrostro *et al.*, 2013, 2015; Yan *et al.*, 2020).

Although ATAC-seq does not provide information about the 3D conformation of the chromatin like chromosome conformation capture methods (3C, 4C, 5C and Hi-C), it can reveal important information about chromatin accessibility, transcription factor occupancy and nucleosome position (Buenrostro *et al.*, 2013; Yan *et al.*, 2020). Furthermore, annotation of ATAC-seq peaks normally shows a distribution where ~30% correspond with promoters, with the majority of the other peaks in introns and intergenic regions (~51%) (Yan *et al.*, 2020). Thus, ATAC-seq can provide a genome-wide study of regions with regulatory potential in focal tissues.

3.1.3 Finding binding sites using ATAC-seq

TFBSs are not only short (Levine, 2010; Stewart, Hannenhalli and Plotkin, 2012), but they are also degenerate and somewhat redundant in sequence and a given TF can recognise different motifs and similar sequences can be bound by different TFs (Crocker *et al.*, 2015; Crocker, Preger-Ben Noon and Stern, 2016). Therefore, a combination of bioinformatic tools and other experimental methods are required to identify enhancers.

ATAC-seq provides information about the accessible chromatin, but this can encompass kilobases of DNA, with many candidates binding sites (Levine, 2010; Stewart, Hannenhalli and Plotkin, 2012; Crocker, Preger-Ben Noon and Stern, 2016). Therefore, in order to better predict binding sites located in accessible chromatin many bioinformatic tools have been developed (Jayaram, Usvyat and Martin, 2016). However there has been a lack of methods developed that are specific to the analysis

of ATAC-seq datasets, and consequently using methods developed for analysis of ChIP-seq or DNase-seq datasets may introduce possible artefacts from not taking the DNA cleavage preferences of Tn5 into account (Li *et al.*, 2019). Nonetheless, there are recently developed methods available that try to take advantage of Tn5 bias in order to find the footprints of TF binding (Li *et al.*, 2019; Bentsen *et al.*, 2020).

ATAC-seq data can also be used to find binding sites in combination with motif analysis tools, such as i-*cis*Target, Homer, and The MEME Suite (Bailey *et al.*, 2009; Heinz *et al.*, 2010; Herrmann *et al.*, 2012). These tools were not initially developed with ATAC-seq datasets in mind, and i-*cis*Target and Homer are mostly used with ChIP-seq datasets (Heinz *et al.*, 2010; Herrmann *et al.*, 2012). However, the MEME suite can allow either discovery of novo motifs or find motif enrichment within peaks of accessible chromatin identified with ATAC-seq (Bailey *et al.*, 2009).

After finding these motifs, they can be compared to transcription factor binding site databases to find which transcription factors could bind to them (Bailey *et al.*, 2009). In this chapter, I combined the analysis of ATAC-seq data and motif analysis to identify candidate TFs that can bind the U4 enhancer (see Chapter 2). I then tested the candidate *caudal* (*cad)* using RNAi.

3.2 Methods

3.2.1 Motif discovery and footprinting

Motif discovery was carried out using MEME and STREME (Bailey *et al.*, 2009; Bailey, 2021). The sequences used were extracted from *D. simulans* (r2.02), *D. mauritiana* (r1.0) using HOMER and were posteriorly aligned using MAFFT with the *D. melanogaster* U4 sequence (Katoh and Standley, 2013).The following parameters were using in MEME: 0-order model of background, motifs distributed in sequences

with any number of repetitions (ANR), sites may be located on either strand, maximum number of motifs 10, minimum and maximum width of motifs 6 -16, minimum number of sites per motif 2. The parameters used in STREME were as follows: Shuffle input sequences, minimum and maximum width of motifs 6 -16, p-value threshold 0.05. Motifs were then compared with JASPAR (non-redundant) database using Tomtom (Gupta *et al.*, 2007; Fornes *et al.*, 2019). Footprinting was carried out using HINT-ATAC, part of the RGT (Regulatory Genomic Toolbox) (Li *et al.*, 2019). Peaks from *D. simulans* ATAC-seq corresponding to U4 were used with genomic bias correction using *D. simulans* (r2.02). HINT-ATAC used the JASPAR insect database to search for binding sites using footprints. TFBSs were filtered using RNA-seq data from developing male genitalia (Hagen *et al.*, 2020).

3.2.2 RNAi Knockdown of *cad*

RNAi was performed against *cad* using the line *y,sc,v,sev;UAS-cadRNAi* (Bloomington Drosophila Stock Center, stock 34702). I used the line U4-GAL4 and *w;*NP6333-GAL4;UAS-Dicer as drivers (Chatterjee *et al.*, 2011). NP6333-GAL4 drives in the leg, wing and genital discs (Stieper *et al.*, 2008; Tanaka *et al.*, 2015). Crosses were carried out at 25 °C on standard medium yeast extract-sucrose. Ten Females and five males were used for each cross and were changed to a new food vial every 2 days. Adult flies were collected on the third day after eclosion and stored in 70% EtOH until dissection. Genitalia and the tibia from T2 legs were dissected and mounted in Hoyer's medium/lactic acid 1:1. Samples were imaged using a Zeiss Axioplan light microscope with 250x magnification for surstyli, posterior lobes and anal plates and 160x magnification for the T2 tibia. Bristles were counted under the microscope while the area for the different structures was measured from images using ImageJ (Schneider, Rasband and Eliceiri, 2012). Whole abdomen images were obtained by

dissecting and mounting in 80% glycerol and imaged using a Zeiss Axiozoom stereomicroscope.

3.2.3 Statistical analysis

A Shapiro-Wilk normality test was performed to test the distribution of datasets. A one-way ANOVA was used to test for differences in normally distributed data, followed by Tukey's test to test for differences between groups. Anal plate area, surstylus area (both RNAi datasets), posterior lobe area from U4-GAL4>UAS-*cad*-RNAi and anal plate bristles number were normally distributed (Shapiro-Wilk normality test p-value > 0.05).

NP6333-GAL>UAS-*cad*-RNAi posterior lobe area (p-value <0.01), NP6333-GAL>UAS-*cad*-RNAi surstylus bristle number (p-value <0.01) and U4-GAL4>UAS-*cad*-RNAi surstylus bristles number (p-value <0.01) were not normally distributed. A Kruskal-Wallis test was therefore used to test for differences in these datasets followed by a Wilcoxon rank sum test to test for differences between groups.

Correlation analysis using a linear regression test was performed to test for a relationship between the tibia and area of genital structures (Supplementary Fig. 3.3). U4-GAL4>UAS-*cad*-RNAi were corrected for body size by dividing the area (mm^2) by the tibia length squared.

All analysis were performed using R studio, R version 4.0.3 (2020-10-10) (R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna) https://www.r-project.org.

3.3 Results

3.3.1 Binding sites in U4

U4 is the most promising enhancer that I found to be recapitulating part of *trn* endogenous expression. Therefore, I focussed on this enhancer to look for binding sites of transcription factors that could regulate *trn*. To do so, I used the peaks of accessible chromatin found within U4 from the ATAC-seq dataset obtained from male genital tissue of *D. mauritiana* and *D. simulans* (see Chapter 4) (Supplementary Fig. 3.1). There are two peaks of accessible chromatin within U4 (see Fig. 4.1 Chapter 4) covering ~300 and ~881 bp of this region (Supplementary Fig. 3.1), I used these peaks to look for the presence of DNA motifs that may have the potential of being bound by a TF using the motif discovery tool Multiple Em for Motif Elicitation (MEME) (Bailey *et al.*, 2009). I used the sequences extracted from the ATAC-seq equivalent to the peaks of accessible chromatin found in *D. mauritiana* and *D. simulans.* I also aligned these sequences with the U4 sequence of *D. melanogaster* and used the corresponding sequence to look for motifs in *D. melanogaster* (Supplementary Fig. 3.1).

I looked for motifs that were between six and sixteen nucleotides because binding sites are normally around ten nucleotides long (Stewart, Hannenhalli and Plotkin, 2012). The motifs that were found in all three species using MEME had an E-value higher than 0.05 (Fig. 3.1), which means that motifs found in the U4 peaks were not significantly enriched compared to the background model used in MEME.

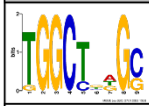
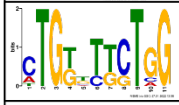
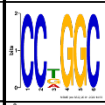
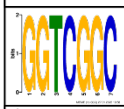
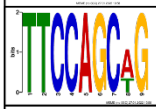
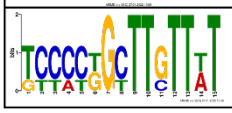
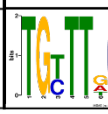
	D. mauritiana			*D. simulans*			*D. melanogaster*		
	Motif	E-value	Sites	Motif	E-value	Sites	Motif	E-value	Sites
1		2.60E+01	10		2.70E+01	10		6.20E+01	10
2		8.60E+01	10		2.70E+01	10		6.40E+02	9
3		1.70E+04	2		3.90E+02	6		1.80E+03	6
4		2.00E+04	3		1.40E+03	9		1.90E+03	9
5		2.60E+04	4		1.80E+04	5		1.00E+04	10

Figure 3.1. Motifs found in U4 accessible chromatin in all three *Drosophila* species. Motifs found using MEME are numbered from 1 to 5. E-value and number of motifs found in each species are on the right of their respective motifs.

However, Motif 2 in *D. simulans* was similar to Motif 1 in *D. melanogaster* and these had the lowest e-value and the highest number of sites found within these sequences (Fig. 3.1). I therefore used the Tomtom tool (Fornes *et al.*, 2019) and I compared this motif with the JASPAR database of TFBSs. This motif correlated significantly with the binding site of Cad (Fig. 3.2). *cad* is highly expressed in RNA-seq data from the genitalia, at around 100 Transcripts per million (TPM) (Hagen *et al.*, 2020) and in addition *cad ISH* experiments of *cad* show a striking expression pattern in the developing anal plates (Vincent *et al.*, 2019). Therefore, *cad* is a good candidate for regulation of U4 in the region surrounding the anal plates.

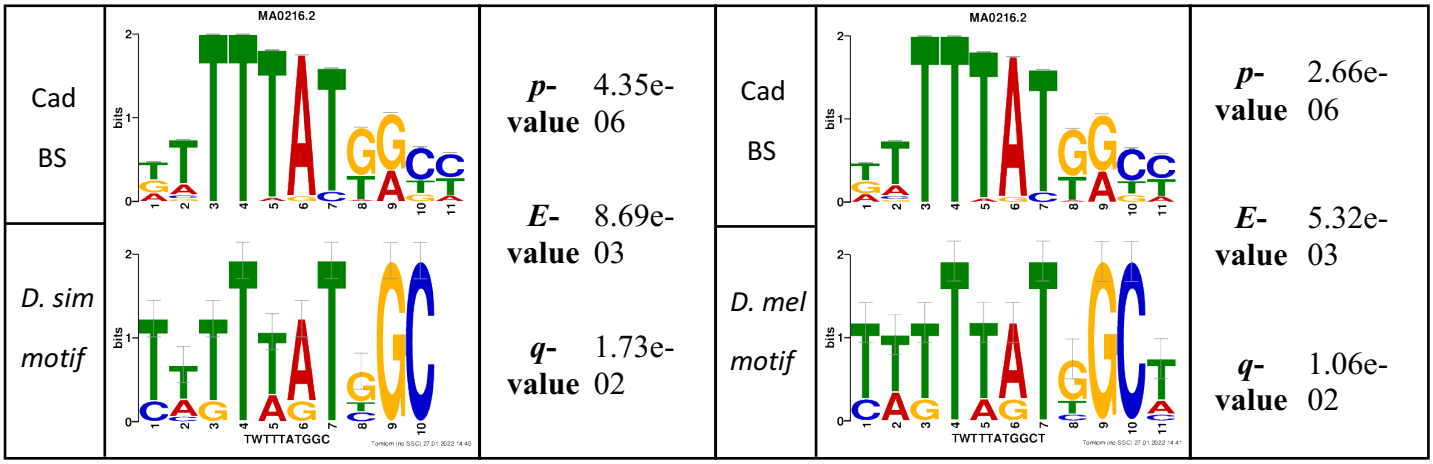


Figure 3.2. Motif comparison with TFBSs from JASPAR. The motif with the lowest E-value that was common between *D. simulans* and *D melanogaster* was aligned with the JASPAR database using Tomtom, and can be seen followed by their p-value, E-value and q-value.

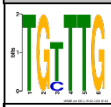
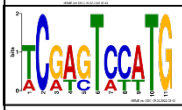
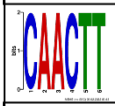
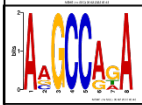
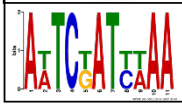
As mentioned above, the peaks that I have used to find motifs can exhibit an intrinsic cleavage bias because of the use of the transposase Tn5, and therefore I applied the bias correcting software HINT-ATAC from RGT (Li *et al.*, 2019).

This software identifies small regions that prevent the TN5 from binding in what would be accessible chromatin, leaving a footprint (Li *et al.*, 2019). Using HINT-ATAC in *D. simulans* accessible chromatin, I identified a total of thirteen TF footprints in the U4 peaks of accessible chromatin (Supplementary Table 3). I then used MEME to look for enriched motifs in these footprints, but no significant binding sites were identified (Fig. 3.3). As an alternative approach, I tested these footprints using STREME (Sensitive, Thorough, Rapid, Enriched Motif Elicitation), which is a more sensitive motif discovery tool that can use the input sequence shuffled as a background model as a control instead of using a background model like MEME (Bailey, 2021).

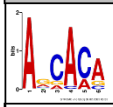
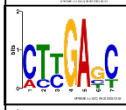
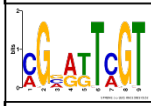
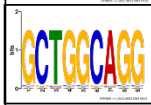
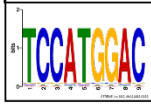
With this approach I obtained five more motifs in the TF footprints from U4 peaks of *D. simulans,* three of which were significantly enriched compared to the control sequence (Fig. 3.3 B).

I then searched for corresponding binding sites in the JASPAR database but only one motif aligned with a TFBSs sequence. Interestingly this motif is predicted to correspond to the binding site sequence for Defective proventriculus (Dve), a transcription factor that is detected in the genitalia RNA-seq data at around 2 TPM (Hagen *et al.*, 2020), and that is involved in midgut and ocelli specification, and wing patterning in *Drosophila* (Nakagoshi *et al.*, 1998; Kölzer *et al.*, 2003; Nakagawa *et al.*, 2011; Yorimitsu, Kiritooshi and Nakagoshi, 2011).

A

	MEME		
	Motif	E-value	Sites
1		3.0E+02	4
2		1.0E+03	4
3		2.4E+03	3
4		1.7E+03	5
5		5.0E+03	2

B

	STREME		
	Motif	E-value	Sites
1		2.6E-03	7
2		2.0E-02	5
3		4.80E-02	4
4		2.4-E01	2
5		2.4E-01	2

C

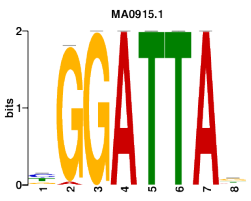
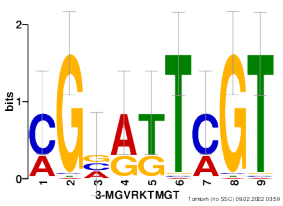
Dve BS		***p*-value** 4.77e-03
STREME Motif		***E*-value** 9.54e+00 ***q*-value** 1.00e+00

Figure 3.3. Motifs found using footprints. (A) Motifs found using MEME with the footprints obtained from HINT-ATAC. (B) Motifs found with STREME using footprints. Only 4 motifs from STREME were significantly found enriched in the footprints. (C) TFBS found to be aligned with Dve binding site using Tomtom.

HINT-ATAC can also be used to predict TFBSs based on the footprints obtained. Using this approach to identify motifs in the footprints found in U4 (Supplementary Table 3) and comparing them with the JASPAR database (Fornes *et al.*, 2019), seven new TFBSs were predicted (Supplementary Table 4) for Mirror (Mirr), Sloppy paired 1(Slp1), Fork head (Fkh), Ladybird early (Lbe), Chorion factor 2 (Cf2), Caupolican (Caup) and Araucan (ara) (Supplementary Table 4). I filtered these TFs encoding genes using RNA-seq data from the developing male genital (Hagen *et al.*, 2020) and only *lbe* was not expressed. The terminalia atlas of transcription factor includes *ara* but it does not show any specific pattern (Vincent *et al.*, 2019).

From the candidates TFs I identified using the various approaches described above, I chose to focus on Cad because its motif was found within the U4 peaks and it has been reported to be expressed in a overlapping pattern to U4 around in the anal plates (Vincent *et al.*, 2019).

3.3.2 *cad* RNAi

It has previously been shown that *cad* is necessary for the correct development of the anal plates, and that the absence of *cad* in clones of cells of the anal segment A10 transforms them in to cells of the genital segment A8 (Mlodzik and Gehring, 1987; Moreno and Morata, 1999). However, the general loss of *cad* on the anal plates and other genital structures was not previously tested. Therefore to investigate if *cad* may regulate *trn* via U4, I first carried out an RNAi experiment to see how knocking down

cad impacts genital development using the NP6333 and U4-GAL4 lines as drivers (Stieper *et al.*, 2008; Chatterjee *et al.*, 2011; Tanaka *et al.*, 2015; Hagen *et al.*, 2019).

3.3.2.1 NP6333>*cad*-RNAi Knockdown

cad knockdown using NP6333 resulted in complete loss of the anal plates (Fig. 3.4 A). It also resulted in the death of the flies at three or four days after their eclosion (data not shown).

I found that other genital structures were also affected by *cad* knockdown using the NP6333 driver, for example the surstylus was apparently duplicated in 2 cases out of 24 (Fig. 3.4 D, E).

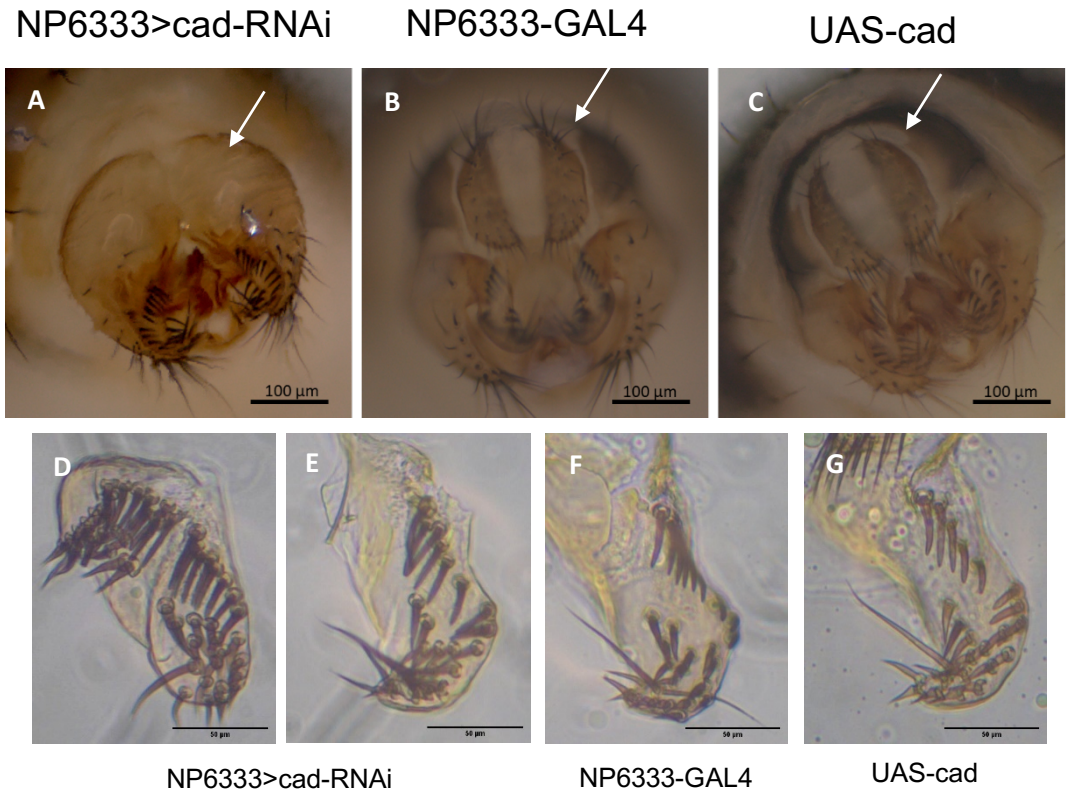


Figure 3.4. Whole terminalia and surstyli from *cad* RNAi knockdown. (A- C) Whole terminalia of the lines used for the *cad* RNAi knockdown with the NP6333 driver. Anal plates are absent in the cad knockdown (A) while present in both controls (B, C). (D-G) Surstyli dissected from the genitalia of the lines used for the *cad* knockdown. The surstyli were duplicated in 2/24 flies after knockdown (E). The surstyli from control flies had a normal shape (F, G). Arrows point at the anal plates.

I then measured the area of the distal part of the surstyli (Fig. 3.5 D), posterior lobe area (Tanaka *et al.*, 2015; Hagen *et al.*, 2019) as well as the T2 tibia length to account for body size. However, I did not find a significant correlation between tibia length and surstyli or posterior lobes size (Supplementary Fig. 3.3 A, B). This is consistent with findings that the genitalia are hypoallometric (Coyne, Rux and David, 1990; Liu, Mercer, *et al.*, 1996; Eberhard, 2009b; Shingleton *et al.*, 2009; Masly *et al.*, 2011) and therefore I proceeded without correcting for body size like in previous studies of genital structures (Masly *et al.*, 2011; Tanaka *et al.*, 2015; Hagen *et al.*, 2019).

For the knockdown of *cad* using the NP6333;UAS-Dicer driver I found that the area of the surstylus was significantly smaller when *cad* was knocked down compared to controls (Fig. 3.5 A). However, the number of bristles found on the surstyli was not significantly different between the controls UAS-*cad*-RNAi and the *cad* knockdown (p-value> 0.05) (Fig. 3.5 B). These results suggest that *cad* knockdown reduces the area of the surstyli but does not affect the number of bristles, although there was a trend towards fewer bristles (Fig. 3.5 B).

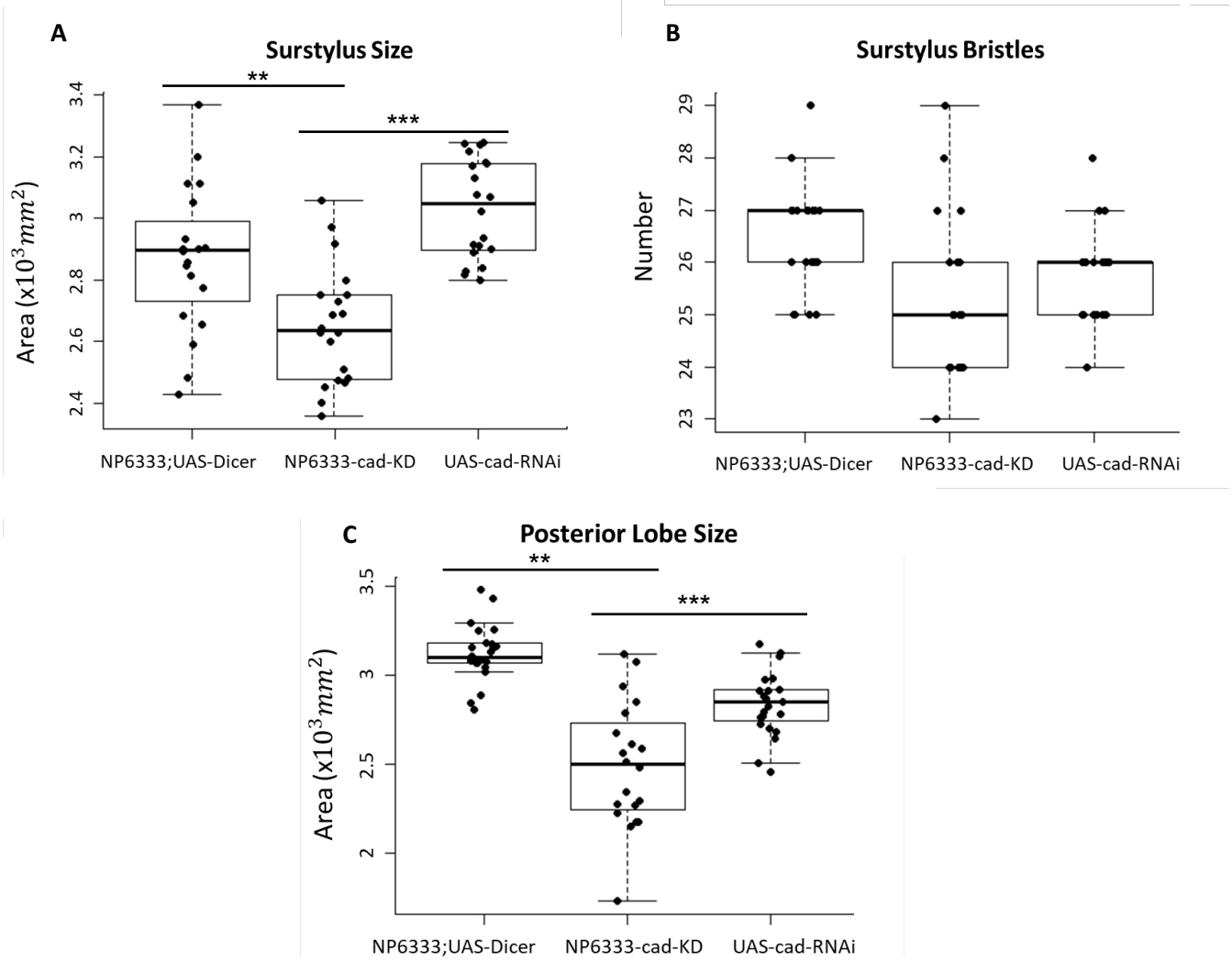


Figure 3.5. Box and whiskers plots of the *cad* RNAi knockdown effects. (A-C) Plots of surstyli area (A), surstyli bristle number (B) and posterior lobes area (C). (A) The surstyli were significantly smaller between *cad* knockdown and both controls (NP6333;UAS-Dicer p-value < 0.01 and UAS-*cad*-RNAi p-value < 0.0001). (B) Posterior lobe area was significantly smaller between *cad* knockdown and both controls < 0.001. (C) Number of surstyli bristles were not significantly different between *cad* knockdown and the UAS-*cad*-RNAi control (p-value > 0.05). Significance codes: 0.001 '***' 0.01 '**' 0.05 '*'. N ≥ 18

cad knockdown also resulted in significantly smaller posterior lobes (Fig. 3.5 C). This suggests *cad* knockdown loss reduces the area of the posterior lobes as well as the surstyli.

It is also worth noting that this *cad* knockdown using this GAL4 driver was variable in terms of consistency, with the observation of duplicated surstyli (Fig. 3.4 D), a posterior lobe with a different shape (Supplementary Fig. 3.2 A) and occasionally what appears

to be a protrusion where the anal plate would normally form (Supplementary data Fig. 3.2 C).

3.3.2.2 *cad*-RNAi Knockdown using U4

I then carried out *cad* RNAi knockdown using the more spatially restricted U4-GAL4 driver to compare with the NP6333 driver results above. In this case, I corrected the area using the tibia length (see Methods) because I found that there was a correlation between tibia length and the areas of anal plates, surstyli and posterior lobes (Supplementary Fig. 3.3 C, D).

Using the U4 driver, I found that the *cad* knockdown did not have a significant effect on anal plate size although there was a trend towards larger anal plates (Fig. 3.6 A). There were significantly more bristles on the anal plates when *cad* was knocked down compared with the controls (Fig. 3.6 B). These results suggest that this *cad* knockdown augment the number of bristles on the anal plate and may increase their area.

In this experiment, the surstyli were not significantly different between the *cad* knockdown and the UAS-*cad*-RNAi control (p-value > 0.05). While this suggests that *cad* knockdown with the U4 driver did not have a significant effect on surstyli area (Fig. 3.6 C), it is worth noting the area of the surstyli of knockdown flies was significantly different from both controls when I did not correct for tibia size (Fig. 3.7 B). However, the number of surstylus bristles was not significantly different between the controls and the *cad* knockdown (Fig. 3.6 D).

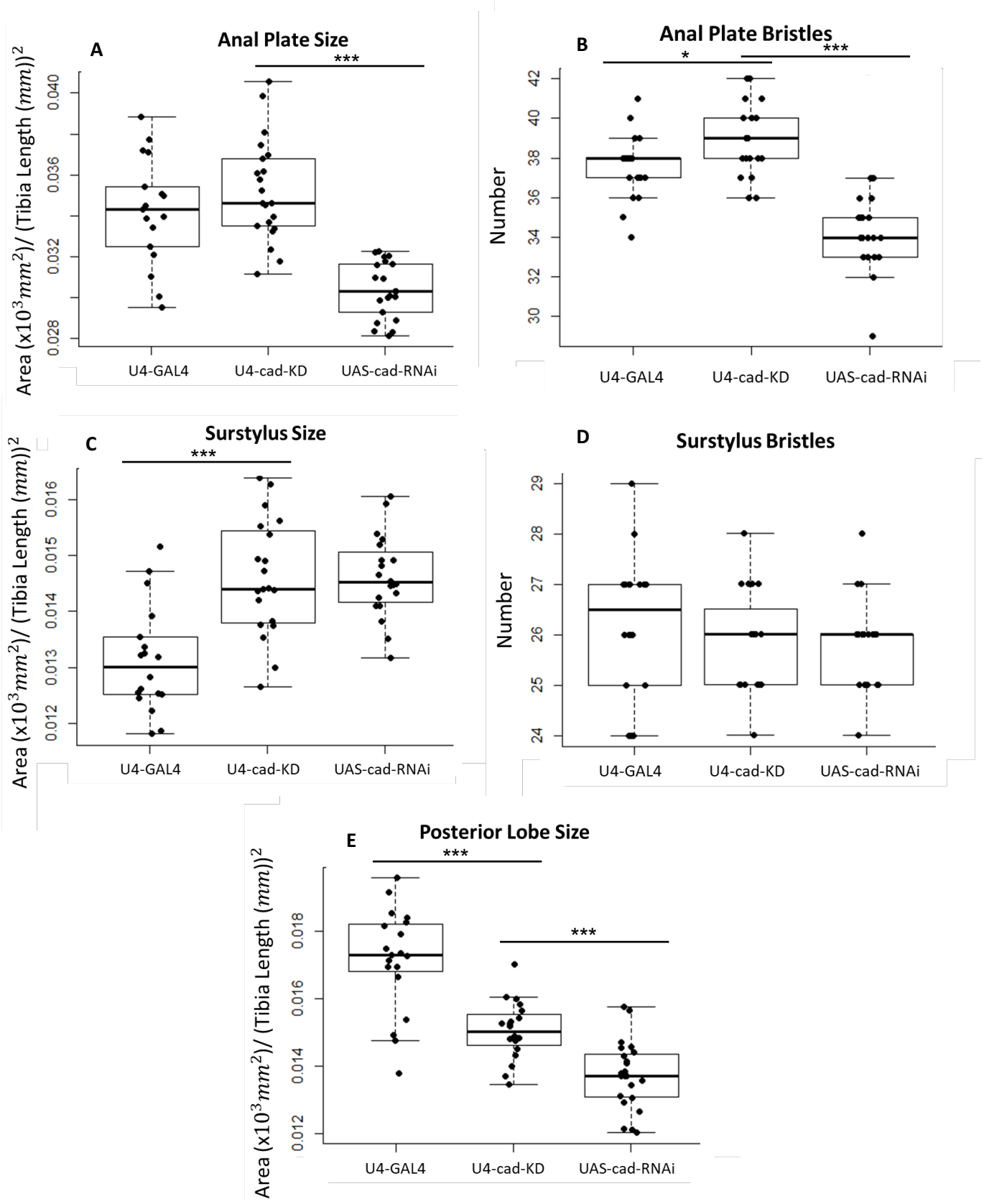


Figure 3.6. Box and whiskers plots of the *cad* knockdown using U4-GAL4. (A, B) Plots of corrected anal plate area (A) and bristle number (B). (C, D) plots of corrected surstyli area (C) and surstyli number. (E) Plot of corrected posterior lobe area. (A) Anal plate area was not significantly different between U4-GAL4 control and *cad* knockdown (p-value > 0.05). (C) Surstyli area was not significantly different between

the *cad* knockdown and the control UAS-*cad*-RNAi (p-value > 0.05). (E) Posterior lobe area was significantly smaller between *cad* knockdown and U4-GAL4 control (p-value < 0.0001) and significantly bigger than the control UAS-*cad*-RNAi (p-value < 0.001). (B) Anal plates had significantly more bristles between the *cad* knockdown and both controls (U4-GAL4 control p-value < 0.05 and UAS-*cad*-RNAi p-value < 0.0001). (D) Number of surstyli bristle was not significantly different between *cad* knockdown and controls (p-value > 0.05). Significance codes: 0.001 '***' 0.01 '**' 0.05 '*'. N ≥ 18

The posterior lobe area of the *cad* knockdown flies was intermediate in size between the controls when considering tibia length or not, and therefore this treatment did not appear to affect the posterior lobe size (Fig. 3.6 E). Taken together the results of *cad* knockdown using the U4-GAL4 driver suggest that *cad* knockdown affects anal plates morphology but had no impact on the surstyli or posterior lobes. These results appear to be inconsistent with the results obtained when using the NP6333 driver.

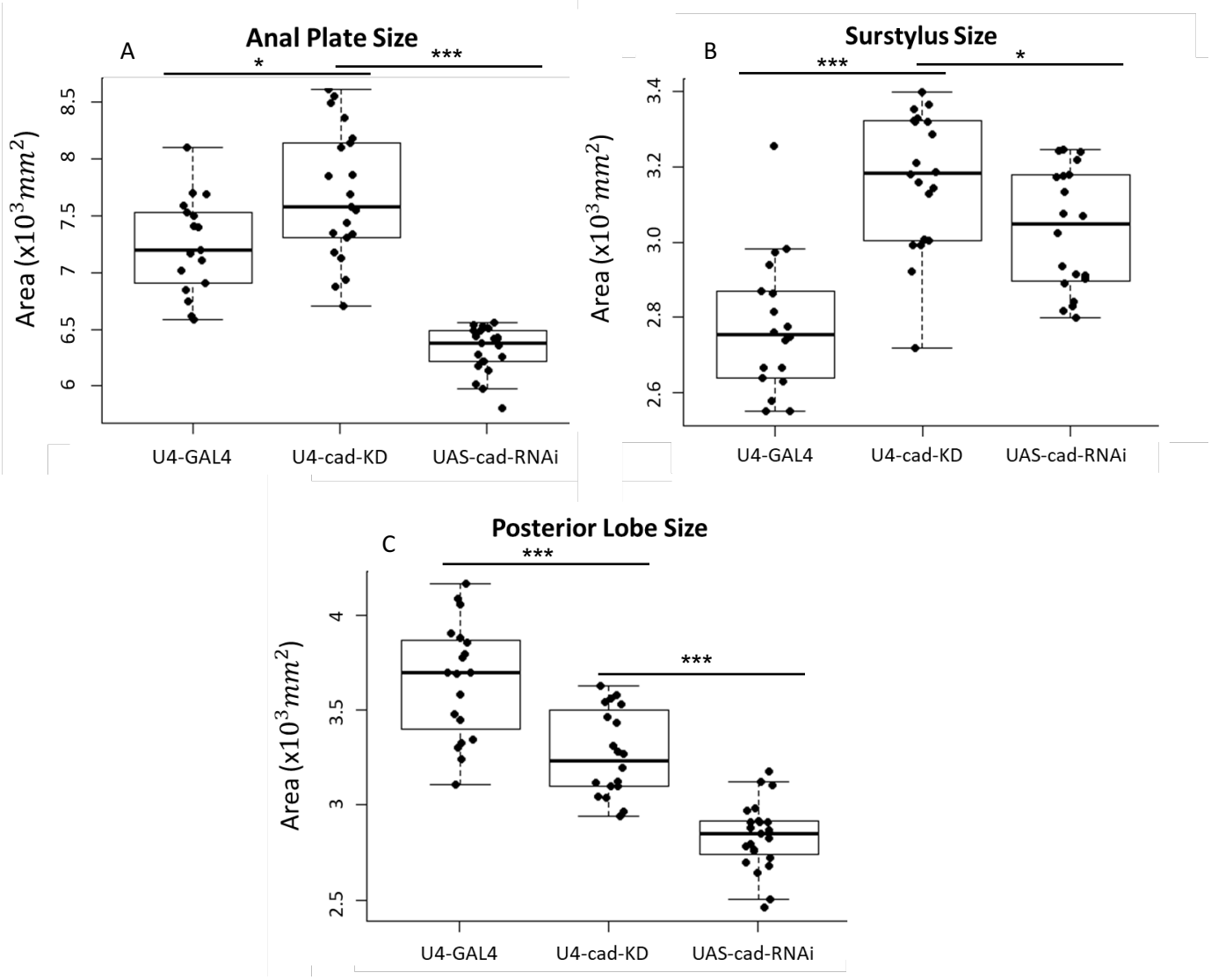


Figure 3.7. Box and whiskers plots of the *cad* knockdown using U4-GAL4 (uncorrected). (A-C) Plots of uncorrected anal plate area (A), surstyli area (B) and posterior lobe area (C). (A) Uncorrected anal plates area was significantly bigger between *cad* knockdown and both controls (U4-GAL4 p-value < 0.05 and UAS-*cad*-RNAi p-value < 0.0001). (B) Uncorrected surstyli area was significantly bigger between *cad* knockdown and both controls (U4-GAL4 p-value < 0.0001 and UAS-*cad*-RNAi p-value <0.05). (C) Uncorrected posterior lobe area was significantly smaller between *cad* knockdown and U4-GAL4 control (p-value < 0.0001) and significantly bigger than the control UAS-*cad*-RNAi (p-value < 0.0001). Significance codes: 0.001 '***' 0.01 '**' 0.05 '*'. N ≥ 18

3.4 Discussion

3.4.1 Binding sites in U4

In this chapter I used chromatin accessibility profiles in combination with various bioinformatic tools/approaches to find several possible TFs that may regulate the U4 sequence.

Using MEME I found no motifs that were more significantly enriched than in the background model. However, this analysis revealed the presence of motifs recognised by Cad (Fig. 3.2) (Bailey *et al.*, 2009). *cad* is a homeobox-containing gene that is expressed in the pupal terminalia at the same timepoints as U4, and its expression pattern appears to overlap with U4 activity at the intersection between the anal plates and the rest of the male genital (Vincent *et al.*, 2019). Furthermore, *cad* is also known for defining the segment A10 and therefore the anal plates and delimiting the boundaries between the segment A10 and A9. For these reasons, despite the motif not being more significantly enriched in the sequence than the model background, I focused on this gene as a potential regulator of *trn.*

As an alternative approach, I corrected for the inherent cleavage bias of the Tn5 for the ATAC-seq data using HINT-ATAC (Li *et al.*, 2019). I identified thirteen footprints and using STREME I found four significantly enriched motifs within them (Li *et al.*, 2019; Bailey, 2021). However, using JASPAR only one motif was found to be related to a known TFBS, in this case Dve (Fig. 3.3). This issue could be due to the limitations of the JASPAR database, because it does not include binding information for every *Drosophila* TF and we may be missing those that bind to other motifs in U4 (Fornes *et al.*, 2019). In any case, these motifs can also be the focus of future experiments to test whether they are indeed binding sites required to regulate *trn*, for example by mutating

them and examining the effect on U4 activity. In the case of *dve*, the expression in the genitalia has not been previously studied, and it would be interesting to investigate further its relation to the genitalia and *trn.*

Using the HINT-ATAC matching tool also identified seven possible TFs that might be binding within the footprints. *lbe* is not expressed in the genitalia according to our RNA-seq data. *mirr* was tested by RNAi knockdown but this did not produce any differences in surstyli, posterior lobe or anal plate size compared to controls (Ridgway *et al.,* in preparation). *ara* is expressed in the developing male genital and although the pattern is not very specific (Vincent *et al.*, 2019) it would be interesting to test its function in the genitalia using RNAi.

The remaining motifs found by STREME could also be tested in the future by removing them to see if the genitalia or *trn* are affected.

3.4.2 *cad* RNAi perturbs genital development

It has previously been described that *cad* defines segment A10, which forms the anal plates of the adult terminalia (Moreno and Morata, 1999). Moreover, clonal null mutants of *cad* result in the loss of tissue identity and transformation of cells to A9 (Moreno and Morata, 1999). The results from the RNAi experiments that I carried out are consistent with these findings. RNAi knockdown of *cad* throughout the developing genitalia using the NP6333 driver resulted in the loss of the anal plates and occasionally duplication of the surstyli (Fig. 3.4).

cad knockdown also reduced the size of the surstyli and posterior lobes (Fig. 3.5). These phenotypes could be due to the regulation of these tissues by *cad* or as an indirect effect of the loss of anal plates perturbing genitalia development more generally. However, it is more likely the latter scenario because *cad* is only expressed

in A10 and the anal plates (Moreno and Morata, 1999; Vincent *et al.*, 2019).

Unfortunately, in the time frame of this thesis I was not able to test the expression of *trn* in the *cad* RNAi knockdown flies. However, the possible differences in *trn* expression that could be found between normal male genitalia and the *cad* RNAi knockdown male genitalia, would be hard to assess due to the massive lose of tissue in these flies.

I also carried out the *cad* RNAi knockdown using U4 as a driver instead of NP6333. In this case, when the area of anal plates, surstyli and posterior lobes were corrected with for tibia length there was no significant differences with the controls (Fig. 3.6 A, C and D). However, I saw a trend towards larger anal plates and surstyli that without the correction for tibia length was significant (Fig. 3.7 A, B).

Curiously, the number of bristles on the anal plates was significantly higher in the knockdown compared with the controls (Fig. 3.6 B) and although, the *cad* knockdown using the U4-GAL4 driver did not affect the area of the genital structures, the trend with this driver is larger anal plates (Fig. 3.6 A), which would be consistent with the increase in number of bristles on the anal plates. This could indicate that *cad* is somehow delimiting *trn* around the anal plates, and that the absence of *cad* at those cells is causing the increase in number of bristles by *trn*.

With the trend seen towards larger anal plates and larger surstyli, but no effect in the posterior lobes or the bristle number on surstyli, there are different possibilities that could explain these results. First, even when knocking down *cad* in the cells with U4 activity located at the intersection between the anal plates and the surstyli and epandrial ventral lobes, this may not strong be enough to perturb their development. Alternatively, a different dissection and measuring method might be needed to assay

this. As U4 drives expression mainly surrounding the anal plates and their boundaries, the size of the base region of the surstyli that connects with the anal plates was not measured directly. This region is in most cases torn in order to separate the surstyli from the anal plates. Therefore, this region of the surstyli was not taken into account when measuring the surstyli area and the effect of *cad* knockdown using the U4 mediated RNAi may be more subtle than broader knockdown using NP6333 (Fig. 3.5 D) of the male genital.

Chapter 4. Evolution of the U4 *trn* enhancer

4.1.1 Chromatin accessibility

Despite extensive phenotypic differences and complexity among animals, for example between sea anemones and vertebrates whose lineages diverged soon after from the common ancestor of all animals, their genetic toolkit is very similar (Ryan *et al.*, 2006; Putnam *et al.*, 2007; Martindale and Hejnol, 2009; Maeso *et al.*, 2013; Simakov *et al.*, 2013). This indicates that the common ancestor of all animals probably already had most of the genes that are present in most subsequent animals (Putnam *et al.*, 2007; Degnan *et al.*, 2009; Harcet *et al.*, 2010; Schwaiger *et al.*, 2014).

Therefore, in order to differentially regulate the fate of cells during development to give rise to diversity, it is thought that there is a differential regulation of these 'toolkit' genes in time and space (Carroll, 2008; Graf and Enver, 2009). A given gene requires many TFs and several enhancers to regulate its precise expression during development (Perry, Boettiger and Levine, 2011; Furlong and Levine, 2018). For example, in the case of gap genes in the *Drosophila* embryo they often require multiple enhancers for even one aspect of their expression pattern (Perry, Boettiger and Levine, 2011; Lagha, Bothma and Levine, 2012; El-Sherif and Levine, 2016).

Moreover, as there are different enhancers controlling different aspects of a given gene, this modularity can readily facilitate the evolution of that gene expression without having pleiotropic effects (Carroll, 2008).

Chromatin accessibility plays a fundamental role in enhancer function and gene expression (Roscito *et al.*, 2018; Bozek *et al.*, 2019). It has been shown that there are differences in chromatin accessibility between different developing tissues and at

different times that correspond to differential enhancer activity and gene expression (Thomas *et al.*, 2011; Roscito *et al.*, 2018; Bozek *et al.*, 2019). These differences in chromatin accessibility can be regulated by different proteins like Zelda, which is a pioneer factor that modifies the chromatin allowing other TFs to bind (Harrison *et al.*, 2011; Foo *et al.*, 2014). Thus, differences in chromatin accessibility contribute to differential gene expression depending on the developmental context due to the different binding sites that will be accessible for TFs to bind in a specific tissue at a specific time. Chromatin accessibility is, therefore one of the most important aspects to consider when we are trying to understand the regulatory landscape of a given gene, as it will provide information of regulatory landscape in the specific context that we are studying for that gene. Analysis of chromatin accessibility also allows us to make predictions about the regulation of the gene by helping to identify potential enhancers and promoters and even where these elements have evolved in function between species (Gao *et al.*, 2018; Peng *et al.*, 2019).

To study the regulatory landscape of a gene, we need to determine which regions of the DNA we are going to investigate. Enhancers may be located far from the gene, for example over 70 kb from their target gene in *Drosophila* (Frankel *et al.*, 2010) and even 1 Mb in mammals (Lettice *et al.*, 2003). However enhancers are frequently located within a short distance from the target gene, which in *Drosophila* is usually around 10 kb (Furlong and Levine, 2018).

4.1.2 Enhancer evolution

As mentioned above genes are generally conserved throughout the animal kingdom, and they form gene regulatory networks that control development (Oliveri and Davinson, 2007; Hinman, Yankura and McCauley, 2009; Maeso *et al.*, 2013; Rebeiz

and Tsiantis, 2017). However, with small changes in regulation genes can evolve new expression patterns in different cells and tissues or even be co-opted to perform similar or new functions, which can give rise to modified or new phenotypes without having large pleiotropic effects (Wray, 2007; Kittelmann *et al.*, 2018).

Enhancers are generally not as conserved as coding sequences because they are composed of TF binding sites and which are usually very short and degenerate and can turnover rapidly (Ludwig *et al.*, 2000; McGregor *et al.*, 2001; Maeso *et al.*, 2013; Burgess and Freeling, 2014). In some cases even when the sequence is not the same the enhancer itself can be conserved (Ludwig *et al.*, 2000; McGregor *et al.*, 2001; Maeso *et al.*, 2013; Buffry, Mendes and McGregor, 2016). However, there have been several studies that have identified conserved non-coding sequences (CNSs) that have been maintained in very distantly related animals such a Nematostella and zebrafish (Panne, Maniatis and Harrison, 2007; Strähle and Rastegar, 2008; Maeso *et al.*, 2013).

Differences in enhancers are thought to be one of the main mechanisms for phenotypic evolution (Wray, 2007; Carroll, 2008). For example, changes in enhancers between closely related *Drosophila* species have been identified that underlie differences in body and wing pigmentation (Gompel *et al.*, 2005; Jeong, Rokas and Carroll, 2006), differences in larval trichome (McGregor *et al.*, 2007; Frankel *et al.*, 2011) and even in copulatory organs (Nagy *et al.*, 2018). Enhancers can also cause differences in phenotype between more distant species such as the loss of limbs in snakes (Roscito *et al.*, 2018) or the evolution of wrist in fishes (Gehrke *et al.*, 2015)

I therefore further studied the sequence and function of the *trn* U4 enhancer in this chapter to investigate if it could contribute to differences seen in *trn* expression during

genital development between *D. mauritiana* and *D. simulans* (Hagen *et al.*, 2019).

4.2 Methods

4.2.1 Cloning

Fly lines used to test enhancer reporters were obtained using ΦC31 integrase transformation. U4D1 was amplified from *D. mauritiana* (strain D1) gDNA and cloned initially into pENTRY/D-TOPO vector. Then, using LR gateway cloning the regions was inserted into the pBPGUw plasmid (pBPGUw was a gift from Gerald Rubin (Addgene plasmid # 17575; http://n2t.net/addgene:17575 ; RRID:Addgene_17575)) (Pfeiffer *et al.*, 2008). Embryos from the stock y[1] w[67c23]; P{y[+t7.7]=CaryP}attP2 (Bloomington stock 8622) were injected with plasmid pBPGUw containing the selected U4D1 region and the ΦC31 integrase helper plasmid pBS130 (pBS130 was a gift from Tom Clandinin (Addgene plasmid # 26290; http://n2t.net/addgene:26290; RRID:Addgene_26290)) (Gohl *et al.*, 2011). Injected F0s were crossed to a balancer line *w*;;TM6B/TM3 and the F1s screened for transgenic individuals marked by the red colour of their eyes. This approach allowed me to compare the activity of the *D. melanogaster* and *D. mauritiana* U4 enhancers when placed at the same genomic location in the former species. Unfortunately, although the *D. simulans* U4 enhancer has already been cloned, it has not been possible to test that enhancer in the time frame of this thesis.

4.2.2 Fly stocks and genetics

All fly stocks were kept at 25 °C on standard medium yeast extract-sucrose. w^+;;U4D1-GAL4/TM6B was crossed to *y, w*;;UAS-GFP for the whole pupae imaging experiments. w^+;;U4-GAL4/TM6B and w^+;;U4D1-GAL4/TM6B were crossed to w^+;Ecad::GFP/CyO;UAS-Redstinger/TM6B for the male genital specific imaging.

4.2.3 Microscope imaging

Prepupae from the crosses were collected and aged at 25 °C until 24 hAPF and 46 hAPF. These time points were selected because they precede important stages of the genitalia development when the surstyli and the epandrial posterior lobe are forming and separating, thus genes involved in their regulation are thought to be active (Glassford *et al.*, 2015). For whole body mounting the pupal case was dissected and then the pupa was mounted on a glass slide in 80% glycerol. Images were taken using a Zeiss Axiozoom stereomicroscope using a GFP filter with the 1x objective at 50x magnification. For pupal genitalia dissections, the posterior part of the pupa was cut off and the genitalia was dissected in PBS 1x and mounted on a glass slide with immunostaining hydro mount liquid. Images were taken using a Zeiss LSM 800 confocal microscope using the 25x objective, 0.2% laser power for DsRed and 1% laser power for GFP.

4.2.4 Accessible chromatin analysis

Using ATAC-seq data generated by Hagen *et al,* unpublished, from developing male genitalia between 30 and 50 hAPF, I converted pre-assembled bam files in to bedgraph files using bamtools. Then I used the R package ‘sushi’ to visualise the chromatin profile from the coordinates corresponding to the region between nearest upstream and downstream genes flanking *trn* in both, *D. simulans* and *D. mauritiana* (Phanstiel, 2021). The *trn* locus ATAC-seq chromatin profiles from eye-antennal imaginal discs were obtained from (Torres-Oliva *et al.*, 2021).

4.2.5 U4 sequence analysis

The UCSC genome viewer was used with the coordinates chr3L:13,114,271-13,118,077 to survey *trn* locus conservation and chr3L:13,100,690-13,103,689 to

survey U4 (Kent *et al.*, 2002). An interspecific polymorphisms analysis was performed using the Pool-seq data from http://www.popoolation.at/pgt/ using the genomes w^{501}:r2.2 and Mauritiana:SRA000000 as reference genomes with the coordinates 3L:12,785,380-12,788,342 and 3L:12,546,409-12,549,370 for U4. A minimum count of two strains polymorphic was used to perform the polymorphisms analysis. Regions containing the fixed differences were blasted using NCBI nucleotide BLAST and coordinates from mau12 line are used to locate fixed differences (Altschul *et al.*, 1990). Trn protein alignment was performed using Clustal Omega at the UniProt webpage (Sievers *et al.*, 2011; The Uniprot Consortium, 2021)

4.3 Results

4.3.1 The *trn* chromatin accessibility profile

Previously Hagen *et al.*, (unpublished) generated ATAC-seq data sets for the developing male genitalia of *D. mauritiana* and *D. simulans* at a time window between 30 and 50 hAPF (hours after pupa formation). To generate a chromatin accessibility profile of *trn* in developing male genitalia from these species at those time points, I analysed this previously assembled ATAC-seq data, using 'BEDtools' and 'Sushi' (see Methods). The resulting chromatin accessibility profiles I generated show several peaks present in the *trn* locus with the largest corresponding to the position of the TSS (Fig. 4.1). These peaks of chromatin accessibility may represent regions with enhancer activity. When compared between the two species, I identified four peaks that differed in height, which means that the region had more reads in that region and is therefore more accessible in one species than in the other (Fig 4.1) (Buenrostro *et al.*, 2015; Yan *et al.*, 2020). Peaks 1, 3 and 4 are found within the candidate regions U1, D2 and D3 respectively, which I tested in Chapter 2, but these sequences from *D. melanogaster* did not appear to have enhancer activity in the developing male genitalia

(Figs. 2.3, 2.7 and 2.8). Differential peak number 2, which is higher in *D. simulans* than in *D. mauritiana* in region U4 (Fig. 4.1), which I found to have enhancer activity in the male genital at both time points in *D. melanogaster* (Fig. 2.5). Note, however, that although this peak is shared between U3 and U4 (Fig.4.1), U3 did not show enhancer activity in the male genital (Fig. 2.4).

However, high peaks in chromatin accessibility profiles generated from a complex tissue such as the developing genitalia do not always necessarily correlate with enhancers because these peaks may have binding sites for TFs that are not available at the developmental stage assayed or there may be repressors for that enhancer at that specific window of time in that tissue (Shlyueva, Stampfel and Stark, 2014). The opposite may also be true, as this data was collected from all the cells from the male genital, if there is a subset of cells with a different chromatin profile, this can be overlooked or masked if there are large numbers of other cells with a different profile (Xi *et al.*, 2007; Jacobs *et al.*, 2018). Nevertheless, the chromatin accessibility profile represents a good guide to help explore regions with enhancer activity and differences in accessibility in the tissues assayed (Bozek *et al.*, 2019). This is also evidenced when compared with a different tissue, as there are also differences between the chromatin profile of the same species but between the eye-antennal imaginal disc and the male genitalia (Supplementary Fig. 4).

4.3.2 U4 in *D. mauritiana*

To investigate the possible functional evolution of U4 I cloned and tested the corresponding U4 enhancer sequence from the strain D1 of *D. mauritiana.* For this I incorporated the *D. mauritiana* U4 into the same line of *D. melanogaster* that I used for *D. melanogaster* U4 and the other regions I tested and inserted this construct into

the same genomic position. Thus, any possible difference observed will be due to the enhancer itself rather than the same genomic background or position.

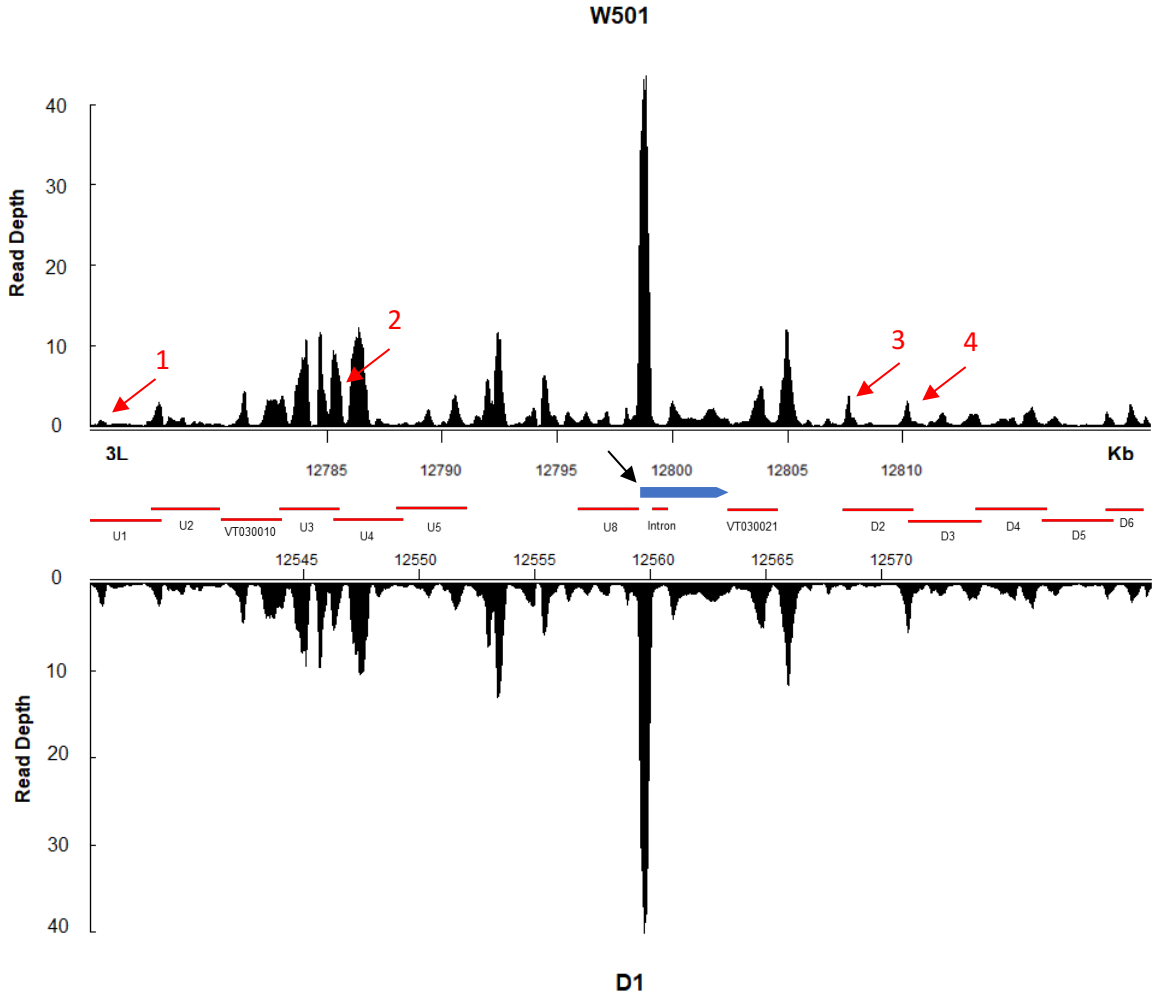


Figure 4.1. Chromatin accessibility profile around the *trn* locus of pupal male genitalia from *D. mauritiana (D1)* and *D. simulans (w501)* generated from ATAC-seq data. Profiles shown opposed one to another from both species for comparison. The *trn* CDS is shown in blue, and the regions tested for enhancer activity in *D. melanogaster* in Chapter 2 are shown in red. Red arrows show where there is a visible difference in the peak height between species, numbered 1-4. A black arrow marks where the TSS of *trn* is located.

The expression driven in the whole pupa by U4D1 is similar to that observed for *D. melanogaster* U4 at both time points and as expected, reporter expression is detected in the developing genitalia (Fig. 4.2). The expression driven by U4D1 seen at 24 hAPF is practically the same as U4 (Fig. 4.2). Even at this low magnification (1x objective

50x magnification), however, the expression in the genitalia looks different from that seen with the *D. melanogaster* enhancer, with U4D1 driving some extra domains of expression in a more anterior position with respect of the expression seen previously with the *D. melanogaster* U4 (Fig. 4.2 B). At 46 hAPF, U4D1 the have a similar expression in the most posterior part of the genitalia, although it shows additional expression at the anterior part of the genitalia (Fig. 4.2 B, D). In addition, at 46 hAPF U4D1 also drives expression in the wing as it was observed in the D. melanogaster U4 enhancer, which further suggests they have broadly similar activity (Fig. 4.2 B, D). I then looked at the male genitalia specifically at higher magnification (25x objective). I observed that U4 from *D. mauritiana* shows an expanded expression compared to the U4 enhancer from *D. melanogaster* (Fig. 4.3 A-C).

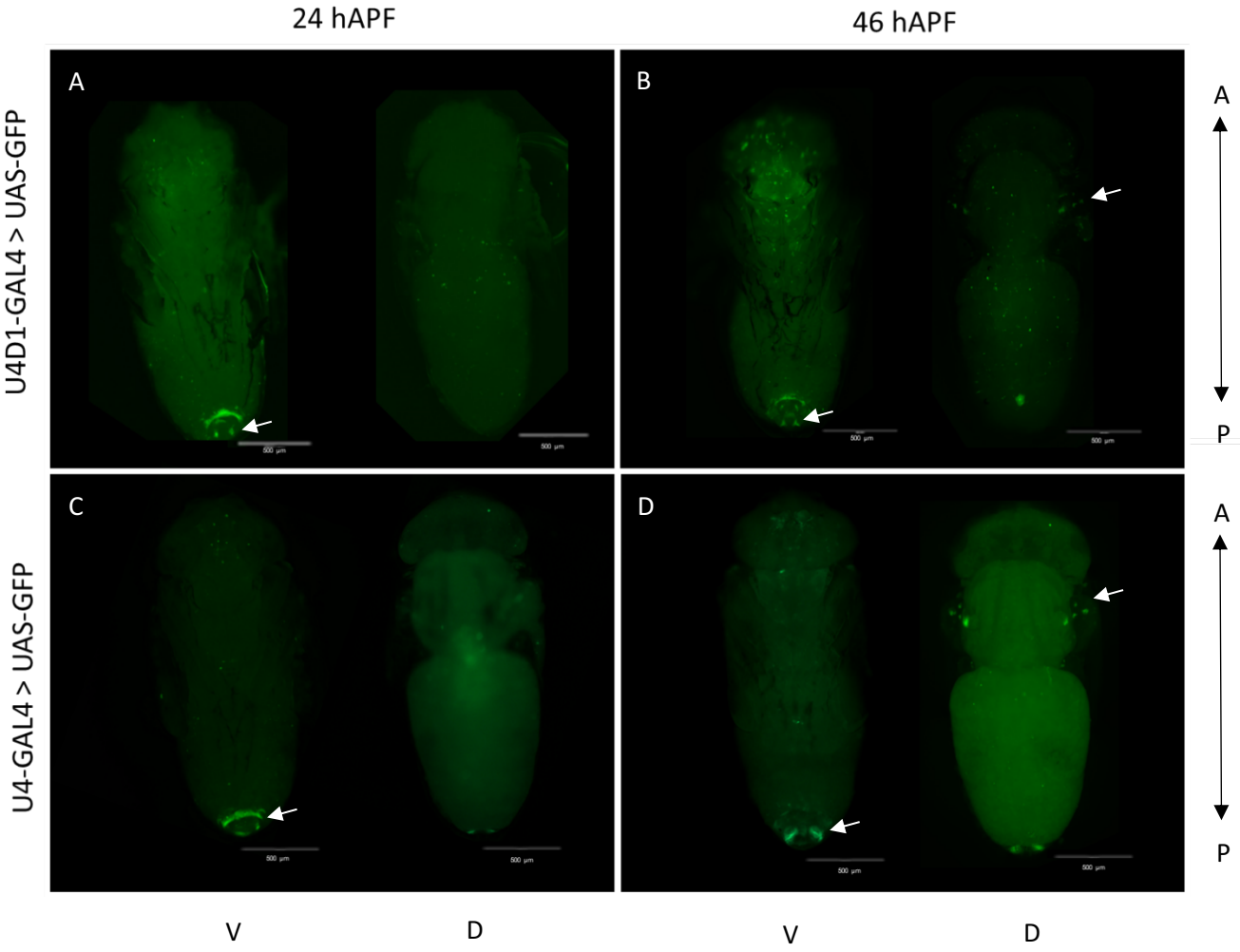


Figure 4.2. Enhancer activity of region U4 from the D1 strain of *D. mauritiana* (U4D1) compared to *D. melanogaster* U4. (A, B) Expression pattern driven by U4D1 at 24 hAPF (A) and 46 hAPF (B) in the whole pupae of males. The ventral and dorsal sides are shown on the left and right respectively. A at 24 hAPF, the GFP expression driven by U4D1 is highly similar to the U4 enhancer from *D. melanogaster*. (B) at 46 hAPF U4D1 drives in the genital as well as in the wings. (C, D) Expression pattern driven by *D. melanogaster* U4 at 24 hAPF (C) and 46 hAPF (D) in the whole pupae of males. Expression driven by U4 was detected in the genitalia at both stages, and at the wings at 46 hAPF (D) as well as weaker expression in the antenna.

U4D1 is not only active in the same places in the developing genitalia as *D. melanogaster* U4, but it drives more strongly in the aedeagal sheath in the phallus (14.3 B, C) and there is also expression in the distal part of the surstyli that was not detected in *D. melanogaster* (Fig. 14.3 B, C). Therefore, there appear to be important differences in the activity of the U4 enhancer between the *D. mauritiana* and *D. melanogaster* that may reflect the evolution of this enhancer and *trn* expression in the former species.

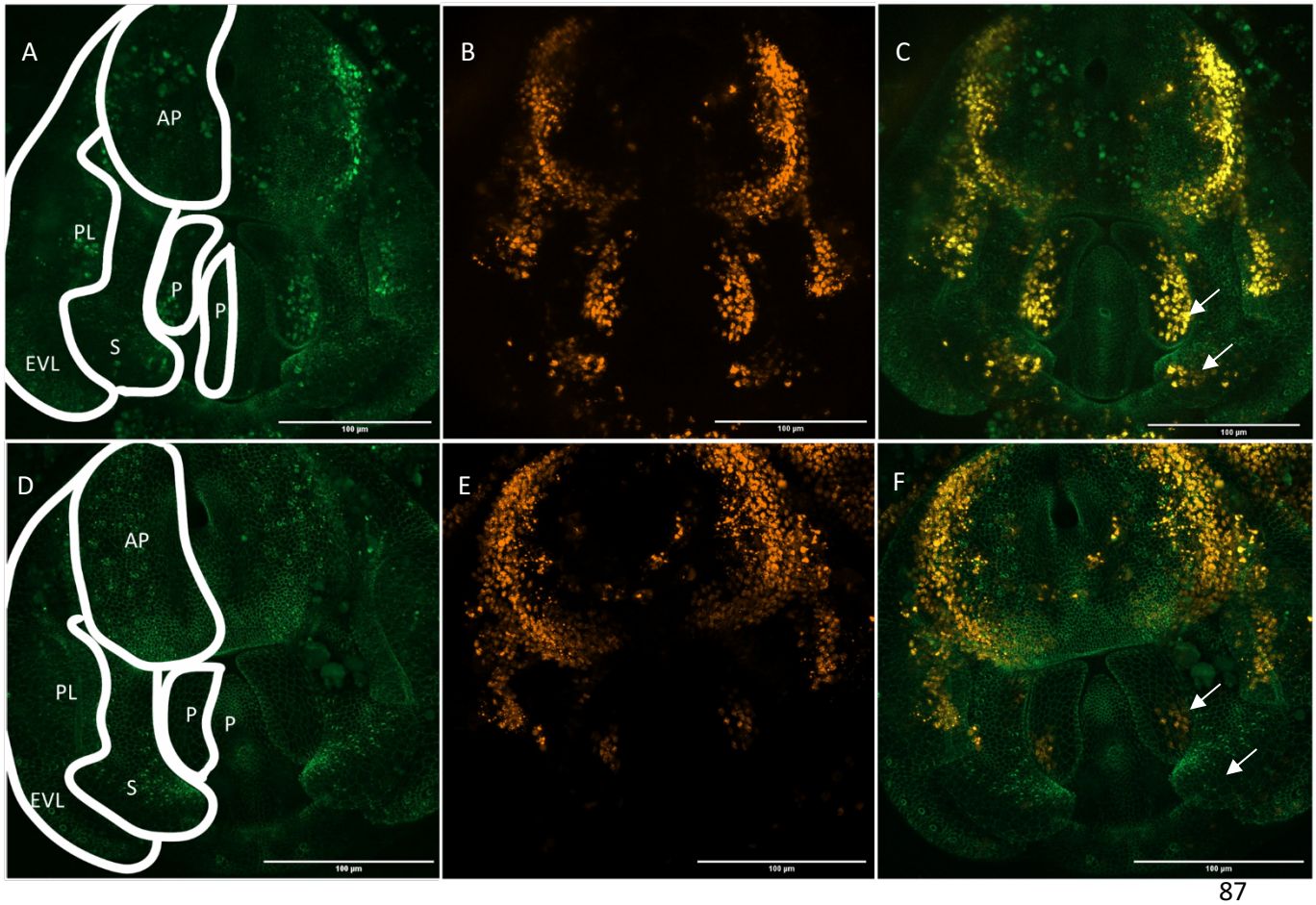


Figure 4.3. Expression driven from *D. melanogaster* U4 and *D. mauritiana* U4. A, B and C Expression driven by U4D1 at 46 hAFP in orange (B and C), and Ecad::GFP in green (A and C). Expression driven by U4 at 46 hAPF in orange (E and F), and Ecad::GFP in green (D and F). Different regions of the male genitalia are outlined in the images A and D, anal plate (AP), surstylus (P), epandrial ventral lobe (EVL), posterior lobe (PL) and phallus (P). White arrows show main differences in expression.

4.3.3 U4 sequence evolution

To further investigate the evolution of the U4 enhancer, I looked at the conservation of U4 sequence specifically among *Drosophila* species using the UCSC genome viewer (Kent *et al.*, 2002).

The nucleotide sequence of the *trn* coding region is well conserved in all *Drosophila* species (Fig. 4.4 A, B) as reflecting conservation of the amino acid sequence that goes from an 80% of similar identity in some species to 99% of similar identity in others (Supplementary Fig. 4.1). In U4 there are some regions that are more conserved than others. Interestingly the regions that are more conserved overlap with the peaks of accessible chromatin found in the ATAC-seq further indicating that they are likely to be functional (Fig. 4.4 C). In one of the peaks there is also a long non-coding RNA, which could also help that DNA region to be more conserved (Fig. 4.4 B). Curiously, there is also another region within U4 (to the right at about 12785.5) that also seems to be conserved but the ATAC-seq did not show any significant peak in that position (Fig. 4.4 B).

To better understand the conservation of the U4 sequence I analysed single nucleotide polymorphisms (SNPs) between *D. simulans* and *D. mauritiana* using Pool-seq data from different strains of *D. simulans* and *D. mauritiana* (see Methods) (Pandey *et al.*, 2011).

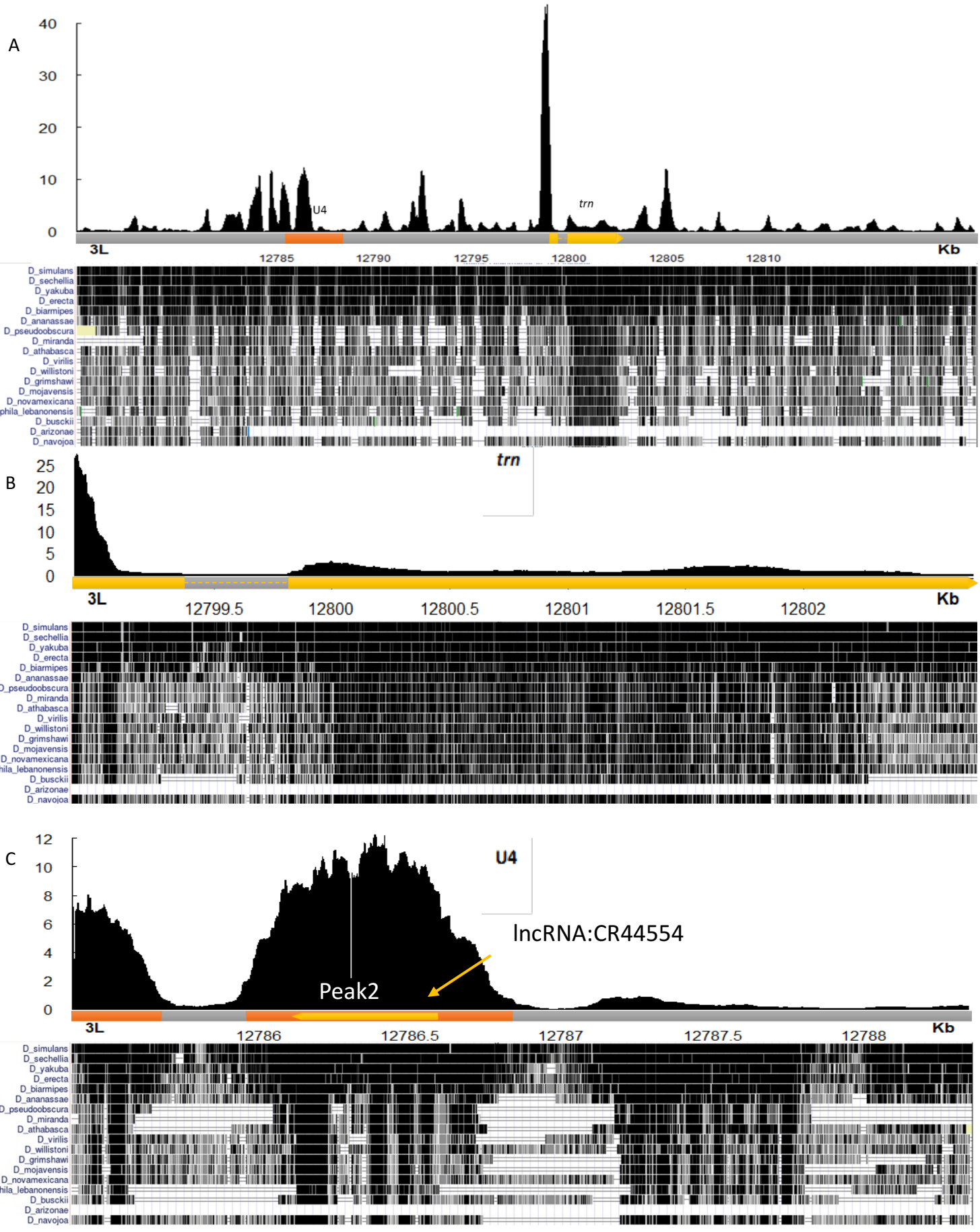

A
40
30
20
10
0
U4
trn
3L
12785
12790
12795
12800
12805
12810
Kb
D_simulans
D_sechellia
D_yakuba
D_erecta
D_biarmipes
D_ananassae
D_pseudoobscura
D_miranda
D_athabasca
D_virilis
D_willistoni
D_grimshawi
D_mojavensis
D_novamexicana
rosophila_lebanonensis
D_busckii
D_arizonae
D_navojoa
B
trn
25
20
15
10
5
0
3L
12799.5
12800
12800.5
12801
12801.5
12802
Kb
C
U4
lncRNA:CR44554
Peak2
12
10
8
6
4
2
0
3L
12786
12786.5
12787
12787.5
12788
Kb

Figure 4.4. sequence comparison among *Drosophila* species. (A-C) *trn* locus (A), *trn* coding sequence and intron (B) and U4 sequence (C) conservation retrieved from UCSC genome viewer (Kent *et al.*, 2002). (A) Chromatin profile of the *trn* locus. *trn* sequence is shown in yellow and U4 sequence in orange. Sequences from *Drosophila* species that match with *D. melanogaster* are shown in black. (B) Chromatin profile of the *trn* CDS and intron. The *trn* sequence is shown in yellow and the intron shown in grey. Sequences from *Drosophila* species that match with *D. melanogaster* are shown in black in the alignment beneath. (C) chromatin profile of the sequence of U4. The U4 sequence is shown in grey, peaks of accessible chromatin are marked in orange and long non-coding RNA is marked in yellow. Sequences from *Drosophila* species that match with *D. melanogaster* are shown in black in the alignment beneath.

I used the entire ~2.9 kb U4 sequence and SNPs were considered when they were found to be polymorphic in at least two different strains other than the one used as reference in order to exclude any possible singletons (Supplementary Table 5) (Pandey *et al.*, 2011).

I found a total of 135 SNPs in *D. simulans* strains and 122 SNPs in the *D. mauritiana* strains. Of these only 34 were polymorphic in both species, leaving a total of 190 SNPs were only polymorphic in only one of the species (Supplementary Table 5).

I then broke down the sequence into smaller regions, taking the peaks of accessible chromatin into account, I found a total of four SNPs for *D. simulans* and eight SNPs for *D. mauritiana* within the first peak, and considering the length of ~ 300 bp it results in a SNP frequency of ~0.013 SNPs/bp and ~0.026 SNPs/bp respectively. In the second peak of ~880 bp, I found a total of 24 SNPs in *D. simulans* and 20 SNPs in *D. mauritiana* (Supplementary Fig. 4.2), with a frequency of ~0.027 SNPs/bp and ~0.023 SNPs/bp respectively. For the regions with no accessible chromatin detected, I found 21 SNPs in *D. simulans* and 18 SNPs for the region of ~243 bp in between the first and the second peak, which results in a frequency of ~0.086 SNP/bp and ~0.074 SNPs/bp respectively. The region after the second peak (including the large, conserved region) I found 86 SNPs in *D. simulans* and 76 SNPs in *D. mauritiana*

(Supplementary Table 5), as this region is ~1535 bp long, it has a frequency of ~0.056 SNPs/bp and ~0.050 SNPs/bp respectively.

The regions with accessible chromatin appear then to be more conserved than the regions without accessible chromatin, consistently with the alignment of other species (Fig 4.4 C).

However, these polymorphisms are not fixed differences between these two species because they can be polymorphic in either *D. simulans* or *D. mauritiana.* I compared the reference sequences used in the Pool-seq data and I have found eleven differences that seem to be fixed between these species (Fig. 4.1 and Supplementary Fig. 3.1 and Table 2).

Table 2. Fixed differences found between *D. simulans* and *D. mauritiana*. Coordinates from the reference genomes w^{501}:r2.2 and mau12 were used. (Note that the *D. mauritiana* genome used for these coordinates is different from those used for the SNPs analysis, but the SNPs are all present in this strain)

Chromosome	Dsim		Dmau	
	Coordinate	Ref Base	Coordinate	Ref Base
Scf_3L/3L	12785572	T	12748397	C
Scf_3L/3L	12785639	A	12748464	T
Scf_3L/3L	12785841	G	12748670	T
Scf_3L/3L	12786002	A	12748833	T
Scf_3L/3L	12786003	C	12748834	G
Scf_3L/3L	12786024	T	12748855	C
Scf_3L/3L	12787461	A	12750260	G
Scf_3L/3L	12787960	G	12750771	A
Scf_3L/3L	12787962	C	12750773	T
Scf_3L/3L	12787963	A	12750774	C
Scf_3L/3L	12788279	A	12751092	T

These nucleotide differences were not found to be polymorphic in neither of the species (Supplementary Fig. 4) and hence, they are possibly fixed between these two species. I also found that these nucleotide changes identified are present between *D. mauritiana* strain mau12 and *D. simulans w501.* However, these fixed differences are not located within the footprints or TFBSs I found in Chapter 3 (Supplementary table 3 and Supplementary Fig. 3.1).

4.4 Discussion

4.4.1 ATAC-seq profile comparison

The ATAC-seq profile comparison did not identify any substantial differences between the studied species, and the differences that were found were mostly not related to enhancers with activity in the male genitalia (Fig. 4.1). Out of the four enhancers regions with differential peaks between *D. mauritiana* and *D. simulans* (U1, U4, D2 and D3), only one can drive expression in the developing male genital (Fig. 2.5), therefore the other three differences in the accessible chromatin from male genital tissue do not seem to be responsible of any differential regulation between species in this tissue. However, there are peaks of accessible chromatin that seem to be different compared to the eye-antennal imaginal disc data. U2 has a peak of accessible chromatin in the eye imaginal disc but not in the male genitalia, and U4 has a peak of accessible chromatin not found in the eye imaginal disc (Supplementary Fig. 4). This is consistent with the enhancer activity seen from these regions (U2 and U4), suggesting that these enhancers are tissue specific (Fig. 2.4 C, D, Fig. 2.6 A, B and Supplementary Fig. 4).

In the case of U4, the differential peak between *D. simulans* and *D. mauritiana* in the male genitalia, is located in the overlapping region between U3 and U4 (Fig. 4.1).

However, U3 is not active in the male genital tissue (Fig. 2.4) and further work is needed, perhaps using shorter versions of U4, to determine whether this peak contributes to the U4 activity and differences between *D. mauritiana* and other species or not.

It is also worth noting that some of the enhancers I identified in *D. melanogaster* in Chapter 2 did not correspond to where chromatin was especially accessible in the genitalia (albeit of *D. mauritiana* and *D. simulans*) for example, U5, D3 and D6 (Fig. 4.1). This could be due to the overrepresentation of cells from parts of the tissue that can mask the signal of smaller subsets of cells (Xi *et al.*, 2007; Jacobs *et al.*, 2018). However, single cell ATAC-seq can overcome these issues, as the chromatin profile resulting from this approach is specific for each cluster of cells (González-Blas *et al.*, 2020).

4.4.2 U4 differential activity

I have found that the expression driven by the U4 sequence form *D. mauritiana* is different from that driven by the equivalent *D. melanogaster* region (Fig. 4.2 and 4.3). The expression is expanded and extended to the distal part of the surstyli and aedeagal sheath of the phallus where endogenous *trn* expression is also observed (Fig. 4.3). This expansion could contribute to differences in the size of the surstylus previously seen between *D. mauritiana* and *D. simulans* (Hagen *et al.*, 2019). However, to test if this expansion of expression in *D. mauritiana* is the responsible for those differences, is also necessary to test the activity of U4 from *D. simulans*.

It is also worth noting that the *D. mauritiana* U4 enhancer was tested in a different species, which could mean that the cellular environment present in this species could be affecting the expression pattern driven by this enhancer.

In other words, the TFs that are normally regulating this enhancer in *D. mauritiana* either by repression or activation, could be less or more highly expressed in the *D. melanogaster* tissue at that specific time, hence resulting in a different expression pattern.

It may also be a result of affinity differences between the TFs of both species (Badis *et al.*, 2009; Payne and Wagner, 2014). Incompatibilities between TFs and their target enhancers between species reflecting co-evolution in their interactions in each species have been observed previously (Shaw *et al.*, 2002). Therefore, while the activity of U4D1 may reflect its functional evolution in *D. mauritiana,* it is possible that any change in the TFs composition in the cellular environment or changes in affinity from these TFs for their TFBSs between species could be responsible of differences seen between U4D1 and U4 (Fig. 4.3). However, the comparison of *svb* enhancers from different species in *D. melanogaster* was previously shown to reflect differences in the activity of *D. sechellia* sequences (McGregor *et al.*, 2007).

4.4.3 U4 Evolution

While the U4 sequence is non-coding, regions with enhancer activity are also expected to be more conserved than a non-coding region without enhancer activity (Ganley and Kobayashi, 2007; Maeso *et al.*, 2013; Buffry, Mendes and McGregor, 2016).

In the case of U4, I found a total of 135 SNPs in *D. simulans* and 122 SNPs in *Drosophila mauritiana* (Supplementary Fig. 4) and most of the SNPs were not shared between the strains of these two species. However, most of these SNPs were found in regions without accessible chromatin, which means that the region of open chromatin is more conserved. The regions with accessible chromatin had lower ratios, for example in *D. simulans* I found ratios of SNPs/nucleotide of ~0.013, ~0.027, while

the regions without accessible chromatin in *D. simulans* had ratios of ~0.086 and ~0.056. Similar results were observed for *D. mauritiana*.

When I looked at the regions conserved among *Drosophila* species (Fig 4.4) the two peaks of accessible were more conserved than the first interpeak region. The second peak of accessible chromatin contains a long non-coding RNA (lncRNA:CR44554), although the function of this lncRNA is not known, this may also explain the conservation of this region (Young *et al.*, 2012).

In the region after the second peak there was another region that seemed to be as highly conserved as the regions with accessible chromatin (Fig. 4.4). As mentioned before, this may potentially be due to a subset of cells being mask by the overall profile of the tissue in the ATAC-seq data.

Apart from the SNPs found, eleven fixed nucleotide differences were identified between the species, which were not polymorphic between *D. simulans* and *D. mauritiana,* implying they are maintained in all the strains sequenced. Although these differences did not fall within footprints or TFBs I detected in Chapter 3, they could still represent potential changes in binding sites for TFs not included in for example the JASPAR database, and therefore could still be responsible of differences in expression for this enhancer (Fig. 4.5). Further investigation is required to understand whether these fixed differences contribute to differences in U4 activity.

Chapter 5. General Discussion

In this thesis I aimed to investigate *trn* regulation and its evolution in *Drosophila* species. I have presented the results of my findings and discussed them in the relevant result chapters. In this chapter I recap my main results and discuss broader implications of my findings and future perspectives.

5.1 Enhancer screening of *trn* locus

The first aim of my thesis was to identify the enhancers that regulate *trn* during the *Drosophila* pupal development. To do that I first further characterised the endogenous expression of *trn* during male genitalia development (Hagen *et al.*, 2019), with a combination of *ISH* and the use of an endogenous *trn* tagged protein. The expression was only studied previously using *ISH*, and even though this is a very useful technique to identify expression pattern of genes in many different tissues, in the developing genitalia the expression is not always clear (Vincent *et al.*, 2019) and differences between samples from the same species can be observed (Fig. 2.1). However, *ISH* combined with imaging the pattern of the endogenous tagged protein can give a more accurate approximation of the pattern of the expression (Fig. 2.2). In terms of *trn* it is now clearer that this gene is expressed in the distal region of developing surstyli as well as proximally at the base.

Although I focused on the male genitalia, this tagged-Trn line can help to further characterise *trn* in other tissues and developmental stages where *trn* is expressed (Zhen Chang *et al.*, 1993; Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Kayoko T. Sakurai *et al.*, 2007). It could also be used to better understand the role of Trn in the genitalia and other many tissues and how this compares to embryos (Paré *et al.*,

2019; Pare and Zallen, 2020) and wings (Milán, Pérez and Cohen, 2005), included further exploring its relative expression to and potential interactions with Caps, Ten-M and other proteins (Blair, 2001; Sakurai *et al.*, 2007; Mao, Kerr and Freeman, 2008; Paré *et al.*, 2019).

It is thought that the spacing between genes in *Drosophila* is positively correlated with the complexity of the regulation of a gene (Nelson, Hersh and Carroll, 2004). The *trn* locus is among the genes with largest intergenic flanking spaces in *D. melanogaster*, with the upstream region of ~25 kb and the downstream region of ~17 kb (Nelson, Hersh and Carroll, 2004; Thurmond *et al.*, 2019; Small and Arnosti, 2020). When screening this locus for enhancers, I identified nine regions with enhancer activity during pupal development. Although only three of them were expressed in the genitalia, the intergenic regions flanking *trn* appear to encode much *cis*-regulatory information.

While I only focused on sequences that are proximal to the *trn* coding sequence, it is thought that around 30% of enhancers may be located more distally (Furlong and Levine, 2018). As mentioned previously, an enhancer of *trn* that drives its embryonic expression and is regulated by Ftz is located ~39 kb from the TSS (Field, Xiang, Anderson, *et al.*, 2016) Therefore the screening I performed on the flanking regions could be extended to more distant DNA, which would require the use of additional techniques such as chromatin capture to identify sequences that interact with the *trn* promoter (Risca and Greenleaf, 2015).

Some of the enhancers I identified showed pleiotropic expression (meaning that I detected activity in different tissues), while others had more specific patterns and some had no detectable activity in pupae (Table 1). Further study of the activity and

regulation of these enhancers will improve our understanding of the regulation of *trn* in many different tissues, for example in the eyes and wings where some of them are active, and other developmental stages such as during the embryogenesis and larval stages.

These enhancers do not all necessarily regulate *trn,* as they could be enhancers of the flanking genes or even more distal genes. This could also be explored further either by removing these enhancers and testing how the expression of *trn* or flanking genes changes and/or using chromatin capture to explore which promoters these regions interact with (Risca and Greenleaf, 2015).

5.2 *Drosophila* male genital enhancers identified

Using enhancer reporters, I successfully identified three enhancers that drive expression in the male genital during pupal development (Fig. 2.13). Region D6, however, may be regulating a different gene because there is no evidence of expression of *trn* that is localised to the bristles and this region is proximal to the downstream flanking gene of *trn*. I therefore hypothesise that this enhancer is probably regulating this gene rather than *trn*.

The other two genital enhancers identified are active in very specific parts of the male genitalia (Fig. 2.13). U4 is active around the anal plates and at the intersection between the anal plates and the rest of the genitalia and is a very good candidate for regulating *trn*. Furthermore, the region at the base of the surstyli is where *trn* expression were found to differ between *D. mauritiana* and *D. simulans* (Hagen *et al.*, 2019). Thus, the function of *trn* in defining the boundaries and distribution of cells in, for example, embryos and wings may also be the case in the male genitalia, and could explain how differences in surstyli size are generated by variation in *trn* expression

(Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Mao, Kerr and Freeman, 2008; Paré *et al.*, 2019). This could be tested further by removing U4 or swapping the U4 sequences between species using CRISPR/Cas9 to determine if *trn* expression changes and the size of the surstyli changes compared with the controls.

In contrast to U4, the activity of the U8 does not fully colocalise with the endogenous expression pattern of *trn* (Fig. 2.13, 2.14 and Supplementary Fig. 2.1 A). Nevertheless, U8 activity is very specific in the primordia of the surstyli and overlaps with *odd paired* (*opa)* expression, a gene expressed in the developing genitalia (Vincent *et al.*, 2019). As I mentioned above, U8 activity could be ectopic because it may be missing important binding sites, which testing a longer version of this enhancer may help to clarify. Deleting this enhancer and examining the effect on *trn* expression could also be worthwhile. Further investigation of U8 regulation like that I carried out for U4 may also reveal new candidate TFs for the development of male genitalia.

5.3 Binding sites in U4

In Chapter 3 I searched for TFs that regulate U4 using the accessible chromatin profile as a guide. Using MEME to study the peaks of accessible chromatin within U4 did not identify any significantly enriched motif compared with the background model (Bailey *et al.*, 2009). However, one of the most enriched motifs in *D. simulans* and *D. melanogaster*, although not significantly was predicted to be a Cad binding site (Fig. 3.1). MEME may not have found this motif to be significantly enriched because it is a very common motif and corresponds to the binding site sequence for many homeodomain TFs, including Abd-B in *D. melanogaster* (Mann, Lelli and Joshi, 2009; Jolma *et al.*, 2013; Weirauch *et al.*, 2014; Nitta *et al.*, 2015).

I then used the HINT-ATAC Footprinting tool from RGT, which improved the results (Fig. 3.2) (Li *et al.*, 2019) and allowed me to find significantly enriched motifs and therefore other candidate TFBSs. However, I could only compare these motifs to known or predicted TFBSs of the database such as JASPAR or *CIS*-BP, which still do not have binding information on all TFs (Dimitrova *et al.*, 2014; Weirauch *et al.*, 2014; Fornes *et al.*, 2019). Recently there have been progress in large scale ChIP-seq assays to identify ~600 TFBSs for human and mouse (Kulakovskiy *et al.*, 2018). This together with new methods to identify footprints and motif discovery tools may improve the TFBSs catalogued in databases (Bailey *et al.*, 2009; Herrmann *et al.*, 2012; Mundade *et al.*, 2014; Li *et al.*, 2019) and ultimately assaying TF binding to enhancer sequences.

5.4 *cad* regulation of *trn* and genital GRN

Although the Cad binding site was not significantly enriched, the motif was still found within the accessible chromatin (Fig. 3.2), and *cad* is expressed in the anal plates where it overlaps with U4 activity and *trn* expression (Vincent *et al.*, 2019).

I therefore assayed *cad* function using RNAi using a genital driver, which resulted in the loss of the anal plates (Fig. 3.3). As previously discussed, this was somewhat expected (Moreno and Morata, 1999), although a transformation of this tissue into genital tissue may have been predicted rather than loss of the anal plates. Curiously, *cad* RNAi also resulted in smaller surstyli and posterior lobes (Fig. 3.3), although this was not observed with U4 driven *cad* RNAi when corrected for tibia length, this knockdown was likely later and more spatially restricted (Fig. 2.12). Furthermore, the difference in the RNAi result using NP6333 versus U4, could be possibly explained as a side effect of the loss of the anal plates when using NP6333. Therefore, since *cad*

is probably not expressed in the surstyli, my results suggest *cad* does not regulate *trn* via the U4 in the size of the surstyli.

Nonetheless, U4 driven RNAi against *cad* did appear to significantly increase the number of anal plate bristles (Fig. 3.6). This effect could be explained by the local delimitation of the anal plate boundaries by *cad,* where *cad* potentially activates or represses *trn*. Hence, the knockdown of *cad* in these cells may result in the loss of *trn* expression around the anal plates or alternatively *trn* being de-repressed in cells where it is not normally expressed, in either scenario resulting in more bristles to develop and slightly larger anal plates. To investigate this, the tagged Trn line could be used together with the *cad* knockdown driven by U4, to see if there is a reduction or expansion in the expression pattern of the Trn protein. In addition, other TFs with similar binding sites to Cad could be tested, especially Abd-B as this TF is expressed in segment A9 (Chen, Christiansen and Baker, 2005; Aspiras, Smith and Angelini, 2011)

RNAi knockdown of *trn* using U4 is also necessary to better understand the role of this gene in anal plate development, especially as Hagen *et al.*, (2019) did not report whether knockdown of *trn* using the NP6333 driver had any effect on the anal plates.

Finally, further analysis of the U4 enhancer region by breaking it down into smaller fragments would help to identify the minimal sequence required to drive expression and narrow the search for TFBSs.

5.5 U4 chromatin profile and activity

The chromatin accessibility profile did not reveal any substantial difference between *D. simulans* and *D. mauritiana* (Fig. 4.1), and the peak I found within U4 with different heights between *D. mauritiana* and *D. simulans* is not likely to contribute to possible

differences in activity as it is shared with U3, which has no detectable activity in the developing genitalia.

I also compared the profile of accessible chromatin in U4 between ATAC-seq generated from the eye-antennal imaginal disc versus the genitalia. This revealed a peak in the male genital that is not present in the eye imaginal disc (Supplementary Fig. 4) and vice versa a peak that is present in the eye but not in the genitalia (Supplementary Fig. 4), which is consistent with the activity detected for these regions (Fig. 2.4 C, D and Fig. 2.6 A, B). These could mark an enhancer specific to the genitalia and an enhancer specific of the eye-antennal imaginal disc respectively and highlights the utility of ATAC-seq data in identifying enhancers. It is clear that improvements in ATAC-seq data acquisition and analysis, such as single-cell ATAC-seq or footprinting (Li *et al.*, 2019; González-Blas *et al.*, 2020; Yan *et al.*, 2020) could provide more insights into the enhancers I identified.

5.6 U4 activity and sequence evolution

In Chapter 4 I found that the enhancer U4, which is active in the developing male genital, appears to have evolved in its activity between *D. mauritiana* and *D. melanogaster* (Fig. 4.3). Although *D. mauritiana* and *D. melanogaster* have similar sizes of surstyli, intriguingly, *D. mauritiana* U4 activity is stronger in the surstyli than in *D. melanogaster* U4.

Testing the *D. simulans* U4 sequence will provide a better insight into the evolution of this enhancer. If the *D. simulans* U4 activity is weaker than that of *D. mauritiana* in the developing surstyli, it would be then more plausible that this enhancer underlies the differences seen in surstyli size between these two species (Hagen *et al.*, 2019).

I analysed the conservation of the U4 sequence between *Drosophila* species, and I found that, as expected, the sequences corresponding to peaks of accessible chromatin were more conserved than those associated with non-accessible chromatin (Ganley and Kobayashi, 2007; Maeso *et al.*, 2013).

I found eleven fixed differences in U4 between strains of *D. mauritiana* and *D. simulans* (Table 2). Although these differences did not overlap with the predicted TFBSs or significantly enriched motifs, they could still represent changes in binding sites and could contribute to differences in U4 activity and *trn* expression. They would definitely be worth testing further if U4 activity differs between *D. mauritiana* and *D. simulans.*

In future experiments, a different approach could be taken and instead of introducing the *D. mauritiana* and *D. simulans* enhancers into *D. melanogaster*, they could be swapped between *D. mauritiana* and *D. simulans* using CRISPR/Cas9 and the effect on *trn* expression and surstyli size, or any other aspect of the genitalia, directly measured. CRISPR/Cas9 could also then be used to modify predicted TFBSs individually and in combination to test for specific TFs that may be regulating U4.

Nevertheless, it remains possible that U4 does not differ in activity or underlies surstyli evolution between *D. mauritiana* and *D. simulans*. It is therefore important that regions U6, U7 and D1 that I was unable to test are investigated for activity in the developing genitalia. If these regions are not able to drive expression in genitalia, then ultimately regions more distal to the *trn* coding region, as mentioned above, would need to be examined with respect to the evolution of *trn* expression and surstyli size between *D. mauritiana and D. simulans.*

5.7 Conclusions

My investigation has contributed to the understanding of the *trn* regulatory landscape in *Drosophila species* in the following key points:

- By identifying several enhancers in the vicinity of *trn* in *D. melanogaster.* One of those enhancers has different activity depending on the species, which can be further investigated to assess the importance of the regulatory sequences in the rapid evolution of male genitalia.
- It has improved the characterisation of *trn* expression in the pupal male genitalia, although we still lack knowledge of its activity during the larval stage. However, seeing that many of the interactions in the imaginal disc are dependent on compartmentalization (Fig 1.3) and *trn* is thought to be involved in that (Blair, 2001; Milán *et al.*, 2001; Milán, Pérez and Cohen, 2005; Kayoko T Sakurai *et al.*, 2007; Mao, Kerr and Freeman, 2008), it is very likely that *trn* is not restricted to the pupal stage.
- While TFBS databases are useful, they still need to be complemented with other techniques. Further analysis of the sequences will be required to find the TFs responsible of regulating these enhancers.
- The evolution of the sequence of U4 shows that the regions within open chromatin are conserved in comparison with regions that are not within open chromatin, which indicates that mutations in these regions are selected against.
- I have found several fixed differences in the U4 sequence between species that can be investigated further, but unfortunately these flies could not be phenotyped to match genetic variation as the population data was obtained from an external source (see methods)

References

Acebes, A. (2003) 'Species-specific effects of single sensillum ablation on mating position in *Drosophila'*, *Journal of Experimental Biology*. The Company of Biologists, 206(17), pp. 3095–3100. doi: 10.1242/jeb.00522.

Ahmed, O. M. *et al.* (2019) 'Evolution of Mechanisms that Control Mating in *Drosophila* Males', *Cell Reports*. Elsevier B.V., 27(9), pp. 2527-2536.e4. doi: 10.1016/j.celrep.2019.04.104.

Alonso, C. R. and Wilkins, A. S. (2005) 'The molecular elements that underlie developmental evolution', *Nature Reviews Genetics*, pp. 709–715. doi: 10.1038/nrg1676.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*. J Mol Biol, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Artero, R. *et al.* (2003) 'Notch and Ras signaling pathway effector genes expressed in fusion competent and founder cells during *Drosophila* myogenesis', *Development*, pp. 6257–6272. doi: 10.1242/dev.00843.

Aspiras, A. C., Smith, F. W. and Angelini, D. R. (2011) 'Sex-specific gene interactions in the patterning of insect genitalia', *Developmental Biology*. Academic Press, 360(2), pp. 369–380.

Atallah, J. *et al.* (2009) 'Developmental constraints and convergent evolution in *Drosophila* sex comb formation', *Evolution and Development*, 11(2), pp. 205–218. doi: 10.1111/j.1525-142X.2009.00320.x.

Badis, G. *et al.* (2009) 'Diversity and Complexity in DNA Recognition by Transcription Factors **', *Science*, 324(5935), pp. 1720–1723. doi: 10.1126/science.1162327.

Bailey, T. L. *et al.* (2009) 'MEME Suite: Tools for motif discovery and searching', *Nucleic Acids Research*, 37(SUPPL. 2). doi: 10.1093/nar/gkp335.

Bailey, T. L. (2021) 'STREME: accurate and versatile sequence motif discovery', *Bioinformatics*. Oxford Academic, 37(18), pp. 2834–2840. doi:

10.1093/BIOINFORMATICS/BTAB203.

Banerji, J., Rusconi, S. and Schaffner, W. (1981) “Expression of a β-globin gene is enhanced by remote SV40 DNA sequences,” *Cell*, 27(2 PART 1), pp. 299–308. Available at: https://doi.org/10.1016/0092-8674(81)90413-X.

Barbieri, M. *et al.* (2012) ‘Complexity of chromatin folding is captured by the strings and binders switch model’, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 109(40), pp. 16173–16178. doi: 10.1073/PNAS.1204799109.

Beira, J. V and Paro, R. (2016) ‘The legacy of *Drosophila* imaginal discs’, *Chromosoma*, 125, pp. 573–592. doi: 10.1007/s00412-016-0595-4.

Bejarano, F. *et al.* (2008) ‘A Gain-of-Function Suppressor Screen for Genes Involved in Dorsal-Ventral Boundary Formation in the *Drosophila* Wing’, *Genetics*, 178, pp. 307–323. doi: 10.1534/genetics.107.081869.

Bentsen, M. *et al.* (2020) ‘ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation’, *Nature Communications*. Nature Publishing Group, 11(1). doi: 10.1038/S41467-020-18035-1.

Blair, S. S. (2001) ‘Cell lineage: Compartments and Capricious’, *Current Biology*, 11, pp. 1017–1021.

Böhme, W. and Ziegler, T. (2009) ‘A review of iguanian and anguimorph lizard genitalia (squamata: Chamaeleonidae; Varanoidea, shinisauridae, xenosauridae, anguidae) and their phylogenetic significance: Comparisons with molecular data sets’, *Journal of Zoological Systematics and Evolutionary Research*, 47(2), pp. 189–202. doi: 10.1111/j.1439-0469.2008.00495.x.

Bouwman, B. A. and de Laat, W. (2011) ‘Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome’, *Genome Biology*, 16, p. 154. doi: 10.1186/s13059-015-0730-1.

Bozek, M. *et al.* (2019) ‘ATAC-seq reveals regional differences in enhancer accessibility during the establishment of spatial coordinates in the *Drosophila* blastoderm’, *Genome Research*. Cold Spring Harbor Laboratory Press, 29(5), pp. 771–783. doi: 10.1101/GR.242362.118.

Broman, K. W. *et al.* (2003) 'R/qtl: QTL mapping in experimental crosses', *BIOINFORMATICS*, 19(7), pp. 889–890. doi: 10.1093/bioinformatics/btg112.

Buenrostro, J. D. *et al.* (2013) 'Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position', *Nature Methods*. NIH Public Access, 10(12), pp. 1213–1218. doi: 10.1038/nmeth.2688.

Buenrostro, J. D. *et al.* (2015) 'ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide.', *Current protocols in molecular biology*. NIH Public Access, 109, pp. 21.29.1–9. doi: 10.1002/0471142727.mb2129s109.

Buffry, A. D., Mendes, C. C. and McGregor, A. P. (2016) 'The Functionality and Evolution of Eukaryotic Transcriptional Enhancers', *Advances in Genetics*. Academic Press, 96, pp. 143–206. doi: 10.1016/BS.ADGEN.2016.08.004.

Bugga, L., Ratnaparkhi, A. and Zinn, K. (2009) 'The Cell Surface Receptor Tartan Is a Potential In Vivo Substrate for the Receptor Tyrosine Phosphatase Ptp52F', *MOLECULAR AND CELLULAR BIOLOGY*, 29(12), pp. 3390–3400. doi: 10.1128/MCB.01764-08.

Burgess, D. and Freeling, M. (2014) 'The Most Deeply Conserved Noncoding Sequences in Plants Serve Similar Functions to Those in Vertebrates Despite Large Differences in Evolutionary Rates', *The Plant Cell*, 26(3), pp. 946–961. Available at: https://www.jstor.org/stable/43190380.

Carroll, S. B. (2008) 'Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution', *Cell*, 134, pp. 25–36. doi: 10.1016/j.cell.2008.06.030.

Casares, F. *et al.* (1997) 'The genital disc of *Drosophila melanogaster*. I. Segmental and compartmental organization', *Development Genes and Evolution*, 207(4), pp. 216–228. doi: 10.1007/s004270050110.

Chang, Zhen *et al.* (1993) 'Molecular and Genetic Characterization of the *Drosophila tartan* Gene', *Developmental Biology*, pp. 315–332. doi: 10.1006/dbio.1993.1310.

Chapman, T. *et al.* (2003) 'Sexual conflict', *Trends in Ecology and Evolution*, 18(1), pp. 41–47. doi: https://doi.org/10.1016/S0169-5347(02)00004-6.

Chatterjee, S. S. *et al.* (2011) 'The female-specific Doublesex isoform regulates pleiotropictranscription factors to pattern genital development in *Drosophila'*, *Development*, 138(6), pp. 1099–1109. doi: 10.1242/dev.055731.

Chen, E. H., Christiansen, A. E. and Baker, B. S. (2005) 'Allocation and specification of the genital disc precursor cells in *Drosophila'*, *Developmental Biology*. Academic Press Inc., 281(2), pp. 270–285. doi: 10.1016/j.ydbio.2005.02.032.

Clutton-Brock, T. H. and Parker, G. A. (1995) 'Sexual coercion in animal societies', *Animal Behaviour*, 49(5), pp. 1345–1365. doi: 10.1006/anbe.1995.0166.

Cong, L. *et al.* (2013) 'Multiplex Genome Engineering Using CRISPR/Cas Systems HHS Public Access', *Science*, 339(6121), pp. 819–823. doi: 10.1126/science.1231143.

Coyne, Jerry A. (1983) 'Genetic Basis of Differences in Genital Morphology Among Three Sibling Species of *Drosophila'*, *Evolution*, 37(6), pp. 1101–1118. doi: 10.2307/2408834.

Coyne, J. A., Rux, J. and David, J. R. (1990) 'Genetics of morphological differences and hybrid sterility between *Drosophila sechellia* and its relatives', *Genet. Res., Camb*, 57, pp. 113–122. doi: 10.1017/S0016672300029177.

Crocker, J. *et al.* (2015) 'Low affinity binding site clusters confer HOX specificity and regulatory robustness', *Cell*, 160(1–2), pp. 191–203. doi: 10.1016/j.cell.2014.11.041.

Crocker, J., Preger-Ben Noon, E. and Stern, D. L. (2016) The Soft Touch: Low-Affinity Transcription Factor Binding Sites in Development and Evolution. 1st edn, *Current Topics in Developmental Biology*. 1st edn. Elsevier Inc. doi: 10.1016/bs.ctdb.2015.11.018.

Degnan, B. M. *et al.* (2009) 'Early evolution of metazoan transcription factors', *Current Opinion in Genetics and Development*, 19(6), pp. 591–599. doi: 10.1016/j.gde.2009.09.008.

Dimitrova, N. *et al.* (2014) 'Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity', *Cell*, 158(6), pp. 1431–1443. doi: 10.1016/j.cell.2014.08.009.Determination.

Dixson, A. F. (1987) 'Observations on the evolution of the genitalia and copulatory

behaviour in male primates', *Journal of Zoology*, 213(3), pp. 423–443. doi: 10.1111/j.1469-7998.1987.tb03718.x.

Eberhard, W. G. (1985) *Sexual Selection and Animal Genitalia*, *Sexual Selection and Animal Genitalia*. Harvard University Press. doi: 10.4159/harvard.9780674330702.

Eberhard, W. G. (2009a) 'Postcopulatory sexual selection: Darwin's omission and its consequences', *PNAS*, 106(1), pp. 10025–10032.

Eberhard, W. G. (2009b) 'STATIC ALLOMETRY AND ANIMAL GENITALIA', *Evolution*. John Wiley & Sons, Ltd (10.1111), 63(1), pp. 48–66. doi: 10.1111/j.1558-5646.2008.00528.x.

Eberhard, W. G. (2010) 'Evolution of genitalia: theories, evidence, and new directions', *Genetica*, 138(1), pp. 5–18. doi: 10.1007/s10709-009-9358-y.

Eberhard, W. G. and Lehmann, G. U. C. (2019) 'Demonstrating sexual selection by cryptic female choice on male genitalia: What is enough?', *Evolution*, 73(12), pp. 2415–2435. doi: 10.1111/evo.13863.

El-Sherif, E. and Levine, M. (2016) 'Shadow Enhancers Mediate Dynamic Shifts of Gap Gene Expression in the *Drosophila* Embryo', *Current Biology*. Elsevier Ltd, 26(9), pp. 1164–1169. doi: 10.1016/j.cub.2016.02.054.

Estrada, B., Casares, F. and Sánchez-Herrero, E. (2003) *Development of the genitalia in Drosophila melanogaster*, *Differentiation*. doi: 10.1046/j.1432-0436.2003.03017.x.

Field, A., Xiang, J., Anderson, W. Ray, *et al.* (2016) 'Activation of Ftz-F1-Responsive Genes through Ftz/Ftz-F1 Dependent Enhancers', *PLOS ONE*. Edited by M. Mallo. Public Library of Science, 11(10), p. e0163128. doi: 10.1371/journal.pone.0163128.

Foo, S. M. *et al.* (2014) 'Zelda potentiates morphogen activity by increasing chromatin accessibility', *Curr Biol*, 24(12), pp. 1341–1346. doi: 10.1016/j.cub.2014.04.032.

Fornes, O. *et al.* (2019) 'JASPAR 2020: update of the open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 48, pp. 87–92. doi: 10.1093/nar/gkz1001.

Foronda, D. *et al.* (2006) 'Requirement of *abdominal-A* and *abdominal-B* in the developing genitalia of *Drosophila* breaks the posterior downregulation rule', *Development*, 133(1), p. 182. doi: 10.1242/dev.02231.

Frankel, N. *et al.* (2010) 'Phenotypic robustness conferred by apparently redundant transcriptional enhancers', *Nature*, 466(7305), pp. 490–493. doi: 10.1038/NATURE09158.

Frankel, N. S. *et al.* (2011) 'Morphological evolution caused by many subtle-effect substitutions in regulatory DNA', *Nature*, 474(7353), pp. 598–603. doi: 10.1038/nature10200.

Frazee, S. R. *et al.* (2021) 'Interspecific introgression reveals a role of male genital morphology during the evolution of reproductive isolation in *Drosophila'*, *Evolution*, 75(5), pp. 989–1002. doi: 10.1111/evo.14169.

Fujisawa, T. *et al.* (2019) 'Genetic basis of species-specific genitalia reveals role in species diversification', *Science Advances*. American Association for the Advancement of Science, 5(6), p. eaav9939. doi: 10.1126/sciadv.aav9939.

Furlong, E. E. M. and Levine, M. (2018) 'Developmental enhancers and chromosome topology', *Science*. American Association for the Advancement of Science, 361(6409), pp. 1341–1345. doi: 10.1126/SCIENCE.AAU0320.

Ganley, A. R. D. and Kobayashi, T. (2007) 'Phylogenetic footprinting to find functional DNA elements', *Methods in Molecular Biology*, pp. 367–379. doi: 10.1385/1-59745-514-8:367.

Gao, L. *et al.* (2018) 'Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution', *Cell*. Elsevier, 173(1), pp. 248-259.e15. doi: 10.1016/j.cell.2018.02.028.

Garrigan, D. *et al.* (2012) 'Genome sequencing reveals complex speciation in the *Drosophila simulans* clade', *Genome Research*, 22(8), pp. 1499–1511. doi: 10.1101/gr.130922.111.

Gehrke, A. R. *et al.* (2015) 'Deep conservation of wrist and digit enhancers in fish', *Proceedings of the National Academy of Sciences*, 112(3), pp. 803–808. doi: 10.1073/pnas.1420208112.

Giniger, E. and Ptashne, M. (1988) ‘Cooperative DNA binding of the yeast transcriptional activator GAL4.’, *Proceedings of the National Academy of Sciences of the United States of America*, 85(2), pp. 382–386. doi: 10.1073/pnas.85.2.382.

Glassford, William J. *et al.* (2015) ‘Co-option of an Ancestral Hox-Regulated Network Underlies a Recently Evolved Morphological Novelty’, *Developmental Cell*. Cell Press, 34(5), pp. 520–531. doi: 10.1016/j.devcel.2015.08.005.

Glassford, William J *et al.* (2015) ‘Co-option of an ancestral Hox-regulated network underlies a recently evolved morphological novelty Graphical Abstract HHS Public Access’, *Dev Cell*, 34(5), pp. 520–531. doi: 10.1016/j.devcel.2015.08.005.

Gohl, D. M. *et al.* (2011) ‘A versatile in vivo system for directed dissection of gene expression patterns’, *Nature Methods*. Nat Methods, 8(3), pp. 231–237. doi: 10.1038/nmeth.1561.

Gompel, N. *et al.* (2005) ‘Chance caught on the wing: *cis*-regulatory evolution and the origin of pigment patterns in *Drosophila’*, *Nature*, pp. 481–487. doi: 10.1038/nature03235.

González-Blas, C. B. *et al.* (2020) ‘Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics’, *Molecular Systems Biology*, 16, p. e9438. doi: 10.1101/2019.12.19.882381.

Gorfinkiel, N., Sánchez, L. and Guerrero, I. (2003) ‘Development of the *Drosophila* genital disc requires interactions between its segmental primordia’, *Development*. The Company of Biologists Ltd, pp. 295–305. doi: 10.1242/dev.00214.

Graf, T. and Enver, T. (2009) ‘Forcing cells to change lineages’, *Nature*, 462(7273), pp. 587–594. doi: 10.1038/NATURE08533.

Graham, P. L. *et al.* (2019) ‘Dynamic expression of *Drosophila* segmental cell surface-encoding genes and their pair-rule regulators’, *Developmental Biology*. Elsevier Inc., 447(2), pp. 147–156. doi: 10.1016/j.ydbio.2019.01.015.

Gupta, S. *et al.* (2007) ‘Quantifying similarity between motifs’, *Genome Biology*. BioMed Central, 8(2), p. R24. doi: 10.1186/gb-2007-8-2-r24.

Hagen, J. F. D. *et al.* (2019) ‘tartan underlies the evolution of *Drosophila* male genital morphology’, *Proceedings of the National Academy of Sciences*. National Academy

of Sciences, 116(38), pp. 19025–19030. doi: 10.1073/pnas.1909829116.

Hagen, J. F. D. *et al.* (2020) 'Unraveling the Genetic Basis for the Rapid Diversification of Male Genitalia between *Drosophila* Species', *Molecular Biology and Evolution*, 38(2), pp. 432–448. doi: 10.1093/molbev/msaa232.

Harcet, M. *et al.* (2010) 'Demosponge EST sequencing reveals a complex genetic toolkit of the simplest metazoans', *Molecular Biology and Evolution*, 27(12), pp. 2747–2756. doi: 10.1093/molbev/msq174.

Hardison, R. C. and Taylor, J. (2012) 'Genomic approaches towards finding *cis*-regulatory modules in animals', *Nature Reviews Genetics*, pp. 469–483. doi: 10.1038/nrg3242.

Harrison, M. M. *et al.* (2011) 'Zelda binding in the early *Drosophila melanogaster* embryo marks regions subsequently activated at the maternal-to-zygotic transition', *PLoS Genetics*, 7(10). doi: 10.1371/JOURNAL.PGEN.1002266.

Heinz, S. *et al.* (2010) 'Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities', *Mol Cell*, 38(4), pp. 576–589. doi: 10.1016/j.molcel.2010.05.004.

Herrmann, C. *et al.* (2012) 'i-*cis*Target: an integrative genomics method for the prediction of regulatory features and *cis*-regulatory modules.', *Nucleic acids research*. Oxford University Press, 40(15), p. e114.

Hinman, V. F., Yankura, K. A. and McCauley, B. S. (2009) 'Evolution of gene regulatory network architectures: Examples of subcircuit conservation and plasticity between classes of echinoderms', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier B.V., 1789(4), pp. 326–332. doi: 10.1016/j.bbagrm.2009.01.004.

Hong, W. *et al.* (2009) 'Leucine-rich repeat transmembrane proteins instruct discrete dendrite targeting in an olfactory map', *Nature Neuroscience*, 12(12), pp. 1542–1550. doi: 10.1038/nn.2442.

Hosken, D. J. *et al.* (2019) 'Penis evolution across species: divergence and diversity', *Nature Reviews Urology*. Springer US, 16(2), pp. 98–106. doi: 10.1038/s41585-018-0112-z.

Hosken, D. J. and Stockley, P. (2004) 'Sexual selection and genital evolution', *Trends in Ecology and Evolution*, 19(2), pp. 87–93. doi: 10.1016/j.tree.2003.11.012.

House, C. M. *et al.* (2013) 'Sexual and Natural Selection Both Influence Male Genital Evolution', *PLoS ONE*, 8(5), p. 63807. doi: 10.1371/journal.pone.0063807.

Jacobs, J. *et al.* (2018) 'The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes', *Nature Genetics*. Nature Publishing Group, 50(7), pp. 1011–1020. doi: 10.1038/s41588-018-0140-x.

Jagadeeshan, S. and Singh, R. S. (2006) 'A time-sequence functional analysis of mating behaviour and genital coupling in *Drosophila*: Role of cryptic female choice and male sex-drive in the evolution of male genitalia', *Journal of Evolutionary Biology*, 19(4), pp. 1058–1070. doi: 10.1111/j.1420-9101.2006.01099.x.

Jayaram, N., Usvyat, D. and Martin, A. C. (2016) 'Evaluating tools for transcription factor binding site prediction', *BMC Bioinformatics*. BioMed Central Ltd., 17(1). doi: 10.1186/S12859-016-1298-9.

Jeong, S., Rokas, A. and Carroll, S. B. (2006) 'Regulation of Body Pigmentation by the Abdominal-B Hox Protein and Its Gain and Loss in *Drosophila* Evolution', *Cell*, 125(7), pp. 1387–1399. doi: 10.1016/j.cell.2006.04.043.

Jolma, A. *et al.* (2013) 'DNA-Binding Specificities of Human Transcription Factors', *Cell*, 152, pp. 327–339. doi: 10.1016/j.cell.2012.12.009.

Kamimura, Y. and Mitsumoto, H. (2011) 'Comparative copulation anatomy of the *Drosophila melanogaster* species complex (Diptera: Drosophilidae)', *Entomological Science*, 14(4), pp. 399–410. doi: 10.1111/j.1479-8298.2011.00467.x.

Katoh, K. and Standley, D. M. (2013) 'Article Fast Track MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30(4), pp. 772–780. doi: 10.1093/molbev/mst010.

Keisman, E. and Baker, B. S. (2001) 'The *Drosophila* sex determination hierarchy modulates *wingless* and *decapentaplegic* signaling to deploy *dachshund* sex-specifically in the genital imaginal disc', *Development*, 128, pp. 1643–1656.

Kent, W. J. *et al.* (2002) 'The Human Genome Browser at UCSC', *Genome*

Research. Cold Spring Harbor Laboratory Press, 12(6), pp. 996–1006. doi: 10.1101/GR.229102.

Kittelmann, S. *et al.* (2018) 'Gene regulatory network architecture in different developmental contexts influences the genetic basis of morphological evolution', *PLOS Genetics*, pp. 1–21. doi: 10.1371/journal.pgen.1007375.

Kölzer, S. *et al.* (2003) 'defective proventriculus is required for pattern formation along the proximodistal axis, cell proliferation and formation of veins in the *Drosophila* wing', *Development*. The Company of Biologists, 130(17), pp. 4135–4147. doi: 10.1242/DEV.00608.

Koshikawa, S. *et al.* (2015) 'Gain of *cis*-regulatory activities underlies novel domains of wingless gene expression in *Drosophila'*, *PNAS*, 112(24), pp. 7524–7529. doi: 10.1073/pnas.1509022112.

Kulakovskiy, I. V. *et al.* (2018) 'HOCOMOCO: Towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D252–D259. doi: 10.1093/nar/gkx1106.

Kvon, Evgeny Z. *et al.* (2014) 'Genome-scale functional characterization of *Drosophila* developmental enhancers in vivo', *Nature*. Nature Publishing Group, 512(1), pp. 91–95. doi: 10.1038/nature13395.

Kvon, E. Z. (2015) 'Using transgenic reporter assays to functionally characterize enhancers in animals', *Genomics*. Elsevier Inc., 106(3), pp. 185–192. doi: 10.1016/j.ygeno.2015.06.007.

Lagha, M., Bothma, J. P. and Levine, M. (2012) 'Mechanisms of transcriptional precision in animal development', *Trends in Genetics*. Elsevier Ltd, 28(8), pp. 409–416. doi: 10.1016/j.tig.2012.03.006.

Lang, M. *et al.* (2014) 'Radiation of the *Drosophila nannoptera* species group in Mexico', *Journal of Evolutionary Biology*, 27(3), pp. 575–584. doi: 10.1111/jeb.12325.

Lettice, L. A. *et al.* (2003) 'A long-range *Shh* enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly', *Human*

Molecular Genetics, 12(14), pp. 1725–1735. doi: 10.1093/hmg/ddg180.

LeVasseur-Viens, H., Polak, M. and Moehring, A. J. (2015) 'No evidence for external genital morphology affecting cryptic female choice and reproductive isolation in *Drosophila'*, *Evolution*, 69(7), pp. 1797–1807. doi: 10.1111/evo.12685.

Levine, M. (2010) 'Transcriptional Enhancers in Animal Development and Evolution', *Current biology : CB*. NIH Public Access, 20(17), p. R754. doi: 10.1016/J.CUB.2010.06.070.

Li, L. M. and Arnosti, D. N. (2011) 'Long-and short-range transcriptional repressors induce distinct chromatin states on repressed genes', *Curr Biol*, 21(5), pp. 406–412. doi: 10.1016/j.cub.2011.01.054.

Li, Z. *et al.* (2019) 'Identification of transcription factor binding sites using ATAC-seq', *Genome Biology*, 20. doi: 10.1186/s13059-019-1642-2.

Lieberman-Aiden, E. *et al.* (2009) 'Comprehensive mapping of long range interactions reveals folding principles of the human genome', *Science (New York, N.Y.)*. NIH Public Access, 326(5950), p. 289. doi: 10.1126/SCIENCE.1181369.

Liu, J., Mercer, John M, *et al.* (1996) 'Genetic analysis of a morphological shape difference in the male genitalia of *Drosophila simulans* and *Drosophila mauritiana*.', *Genetics*. Genetics Society of America, 142(4), pp. 1129–45. Available at: http://www.ncbi.nlm.nih.gov/pubmed/8846893.

Lopes, R., Korkmaz, G. and Agami, R. (2016) 'Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers', *Molecular Cell Biology*, 17, pp. 597–604. doi: 10.1038/nrm.2016.79.

López-Rivera, F. *et al.* (2020) 'A Mutation in the *Drosophila melanogaster eve* Stripe 2 Minimal Enhancer Is Buffered by Flanking Sequences', *G3 Genes|Genomes|Genetics*. Genetics Society of America, 10(12), pp. 4473–4482. doi: 10.1534/g3.120.401777.

Ludwig, M. Z. *et al.* (2000) 'Evidence for stabilizing selection in a eukaryotic enhancer element', *Nature*, 403(6769), pp. 564–567. doi: 10.1038/35000615.

Maeso, I. *et al.* (2013) 'Deep conservation of *cis*-regulatory elements in metazoans', *Philosophical Transactions of the Royal Society B: Biological Sciences*. The Royal

Society, 368(1632), p. 20130020. doi: 10.1098/RSTB.2013.0020.

Mali, P. *et al.* (2013) 'RNA-Guided Human Genome Engineering via Cas9', *Science*, 339(6121), pp. 823–826. doi: 10.1126/science.1232033.

Mann, R. S., Lelli, K. M. and Joshi, R. (2009) 'Hox Specificity: Unique Roles for Cofactors and Collaborators', *Curr Top Dev Biol*, 88, pp. 63–101. doi: 10.1016/S0070-2153(09)88003-4.

Mao, Y., Kerr, M. and Freeman, M. (2008) 'Modulation of *Drosophila* Retinal Epithelial Integrity by the Adhesion Proteins Capricious and Tartan', *PLoS ONE*. Edited by P. Callaerts. Public Library of Science, 3(3), p. e1827. doi: 10.1371/journal.pone.0001827.

Martindale, M. Q. and Hejnol, A. (2009) 'Developmental Cell Perspective A Developmental Perspective: Changes in the Position of the Blastopore during Bilaterian Evolution', *Developmental Cell*, 17, pp. 162–174. doi: 10.1016/j.devcel.2009.07.024.

Masly, J. P. *et al.* (2011) 'The Genetic Basis of Rapidly Evolving Male Genital Morphology in *Drosophila'*, *Genetics*, 189, pp. 357–374. doi: 10.1534/genetics.111.130815.

Masly, J. P. and Kamimura, Y. (2014) 'Asymmetric mismatch in strain-specific genital morphology causes increased harm to *Drosophila* females', *Evolution*, 68(8), pp. 2401–2411. doi: 10.1111/evo.12436.

Mattei, A. L. *et al.* (2015) 'Integrated 3D view of postmating responses by the *Drosophila melanogaster* female reproductive tract, obtained by micro-computed tomography scanning', *Proceedings of the National Academy of Sciences of the United States of America*, 112(27), pp. 8475–8480. doi: 10.1073/pnas.1505797112.

McGregor, A. P. *et al.* (2001) 'Rapid restructuring of bicoid-dependent hunchback promoters within and between Dipteran species: Implications for molecular coevolution', *Evolution and Development*, 3(6), pp. 397–407. doi: 10.1046/j.1525-142X.2001.01043.x.

McGregor, A. P. *et al.* (2007) 'Morphological evolution through multiple *cis*-regulatory mutations at a single gene', *Nature*, 448(7153), pp. 587–590. doi:

10.1038/nature05988.

Milán, M. *et al.* (2001) 'The LRR proteins capricious and Tartan mediate cell interactions during DV boundary formation in the *Drosophila* wing.', *Cell*. Elsevier, 106(6), pp. 785–94. doi: 10.1016/S0092-8674(01)00489-5.

Milán, M., Pérez, L. and Cohen, S. M. (2002) 'Short-Range Cell Interactions and Cell Survival in the *Drosophila* Wing in small clusters of cells in the developing wing disc, suggesting that local cell interactions may play an im-portant role in the decision of cells to die (Milá n et al', *Developmental Cell*, 2, pp. 797–805.

Milán, M., Pérez, L. and Cohen, S. M. (2005) 'Boundary formation in the *Drosophila* wing: Functional dissection of Capricious and Tartan', *Developmental Dynamics*. John Wiley & Sons, Ltd, 233(3), pp. 804–810. doi: 10.1002/dvdy.20384.

MingKuan, S. and Wei, X. (2012) 'Cell adhesion molecules in *Drosophila* synapse development and function', *Sci China Life Sci*, 55(1), pp. 20–26. doi: 10.1007/s11427-012-4273-3.

Mlodzik, M. and Gehring, W. J. (1987) 'Expression of the *caudal* gene in the germ line of *Drosophila*: Formation of an RNA and protein gradient during early embryogenesis', *Cell*. Elsevier, 48(3), pp. 465–478. doi: 10.1016/0092-8674(87)90197-8.

Moreno, E. and Morata, G. (1999) '*caudal* is the Hox gene that specifies the most posterior Drosophile segment', *Nature*, 400(6747), pp. 873–877. doi: 10.1038/23709.

Mundade, R. *et al.* (2014) 'Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond', *Cell Cycle*, 13(18), pp. 2847–2852. doi: 10.4161/15384101.2014.949201.

Mutanen, M., Kaitala, A. and Mönkkönen, M. (2006) 'Genital variation within and between three closely related *Euxoa* moth species: Testing the lock-and-key hypothesis', *Journal of Zoology*, 268(2), pp. 109–119. doi: 10.1111/j.1469-7998.2005.00029.x.

Nagy, O. *et al.* (2018) 'Correlated Evolution of Two Copulatory Organs via a Single *cis*-Regulatory Nucleotide Change', *Current Biology*, 28, pp. 3450–3457. doi: 10.1016/j.cub.2018.08.047.

Nakagawa, Y. *et al.* (2011) 'Spatial and temporal requirement of Defective proventriculus activity during *Drosophila* midgut development', *Mechanisms of Development*. Elsevier, 128(5–6), pp. 258–267. doi: 10.1016/J.MOD.2011.02.003.

Nakagoshi, H. *et al.* (1998) 'A novel homeobox gene mediates the Dpp signal to establish functional specificity within target cells', *Genes and Development*, 12(17), pp. 2724–2734. doi: 10.1101/gad.12.17.2724.

Nelson, C. E., Hersh, B. M. and Carroll, S. B. (2004) 'The regulatory content of intergenic DNA shapes genome architecture', *Genome Biology*, 5, p. R25. Available at: http://genomebiology.com/2004/5/4/R25.

Nitta, K. R. *et al.* (2015) 'Conservation of transcription factor binding specificities across 600 million years of bilateria evolution', *eLIFE*, 4, p. e04837. doi: 10.7554/eLife.04837.001.

Oliveri, P. and Davinson, E. H. (2007) 'Built to run, not fail', *Science*, 315, pp. 1510–1511.

Pandey, R. V. *et al.* (2011) 'PoPoolation DB: a user-friendly web-based database for the retrieval of natural polymorphisms in *Drosophila'*. doi: 10.1186/1471-2156-12-27.

Panne, D., Maniatis, T. and Harrison, S. C. (2007) 'An Atomic Model of the *Interferon-b* Enhanceosome', *Cell*, 129, pp. 1111–1123. doi: 10.1016/j.cell.2007.05.019.

Paré, Adam C. *et al.* (2019) 'An LRR Receptor-Teneurin System Directs Planar Polarity at Compartment Boundaries', *Developmental Cell*, 51(2), pp. 208-221.e6. doi: 10.1016/j.devcel.2019.08.003.

Paré, Adam C *et al.* (2019) 'An LRR Receptor-Teneurin System Directs Planar Polarity at Compartment Boundaries', *Developmental Cell*, 51. doi: 10.1016/j.devcel.2019.08.003.

Pare, A. C. and Zallen, J. A. (2020) 'Cellular, molecular, and biophysical control of epithelial cell intercalation', *Current topics in developmental biology*, pp. 167–193.

Paulson, D.R. (1974) "Reproductive isolation in damselflies," *Systematic Biology*, 23(1), pp. 40–49. Available at: https://doi.org/10.1093/sysbio/23.1.40.

Payne, J. L. and Wagner, A. (2014) 'The robustness and evolvability of transcription factor binding sites', *Science*. American Association for the Advancement of Science, 343(6173), pp. 875–877. doi: 10.1126/SCIENCE.1249046.

Peng, P. C. *et al.* (2019) 'The role of chromatin accessibility in *cis*-regulatory evolution', *Genome Biology and Evolution*, 11(7), pp. 1813–1828. doi: 10.1093/gbe/evz103.

Perry, M. W., Boettiger, A. N. and Levine, M. (2011) 'Multiple enhancers ensure precision of gap gene-expression patterns in the *Drosophila* embryo', *PNAS*, 108(33), pp. 13570–13575. doi: 10.1073/pnas.1109873108.

Pfeiffer, B. D. *et al.* (2008) 'Tools for neuroanatomy and neurogenetics in *Drosophila'*, *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 105(28), pp. 9715–9720. doi: 10.1073/pnas.0803697105.

Phanstiel, D. H. (2021) 'Sushi: Tools for visualizing genomics data. R package version 1.32.0.'

Putnam, N. H. *et al.* (2007) 'Repertoire and Genomic Organization Sea Anemone Genome Reveals Ancestral Eumetazoan Gene', *Science*, 317, pp. 86–94. doi: 10.1126/science.1139158.

Raghavan, S. and White, R. A. H. (1997) 'Connectin Mediates Adhesion in *Drosophila'*, *Neuron*, 18, pp. 873–880.

Rebeiz, M. and Tsiantis, M. (2017) 'Enhancer evolution and the origins of morphological novelty', *Current Opinion in Genetics and Development*. Elsevier Ltd, 45, pp. 115–123. doi: 10.1016/j.gde.2017.04.006.

Rice, G. *et al.* (2019) 'A standardized nomenclature and atlas of the male terminalia of *Drosophila melanogaster*', *Fly*. Taylor & Francis, 13(1–4), pp. 51–64. doi: 10.1080/19336934.2019.1653733.

Risca, V. I. and Greenleaf, W. J. (2015) 'Unraveling the 3D genome: genomics tools for multiscale exploration', *Trends in Genetics xx*, pp. 1–16. doi: 10.1016/j.tig.2015.03.010.

Ritchie, M. G. (2007) 'Sexual Selection and Speciation', *Annual Review of Ecology,*

Evolution, and Systematics, 38(1), pp. 79–102. doi: 10.1146/annurev.ecolsys.38.091206.095733.

Rivera, J. *et al.* (2019) 'REDfly: the transcriptional regulatory element database for *Drosophila'*, *Nucleic Acids Research*, 47. doi: 10.1093/nar/gky957.

Robertson, H. M. (1988) 'Mating Asymmetries and Phylogeny in the *Drosophila melanogaster* Species Complex', *Pacific Science*, 42, pp. 72–80.

Roscito, J. G. *et al.* (2018) 'Phenotype loss is associated with widespread divergence of the gene regulatory landscape in evolution', *Nature Communications*, 9, p. 4737. doi: 10.1038/s41467-018-07122-z.

Rothberg, J. M. *et al.* (1990) 'slit: an extracellular protein necessary for development of midline glia and commissural axon pathways contains both EGF and LRR domains', *Genes & development*, 4, pp. 2169–2187.

Ryan, J. F. *et al.* (2006) 'The cnidarian-bilaterian ancestor possessed at least 56 homeoboxes: evidence from the starlet sea anemone, Nematostella vectensis', *Genome Biology*, 7(7), p. 64. doi: 10.1186/gb-2006-7-7-r64.

Sakurai, Kayoko T. *et al.* (2007) 'Differential control of cell affinity required for progression and refinement of cell boundary during *Drosophila* leg segmentation', *Developmental Biology*, 309(1), pp. 126–136. doi: 10.1016/j.ydbio.2007.07.001.

Sánchez, L. *et al.* (1997) 'The genital disc of *Drosophila melanogaster*. II. Role of the genes *hedgehog*, *decapentaplegic* and *wingless'*, *Development Genes and Evolution*, 207(4), pp. 229–241. doi: 10.1007/s004270050111.

Sánchez, L. and Guerrero, I. (2001) 'The development of the *Drosophila* genital disc', *BioEssays*, 23(8), pp. 698–707. doi: 10.1002/bies.1099.

Sanyal, A. *et al.* (2012) 'The long-range interaction landscape of gene promoters ENCODE Encyclopedia of DNA Elements', *Nature*, 489, pp. 109–115. doi: 10.1038/nature11279.

Sawamura, K., Davis, A. W. and Wu, C.-I. (2000) 'Genetic analysis of speciation by means of introgression into *Drosophila melanogaster*', *PNAS*, 97(6), pp. 2652–2655. Available at: www.pnas.orgcgidoi10.1073pnas.050558597.

Schneider, C. A., Rasband, W. S. and Eliceiri, K. W. (2012) 'NIH Image to ImageJ: 25 years of Image Analysis', *Nature methods*. NIH Public Access, 9(7), p. 671. doi: 10.1038/NMETH.2089.

Schwaiger, M. *et al.* (2014) 'Evolutionary conservation of the eumetazoan gene regulatory landscape', *Genome Research*, 24(4), pp. 639–650. doi: 10.1101/gr.162529.113.

Seeholzer, L. F. *et al.* (2018) 'Evolution of a central neural circuit underlies *Drosophila* mate preferences', 559, pp. 564–573. doi: 10.1038/s41586-018-0322-9.

Shaner, N. C. *et al.* (2004) 'Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma sp.* red fluorescent protein', *Nature Biotechnology*. Nature Publishing Group, 22(12), pp. 1567–1572. doi: 10.1038/nbt1037.

Shaner, N. C. *et al.* (2013) 'A bright monomeric green fluorescent protein derived from *Branchiostoma lanceolatum*', *Nature Methods*, 10(5), pp. 407–409. doi: 10.1038/nmeth.2413.

Shapiro, A. M. and Porter, A. H. (1989) 'The Lock-And-Key Hypothesis: Evolutionary And Biosystematic Interpretation Of Insect Genitalia', *Annual Review of Entomology*, 34(1), pp. 231–245. doi: 10.1146/annurev.ento.34.1.231.

Shaw, P. J. *et al.* (2002) 'Coevolution in bicoid-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera', *Evolution and Development*, 4(4), pp. 265–277. doi: 10.1046/j.1525-142X.2002.02016.x.

Shingleton, A. W. *et al.* (2009) 'Many ways to be small: different environmental regulators of size generate distinct scaling relationships in *Drosophila melanogaster*', *Proc. R. Soc. B*, 276, pp. 2625–2633. doi: 10.1098/rspb.2008.1796.

Shishido, E., Takeichi, M. and Nose, A. (1998) '*Drosophila* synapse formation: Regulation by transmembrane protein with leu-rich repeats, capricious', *Science*, 280(5372), pp. 2118–2121. doi: 10.1126/science.280.5372.2118.

Shlyueva, D., Stampfel, G. and Stark, A. (2014) 'Transcriptional enhancers: From properties to genome-wide predictions', *Nature Reviews Genetics*. Nature Publishing Group, 15(4), pp. 272–286. doi: 10.1038/nrg3682.

Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple

sequence alignments using Clustal Omega', *Molecular Systems Biology*, 7(539). doi: 10.1038/msb.2011.75.

Simakov, O. *et al.* (2013) 'Insights into bilaterian evolution from three spiralian genomes', *Nature*, 493, pp. 526–531. doi: 10.1038/nature11696.

Simmons, L. W. (2014) 'Sexual selection and genital evolution', *Austral Entomology*, 53(1), pp. 1–17. doi: 10.1111/aen.12053.

Slattery, M. *et al.* (2011) 'Cofactor binding evokes latent differences in DNA binding specificity between Hox proteins', *Cell*, 147(6), pp. 1270–1282. doi: 10.1016/j.cell.2011.10.053.

Small, S. and Arnosti, D. N. (2020) 'Transcriptional Enhancers in *Drosophila'*, *Genetics*, 216, pp. 1–26. doi: 10.1534/genetics.120.301370.

Stewart, A. J., Hannenhalli, S. and Plotkin, J. B. (2012) 'Why Transcription Factor Binding Sites Are Ten Nucleotides Long', *Genetics*, 192, pp. 973–985. doi: 10.1534/genetics.112.143370.

Stieper, B. C. *et al.* (2008) 'Imaginal discs regulate developmental timing in *Drosophila melanogaster*', *Developmental biology*. Dev Biol, 321(1), pp. 18–26. doi: 10.1016/J.YDBIO.2008.05.556.

Strähle, U. and Rastegar, S. (2008) 'Conserved non-coding sequences and transcriptional regulation', *Brain Research Bulletin*, 75(2–4), pp. 225–230. doi: 10.1016/j.brainresbull.2007.11.010.

Suryamohan, K. and Halfon, M. S. (2015) 'Identifying transcriptional *cis*-regulatory modules in animal genomes', *Wiley Interdisciplinary Reviews: Developmental Biology*, 4(2), pp. 59–84. doi: 10.1002/wdev.168.

Takahara, B. and Takahashi, K. H. (2015) 'Genome-Wide Association Study on Male Genital Shape and Size in *Drosophila melanogaster*', *PLoS ONE*, 10(7), p. e0132846. doi: 10.1371/journal.pone.0132846.

Tanaka, K. M. *et al.* (2015) 'Genetic architecture and functional characterization of genes underlying the rapid diversification of male external genitalia between *Drosophila simulans* and *Drosophila mauritiana.*', *Genetics*. Genetics, 200(1), pp. 357–69. doi: 10.1534/genetics.114.174045.

The Uniprot Consortium (2021) ‘UniProt: the universal protein knowledgebase in 2021’, *Nucleic Acids Research*, 49, pp. D480–D489.

Thomas, S. *et al.* (2011) ‘Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development’, *Genome Biology*, 12, p. R43. doi: 10.1186/gb-2011-12-5-r43.

Thurmond, J. *et al.* (2019) ‘FlyBase 2.0: the next generation.’, *Nucleic acids research*. Oxford University Press, 47(D1), pp. D759–D765. doi: 10.1093/nar/gky1003.

Torres-Oliva, M. *et al.* (2021) ‘Differences in orthodenticle expression promote ommatidial size variation between *Drosophila* species’, *bioRxiv*, p. 2021.03.17.435774. Available at: https://doi.org/10.1101/2021.03.17.435774.

True, J. R. *et al.* (1997) ‘Quantitative genetic analysis of divergence in male secondary sexual traits between *Drosophila simulans* and *Drosophila mauritiana*’, *Evolution*. Wiley-Blackwell, 51(3), pp. 816–832. doi: 10.1111/j.1558-5646.1997.tb03664.x.

True, J. R. (2008) ‘Combing evolution’, *Evolution and Development*, 10(4), pp. 400–402. doi: 10.1111/j.1525-142X.2008.00250.x.

Tsuboi, A. *et al.* (2017) ‘Inference of Cell Mechanics in Heterogeneous Epithelial Tissue Based on Multivariate Clone Shape Quantification’, *Frontiers in Cell and Developmental Biology*, 5. doi: 10.3389/fcell.2017.00068.

Vincent, B. J. *et al.* (2019) ‘An Atlas of Transcription Factors Expressed in Male Pupal Terminalia of *Drosophila melanogaster*’, *G3 Genes*|*Genomes*|*Genetics*, 9, pp. 3961–3972. doi: 10.1534/g3.119.400788.

Wang, J. and Dahmann, C. (2020) ‘Establishing compartment boundaries in *Drosophila* wing imaginal discs_ An interplay between selector genes, signaling pathways and cell mechanics’, *Seminars in Cell and Developmental Biology*, 107, pp. 161–169. Available at: https://reader.elsevier.com/reader/sd/pii/S1084952119302393?token=9717DF74BBFAC1897B5BFE07F0E9B57691ED02698CC86B41C709CFB24ADFDE7A489FF42D9D40AB26EAECD75AD445A57E&originRegion=eu-west-

1&originCreation=20211027162110.

Weirauch, M. T. *et al.* (2014) ‘Determination and inference of eukaryotic transcription factor sequence specificity.’, *Cell*. NIH Public Access, 158(6), pp. 1431–1443.

Wray, A. G. (2007) ‘The evolutionary significance of *cis*-regulatory mutations’, *Nature*, 8(3), pp. 206–216.

Xi, H. *et al.* (2007) ‘Identification and Characterization of Cell Type-Specific and Ubiquitous Chromatin Regulatory Structures in the Human Genome’, *PLOS Genetics*, 3(8), p. e136. doi: 10.1371/journal.pgen.0030136.

Yan, F. *et al.* (2020) ‘From reads to insight: A hitchhiker’s guide to ATAC-seq data analysis’, *Genome Biology*, 21. doi: 10.1186/s13059-020-1929-3.

Yorimitsu, T., Kiritooshi, N. and Nakagoshi, H. (2011) ‘Defective proventriculus specifies the ocellar region in the *Drosophila* head’, *Developmental Biology*. Academic Press, 356(2), pp. 598–607. doi: 10.1016/J.YDBIO.2011.06.015.

Young, R. S. *et al.* (2012) ‘Identification and Properties of 1,119 Candidate LincRNA Loci in the *Drosophila melanogaster* Genome’, *Genome Biology and Evolution*, 4(4), pp. 427–442. doi: 10.1093/gbe/evs020.

Supplementary Material

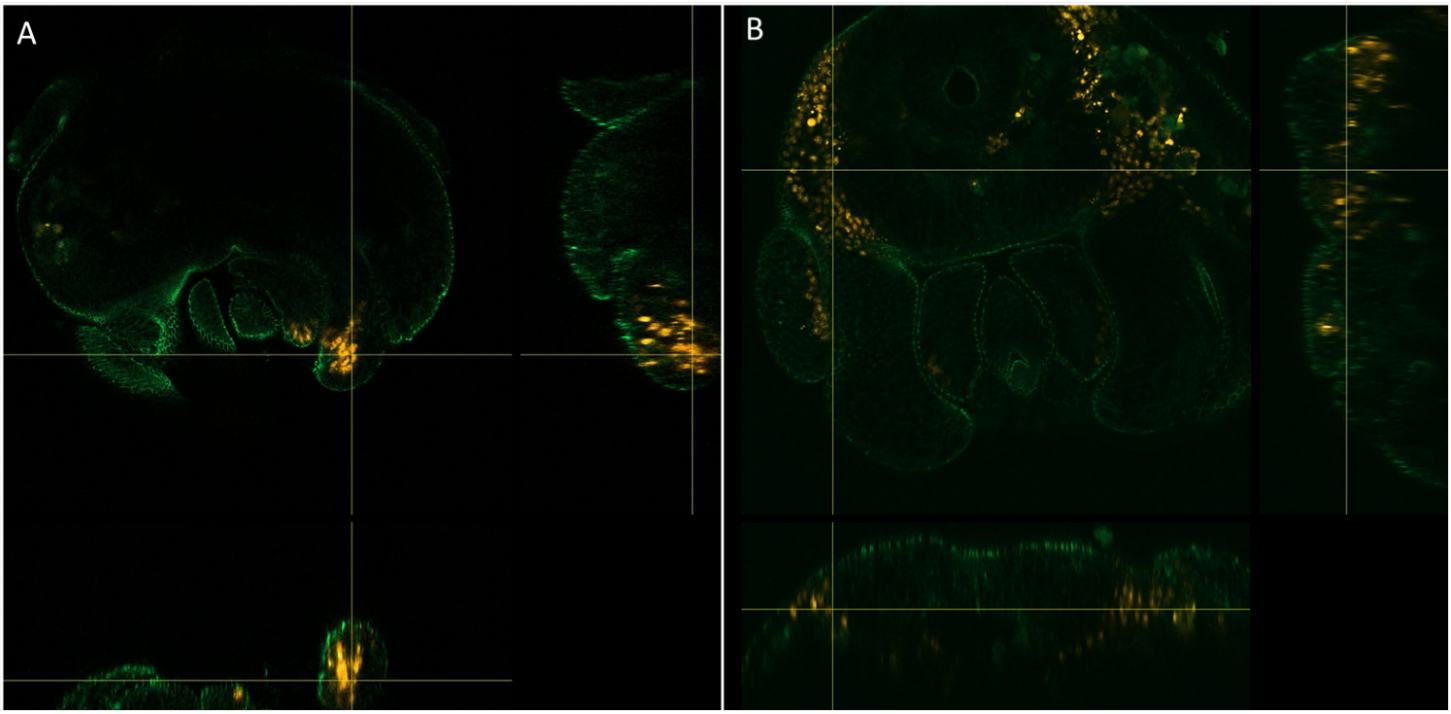
Chapter 2

Supplementary Table 1. Enhancer candidate primers. Primer used for the amplification and posterior cloning of the enhancer candidates. All Fwd primers carried an extra 5'- CCAC -3' sequence for pENTRY/D-Topo cloning.

Candidate	Fwd	Rev
U1	5'- TTGTAAGGCAGCCTACTGGAT -3'	5'- GCGATGTTGGTATGGGGTCT -3'
U2	5'- CAACGAGCGGGTGCATTAAAA -3'	5'- TGATCATCGAAGATGGATGCGA -3'
U3	5'- GGTCTTGTGCCCAGAGCTTC -3'	5'- CAATGAATACTTTGTACCCCTTTGC -3'
U4	5'- GAACTTGGATCTGGGCTCTG -3'	5'- AAAACTAAAATTCCAGTACCCCACT -3'
U5	5'- TTTTCTTCTTCAATGCCCGACC -3'	5'- CAATCACAAAACTCATTTGCGGC -3'
U8	5'- CGTCGCTTATGAGTCGGACA -3'	5'- CTCGATGGAGAACTCAGTGTG -3'
trn intron	5'- GGATAAGTTTCACTTGTGGCATACA -3'	5'- TAGTGCTGCGGGTAGACAAAC -3'
D2	5'- CAGACTCGCACTTACTCATACG -3'	5'- CTAATAGTGGAGGCTCGGGATG -3'
D3	5'- TTCCCATGAACCTACCGTCG -3'	5'- ATTGCACAAGAGATTCAAGGG -3'
D4	5'- TGCCCAAAATAAAGTTCGATGGC -3'	5'- GCCTCTTCTATTTGTCTGTGCC -3'
D5	5'- TAGAAACTTAGCACCGGCGAC -3'	5'- GCGAACTAAGTTTTCTTATCGGC -3'
D6	5'- GTAGAATTTATTGCAGACCCGC -3'	5'- ATTGCCCGCATTGTAAAAGG -3'

Supplementary Table 2. Coordinates for enhancer candidates. Coordinates obtained from flybase.org *D. melanogaster* genome version r6.30.

D. melanogaster	Coordinates
U1	3L: 13,09,005 - 13,093,004
U2	3L: 13,092,685 - 13,095,691
VT030010	3L: 13,095,743 - 13,098,334
U3	3L: 13,098,317 - 13,100,976
U4	3L: 13,100,690 - 13,103,689
U5	3L: 13,103,455 - 13,106,444
U8	3L: 13,111,400 - 13,113,999
trn intron	3L: 13,114,746 - 13,115,193
VT030021	3L: 13,118,026 - 13,120,128
D2	3L: 13,122,950 - 13,125,953
D3	3L: 13,125,783 - 13,128,815
D4	3L: 13,128,618 - 13,131,652
D5	3L: 13,131,500 - 13,134,492
D6	3L: 13,134,226 - 13,135,927



Supplementary Figure 2.1. Orthogonal views of U4 and U8 expression patterns. (A) Orthogonal view of expression pattern driven by U8 at 46 hAPF. Expression pattern driven by U8 in the surstylus is localised in the internal part. (B) Orthogonal view of expression pattern driven by U4 at 46 hAPF. Expression driven by U4 is mainly localised in the intersection between the anal plates and the epandrial ventral lobes and surstyli, surrounding the anal plates.

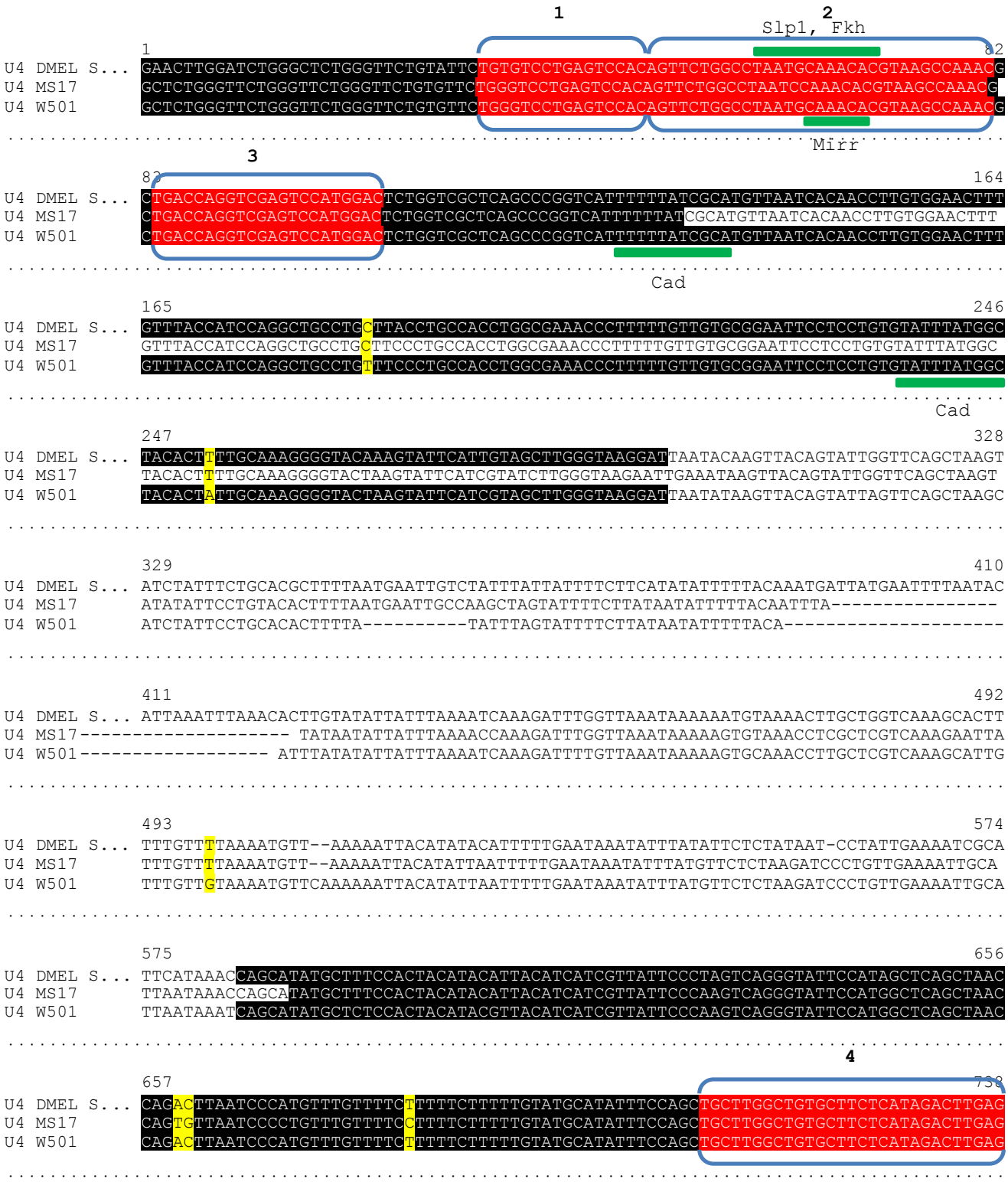
Chapter 3

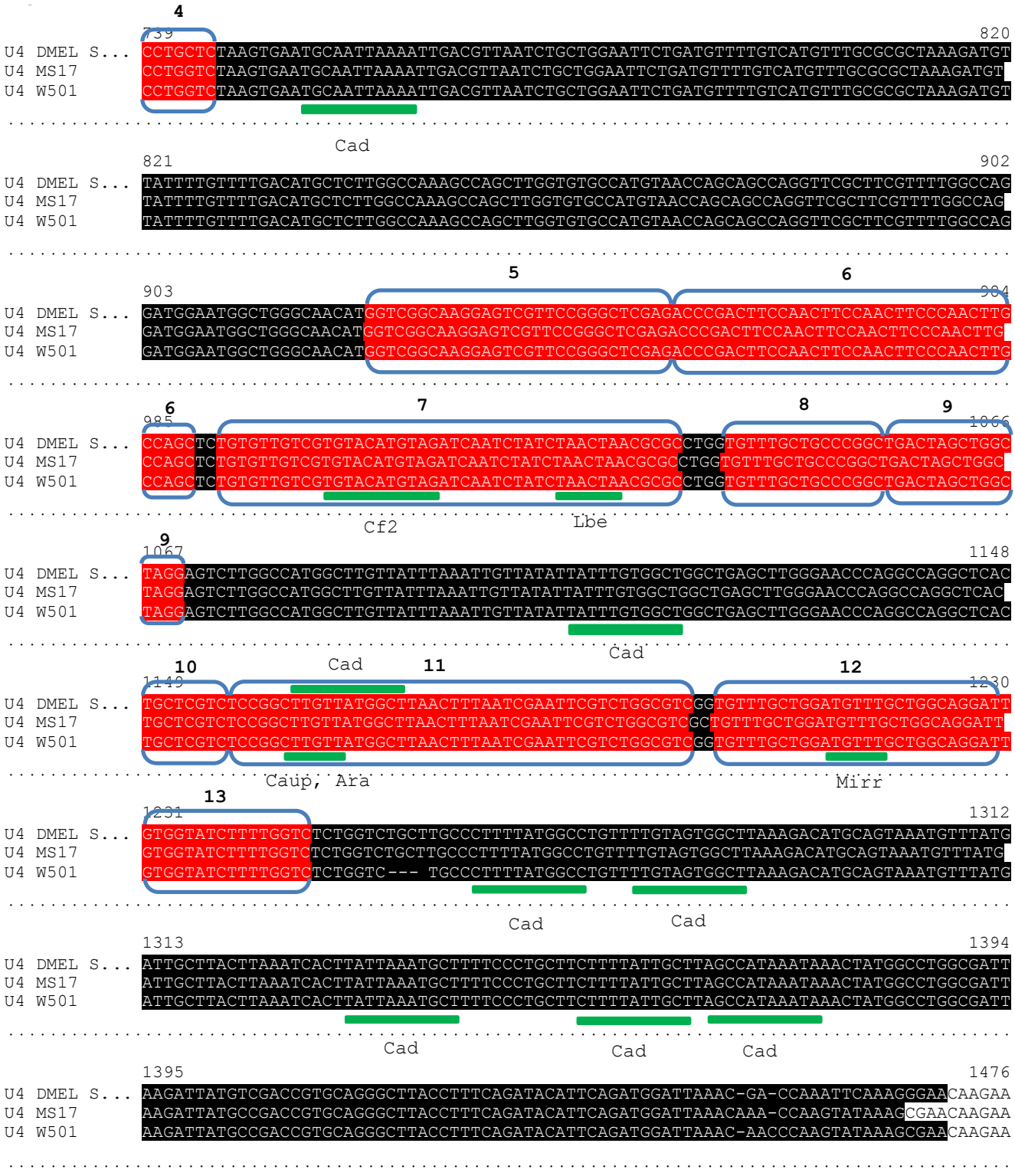
Supplementary Table 3. Transcription factor footprints of U4. Footprints found using HINT-ATACseq with the accessible chromatin peaks within U4. Chromosome location in the genome used is delimited between the Chr start and Chr end.

Chromosome	Chr start	Chr end	Sequence
Scf_3L	12785419	12785434	TGGGTCCTGAGTCCAC
Scf_3L	12785467	12785467	AGTTCTGGCCTAATGCAAACACGTAAGCCAAAC
Scf_3L	12785470	12785491	TGACCAGGTCGAGTCCATGGAC
Scf_3L	12786052	12786087	TGCTTGGCTGTGCTTCTCATAGACTTGAGCCTGGTC
Scf_3L	12786266	12786294	GGTCGGCAAGGAGTCGTTCCGGGCTCGAG
Scf_3L	12786295	12786331	ACCCGACTTCCAACTTCCAACTTCCCAACTTGCCAGC
Scf_3L	12786334	12786377	TGTGTTGTCGTGTACATGTAGATCAATCTATCTAACTAACGCGC
Scf_3L	12786382	12786396	TGTTTGCTGCCCGGC
Scf_3L	12786397	12786412	TGACTAGCTGGCTAGG
Scf_3L	12786491	12786498	TGCTCGTC
Scf_3L	12786499	12786542	TCCGGCTTGTTATGGCTTAACTTTAATCGAATTCGTCTGGCGTC
Scf_3L	12786545	12786571	TGTTTGCTGGATGTTTGCTGGCAGGAT
Scf_3L	12786572	12786588	TGTGGTATCTTTTGGTC

Supplementary Table 4. Transcription factor BSs. Binding sites found using HINT-ATACseq with the footprints found in U4. Chromosome location in the genome used is delimited between the Chr start and Chr end. Transcription factor

Chromosome	Chr start	Chr end	Binding Site
Scf_3L	12785450	12785455	mirr
Scf_3L	12785445	12785456	slp1
Scf_3L	12785444	12785455	fkh
Scf_3L	12786365	12786371	lbe
Scf_3L	12786344	12786354	Cf2
Scf_3L	12786505	12786510	caup
Scf_3L	12786505	12786510	ara
Scf_3L	12786555	12786560	mirr


1
2
Slp1, Fkh
Mirr
3
Cad
Cad
4
U4 DMEL S...
U4 MS17
U4 W501


4
739
820
U4 DMEL S...
U4 MS17
U4 W501
Cad
821
902
5
6
903
904
7
8
9
985
1066
Cf2
Lbe
1067
1148
Cad
Cad
10
11
12
1149
1230
Caup, Ara
Mirr
13
1231
1312
Cad
Cad
1313
1394
Cad
Cad
Cad
1395
1476

```
            1477                                                                1558
U4 DMEL S... TAAAAATGTTGGAATTCTGCATTATACAATCTTGAAAAAAAAAG-TTTGCCCTAACAAAGCATTTATTTGATATACTAAACT
U4 MS17      TAAAAATGTTGGATTT---CATTATAAGATCTTGGAAAAAAAAACTTTGCCCTAACAAAGCATTTATTTGATATACTAAAGT
U4 W501      TAAAAATGTTGGATTTGTGCATTATAAGATCTTGGAAAAAAAAC-TTTTCCCTAACAAAGCATTTATTTGATATACTAAACT

.....

            1559                                                                1640
U4 DMEL S... GAATTTAGAATCTCACATACTTCGACAACAAGATATCAATCAGAAACTTTTGAAATTGTATCTTAAAAAT------------
U4 MS17      GAATTTAGGATCTCACATATTTCGACTACAAGTTATCAATCCGAAACTATTGACATTGTATCTTTATCTT------------
U4 W501      GAAATTAGGATCTCACATATTTCGACTAGAAGTTATCAATCAGAAACTATTGAAATTGTATCCTAGAAATCCTAATATATCC

.....

            1641                                                                1722
U4 DMEL S...---- GAAAAAAAAAAAAACAATTTTCCCATTGCTTTTCTAATTATATATAAAATATTTTTCCAGCCAGGTTTTGTAGCTCCA
U4 MS17      ----TAAAATCGAAA---------- CCAATGCTTTTCTAATAATATATTAAATATTTTCCCAGCCAGGTTTTGTAGCTCCA
U4 W501      TAAAAATTAATAATAATACAATTTACTTAATGCTTTTCTAATAATATATTAAATATTTTCCCAGCCAGGTTTTGTAGCTCCA

.....

            1723                                                                1804
U4 DMEL S... TTTAGTAAGTCTCATTTGGCGCAAAACTGAGTACATTTGTATTGCCCTGGAGACCATTCGTGGACCACACA-----------
U4 MS17      TTTAGTAAGTCTTATTTGGAGCAAAGCAAAGTACATTTGTATTGCTCTGGAGACCATTCGTGGACCACACACTTCACTCTGG
U4 W501      TTTAGTAAGTCTTATTTGGAGCAAATCAGAGTACATTTGTATTGCTCTGGAGGCCATTCGTGGACCACACACTTCACTCTGG

.....

            1805                                                                1886
U4 DMEL S... CTGCCAAAAAAGGCGATCACAAAAACTGGAGGAAAAGAAAGCAAAGCTGAGCAGCTATAAATCTCAAGCTACTATCTCAATG
U4 MS17      CTGCCAAAAAAGGCGATCACAAAAACTGTAGGAAAAGAAAGCAGAGCTGAGCAGCTATAAATCTCAAGCTACTATCTCAATG
U4 W501      CTGCCAAAAAAGGCGATCACAAAAACTGGAGGAAAAGAAAGCAGAGCTGAGCAGCTATAAATCTCAAGCTACTATCTCAATG

.....

            1887                                                                1968
U4 DMEL S... CACACACACTTTTAGCCATTATTACTGCAACATTAAAACACCCAAATGGCCAGGCTTTTGTTGGCAAATCCCAAAGCCAGAT
U4 MS17      CACACACACTTTTAGCCATTATTACTGCAACATTAAGACACCCAAATGGCCAGGCTTTTGTTGGCAGATCCCAAAGCCAGAT
U4 W501      CACACACACTTTTAGCCATTATTACTGCAACATTAAGACACCCAAATGGCCAGGCTTTTGTTGGCAGATCCCAAAGCCAGAT

.....

            1969                                                                2050
U4 DMEL S... GCCAGGCCATAACTGCCATTAACATTGGCACTTAATTTACTTTTAGCCAAAAACCGAAAACCAAAACAAACGCAAACAGACA
U4 MS17      GCCAGGCCATAACTGCCATTAACATTGGCACTTAATTTACTTTTAGCCAAAAACCGAAAACCAAAACAAACGCAAACAGACA
U4 W501      GCCAGGCCATAACTGCCATTAACATTGGCACTTAATTTACTTTTAGCCAAAAACCGAAAACCAAAACAAACGCAAACAGACA

.....

            2051                                                                2132
U4 DMEL S... AACAGAAGCGAATACACAACACACAAAATGACGCACTGCTATCTTATTATATTTGGCCAACAATAAGGAGCCACGAAAAAAT
U4 MS17      AACAGAAGCGAATACACAACACACAAAATGACGCACTGCTATCTTATTATATTTGGCCAACAATAAGGAGCCACGAAAAAAT
U4 W501      AACAGAAGCGACTACACAACACACAAAATGACGCACTGCTATCTTATTATATTTGGCCAACAATAAGGAGCCACAAAAAAAT

.....

            2133                                                                2214
U4 DMEL S... ACAATGAAAAATAAAAATATATACCAGGGGAACAACAATTAAAACCATTTTCAAGCAAGCAAACAAATGGCGCAGAATGGGA
U4 MS17      ACAATGAAAAATAAAAATATATACCAGGGGAACAACAATTAAAACCATTTTCAAGCAAGCAAACAAATGGCGCAGAATGGGA
U4 W501      ACAATGAAAAATAAAAATATATACCAGGGGAACAACAATTAAAACCATTTTCAAGCAAGCAAACAAATGGCGCAGAATGGGA

.....
```

```
            2215                                                                                2296
U4 DMEL S... AAAAGTCGCATTCCCATTGTCCGGCAAACCTCCTGCGCTTATTTATTGGCGTCCTGTCCAGTTTGGACCATTGTTTGTTGTC
U4 MS17      AAAAGTCGCATTCCCATTGTCCGGCAAACCTCCTGCGCTTATTTATTGGCGTCCTGTCCAGTTTGGACCATTGTTTGCTGTC
U4 W501      AAAAGTCGCATTCCCATTGTCCGGCAAACCTCCTGCGCTTATTTATTGGCGTCCTGTCCAGTTTGGACCATTGTTTGCTGTC
.....................................................................................................

            2297                                                                                2378
U4 DMEL S... CGGGGGGAAACGAGGCCAAGAAAAGGAAAGTAAACCAGGCAAAAAGGCCACCAAATAGGATAGAAAGAAAGTAGCTCAATCT
U4 MS17      GGGGGGGAATCGAGGCCAAGAAAAGGAAAGTAAACCAGGCAAAAAGGCCACCAAATAGGATAGAAAGAAAGTAGCTCAATCT
U4 W501      CGGGGGGAAACGAGGCCAAGAAAAGGAAAGTAAACCAGGCAAAAAGGCCACCAAATAGGATAGAAAGAAAGTAGCTCAATCT
.....................................................................................................

            2379                                                                                2460
U4 DMEL S... CCTTCGTTCTCTCGATATCTTAACAATTTTCAATTTTAGCTTTTGTTCTCGAAAGCAAACAATTCGAGAAACAAGATGAAAG
U4 MS17      CCTTCGTTCTCTCGATATCTTAACAATTTTCAATTTTAGCTTTTGTTCTCGAAAGCAAACAATTCGAGAAACAAGCTGAAAG
U4 W501      CCTTCGTTCTCTCGATATCTTAACAATTTTCAATTTTAGCTTTTGTTCTCGAAAGCAAACAATTCGAGAAACAATCTGAAAG
.....................................................................................................

            2461                                                                                2542
U4 DMEL S... CAAAGCGGTTGGAACTGAAGGATACCCTAT--AA--AAATGTGCGGTACATAAACAAAACTTTTTAATTTGTTTAAAGTCTT
U4 MS17      CAAAGCGGTCCGAACTGAATGATACCCTAT--AAACAAATGTGCGGTACCTAACCAACTTTTTTGAATTTACTTAAAGACTT
U4 W501      CAAAGCGGTCGGAACTGAATGATACCCTATAAAA--AAATGTGCGGTACCTAACCAACTTTTTTGAATTTGCTTAAAGACTT
.....................................................................................................

            2543                                                                                2624
U4 DMEL S... ATGCTTGATGCTCATTATGG---------- AAATCATACCCACATTCTACTATGCTGAATTCGAATTATACCATAAACTAT
U4 MS17      ATTATTTATGTTCATTAGAGAAATACATATAAAATCTCACTTACATTCTACTATGTTGCTTTCGAATTATACTATAAACTAT
U4 W501      -CATTAGAGAAATACATATA---------- AAATCTCACTTACATTCTACTATGTTGCTTTCAAATTATATTATAAACTAT
.....................................................................................................

            2625                                                                                2706
U4 DMEL S... ATTTAACTTAAGCACCAAATTCAGTTTTTAAAATAATTTAATTTAAAACGAAGTTATTTTATAGACGCTATATACCCTATAA
U4 MS17      -----ACATAAACACTAAATTCAGTTTTTAAAATAATT---------- GAATTATTTCATAGACCCTATATACCCTATAA
U4 W501      -----ACATAAACACTAAGTCAAGTTTTTAAAATAATT---------- GAATTATTTCATAGACCCTATATACCCAATAA
.....................................................................................................

            2707                                                                                2788
U4 DMEL S... CTAAGCGGTACGATTCCAGGGGGACAACTAAATCTAAGTGCCTTTATCCTCATCATCGCGACCTTCTTCAATGCTCGTATAC
U4 MS17      CTAAGCGGTGCGATTCCAGGGGGACAACTAAATCTAAGTGCCTTTATCCTCATCATCACGACCTTCTTCAATGTTCGTATAC
U4 W501      CTAAGCGGTACGATTCCAGGGGGACAACTAAATCTAAGTGCCTTTATCCTCATCATCACGACCTTCTTCACTGTTCGTATAC
.....................................................................................................

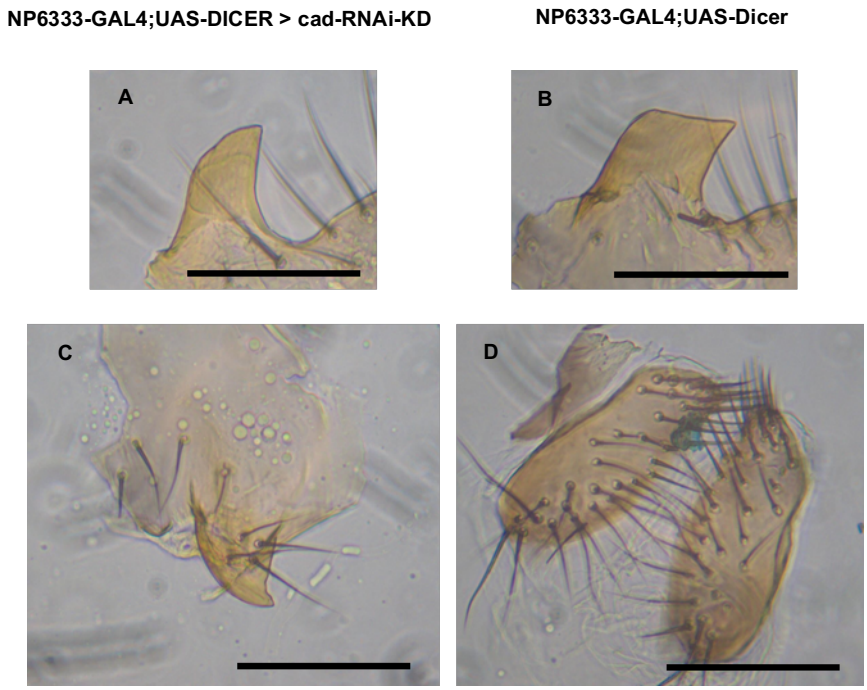
            2789                                                                                2870
U4 DMEL S... GCATTCTGTGGTAGGTCCTTTTCTTCTTCAATGACCGACCCATCACTTGATTTTAATTTAATTTGATTGCATTTCTAAAGCA
U4 MS17      GCATTCTGCGGTAGGTCCTT-- TTTTTCAATGCCCGACCCATCACTTGATTTTAATTTAATTTGATTGCATTTCTAAAGCA
U4 W501      GCATTCTGCGGTAGGTCCTTTT--- TTCAATGCCCGACCCATCACTTGATTTTAATTTAATTTGATTGCATTTCTAAAGCA
.....................................................................................................

            2871                                                                                2952
U4 DMEL S... ATTTCAATTCTACTTTAATTCCTATTTTTGAACCCATAAATAATTGTCTAAGCGCACAGGCGTGAATCGCTTTACCCCTTTG
U4 MS17      ATTTCAATTCTACTTTAATTCCTATTTTTGAACCCATAAATAATTGTCTAAGCGCACGGGCGTGAATCGCTTTACCCCTTTG
U4 W501      ATTTCAATTCTACTTTAATTCCTATTTTTGAACCCATAAATAATTGTCTAAGCGCACGGGCGTGAATCGCTTTACCCCTTTG
.....................................................................................................

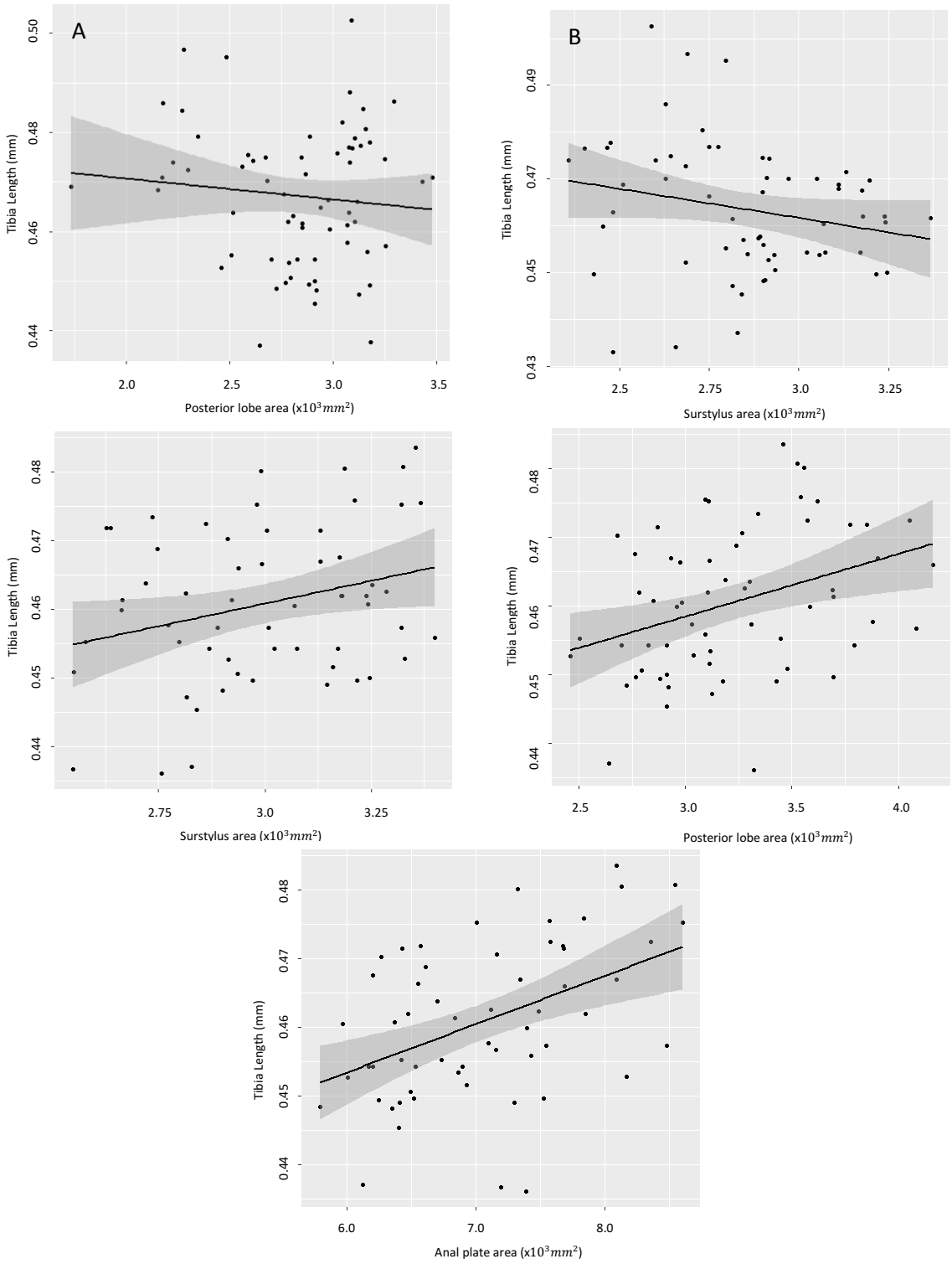
            2953                                                                                3034
U4 DMEL S... TTTGTGCTGCCCGAAATAAAACTCATAATTTTTGTCAAATTCAAAATGGTACTTCGCTATATATAGTGGGGTACTGGAATTT
U4 MS17      TTTGTGCTGCCCGAAATAAAACTCATAATTTTTGTCAAATTCAAAATGGTACTTCGCTATATATAGTGGGGTACTGGAATTT
U4 W501      TTTGTGCTGCCCGAAATAAAACTCAAAATTTTTGTCAAATTCAAAATGGTACTTCGCTATATATAGTGGGGTACTGGAATTT
.....................................................................................................

            3035
U4 DMEL S... TAGTTTT
U4 MS17      TAGTTTT
U4 W501      TAGTTTT
.....................................................................................................
```

Supplementary Figure 3.1. U4 alignment for *D. melanogaster, D. mauritiana* (MS17) and *D. simulans* (w^{501}). Alignment of U4 sequence, highlighted in black are the peaks of accessible chromatin used to look for motifs. Highlighted in red, and by blue brackets are the footprints found using HINT-ATAC. Highlighted in yellow are the fixed differences between *D. mauritiana and D. simulans.* TFBSs are marked with a green bar. Note that the first peak of accessible chromatin is shorter in *D. mauritiana* (MS17) than in *D. simulans* (w^{501}).



Supplementary Figure 3.2. Posterior lobes and anal plates from *cad* KD and NP6333-GAL4;UAS-Dicer. (A, B) Posterior lobes dissected from NP6333-GAL4;UAS-DICER (A) and NP6333-GAL4;UAS-Dicer (B). (C-D) Anal plates from dissected from NP6333-GAL4;UAS-DICER (A) and NP6333-GAL4;UAS-Dicer (B). Black scale equals to 0.1 mm in all panels.



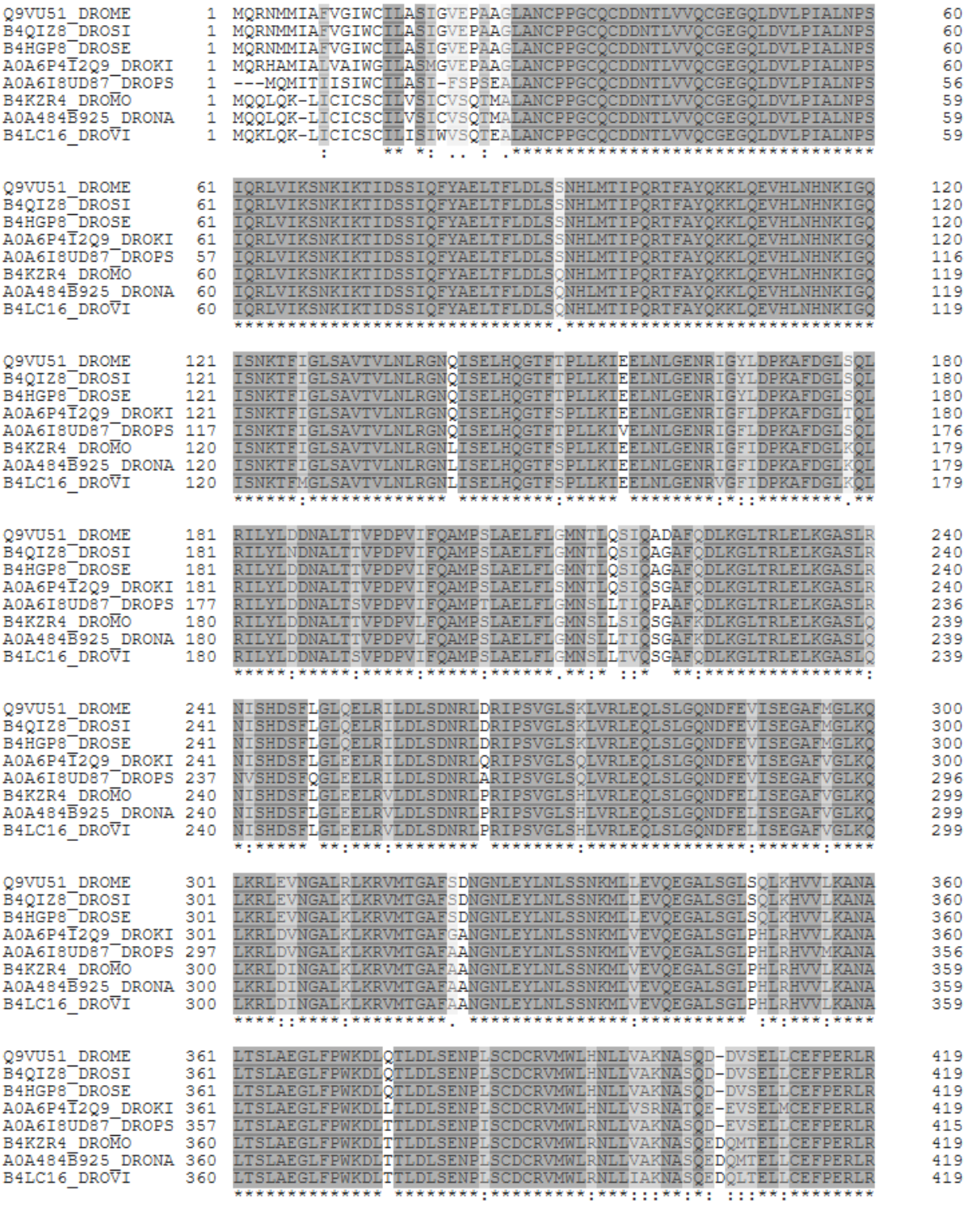
Supplementary Figure 3.3. Correlation between areas measured and tibia lenght. (A, B) Regression analysis performed for posterior lobes and surstyli from NP6333-GAL4;UAS-DICER > cad-RNAi-KD. (C - E) Regression analysis performed for posterior lobes and surstyli from U4-GAL4 > cad-RNAi-KD. There was no correlation found in posterior lobe area vs tibia length (p-value > 0.05) (A) and surstylus area vs tibia length (p-value > 0.05) (A). Correlation was found between surstylus area (C) (p-value < 0.05), posterior lobe area (D) (p-value < 0.01) and anal plate area (E) (p-value < 0.001) vs tibia length.

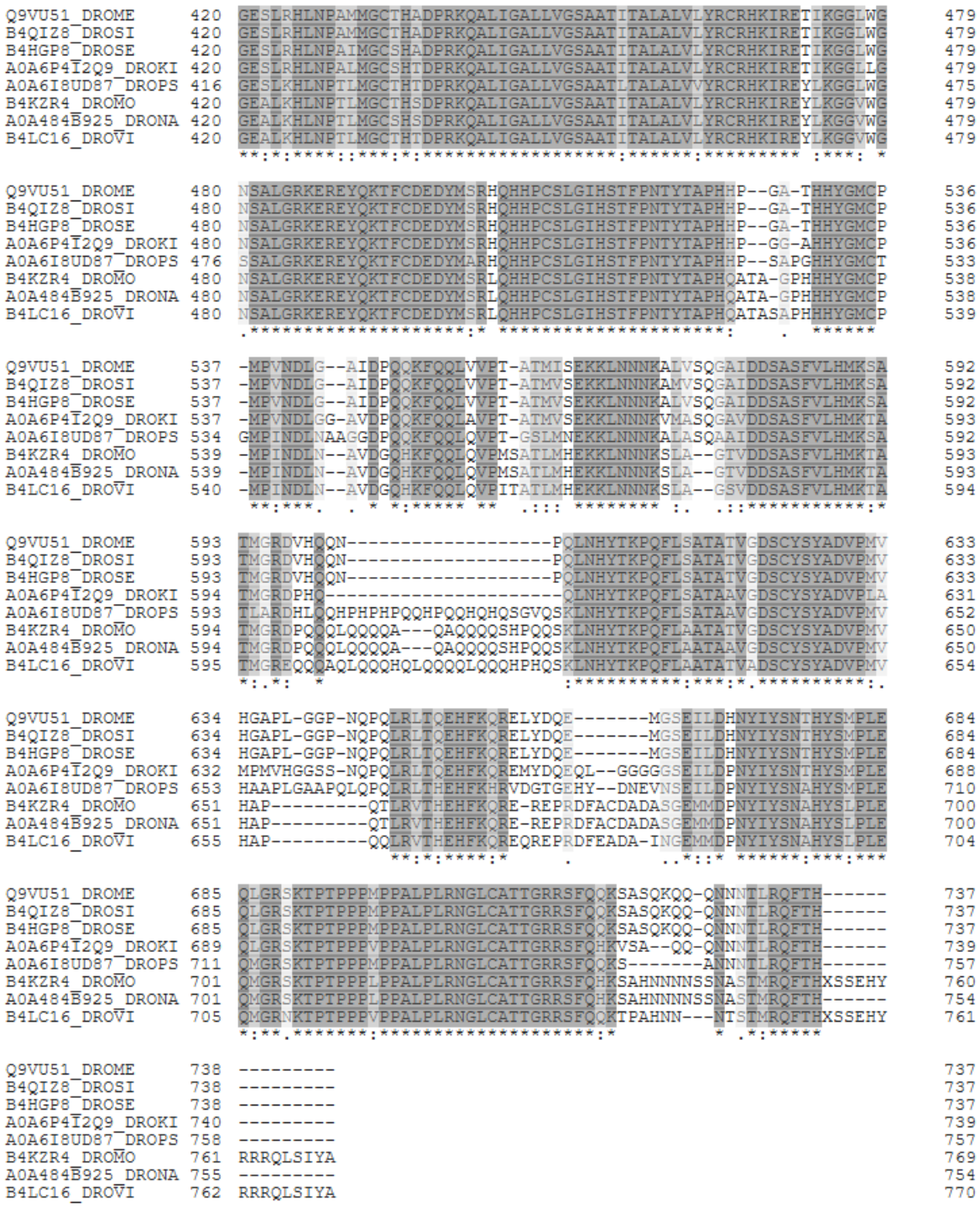
Chapter 4

Chromosome	Dsim						Dmau					
	Coordinate	Ref Base	A	T	C	G	Coordinate	Ref Base	A	T	C	G
Scf_3L/3L	12785400	G	0	0	0	25	12546429	G	0	2	0	12
Scf_3L/3L	12785407	G	0	3	0	26	12546436	G	0	5	0	8
Scf_3L/3L	12785466	A	24	0	0	0	12546495	A	44	0	0	2
Scf_3L/3L	12785505	C	0	0	23	0	12546534	C	0	2	39	0
Scf_3L/3L	12785589	G	0	0	0	30	12546618	G	9	0	0	26
Scf_3L/3L	12785617	C	5	0	28	0	12546646	C	0	0	26	0
Scf_3L/3L	12785649	G	0	0	0	21	12546678	G	2	0	0	18
Scf_3L/3L	12785653	C	0	0	16	6	12546682	C	0	0	23	0
Scf_3L/3L	12785654	T	12	9	0	0	12546683	T	0	24	0	0
Scf_3L/3L	12785669	G	0	0	1	18	12546698	T	0	3	2	16
Scf_3L/3L	12785680	G	0	0	0	16	12546709	A	3	0	0	17
Scf_3L/3L	12785683	T	0	17	0	0	12546712	T	2	19	0	0
Scf_3L/3L	12785684	A	15	0	0	0	12546713	G	13	0	0	7
Scf_3L/3L	12785686	T	0	15	1	0	12546715	A	6	13	0	0
Scf_3L/3L	12785702	A	16	0	0	0	12546731	G	17	0	0	2
Scf_3L/3L	12785706	C	0	0	18	0	12546735	C	4	0	15	0
Scf_3L/3L	12785714	C	0	3	14	0	12546743	T	0	17	0	0
Scf_3L/3L	12785717	C	2	0	15	0	12546746	A	13	0	3	0
Scf_3L/3L	12785724	T	0	18	0	0	12546753	T	0	7	1	3
Scf_3L/3L	12785726	C	0	1	16	0	12546755	T	0	5	5	0
Scf_3L/3L	12785748	T	0	23	0	0	12546787	T	0	17	3	0
Scf_3L/3L	12785761	T	4	21	0	0	12546800	T	0	19	0	0
Scf_3L/3L	12785777	A	27	0	0	3	12546819	A	17	0	0	0
Scf_3L/3L	12785784	A	22	0	0	3	12546826	A	16	0	0	1
Scf_3L/3L	12785785	T	2	23	0	0	12546827	C	0	11	4	0
Scf_3L/3L	12785789	A	26	0	0	0	12546831	A	15	0	0	2
Scf_3L/3L	12785795	T	0	16	0	9	12546837	G	0	11	0	6
Scf_3L/3L	12785805	A	21	0	0	2	12546847	A	13	0	0	0
Scf_3L/3L	12785808	G	2	0	0	24	12546850	G	0	0	0	14
Scf_3L/3L	12785811	C	0	19	5	1	12546853	T	0	14	0	0
Scf_3L/3L	12785815	C	3	0	19	4	12546857	C	0	0	13	0
Scf_3L/3L	12785818	T	1	24	0	0	12546860	C	0	8	4	0
Scf_3L/3L	12785826	A	22	0	0	2	12546868	A	14	0	0	0
Scf_3L/3L	12785830	C	0	0	24	0	12546872	A	3	0	10	0
Scf_3L/3L	12785832	T	0	19	5	0	12546874	T	0	15	0	0
Scf_3L/3L	12785834	G	16	0	0	8	12546876	A	15	0	0	0
Scf_3L/3L	12785842	T	4	17	0	0	12546884	T	0	17	0	0
Scf_3L/3L	12785863	T	0	19	3	0	12546903	T	0	22	0	0
Scf_3L/3L	12785867	A	23	0	0	0	12546907	A	20	0	3	0
Scf_3L/3L	12785874	G	5	0	0	16	12546914	G	0	0	0	27
Scf_3L/3L	12785904	G	2	0	0	18	12546944	G	14	0	0	20
Scf_3L/3L	12785913	T	0	14	4	0	12546953	T	0	35	6	0
Scf_3L/3L	12785922	A	17	3	0	0	12546962	A	31	11	0	0
Scf_3L/3L	12785925	T	0	8	12	0	12546965	C	0	1	40	0
Scf_3L/3L	12785930	A	23	0	0	0	12546970	A	40	0	0	2
Scf_3L/3L	12785933	T	0	24	0	0	12546973	T	0	41	2	0
Scf_3L/3L	12785937	C	0	10	15	0	12546977	T	0	42	0	0
Scf_3L/3L	12785949	C	0	0	28	0	12546989	C	2	0	36	0
Scf_3L/3L	12785950	G	28	0	0	0	12546990	A	32	0	4	0
Scf_3L/3L	12785952	T	2	31	0	0	12546992	T	0	37	0	0
Scf_3L/3L	12785960	C	2	0	30	0	12547000	C	0	0	35	0
Scf_3L/3L	12785964	A	31	0	0	0	12547004	A	33	4	0	0
Scf_3L/3L	12785969	C	0	0	31	0	12547009	C	0	0	34	2
Scf_3L/3L	12785972	G	2	0	0	30	12547012	G	0	0	0	34
Scf_3L/3L	12785973	T	2	30	0	0	12547013	T	0	34	0	0
Scf_3L/3L	12785984	C	0	2	31	0	12547024	C	0	0	31	0
Scf_3L/3L	12785985	A	33	2	0	0	12547025	A	31	0	0	0
Scf_3L/3L	12785989	C	0	0	37	0	12547029	C	0	4	27	0
Scf_3L/3L	12786012	A	36	0	0	0	12547052	C	23	0	2	0
Scf_3L/3L	12786048	C	0	0	37	0	12547088	C	0	2	39	0
Scf_3L/3L	12786152	G	0	0	0	37	12547192	G	2	0	0	41
Scf_3L/3L	12786186	G	0	0	0	36	12547226	G	2	0	0	47
Scf_3L/3L	12786231	T	0	38	2	0	12547271	T	0	42	0	0
Scf_3L/3L	12786238	T	0	34	0	3	12547278	T	0	45	0	0
Scf_3L/3L	12786268	T	5	25	0	0	12547308	T	16	13	0	0
Scf_3L/3L	12786272	C	0	2	35	0	12547312	C	0	0	30	0
Scf_3L/3L	12786288	G	0	2	0	29	12547328	G	0	0	0	33
Scf_3L/3L	12786391	C	0	3	33	0	12547431	C	0	0	34	0
Scf_3L/3L	12786393	C	2	1	32	0	12547433	C	0	0	34	0
Scf_3L/3L	12786407	G	2	0	0	26	12547447	G	0	0	0	37
Scf_3L/3L	12786544	G	0	0	0	31	12547584	C	0	0	19	6
Scf_3L/3L	12786546	G	0	0	0	32	12547586	G	4	1	0	23
Scf_3L/3L	12786636	C	0	0	30	0	12547680	C	3	0	24	0
Scf_3L/3L	12786679	C	0	0	29	0	12547723	C	2	0	33	0
Scf_3L/3L	12786686	C	0	0	32	0	12547730	C	3	0	36	0
Scf_3L/3L	12786692	C	0	4	26	0	12547736	C	0	0	40	0
Scf_3L/3L	12786704	A	33	0	0	2	12547748	A	38	0	0	0
Scf_3L/3L	12786722	G	0	0	0	31	12547766	G	18	0	0	22
Scf_3L/3L	12786739	A	29	3	0	0	12547783	A	47	0	0	0
Scf_3L/3L	12786742	C	0	6	25	0	12547786	C	0	0	46	0
Scf_3L/3L	12786748	G	2	0	0	29	12547792	G	0	0	0	45
Scf_3L/3L	12786760	C	3	0	23	0	12547804	C	0	0	47	0
Scf_3L/3L	12786761	C	0	0	26	0	12547805	C	0	6	41	0
Scf_3L/3L	12786771	C	0	4	24	0	12547815	C	0	0	53	0
Scf_3L/3L	12786780	G	0	1	8	17	12547824	G	0	0	0	49
Scf_3L/3L	12786791	C	0	0	25	0	12547835	A	33	0	11	0
Scf_3L/3L	12786804	C	0	0	25	3	12547848	C	0	0	47	0
Scf_3L/3L	12786808	C	5	0	21	0	12547852	C	0	7	44	0
Scf_3L/3L	12786814	T	2	28	0	0	12547858	T	0	45	3	0
Scf_3L/3L	12786816	A	31	2	0	0	12547860	A	45	0	0	0
Scf_3L/3L	12786817	A	30	0	0	2	12547861	A	45	0	0	0
Scf_3L/3L	12786820	T	0	29	0	2	12547864	T	0	44	0	0
Scf_3L/3L	12786821	G	0	2	0	30	12547865	G	0	0	0	44
Scf_3L/3L	12786822	T	0	32	0	0	12547866	T	0	39	2	0
Scf_3L/3L	12786835	T	0	13	19	0	12547876	T	1	35	3	0
Scf_3L/3L	12786851	A	8	0	0	8	12547893	A	7	0	0	0
Scf_3L/3L	12786857	C	3	0	27	0	12547899	C	0	0	34	0
Scf_3L/3L	12786858	T	3	27	0	0	12547900	T	0	37	0	0
Scf_3L/3L	12786861	T	0	12	0	20	12547903	G	0	4	0	34
Scf_3L/3L	12786862	C	0	0	28	5	12547904	C	0	0	36	0
Scf_3L/3L	12786866	A	29	4	0	0	12547908	A	42	0	0	0
Scf_3L/3L	12786870	A	27	0	0	5	12547912	A	41	0	0	0
Scf_3L/3L	12786872	G	0	0	0	34	12547914	G	0	8	0	34
Scf_3L/3L	12786874	A	31	6	0	0	12547916	A	40	0	0	0
Scf_3L/3L	12786876	T	4	31	0	0	12547918	T	0	41	0	0
Scf_3L/3L	12786881	T	0	30	0	8	12547923	T	0	42	0	0
Scf_3L/3L	12786882	G	0	0	8	31	12547924	G	0	0	0	42
Scf_3L/3L	12786893	C	0	0	35	0	12547935	G	0	4	11	27
Scf_3L/3L	12786895	G	0	1	0	35	12547937	G	0	3	0	37
Scf_3L/3L	12786898	A	2	31	0	0	12547940	T	3	36	0	0

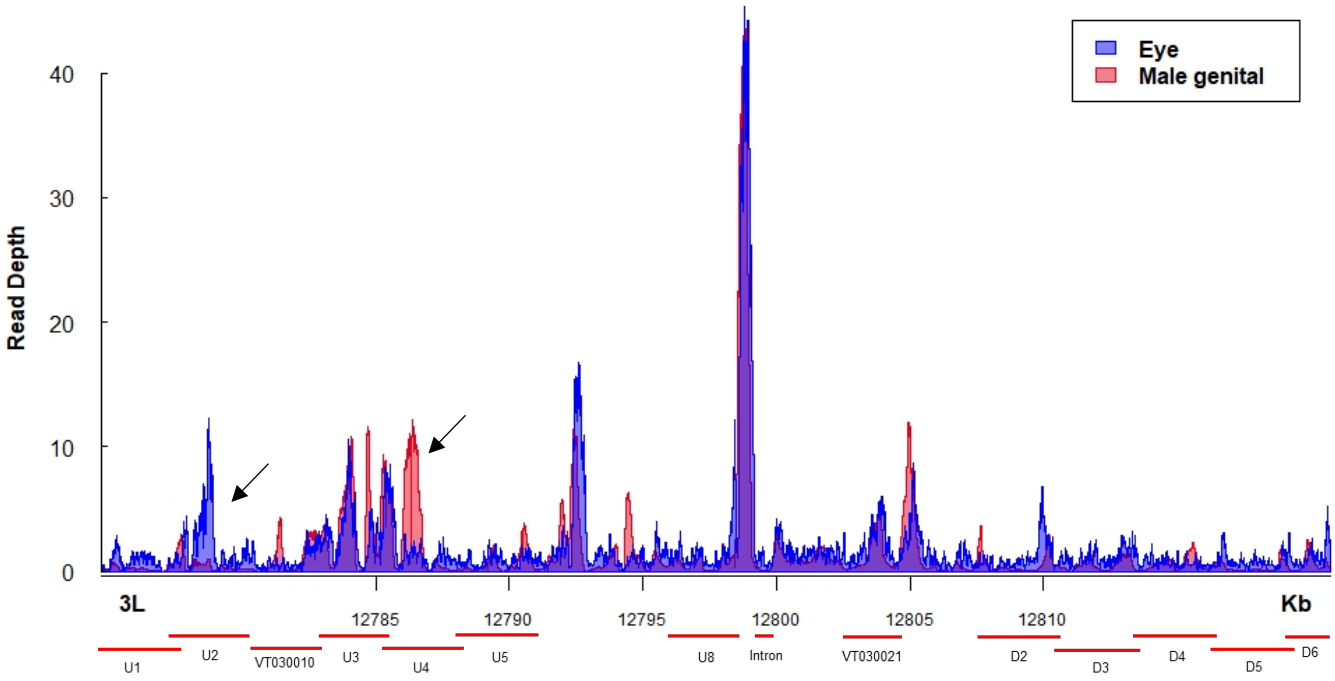
Chromosome	Dsim						Dmau					
	Coordinate	Ref Base	A	T	C	G	Coordinate	Ref Base	A	T	C	G
Scf_3L/3L	12786920	C	10	0	23	0	12547962	C	0	0	44	0
Scf_3L/3L	12786921	T	0	32	0	0	12547963	T	6	37	0	0
Scf_3L/3L	12786923	G	1	0	26	6	12547965	C	0	0	31	13
Scf_3L/3L	12786935	C	0	2	33	0	12547977	C	0	0	48	0
Scf_3L/3L	12786936	A	17	0	16	0	12547978	C	11	0	36	0
Scf_3L/3L	12786939	A	33	0	0	0	12547981	A	46	2	0	0
Scf_3L/3L	12786940	A	25	9	0	0	12547982	A	38	11	0	0
Scf_3L/3L	12786946	G	2	0	0	29	12547988	G	1	0	0	46
Scf_3L/3L	12786948	A	31	0	0	0	12547990	C	16	0	32	0
Scf_3L/NA	12786972	T	2	2	0	0		NA				
Scf_3L/NA	12786979	A	5	0	2	0		NA				
Scf_3L/NA	12786985	A	4	0	0	3		NA				
Scf_3L/NA	12786987	T	3	3	0	0		NA				
Scf_3L/NA	12786990	T	3	3	0	0		NA				
Scf_3L/NA	12786998	T	0	15	3	0		NA				
Scf_3L/3L	12787010	T	0	27	0	0	12548025	T	4	30	0	0
Scf_3L/3L	12787014	C	0	0	29	0	12548029	C	0	2	39	0
Scf_3L/3L	12787030	T	0	7	25	0	12548045	T	0	34	5	0
Scf_3L/3L	12787043	A	28	5	0	0	12548058	A	35	0	0	0
Scf_3L/3L	12787044	G	0	0	0	34	12548059	G	0	2	0	31
Scf_3L/3L	12787048	T	0	30	0	3	12548063	T	0	34	0	0
Scf_3L/3L	12787056	C	0	2	33	0	12548071	C	0	0	33	0
Scf_3L/3L	12787057	C	0	3	32	0	12548072	C	0	0	34	0
Scf_3L/3L	12787060	T	0	36	0	0	12548075	T	0	30	0	2
Scf_3L/3L	12787065	A	31	6	0	0	12548080	A	34	0	0	0
Scf_3L/3L	12787070	T	0	33	1	0	12548085	T	0	27	4	1
Scf_3L/3L	12787074	T	3	32	0	0	12548089	T	0	26	0	0
Scf_3L/3L	12787077	G	0	0	3	29	12548092	G	0	0	0	27
Scf_3L/3L	12787084	T	13	0	0	19	12548099	G	4	0	0	25
Scf_3L/3L	12787087	G	21	0	7	4	12548102	A	26	0	0	4
Scf_3L/3L	12787088	A	26	2	0	3	12548103	A	31	0	0	0
Scf_3L/3L	12787098	T	2	33	0	0	12548113	T	0	29	0	0
Scf_3L/3L	12787110	G	0	2	0	34	12548125	G	0	0	0	34
Scf_3L/3L	12787111	G	30	0	0	2	12548126	A	34	0	0	0
Scf_3L/3L	12787135	C	0	0	25	7	12548150	C	0	0	37	0
Scf_3L/3L	12787140	G	0	0	0	29	12548155	G	2	0	0	36
Scf_3L/3L	12787160	C	0	0	32	0	12548175	C	0	0	40	2
Scf_3L/3L	12787169	G	0	0	0	35	12548184	T	0	22	0	20
Scf_3L/3L	12787171	G	0	0	0	37	12548186	G	1	2	0	44
Scf_3L/3L	12787184	G	1	2	0	35	12548199	G	0	0	2	45
Scf_3L/3L	12787395	C	0	0	34	2	12548410	C	0	0	47	0
Scf_3L/3L	12787398	C	5	0	32	0	12548413	A	39	0	11	0
Scf_3L/3L	12787454	G	3	0	0	33	12548469	G	0	0	0	42
Scf_3L/3L	12787541	C	0	0	36	0	12548556	C	0	4	44	0
Scf_3L/3L	12787542	A	33	0	0	2	12548557	A	49	0	0	0
Scf_3L/3L	12787556	T	3	27	0	0	12548571	T	0	48	0	0
Scf_3L/3L	12787576	A	24	0	2	0	12548591	A	38	0	1	0
Scf_3L/3L	12787583	C	0	2	23	0	12548598	C	0	0	42	0
Scf_3L/3L	12787633	C	0	0	26	0	12548648	G	3	0	26	0
Scf_3L/3L	12787643	C	0	0	29	0	12548658	C	0	0	34	3
Scf_3L/3L	12787698	A	17	0	0	15	12548713	A	41	0	1	0
Scf_3L/3L	12787716	C	0	0	27	0	12548731	C	33	0	9	0
Scf_3L/3L	12787765	G	1	0	0	23	12548780	G	2	0	0	36
Scf_3L/3L	12787789	T	0	0	0	26	12548804	G	4	0	0	30
Scf_3L/3L	12787793	A	23	0	1	3	12548808	A	30	4	0	0
Scf_3L/3L	12787795	A	21	0	0	5	12548810	A	35	0	0	0
Scf_3L/3L	12787800	A	23	2	0	2	12548815	A	30	0	0	1
Scf_3L/3L	12787807	G	0	0	4	26	12548822	C	0	0	19	5
Scf_3L/3L	12787808	G	2	0	0	27	12548823	G	0	0	0	24
Scf_3L/3L	12787812	T	2	22	4	0	12548827	T	0	24	0	0
Scf_3L/3L	12787829	A	18	0	0	2	12548844	A	25	0	0	0
Scf_3L/3L	12787830	A	20	0	1	0	12548845	C	6	12	8	0
Scf_3L/3L	12787832	A	19	3	0	0	12548847	A	24	0	0	0
Scf_3L/3L	12787837	G	4	0	0	17	12548852	G	6	0	0	20
Scf_3L/3L	12787842	A	7	0	0	11	12548857	A	21	0	0	4
Scf_3L/3L	12787844	C	0	11	6	0	12548859	C	0	4	23	0
Scf_3L/3L	12787845	T	1	5	11	0	12548860	T	0	22	4	0
Scf_3L/3L	12787853	T	0	14	0	3	12548868	T	0	29	0	0
Scf_3L/3L	12787854	T	0	17	0	0	12548869	T	0	28	2	0
Scf_3L/3L	12787865	G	3	0	0	12	12548880	A	28	0	0	4
Scf_3L/3L	12787866	C	0	8	6	0	12548881	C	0	4	28	0
Scf_3L/3L	12787872	G	0	0	0	12	12548887	G	1	0	5	32
Scf_3L/3L	12787873	A	4	1	0	0	12548888	A	31	0	5	0
Scf_3L/3L	12787875	T	0	4	0	0	12548902	T	0	29	5	0
Scf_3L/3L	12787891	A	6	0	0	0	12548918	A	32	3	0	1
Scf_3L/3L	12787901	T	0	6	0	0	12548928	T	0	32	3	3
Scf_3L/3L	12787902	C	0	0	6	0	12548929	C	1	3	34	0
Scf_3L/3L	12787905	T	0	7	0	0	12548932	T	3	33	0	0
Scf_3L/3L	12787906	T	0	7	0	0	12548933	T	0	35	3	0
Scf_3L/3L	12787908	C	0	0	26	0	12548935	C	2	0	37	0
Scf_3L/3L	12787915	C	0	0	24	2	12548942	C	0	0	38	0
Scf_3L/3L	12787916	T	0	26	0	0	12548943	T	4	34	0	0
Scf_3L/3L	12787918	T	0	26	0	0	12548945	T	0	35	3	0
Scf_3L/3L	12787922	G	0	6	0	20	12548949	G	0	0	0	40
Scf_3L/3L	12787927	C	1	0	24	0	12548954	C	0	3	38	0
Scf_3L/3L	12787928	A	0	0	0	24	12548955	G	2	0	0	39
Scf_3L/3L	12787933	A	22	0	0	6	12548960	A	41	0	0	0
Scf_3L/3L	12787936	T	18	0	7	0	12548963	C	0	0	39	0
Scf_3L/3L	12787940	A	22	0	0	4	12548967	A	39	0	0	0
Scf_3L/3L	12787946	T	0	26	0	0	12548973	T	5	37	0	0
Scf_3L/3L	12787949	A	17	6	0	0	12548976	A	6	31	0	0
Scf_3L/3L	12787953	A	12	4	0	0	12548980	A	35	3	0	0
Scf_3L/3L	12787962	C	0	4	6	0	12548989	T	0	37	2	0
Scf_3L/3L	12787966	T	0	11	0	0	12548993	T	2	36	0	0
Scf_3L/3L	12787971	A	9	0	0	0	12548998	A	28	0	0	10
Scf_3L/3L	12787977	A	8	0	0	0	12549004	A	36	2	0	0
Scf_3L/3L	12787986	T	0	21	0	0	12549013	T	0	29	8	0
Scf_3L/3L	12787993	G	0	2	0	20	12549020	G	0	0	0	39
Scf_3L/3L	12787996	C	0	0	22	0	12549023	C	0	3	31	0
Scf_3L/3L	12787997	C	0	0	22	2	12549024	C	0	0	39	0
Scf_3L/3L	12788003	A	29	0	0	0	12549030	A	39	2	0	0
Scf_3L/3L	12788007	A	10	14	3	0	12549034	T	1	41	0	1
Scf_3L/3L	12788008	A	28	1	0	0	12549035	A	37	4	0	0
Scf_3L/3L	12788012	C	2	0	27	0	12549039	C	0	0	42	0
Scf_3L/3L	12788017	C	0	0	31	0	12549044	C	0	3	39	0
Scf_3L/3L	12788031	G	4	0	0	24	12549058	G	1	0	0	43
Scf_3L/3L	12788082	C	12	0	12	0	12549109	A	6	0	27	0
Scf_3L/3L	12788090	A	21	0	0	0	12549117	A	28	0	0	2
Scf_3L/3L	12788122	G	0	0	0	19	12549150	G	1	0	2	25
Scf_3L/3L	12788166	A	31	0	0	0	12549194	A	27	2	0	0
Scf_3L/3L	12788205	C	0	0	28	0	12549233	C	0	3	33	0
Scf_3L/3L	12788239	C	0	2	25	0	12549267	C	0	1	31	0
Scf_3L/3L	12788260	C	0	2	20	0	12549288	C	0	0	29	0
Scf_3L/3L	12788308	C	0	0	26	0	12549336	C	0	4	15	0

Supplementary Table 5. Table of polymorphism among *D. simulans* and *D. mauritiana* strains. Table containing the SNPs found with popoolation2. Cells where there is at least one strain with the nucleotide present in that coordinate are coloured in yellow (*D. simulans*) and green (*D. mauritiana).*





Supplementary Figure 4. 1. Trn protein alignment with other *Drosophila* species. Alignment of Trn between *D. melanogaster, D. simulans, D. sechellia, D. kikkawai, D. pseudoscura, D. mojavensis, D. navojoa* and *D. virilis.*



Supplementary Figure 4. 2. Chromatin accessibility profile around the *trn* locus of pupal male genitalia from *D. simulans (w501)* and eye imaginal disc from *D. simulans (yvf)* generated from ATAC-seq data. Profiles shown overlapping one to another from both tissues for comparison. Eye imaginal disc is shown in blue and male genital in red. Arrows point the two main differences seen in accessible chromatin. Eye imaginal disc ATAC-seq data obtained from Torres-Oliva *et al.*, (2021).