Genome-Wide Regulatory Interactions in the Early Stages of *Drosophila* Speciation

By

Alwyn Clark Go

A thesis submitted to the Faculty of Graduate Studies of

The University of Winnipeg

In partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Bioscience, Technology, and Public Policy

Department of Biology

The University of Winnipeg

Winnipeg, Manitoba, Canada

2020

ABSTRACT

Speciation occurs when reproductive barriers prevent the exchange of genetic information between individuals. A common form of reproductive barrier between species capable of interbreeding is hybrid sterility. Genomic incompatibilities between the divergent genomes of different species contribute to a reduction in hybrid fitness. These incompatibilities continue to accumulate after speciation, therefore, young divergent taxa with incomplete reproductive isolation are important in understating the genetics leading to speciation. Here, I use two *Drosophila* subspecies pairs. The first is *D. willistoni* consisting of *D. w. willistoni* and *D. w. winge*. The second subspecies pair is *D. pseudoobscura*, which is composed of *D. p. pseudoobscura* and *D. p. bogotana*. Both subspecies pairs are at the early stages of speciation and show incomplete reproductive isolation through unidirectional hybrid male sterility. In this thesis, I performed an exploratory survey of genome-wide expression analysis using RNA-sequencing on *D. willistoni* and determined the extent of regulatory divergence between the subspecies using allele-specific expression analysis. I found that misexpressed genes showed a degree of tissue-specificity and that the sterile male hybrids had a higher proportion of misexpressed genes in the testes relative to the fertile hybrids. The analysis of regulatory divergence between this subspecies pair found a large (66-70%) proportion of genes with conserved regulatory elements. Of the genes showing evidence or regulatory divergence between subspecies, *cis*-regulatory divergence was more common than other types. In the *D. pseudoobscura* subspecies pair, I compared sequence and expression divergence and found no support for directional selection driving gene misexpression in their hybrids. Allele-specific expression analysis revealed that compensatory *cis-trans* mutations partly explained gene misexpression in the hybrids. The remaining hybrid misexpression occurs due to interacting gene networks or possible co-option of *cis*-regulatory elements by divergent *trans*-acting factors. Overall, the results of this thesis highlight the role of regulatory interactions in a hybrid genome and how these interactions could lead to hybrid breakdown by disrupting gene interaction networks.

ACKNOWLEDGEMENTS

The work done in this thesis would not have been possible without all the support I received throughout my graduate career. I would like to extend my sincerest thanks to everyone involved.

First and foremost, I would like to thank my supervisor, Dr. Alberto Civetta. My success as a graduate student is largely attributable to your mentorship and guidance. Thank you for giving me the opportunity to develop different skills during my graduate program. I started in your lab over three years ago and since then my enthusiasm for science has only grown. Your continued support and encouragement made this thesis possible.

Thank you to my committee members, Drs. Jens Franck and Ken Jeffries. I appreciate the time you spent in providing advice and feedback on my thesis. Dr. Franck, I always enjoyed our little chats throughout the university. Dr. Jeffries, your early advice and pointers on the basics of RNA-sequencing helped me at the start of my Master's when I was teaching myself the basics.

I would also like to thank our collaborator, Dr. Jose Ranz. What I initially thought would be a small addition to my thesis ended up snowballing into an interesting project. The data you shared helped me gain a deeper understanding of the versatility of RNA-sequencing and its many applications.

To the members of the Fly lab, both past and present. You all have made doing a Master's less stressful and actually enjoyable. I would especially like to thank the members during my first and last years: Doaa, Gurman, Karan, Hunter, Bahar, and Lauren. You were the first people I missed when we all started working from home. Each and everyone of you played a role in this thesis, even if it's as small as tolerating the really bad jokes I have made throughout the years which no doubt tends to fly over your heads.

Table of Contents

List of Tables

List of Figures

Chapter 1: General Introduction

Organisms belonging to the same species are characterised by their ability to exchange genetic information through interbreeding. Speciation is a process that occurs when this free exchange of genetic information is inhibited through the formation of reproductively isolating barriers (Dobzhansky 1937). This leads to an increase in biodiversity making it an active area of research among biologists. The reproductive barriers that isolate species can be broadly classified as prezygotic and postzygotic (Mayr 1970). Prezygotic isolation occurs before fertilisation and can be premating such as differences in mating rituals or they can be post-mating-prezygotic which acts after mating but before the formation of a zygote. This generally involves incompatibilities between gametes (Price 1997; Howard et al. 1998, 1999). Postzygotic isolation comes in the form of hybrid dysfunction, commonly manifesting itself as hybrid male sterility or inviability (Coyne and Orr 2004). The reduction of fitness in the hybrids serves as a reproductive barrier that prevents further gene flow between nascent species.
Reproductive barriers do not often evolve immediately. In *Drosophila*, premating isolation has been shown to evolve the fastest with postzygotic isolation evolving the slowest (Coyne and Orr 1989, 1997; Turissini et al. 2018). However, different types of isolation can be important to speciation and the average rates of their evolution does not necessarily indicate that premating isolation is more relevant to speciation. For example, among Hawaiian species of *Drosophila*, sympatric species experience premating isolation while allopatric species experience postzygotic isolation (Carson 1989; Kang et al. 2017). This suggests that the opportunities for interbreeding among sympatric species limits the development of severe reproductive isolation such as hybrid sterility or inviability (Kisel

and Barraclough 2010). Reproductively isolating barriers also do not occur immediately resulting in the creation of new species. Instead, the early stages of speciation often allow partial exchange of genetic material between nascent species. Haldane's rule is an example of partial reproductive isolation at the early stages of speciation. The rule states that when only one sex is inviable or sterile in hybrids between closely related species, that sex is often the heterogametic sex (*i.e.* XY or ZW) (Haldane 1922). This rule applies to a wide range of taxa including *Drosophila* where early signs of postzygotic isolation often occur in the form of hybrid male sterility.

The Origin of Hybrid Incompatibilities

When species hybridise, two divergent genomes are forced to interact with each other leading to misregulated gene expression driven by genetic incompatibilities. These genetic incompatibilities are typically regulatory dysfunctions that can lead to transgressive gene expression in hybrids (*i.e.* expression levels above or below levels found in the parental species). Transgressive expression has been associated with hybrid sterility in *Drosophila* (Michalak and Noor 2003; Ranz et al. 2004; Moehring et al. 2007; Gomes and Civetta 2015). Gene regulation relies on the proper interactions between co-adapted *cis*- and *trans*-regulatory elements. *Cis*-regulatory elements, such as promoters and enhancers, are segments of non-coding DNA that act as binding sites for *trans*-factors (e.g. transcription factors). Promoters are consensus sequences, like the TATA box, found upstream and proximal to the transcription start site of the gene they regulate. They initiate transcription by serving as binding sites for the RNA polymerase II complex and general transcription factors. Enhancers, on the other hand, are usually more distal

and may be located up to several kilobases upstream or downstream from the gene they regulate. They affect the rate of transcription by remodeling chromatin structure through interactions with the general transcription complex or other transcription factors (Kadauke and Blobel 2009). Divergent species evolve slightly different fine-tuned interactions between *cis*-regulatory elements and *trans*-acting factors that keep gene expression regulated, but such interactions can be disturbed in hybrids, resulting in gene misexpression.

The Bateson-Dobzhansky-Muller model (Dobzhansky 1937; Muller 1942; Orr 1996) explains how regulatory elements can function normally in pure species but become incompatible in a hybrid genetic background. A brief description of the model follows. Consider a species with an *AA* genotype at one locus and the *BB* genotype at another (*AA BB*). When this species is divided into two separate populations, one population may experience an *A* to *a* mutation while the other undergoes a *B* to *b* mutation. Both *a* and *b* alleles are either neutral and fixed by genetic drift or provide a fitness advantage to their respective populations and become fixed by selection. The two populations will therefore now have *aa BB* and *AA bb* genotypes respectively. While the *a*-*B* and *A*-*b* alleles are compatible, the *a*-*b* interaction is untested in a common genetic background. When hybrids between the two populations are formed through interbreeding, the *a* and *b* alleles are brought together in a common genome (*Aa Bb*). This novel interaction between the two alleles may lead to regulatory incompatibilities and gene misexpression in the hybrids. This simplified description of the model assumes the interaction between two loci but can be expanded to include the more common multi-loci system of divergent interactions when different species hybridise. The adaptive changes

that occur within different populations or species can lead to multiple untested interactions that might become incompatible in hybrids.

Fast Male Regulatory Divergence and Gene Misregulation

Genes with reproductive functions are often rapidly evolving (Civetta and Singh 1995; Civetta and Singh 1998; Haerty et al. 2007). The rapid evolution of these genes has been attributed to adaptive evolution (Swanson and Vacquier 2002). The analysis of rates of molecular evolution in *Drosophila* from the comparison of genomes from 12 different species, revealed that genes with sex and reproductive related functions have faster rates of sequence evolution, with some of those genes evolving under the influence of positive selection (Haerty et al. 2007). This trend is especially pronounced for male-biased genes which are more likely to experience expression divergence between species than female- or non-biased genes (Meiklejohn et al. 2003; Ranz et al. 2003; Assis et al. 2012; Harrison et al. 2015). Among these male-biased genes, those primarily or only expressed in the testes and accessory glands with functions related to spermatogenesis or the production of seminal fluids accumulated nonsynonymous nucleotide substitutions at a greater rate across lineages than other classes of genes suggesting changes in the direction of modifying protein function (Haerty et al. 2007). The narrow breadth of expression for these genes makes them more susceptible to faster rates of divergence compared to more pleiotropic genes (Meiklejohn et al. 2003; Zhang and Parsch 2005; Zhang et al. 2007; Assis et al. 2012). It is therefore possible that the rapid divergence of male-biased genes could lead to regulatory incompatibilities in hybrids. In an introgression analysis between

D. simulans and *D. mauritiana*, Ferguson et al. (2013) found support for rapid male regulatory divergence as a driver of misexpression for spermatogenesis genes in hybrids.

Analyses on *Drosophila* hybrids have shown that male-biased genes are disproportionately misexpressed in hybrids. In the *D. melanogaster* group, interspecific crosses between *D. melanogaster* females and *D. simulans* males produce inviable male offspring and sterile female hybrids. Gene expression analysis on the female hybrids found an overrepresentation of overexpressed male-biased genes which was attributed to a breakdown in the regulatory elements that normally supress the expression of these male-biased genes in females (Ranz et al. 2004). In sterile male hybrids between *D. simulans* and *D. mauritiana*, a microarray analysis revealed that male-biased genes with functions related to spermatogenesis or male-specific phenotypes were also more likely to be misexpressed in the sterile male hybrids than other classes of genes (Michalak and Noor 2003). Using testes-specific RNA, Haerty and Singh (2006) also found that male-biased genes, particularly those with sex-related functions, were predominantly misexpressed in sterile male hybrids between species of the *D. melanogaster* complex. Using a sperm-specific transcript array developed for the *D. simulans* species clade, Moehring et al. (2007) found an enrichment of misexpressed genes involved in the late stages of spermatogenesis among sterile male hybrids of this clade. These studies suggest that the rapid divergence of male-biased genes, especially those involved in spermatogenesis, have a major contribution to hybrid dysfunction between species of *Drosophila*. However, a caveat is that often these studies did not examine gene expression at the reproductive tissue itself but rather across the whole fly. Tissue-specific

assays have found that early stage genes of spermatogenesis also undergo previously undetected patterns of misexpression (Sundararajan and Civetta 2011).

Defects in spermatogenesis is indeed common in *Drosophila* hybrids where the sterility phenotype is often due to abnormalities in sperm production. For example, sterile hybrids between *D. simulans* and *D. mauritiana* failed to produce individualised sperm at best (Kulathinal and Singh 1998), while hybrids between *D. yakuba* and *D. santomea* failed to produce motile sperm (Moehring et al. 2006). In hybrids between species of the Hawaiian picture-wing clade, *D. planitibia* and *D. silvestris*, defects in spermatogenesis are more severe where the production of sperm is completely absent (Brill et al. 2016). In the *D. pseudoobscura* subspecies pair which exhibit unidirectional hybrid male sterility (*i.e.* the fertility of the hybrid male is dependent on the maternal species), sterile hybrid males suffered from the production of immotile sperm while fertile hybrids experienced no known sperm defects (Prakash 1972, Gomes and Civetta 2014). Interestingly, a relatively recently described pair of *D. willistoni* subspecies (Mardiros et al. 2016) showed no defects in sperm development but rather subtle atrophies in testes tissue development that prevented sperm transfer during copulation (Davis et al. 2020). A genome-wide expression analysis on the *D. pseudoobscura* subspecies pair and their reciprocal male hybrids found a significantly higher proportion of misregulated genes in the sterile F_1 male hybrids relative to the fertile F_1 male hybrids (Gomes and Civetta 2015). This finding is interesting given that the only difference between these two hybrids is the composition of their sex chromosomes therefore, implicating a role for the X-chromosome on the formation of hybrid incompatibilities.

Fast Evolution of the X-chromosome

In sterile male hybrids, the X-chromosome has been observed to play a disproportionate role in the formation of hybrid incompatibilities. This large "X-effect" is in part due to the hemizygosity of genes on the X-chromosome in males. Evidence for this was provided by introgression analyses of *D. mauritiana* chromosomes into an otherwise *D. sechellia* genetic background. The authors of that study found that X-linked introgressions were more likely to cause hybrid male sterility (60%) while autosomal introgressions of the same size were less likely to do so (18%) (Masly and Presgraves 2007). The hemizygosity of the X-chromosome in males exposes the full effects of X-linked genes. This allows the evolution of X-linked genes to be more rapid than autosomal genes especially when the new mutations are beneficial and recessive (Charlesworth et al. 1987). X-linked genes not only tend to evolve more rapidly at the sequence level, but an analysis of gene expression levels across six species of *Drosophila* also found more rapid rates of interspecific expression divergence for X-linked genes relative to autosomal genes (Meisel et al. 2012) suggesting that X-linked genes enhance divergence between species. The rapid evolution of X-linked genes can also cause X-autosomal incompatibilities in the hybrid genome. Consistent with this, an analysis of *D. santomea*, *D. yakuba*, and their F_1 sterile male hybrids found that hybrid misexpression is more frequent among autosomal genes, which is likely facilitated by rapidly evolving X-linked *trans*-acting factors (Llopart 2012). Introgression analyses show that the X-chromosome is a hotspot for hybrid male sterility factors (Tao et al. 2003; Masly and Presgraves 2007) and some major hybrid male sterility genes residing on the X-chromosome have been identified in *Drosophila*.

The first of these genes identified was *Odysseus-site homeobox* (*OdsH*) (Ting et al. 1998). This gene lies within the *Odysseus* locus which was found to have a major sterility effect in hybrids between *D. simulans* and *D. mauritiana* (Perez et al. 1993). Hybrids males that carried the *D. simulans* allele for *OdsH* were fertile while those with the *D. mauritiana* allele were sterile (Ting et al. 1998). *OdsH* encodes a transcription factor with a homeobox DNA-binding motif. Interestingly, while homeoboxes are typically highly conserved due to their function in DNA-binding, the homeobox domain of *OdsH* is rapidly evolving relative to the rest of the protein-coding gene and has acquired 15 amino acid replacements within the approximately 250,000 years of divergence between *D. simulans* and *D. mauritiana* (Ting et al. 1998). A duplication event from the *unc4* transcription factor resulted in *OdsH* which is now exclusively expressed in the testes (Ting et al. 1998). This acquisition of a male-biased function likely contributed to its rapid evolution. Aberrant binding of the *OdsH D. mauritiana* protein on the *D. simulans* Y-chromosome and 4th autosome alters chromatin morphology and causes sterility (Bayes and Malik 2009; Phadnis and Malik 2013).

Another major hybrid male sterility gene is *Overdrive* (*Ovd*). Discovered in the *D. pseudoobscura* subspecies pair, hybrid males with the *D. p. pseudoobscura* allele for *Ovd* are fertile while those with the *D. p. bogotana* allele are sterile. *Ovd* is also located on the X-chromosome and is predicted to encode a protein with a Myb/SANT-like Adf-1 (MADF) DNA-binding domain (Phadnis and Orr 2009). This DNA-binding domain is similar to the one found on the Adf-1 transcription factor which is responsible for the activation of a diverse group of genes (Cutler et al. 1998). *Ovd* is expressed in the testes and sequence analysis found that it has undergone a rapid rate of evolution accumulating

seven non-synonymous and five synonymous fixed nucleotide changes in its relatively short (591 bp) coding region (Phadnis and Orr 2009). *Ovd* exerts its sterility effect by acting in *trans* and interacting with genetic targets found in the 2nd and 3rd autosomes (Phadnis 2011).

Taken together, the adaptive evolution acting on male-biased genes as well as the environment of the X-chromosome has led to the rapid divergence of male-biased genes (Llopart 2012; Llopart et al. 2018). Despite the characterisation of rapidly evolving *trans*-factors on the X-chromosome like *Ovd* and the manifestation of hybrid male sterility, genome-wide expression level analysis showed minimal expression divergence between *D. p. pseudoobscura* and *D. p. bogotana* (Gomes and Civetta 2015). Suggesting that stabilising selection which act to maintain similar levels of expression between species may favour changes that help keep the norm within species but cause regulatory incompatibilities and misexpression in hybrids.

Regulatory Divergence and Compensatory Mutations

An analysis of gene expression levels across seven species of *Drosophila* that span roughly 42 million years of divergence found that the divergence in gene expression levels between these seven species is not proportional to the amount of time that separates them (Bedford and Hartl 2009). This suggests that neutral evolution acting on gene expression is unlikely. Instead, the authors of the study found that gene expression divergence rapidly reaches a saturation point in evolutionary time caused by stabilising selection that preserves optimum levels of gene expression between species preventing further variation in gene expression. Although divergence in gene expression levels

between species tend to be fairly limited, the regulatory networks behind them were not necessarily conserved. The process of developmental system drift shows that natural selection allows the divergence of regulatory networks if the underlying phenotype (e.g. gene expression) is conserved (True and Haag 2001). Gene expression can be conserved despite the divergence of their regulatory elements through lineage specific co-evolution between the *cis*- and *trans*-elements that regulate them. In this situation, a detrimental mutation in a *cis*-regulatory element is compensated for by a change in its *trans* interacting partner, or vice versa, thereby stabilising overall gene expression levels (Figure *1*A). The fixation of these regulatory elements could explain how gene expression divergence is limited across different lineages despite sequence divergence.

The divergence of *cis*- and *trans*-regulatory elements between species can be inferred through the measurement of species-specific allele expression in an interspecific F_1 hybrid genetic background (Wittkopp et al. 2004). This allows the identification of *cis*-only and *trans*-only regulatory divergence that cause gene expression differences between parental species as well as compensatory *cis*-*trans* mutations that preserve gene expression levels between species. Since *cis*-regulatory elements affect gene expression in an allele specific manner, *cis*-only divergence between species is seen when differences in parental allele expressions are observed in the F_1 hybrid (Figure *1*B). On the other hand, the two alleles in a hybrid background are in a common *trans*-acting environment and are therefore equally affected by *trans*-acting factors. *Trans*-only regulatory divergence between species can be inferred when the hybrid shows equal expression of parental alleles despite the gene showing differential expression between species (Figure *1*C). Unlike *cis*-only and *trans*-only regulatory divergence that cause

differences in gene expression between species, compensatory *cis-trans* mutations mask regulatory divergence between species by maintaining similar levels of gene expression. However, interactions between these divergent regulatory elements lead to incompatibilities and are detected through differences in species-specific allele expression in the F_1 hybrid (Figure *1*A).

This approach of using allele specific expression in interspecific F_1 hybrids has been used to study patterns of regulatory divergence between species of *Drosophila*. *Cis*-regulatory divergence was found to have a bigger contribution to gene expression differences between species. This was observed in analyses between *D. melanogaster* and *D. simulans* (Wittkopp et al. 2004, 2008; Graze et al. 2009), *D. simulans* and *D. sechellia* (Coolon et al. 2014), *D. p. pseudoobscura* and *D. p. bogotana* (Gomes and Civetta 2015), as well as *D. silvestris* and *D. planitibia* (Brill et al. 2016). In contrast, an analysis between *D. melanogaster* and *D. sechellia* found more *trans*-regulatory divergence than *cis*-regulatory divergence (McManus et al. 2010), though McManus et al. (2010) suggested that the unexpected pattern of regulatory divergence between *D. melanogaster* and *D. sechellia* may reflect the unique evolutionary history of *D. sechellia* which allows natural selection to act less efficiently. The overall larger contribution of *cis*- rather than *trans*-regulatory divergence may be due to the nature of *cis*- and *trans*-regulatory elements. *Trans*-acting factors are more pleiotropic and interact with multiple genes, mutations on these elements will therefore have a higher likelihood of being detrimental. On the other hand, multiple *cis*-regulatory elements usually control the regulation of one gene, changes in one of these *cis* elements may therefore be more tolerable (Wittkopp and Kalay 2012).

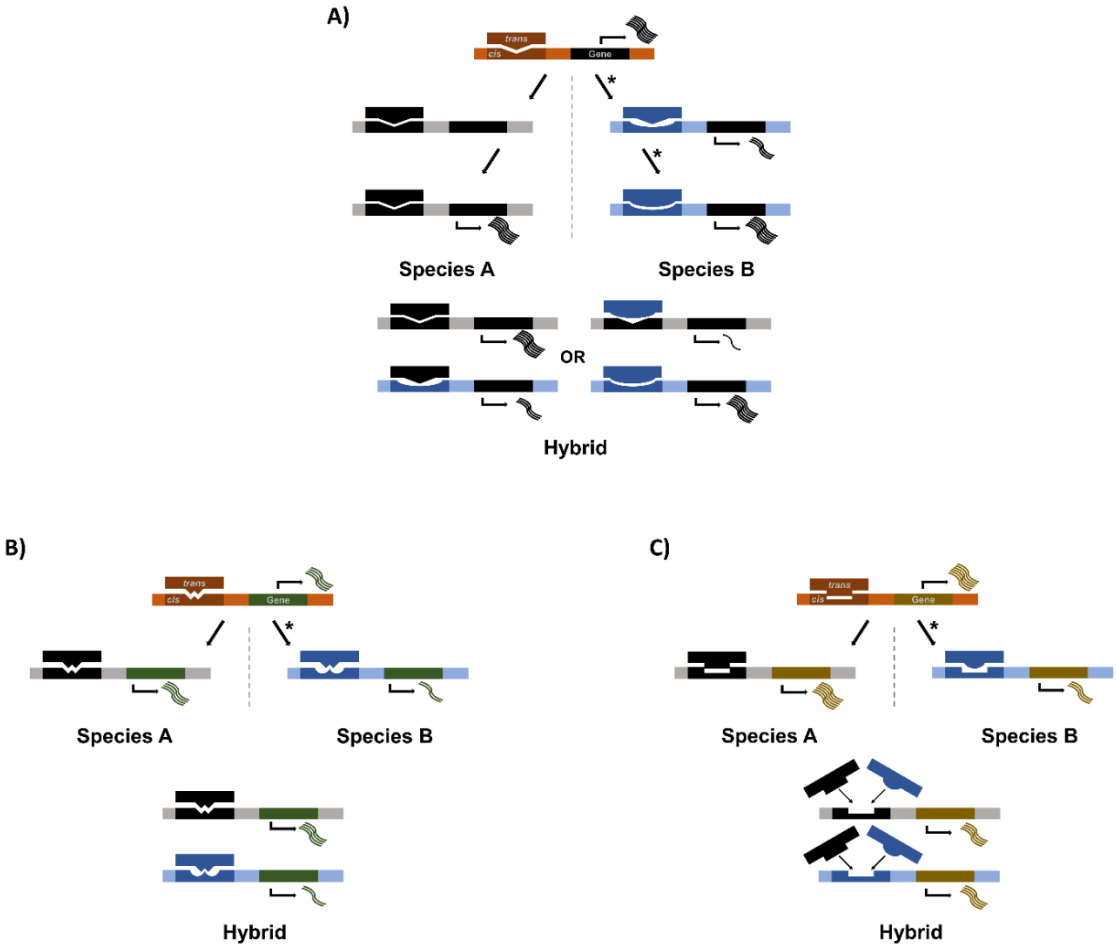


Figure 1: Schematic representation of *cis*- and *trans*-regulatory divergence and the resulting interaction in the F1 hybrid. Bars represent the gene region with grey bars representing species A, blue bars species B, and orange bars the ancestral gene. Asterisks denote lineage-specific changes in regulatory elements. A) shows *cis*-*trans* compensatory divergence in which an initial change in *cis* for species B is compensated for by a subsequent change in *trans* bringing expression levels back to similar levels between species. The competing regulatory interactions present in the hybrid leads to overall gene misexpression. B) represents a *cis*-only divergence between species. C) shows a *trans*-only change between species where competing divergent *trans*-factors in the hybrid background result in misexpression that affects both alleles equally.

Since *cis*-regulatory changes are more tolerable, they have been shown to accumulate linearly with divergence time between species. This was observed between *D. melanogaster* and *D. simulans* as well as between *D. simulans* and *D. sechellia* (Coolon et al. 2014). Although *cis*-regulatory divergence increased, total gene expression between these species did not. This suggests that *trans*-regulatory factors arise to compensate for gene expression changes caused by *cis*-regulatory differences. In support of this, 87% and 73% of genes showing *cis*- and *trans*-regulatory divergence between *D. melanogaster* and *D. simulans* as well as *D. simulans* and *D. sechellia*, respectively, were compensatory in nature (Coolon et al. 2014). Although compensatory *cis*-*trans* mutations restore gene expression back to optimal levels between species, novel interactions between these divergent elements lead to misregulation in the hybrids (Wittkopp et al 2004; Landry et al. 2005; McManus et al. 2010). The misregulation of gene expression in the hybrids leads to a lost in fitness such as sterility which acts as a form of postzygotic reproductive isolation that prevents further gene flow between nascent species.

The Early Stages of Speciation

Regulatory incompatibilities responsible for hybrid misregulation continue to accumulate between species even after speciation is complete. Evidence for this "snowball" effect has been found in *Drosophila* where the number of genes involved in postzygotic incompatibilities increases exponentially with divergence time between species (Matute et al. 2010). This makes it hard to disentangle between the genetic incompatibilities responsible for speciation from those that arise after speciation is complete. This highlights the importance of young species pairs at the early stages of

speciation where reproductive isolation is incomplete. The *Drosophila pseudoobscura* subspecies pair and *D. willistoni* subspecies pair are such examples.

The *D. pseudoobscura* subspecies pair consists of *D. p. pseudoobscura* which is distributed across western North America (Dobzhansky 1937) and *D. p. bogotana* found in the elevated regions of Bogota, Colombia (Dobzhansky et al. 1964). The two subspecies are in allopatry, separated by roughly 2000 km, and diverged between 150,000 and 230,000 years ago (Schaeffer and Miller 1991; Wang et al. 1997). The subspecies pair can freely interbreed with each other though differences in mating behaviours and cuticular hydrocarbons have been identified (Kim et al. 2012). Reflecting their recent divergence, the *D. pseudoobscura* subspecies pair show incomplete reproductive isolation and exhibit an early form of postzygotic reproductive isolation through unidirectional hybrid male sterility. Crosses with *D. p. bogotana* mothers and *D. p. pseudoobscura* fathers produce sterile hybrid males while hybrid males with *D. p. pseudoobscura* mothers are fertile (Prakash 1972). Hybrid females in both directions of the cross are fertile. X-autosomal incompatibilities between the *D. p. bogotana* X-chromosome and *D. p. pseudoobscura* autosomes are responsible for hybrid sterility (Orr and Irving 2001; Phadnis and Orr 2009; Phadnis 2011) which is manifested with the production of immotile sperm (Gomes and Civetta 2014).

Similar to the *D. pseudoobscura* subspecies pair, the *D. willistoni* subspecies group is also at the early stages of speciation. The subspecies group was recently suggested to have three members, *D. w. quechua* (narrowly distributed west of the Andes around Lima, Peru), *D. w. willistoni* (found south of the American mainland, Mexico, and the Carribean islands), and *D. w. winge* (distributed across much of the south American

continent west of the Andes) (Mardiros et al. 2016). No formal estimation has been made on the divergence between these subspecies but allozyme analyses between *D. w. quechua* and *D. w. willistoni* suggests a divergence time similar to the *D. pseudoobscura* subspecies pair (Ayala and Tracey 1973; Ayala and Dobzhansky 1974; Ayala et al. 1974). Between *D. w. willistoni* and *D. w. winge*, no fixed premating isolation has been observed (Davis et al. 2020) and a haplotype network analysis found limited evidence of genetic differentiation and a high degree of gene flow between the subspecies (Mardiros et al. 2016). Despite this, the subspecies pair also show unidirectional hybrid male sterility wherein only hybrid males with *D. w. willistoni* mothers are sterile (Gomes and Civetta 2014; Civetta and Gaudreau 2015; Mardiros et al. 2016). Unlike sterile hybrids of the *D. pseudoobscura* subspecies pair, sterile hybrid males between the *D. willistoni* subspecies produce normal motile sperm but an abnormal bulge at the basal end of the testes prevents sperm transfer into the female reproductive tract (Gomes and Civetta 2014; Civetta and Gaudreau 2015; Davis et al. 2020).

The advantage of species pairs that exhibit unidirectional hybrid male sterility is the availability of both sterile and fertile F_1 hybrids. This allows the identification of genes linked to sterility and the formation of reproductive isolation as those uniquely misexpressed in the sterile hybrids.

RNA-Sequencing as a Tool for Measuring Gene Expression

An early method used for genome-wide measurements of gene expression levels was DNA microarrays (Michalak and Noor 2003; Moehring et al. 2007). This technology

relies on the complementarity of experimental cDNA transcripts with known DNA molecules attached on a slide. These DNA molecules act as a probe and hybridise through Watson-Crick base pairing with experimental cDNA transcripts which are labeled with a fluorescent dye. The resulting intensity of the fluorescent signals from probe hybridisation are then used to infer transcript abundance which serves as a proxy for gene expression. Since this technique relies on the hybridisation of the sample cDNA with known probes, it is limited by the availability of known sequences from a genome assembly. Furthermore, since the probes are usually designed using sequence information from one species, sequence bias is introduced when measuring the expression profile of other species or F_1 hybrids (Gilad et al. 2005).

The limitations of DNA microarray are improved upon by RNA-sequencing which was made more accessible by recent developments in high-throughput DNA-sequencing technologies such as the Illumina platform. Briefly, the RNA-sequencing method begins with the extraction of total RNA, followed by ribosomal RNA depletion, cDNA synthesis through reverse transcription, and library preparation. The prepared library is then sequenced using high-throughput platforms like Illumina. After sequencing, the reads are aligned to a reference genome, or a *de novo* genome constructed with RNA-sequence data, and the expression level of a gene can be estimated based on its read counts (Wang et al. 2009; McManus et al. 2010; Gomes and Civetta 2015). RNA-sequencing improves upon microarrays especially in the measurement of lowly expressed genes and by limiting the bias in measuring gene expression levels between species. Reads from RNA-sequencing can also be assigned to a species of origin using fixed single nucleotide polymorphisms (SNPs) allowing the measurement of allele

specific expression and the identification of divergent regulatory elements in F_1 hybrids (McManus et al. 2010, Gomes and Civetta 2015). This makes RNA-sequencing a versatile tool for the identification of genome-wide interactions and regulatory divergence that lead to the formation of new species.

Objectives

In this thesis, I take advantage of RNA-sequencing and *Drosophila* subspecies pairs at the early stages of speciation. In the first chapter, I perform a genome-wide exploratory survey on *D. w. willistoni* and *D. w. winge*. Using RNA sequences extracted from the testes, accessory glands, and ovaries of the *D. willistoni* subspecies pair and their reciprocal F_1 hybrids, I identified tissue-specific genes and determined whether these genes are more likely to be misregulated in the hybrids. I also determined the extent of regulatory divergence between the *D. willistoni* subspecies pair through the use of allele specific expression analysis.

In the second chapter, I used previously published transcriptomics data for *D. p. pseudoobscura, D. p. bogotana* and their reciprocal F_1 hybrids (Gomes and Civetta 2015) in conjunction with a more recent genome assembly to investigate the basis of transgressive gene expression in the hybrids. I determined whether directional selection plays a role in hybrid misexpression and the role of compensatory *cis-trans* mutations in the early stages of speciation. I further propose alternative models that might trigger gene misexpression in interspecies hybrids.

REFERENCES

Assis, R., Zhou, Q., and Bachtrog, D. (2012). Sex-biased transcriptome evolution in drosophila. *Genome Biol. Evol.* 4, 1189–1200. doi:10.1093/gbe/evs093.

Ayala, F. J., and Tracey, M. L. (1973). Enzyme variability in the Drosophila willistoni group: VIII. Genetic differentiation and reproductive isolation between two subspecies. *J. Hered.* doi:10.1093/oxfordjournals.jhered.a108367.

Ayala, F.J. & Dobzhansky, T.H. (1974). A new subspecies of Drosophila pseudoobscura (Diptera: Drosophilidae). *Pan-Pac. Entomol.* 50: 211–219.

Ayala, F. J., Tracey, M. L., Hedgecock, D., and Richmond, R. C. (1974). Genetic Differentiation During the Speciation Process in Drosophila. *Evolution*. doi:10.2307/2407283.

Bayes, J. J., and Malik, H. S. (2009). Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. *Science (80-.).* 326, 1538–1541. doi:10.1126/science.1181756.

Bedford, T., and Hartl, D. L. (2009). Optimization of gene expression by natural selection. *Proc. Natl. Acad. Sci. U. S. A.* 106, 1133–1138. doi:10.1073/pnas.0812009106.

Brill, E., Kang, L., Michalak, K., Michalak, P., and Price, D. K. (2016). Hybrid sterility and evolution in Hawaiian *Drosophila*: Differential gene and allele-specific expression analysis of backcross males. *Heredity (Edinb).* 117, 100–108. doi:10.1038/hdy.2016.31.

Bush, G. L., Chilcote, C. a, Smith, D. C., Berlocher, S. H., Bennett, E. W., Vet, L. E. M., et al. (2009). Evolution of the Drosophila Nuclear. *Nat. Hist.* 7, 779–782.

Carson, H. L., Kaneshiro, K. Y., and Val, F. C. (1989). Natural hybridization between the sympatric Hawaiian species Drosophila silvestris and Drosophila heteroneura. *Evolution*. doi:10.1111/j.1558-5646.1989.tb04217.x.

Charlesworth, B., Coyne, J. A., and Barton, N. H. (1987). The relative rates of evolution of sex chromosomes and autosomes. *Am. Nat.* doi:10.1086/284701.

Civetta, A., and Gaudreau, C. (2015). Hybrid male sterility between Drosophila willistoni species is caused by male failure to transfer sperm during copulation. *BMC Evol. Biol.* 15, 1–8. doi:10.1186/s12862-015-0355-8.

Civetta, A., and Singh, R. S. (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. *J. Mol. Evol.* doi:10.1007/BF00173190.

Civetta, A., and Singh, R. S. (1998). Sex-related genes, directional sexual selection, and speciation. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a025994.

Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in *Drosophila*. *Genome Res.* 24, 797–808. doi:10.1101/gr.163014.113.

Coyne, J. A., and Orr, H. A. (1989). Patterns of Speciation in *Drosophila*. *Evolution* 43, 362. doi:10.2307/2409213.

Coyne, J.A., & Orr, H.A. 2004. Speciation. Sinauer Associates, Inc., Sunderland, MA.

Cutler, G., Perry, K. M., and Tjian, R. (1998). Adf-1 Is a Nonmodular Transcription Factor That Contains a TAF-Binding Myb-Like Motif. *Mol. Cell. Biol.* doi:10.1128/mcb.18.4.2252.

Davis, H., Sosulski, N., and Civetta, A. (2020). Reproductive isolation caused by azoospermia in sterile male hybrids of Drosophila. *Ecol. Evol.* 10, 5922–5931. doi:10.1002/ece3.6329.

Dobzhansky T. (1937). Genetics and the origin of species. In Columbia biological series. New York: Columbia University Press.

Dobzhansky, T., Anderson, W. W., Pavlovsky, O., Spassky, B., and Wills, C. J. (1964). Genetics of Natural Populations. XXXV. A Progress Report on Genetic Changes in Populations of Drosophila pseudoobscura in the American Southwest. *Evolution*. doi:10.2307/2406389.

Ferguson, J., Gomes, S., and Civetta, A. (2013). Rapid Male-Specific Regulatory Divergence and Down Regulation of Spermatogenesis Genes in Drosophila Species Hybrids. *PLoS One*. doi:10.1371/journal.pone.0061575.

Gilad, Y., Rifkin, S. A., Bertone, P., Gerstein, M., and White, K. P. (2005). Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles. *Genome Res.* 15, 674–680. doi:10.1101/gr.3335705.

Gomes, S., and Civetta, A. (2014). Misregulation of spermatogenesis genes in *Drosophila* hybrids is lineage-specific and driven by the combined effects of sterility and fast male regulatory divergence. *J. Evol. Biol.* 27, 1775–1783. doi:10.1111/jeb.12428.

Gomes, S., and Civetta, A. (2015). Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci. Rep.* 5, 11976. doi:10.1038/srep11976.

Graze, R. M., McIntyre, L. M., Main, B. J., Wayne, M. L., and Nuzhdin, S. V. (2009). Regulatory divergence in *Drosophila melanogaster* and *D. simulans*, a genomewide analysis of allele-specific expression. *Genetics* 183, 547–561. doi:10.1534/genetics.109.105957.

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ram, K. R., Sirot, L. K., et al. (2007). Evolution in the fast lane: Rapidly evolving sex-related genes in Drosophila. *Genetics* 177, 1321–1335. doi:10.1534/genetics.107.078865.

Haerty, W., and Singh, R. S. (2006). Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. *Mol. Biol. Evol.* 23, 1707–1714. doi:10.1093/molbev/msl033.

Haldane, J. B. S. (1922). Sex ratio and unisexual sterility in hybrid animals. *J. Genet.* doi:10.1007/BF02983075.

Harrison, P. W., Wright, A. E., Zimmer, F., Dean, R., Montgomery, S. H., Pointer, M. A., et al. (2015). Sexual selection drives evolution and rapid turnover of male gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 112, 4393–4398. doi:10.1073/pnas.1501339112.

Howard, D. J. (1999). Conspecific sperm and pollen precedence and speciation. *Annu. Rev. Ecol. Syst.* doi:10.1146/annurev.ecolsys.30.1.109.

Howard, D. J., Gregory, P. G., Chu, J., and Cain, M. L. (1998). Conspecific sperm precedence is an effective barrier to hybridization between closely related species. *Evolution*. doi:10.1111/j.1558-5646.1998.tb01650.x.

Kadauke, S., and Blobel, G. A. (2009). Chromatin loops in gene regulation. *Biochim. Biophys. Acta - Gene Regul. Mech.* 1789, 17–25. doi:10.1016/j.bbagrm.2008.07.002.

Kang, L., Garner, H. R., Price, D. K., and Michalak, P. (2017). A Test for Gene Flow among Sympatric and Allopatric Hawaiian Picture-Winged Drosophila. *J. Mol. Evol.* 84, 259–266. doi:10.1007/s00239-017-9795-7.

Kim, Y. K., Ruiz-García, M., Alvarez, D., Phillips, D. R., and Anderson, W. W. (2012). Sexual isolation between North American and Bogota strains of Drosophila pseudoobscura. *Behav. Genet.* 42, 472–482. doi:10.1007/s10519-011-9517-7.

Kisel, Y., and Timothy, T. G. (2010). Speciation has a spatial scale that depends on levels of gene flow. *Am. Nat.* doi:10.1086/650369.

Kulathinal, R., and Singh, R. S. (1998). Cytological characterization of premeiotic versus postmeiotic defects producing hybrid male sterility among sibling species of the *Drosophila melanogaster* complex. *Evolution* 52, 1067. doi:10.2307/2411237.

Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G., and Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. *Genetics* 171, 1813–1822. doi:10.1534/genetics.105.047449.

Llopart, A. (2012). The rapid evolution of X-linked male-biased gene expression and the large-X effect in *Drosophila yakuba, D. santomea*, and their hybrids. *Mol. Biol. Evol.* 29, 3873–3886. doi:10.1093/molbev/mss190.

Llopart, A., Brud, E., Pettie, N., and Comeron, J. M. (2018). Support for the dominance theory in drosophila transcriptomes. *Genetics* 210, 703–718. doi:10.1534/genetics.118.301229.

Mardiros, X. B., Park, R., Clifton, B., Grewal, G., Khizar, A. K., Markow, T. A., et al. (2016). Postmating Reproductive isolation between strains of Drosophila willistoni. *Fly (Austin).* 10, 162–171. doi:10.1080/19336934.2016.1197448.

Masly, J. P., and Presgraves, D. C. (2007). High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biol.* 5, 1890–1898. doi:10.1371/journal.pbio.0050243.

Matute, D. R., Butler, I. A., Turissini, D. A., and Coyne, J. A. (2010). A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science (80-.).* 329, 1518–1521. doi:10.1126/science.1193440.

Mayr, E. (1970). Populations, species, and evolution: an abridgment of animal species and evolution (Vol. 19). Harvard University Press.

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi:10.1101/gr.102491.109.

Meiklejohn, C. D., Parsch, J., Ranz, J. M., and Hartl, D. L. (2003). Rapid evolution of male-biased gene expression in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9894–9899. doi:10.1073/pnas.1630690100.

Meisel, R. P., Malone, J. H., and Clark, A. G. (2012). Faster-X Evolution of Gene Expression in Drosophila. *PLoS Genet.* 8. doi:10.1371/journal.pgen.1003013.

Michalak, P., and Noor, M. A. F. (2003). Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol. Biol. Evol.* 20, 1070–1076. doi:10.1093/molbev/msg119.

Moehring, A. J., Llopart, A., Elwyn, S., Coyne, J. A., and Mackay, T. F. C. (2006). The genetic basis of postzygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to hybrid male sterility. *Genetics* 173, 225–233. doi:10.1534/genetics.105.052985.

Moehring, A. J., Teeter, K. C., and Noor, M. A. F. (2007). Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Mol. Biol. Evol.* 24, 137–145. doi:10.1093/molbev/msl142.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biol. Symp.*

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics*.

Orr, H. A., and Irving, S. (2001). Complex epistasis and the genetic basis of hybrid sterility in the *Drosophila pseudoobscura* Bogota-USA hybridization. *Genetics* 158, 1089–1100.

Perez, D. E., Wu, C. I., Johnson, N. A., and Wu, M. L. (1993). Genetics of reproductive isolation in the Drosophila simulans clade: DNA marker-assisted mapping and characterization of a hybrid-male sterility gene, Odysseus (Ods). *Genetics*.

Phadnis, N. (2011). Genetic architecture of male sterility and segregation distortion in *Drosophila pseudoobscura* bogota-USA hybrids. *Genetics* 189, 1001–1009. doi:10.1534/genetics.111.132324.

Phadnis, N., and Malik, H. S. (2013). "The molecular and evolutionary basis of hybrid sterility: From Odysseus to Overdrive," in *Speciation: Natural Processes, Genetics and Biodiversity*.

Phadnis, N., and Orr, H. A. (2009). A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science (80-.).* 323, 376–379. doi:10.1126/science.1163934.

Prakash, S. (1972). Origin of reproductive isolation in the absence of apparent genic differentiation in a geographic isolate of Drosophila pseudoobscura. *Genetics* 72, 143–155.

Price, C. S. C. (1997). Conspecific sperm precedence in Drosophila. *Nature*. doi:10.1038/41753.

Ranz, J. M., Namgyal, K., Gibson, G., and Hartl, D. L. (2004). Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* 14, 373–379. doi:10.1101/gr.2019804.

Ranz, J. M., Ponce, A. R., Hartl, D. L., and Nurminsky, D. (2003). Origin and evolution of a new gene expressed in the Drosophila sperm axoneme. *Genetica* 118, 233–244. doi:10.1023/A:1024186516554.

Schaeffer, S. W., and Miller, E. L. (1991). Nucleotide sequence analysis of Adh genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U. S. A.* 88, 6097–6101. doi:10.1073/pnas.88.14.6097.

Sundararajan, V., and Civetta, A. (2011). Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in drosophila sterile hybrids. *J. Mol. Evol.* doi:10.1007/s00239-010-9404-5.

Swanson, W. J., and Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* doi:10.1038/nrg733.

Tao, Y., Zeng, Z. B., Li, J., Hartl, D. L., and Laurie, C. C. (2003). Genetic dissection of hybrid incompatibilities between Drosophila simulans and D. mauritiana. II. Mapping hybrid male sterility loci on the third chromosome. *Genetics*.

Ting, C. T., Tsaur, S. C., Wu, M. L., and Wu, C. I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science (80-.).* 282, 1501–1504. doi:10.1126/science.282.5393.1501.

True, J. R., and Haag, E. S. (2001). Developmental system drift and flexibility in evolutionary trajectories. *Evol. Dev.* 3, 109–119. doi:10.1046/j.1525-142X.2001.003002109.x.

Turissini, D. A., McGirr, J. A., Patel, S. S., David, J. R., and Matute, D. R. (2018). The rate of evolution of postmating-prezygotic reproductive isolation in *Drosophila*. *Mol. Biol. Evol.* 35, 312–334. doi:10.1093/molbev/msx271.

Wang, R. L., Wakeley, J., and Hey, J. (1997). Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. *Genetics* 147, 1091–1106.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. doi:10.1038/nature02698.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40, 346–350. doi:10.1038/ng.77.

Wittkopp, P. J., and Kalay, G. (2012). Cis-regulatory elements: Molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13, 59–69. doi:10.1038/nrg3095.

Zhang, Y., Sturgill, D., Parisi, M., Kumar, S., and Oliver, B. (2007). Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature* 450, 233–237. doi:10.1038/nature06323.

Zhang, Z., and Parsch, J. (2005). Positive correlation between evolutionary rate and recombination rate in Drosophila genes with male-biased expression. *Mol. Biol. Evol.* doi:10.1093/molbev/msi189.

Chapter 2: Genome-wide identification of regulatory interactions responsible for sterility in male hybrids between *D. willistoni willistoni* and *D. w. winge*

Alwyn C. Go and Alberto Civetta

Department of Biology, The University of Winnipeg

In collaboration with:

Jose M. Ranz

Department of Ecology and Evolutionary Biology, University of California, Irvine

ABSTRACT

Species pairs in the early stages of speciation provide an excellent opportunity to understand the genetic mechanisms behind reproductive isolation. Here we use two subspecies of *Drosophila willistoni*: *D. w. willistoni* and *D. w. winge*. This recently diverged subspecies pair show an early form of postzygotic reproductive isolation through unidirectional hybrid male sterility. Using RNA-sequencing, we identified genes specifically expressed in the testes, accessory glands, and ovaries of the subspecies pair and their reciprocal F_1 hybrids. We found a higher proportion of uniquely misexpressed genes in the sterile hybrid relative to the fertile hybrid and that these misexpressed genes showed a high degree of tissue-specificity. Consistent with the nature of the sterility phenotype, the testes had the largest proportion of misexpressed genes. We further performed an allele-specific expression analysis to determine the extent of regulatory divergence between subspecies and found a surprisingly large proportion of genes with conserved regulatory elements. Of those with divergent regulatory elements, *cis*-regulatory divergence was more common than *trans*- or *cis-trans* divergence. Among the misexpressed genes in the sterile hybrids, we found a significant protein-protein interaction networks suggesting that limited levels of regulatory divergence may be enough to cause hybrid breakdown if they disrupt the expression of closely interacting genes.

INTRODUCTION

The *D. willistoni* subspecies group presents a unique opportunity in identifying how post-zygotic reproductive isolation could develop in the presence of gene flow during the early stages of speciation. The group was recently suggested to have three subspecies, *D. w. quechua* (narrowly distributed around Lima, Peru west of the Andes), *D. w. willistoni* (found south of the American mainland, Mexico, and the Caribbean islands), and *D. w. winge* (distributed across much of the south American continent east of the Andes) (Mardiros et al. 2016). No formal analysis has been performed to estimate the divergence between the subspecies however, allozyme analyses between *D. w. quechua* and *D. w. willistoni* suggests a divergence time of at least 0.25 mya (Ayala and Tracey 1973; Ayala and Dobzhanksy 1974; Ayala et al. 1974). We focus on *D. w. willistoni* and *D. w. winge*. No fixed premating isolation has been found between them (Davis et al. 2020) and a haplotype analysis showed no evidence of genetic differentiation and a high degree of intermingling (Mardiros et al. 2016). Despite this, the subspecies pair show unidirectional hybrid male sterility wherein hybrid males with *D. w. willistoni* mothers are sterile while those with *D. w. winge* mothers are fertile. Females in either direction of the cross are fertile. Unlike other sterile male hybrids between *Drosophila* species that show abnormalities in sperm development (Prakash 1972; Orr 1989; Snook 1998; Gomes and Civetta 2014; Brill et al. 2016), sterile male hybrids of the *D. willistoni* subspecies pair produce motile sperm but an abnormal bulge in the basal end of the testes prevents sperm transfer into the female reproductive tract (Gomes and Civetta 2014; Civetta and Gaudreau 2015; Davis et al. 2020). This milder form of the

sterility phenotype could reflect the early divergence between the subspecies and the presence of gene flow.

When species hybridise, two divergent genomes are forced to interact with each other resulting in incompatibilities and misregulated gene expression. This regulatory dysfunction can lead to sterility through transgressive gene expression (*i.e.* expression levels above or below levels found in the parental species). Studies in *Drosophila* have found large proportions of transgressive gene expression in sterile interspecific hybrids (Michalak and Noor 2003; Ranz et al. 2004; Moehring et al. 2007; Gomes and Civetta 2015). Transgressive expression has also been disproportionately observed in male-biased genes (Michalak and Noor 2003; Ranz et al. 2004). This can be the consequence of rapidly diverging genes between species as tissue-specific genes and those with narrow breadths of expression have been found to experience faster rates of sequence evolution (Duret and Mouchiroud 2000; Zhang and Li 2004; Liao et al. 2006). Among these genes, those with male-biased expression showed greater rates of evolution compared to female-biased genes or non-biased genes (Mank et al. 2008; Meisel 2011; Assis et al. 2012). It is therefore reasonable to assume that the rapid evolution of genes between species as well as the regulatory elements that drive their expression could lead to transgressive expression in a hybrid background.

Gene regulation relies on the proper interactions between co-adapted *cis*- and *trans*-regulatory elements. In their simplest form, *cis*-regulatory elements are regions of non-coding DNA such as promoters or enhancers that act as binding sites for *trans*-acting (transcription) factors to regulate gene expression. The Bateson-Dobzhansky-Muller model (Dobzhansky 1937; Muller 1942; Orr 1996) explains how genes that function

normally in the genome of a pure-species become misexpressed in a hybrid genome. Regulatory divergence between parental species can be inferred using interspecific F_1 hybrids. Differences in transcript abundance between two alleles in the F_1 hybrid suggest changes in *cis* since these alleles are in a common *trans*-acting environment (Cowles et al. 2002). On the other hand, if the two alleles show the same level of transcript abundance, regulatory differences between the parental species are in *trans* (Wittkopp et al. 2004). This approach has been used to identify genome wide regulatory divergence between species of *Drosophila* (McManus et al. 2010; Coolon et al. 2014; Gomes and Civetta 2015).

As species continue to diverge, incompatibilities in the regulatory elements between them continue to accumulate (Orr 1995; Orr and Turelli 2001). Evidence for this "snowball" effect has been found in *Drosophila* (Matute et al. 2010) making it hard to disentangle between regulatory incompatibilities that arose after speciation from the incompatibilities that lead up to speciation. This highlights the importance of species pairs in the early stages of speciation. Here we use the *D. w. willistoni* and *D. w. winge* subspecies pair (referred to as Guadeloupe and Uruguay respectively hereafter). Using a *de novo* assembly of the *D. willistoni* genome and annotation, we performed a genome-wide survey on the subspecies pair and their sterile and fertile hybrids with RNA sequences obtained from the testes, accessory glands, and ovaries in an attempt to detect the regulatory differences between species that may be associated with sterility. We found that genes expressed in the testes had the most differential expression between subspecies suggesting tissue-specific patterns of regulation. The sterile hybrids had a larger proportion of transgressive expression than the fertile hybrids. Consistent with the

sterility phenotype, the testes had the largest proportion of transgressive expression in the hybrids with evidence of misregulation for genes involved in epidermal growth. Lastly, an allele-specific expression analysis revealed limited evidence of regulatory divergence between the subspecies, likely a consequence of gene flow. However, hybrid dysfunction can still occur if regulatory divergence causes misregulation of genes involved in a common network.

MATERIALS AND METHODS

RNA Sequencing

RNA extraction and sequencing were done by collaborators at the University of California Irvine Genomics High Throughput Facility. Briefly, total RNA was extracted using Trizol and purified with RNeasy Mini kit. RNA yield, purity, and integrity were evaluated using a Qubit 2.0 Fluorometer, a NanoDrop 8000 Spectrophotometer, or with a BioAnalyzer (Agilent Technologies Inc.) using the RNA 6000 Pico or RNA 6000 Nano kits. For gene annotation, one whole-body naïve male and one whole-body virgin female stranded, non-poly(A) enriched libraries were constructed using the TruSeq Stranded Total RNA Library Prep Kit (Illumina), and with Ribo-Zero Gold Set A (Epicenter). For assessing differences in gene expression among different *D. willistoni* subspecies and their hybrids, 36 (3 biological replicates × 4 genotypes × 3 tissues) stranded poly(A) enriched libraries were prepared with the TruSeq RNA Library prep kit v2 (Illumina). The cDNAs of all the libraries for each particular tissue were multiplexed and 100 bp paired-end sequenced over one line per tissue sample on an Illumina HiSeq 2500 instrument.

Differential gene expression analysis

Quality checks of the raw RNA-sequencing paired-end reads were performed using FastQC (Andrews 2010). Following this, the reads were processed using Trimmomatic (Bolger et al. 2014) to exclude reads with an average quality below a Phred score of 28 and a final length shorter than 36 bp. The processed paired-end reads were then mapped to a SoftMasked *de novo* genome assembly of the standard strain for *D.*

willistoni using STAR (Dobin et al. 2013) under default settings. Read counting was performed for each gene model using featureCounts (Liao et al. 2014) with the reversely stranded (-s 2) and fragment counting (-p) parameters and the *de novo* genome annotation serving as a guide.

Pairwise differential expression analysis across the parental subspecies and their hybrids for each of the three tissues (accessory glands, testes, and ovaries) were performed using both DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2010). For the analysis with edgeR, a minimum count-per-million (CPM) value was used for filtering rather than absolute read counts to avoid bias towards genes expressed in larger libraries (Chen et al. 2016). A cut-off value equivalent to at least 10 counts was used (Chen et al. 2016). Due to the differing library sizes across the tissues, a cut-off value of 1.5, 2.0, and 1.0 CPM was used for the testes, accessory glands, and ovaries respectively. Per gene counts for each sample were normalised using the TMM method (Robinson and Oshlack 2010). Further, in the analysis with DESeq2, per gene read counts were normalised with the default method and the independent filtering method was performed. Briefly, the independent filtering method increases the detection of significantly differentially expressed genes by automatically determining a threshold value, based on the mean of normalised counts over all samples, to filter lowly expressed genes. The local fit type was used for both the accessory glands and testes analyses while the parametric fit type was used for the ovaries analysis. A $\log_2$ fold-change threshold of 0.5 was applied to the results of both edgeR and DESeq2 to increase true positive rate (Schurch et al. 2016) and the consensus list of differentially expressed genes between both tools was

used for downstream analyses. All tools used for the differential gene expression analysis were ran on UseGalaxy (http://usegalaxy.org).

Identification of tissue-specific genes in the parental subspecies

We identified genes with tissue-specific expression in the parental subspecies using three different criteria. First, we performed independent differential expression analyses for tissue-specific samples in the parental subspecies using both edgeR (Robinson et al. 2010) and DESeq2 (Love et al. 2014) to find genes differentially expressed between tissues. Next, we required that differentially expressed genes showed at least a two-fold or higher expression in the focal tissue relative to the others. As the last criteria, we used the following formula to calculate a tissue-specificity score, τ, for each gene:

$$\tau = \frac{\sum_{i=1}^{N}(1-\hat{x}_i)}{N-1}\,;\, \hat{x}_i = \frac{x_i}{max(x_i)}.$$

Where N is three, the number of tissues (testes, ovaries, and accessory glands), and x_i is the expression of the gene in tissue i (Yanai et al. 2005). τ ranges from 0 to 1 with higher tissue-specificity represented by greater values. For context, a gene only expressed in one tissue will have a τ score of 1. We applied a threshold of $\tau \geq 0.9$ to only retain genes with high tissue-specificity (Larracuente et al. 2008; Assis et al. 2012).

Only genes that met all three criteria in both parental subspecies were considered tissue-specific.

Allele specific expression analysis

To determine the extent of *cis*- and *trans*-regulatory incompatibilities between the parental subspecies, we identified fixed species-specific single nucleotide polymorphisms (SNPs) and their relative allele-specific expression in the hybrids. SNPs between the parental subspecies were identified from their mapped reads using Naïve variant caller followed by processing with the Variant annotator (Blankenberg et al. 2014). SNPs were considered fixed in each parental subspecies if each parent had a single different allele and at least 3 supporting reads. Allele-specific expression in the hybrids was measured by first assigning their RNA-seq reads to a parent of origin based on the identity of the allele at fixed SNP positions in each parent. Reads with fixed SNPs mapping to a single gene were summed and any gene with less than 20 mapped reads from both parental subspecies combined were discarded from further analysis (McManus et al. 2010, Gomes and Civetta 2015). SNP counts for each gene were adjusted to account for differences in sequencing depth between samples and those with zero SNP counts were given a value of 1 to allow for statistical testing. Since male hybrids are hemizygous for the X chromosome, Guadeloupe and Uruguay alleles were inferred using sterile F_1 male (H4) and fertile F_1 male (H3) hybrids respectively for X-linked genes. Significant differences in gene expression between the parental subspecies and between alleles in the hybrids were determined using a binomial exact test. To keep consistency with the differential expression analysis using RNA-seq, a similar $\log_2$ fold-change threshold of 0.5 was applied in addition to the binomial exact test before considering SNP counts between the parental subspecies or between allele expression in the hybrids differentially expressed. To detect significant differences between the ratio of parental SNP counts to the ratio of

each parental allele in the hybrids, the Fisher's exact test was used. FDR corrected q-values were used for the binomial exact test and Fisher's exact test with a significance threshold of 0.5%. Types of regulatory divergence driving gene expression in the hybrids were categorised using the patterns of allele expression summarised in Table 2.1 (McManus et al. 2010, Gomes and Civetta 2015).

Table 2.1: Categories of regulatory divergence and their patterns of allelic expression. G = Guadeloupe, U = Uruguay, H_G = Guadeloupe allele in F_1 hybrid, H_U = Uruguay allele in F_1 hybrid. NS = non-significant differences in SNP counts, S = significant differences in SNP counts.

Regulatory Divergence	**G *vs.* U**	**H_G *vs.* H_U**	**G/U *vs.* H_G/H_U**
Conserved	NS	NS	NS
cis-only	S	S	NS
trans-only	S	NS	S
Compensatory	NS	S	S
cis and *trans*	S	S	S

RESULTS

Transcriptome sequencing

The transcriptome of the testes, accessory glands, and ovaries of the parental subspecies (Guadeloupe and Uruguay) as well as fertile and sterile hybrids (H3 and H4 respectively) were sequenced at the University of California Irvine Genomics High Throughput Facility. Three biological replicates were used for each genotype and tissue combination. Overall, 447 million reads were generated from all samples combined (Table S1). Samples from the ovaries showed the highest proportion of uniquely mapped reads while the accessory gland samples showed the lowest, with the differences across tissues being statistically significant (Table S2; nonparametric pair-wise Steel-Dwass test; ovaries vs testes, P_{adj}=0.0138; ovaries vs accessory glands, P_{adj}=0.0138; testes vs accessory glands, P_{adj}=0.0141). This could be the result of incomplete ribosomal depletion in the accessory gland and testes samples, an unanticipated biological difference across tissues, or both (Supplementary Text S1). Importantly, the similarities in the per tissue mapping proportions between the parental subspecies, Guadeloupe and Uruguay, suggests no mapping bias towards the reference genome (Table S2) (Kruskal Wallis; Ovaries, P=0.507 ; Accessory Glands, P=0.0495; Testes, P=0.827). Furthermore, principal component analyses (PCA) largely corroborated the expected grouping of the sequenced samples where samples from different tissues grouped apart from each other as expected (Figure 2.1A). Within tissues, the two parental subspecies grouped separately while the hybrids showed a pattern of additivity (*i.e.* intermediate between those of the subspecies) or dominance (*i.e.* substantially closer to one of two subspecies) (Figure 2.1B).

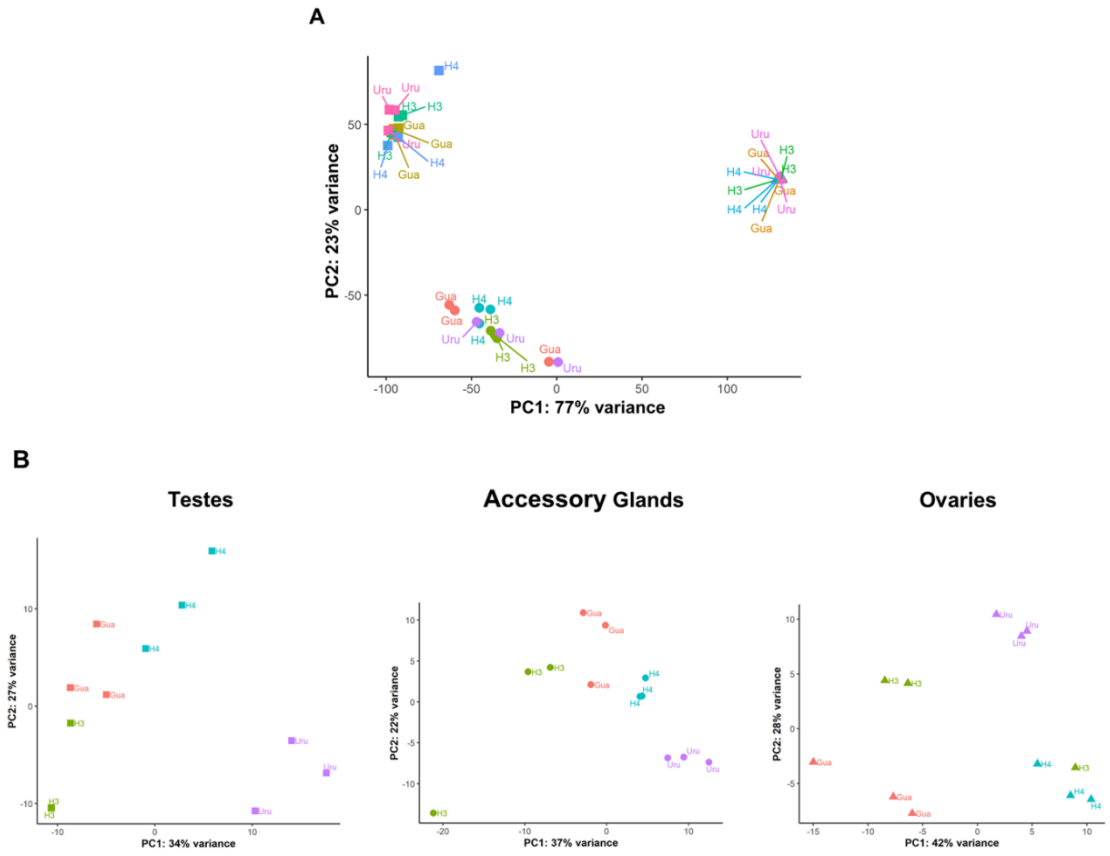


Figure 2.1: Principle Component Analysis (PCA) plot of the different RNA-sequenced samples. (A) Global and (B) per-tissue analysis. Testes (■); accessory glands (●); and ovaries (▲). Each point represents a sample, with biological replicates from the same genotype (i.e. parental subspecies or their hybrid progeny) sharing the same colour. The fraction of variance explained by each component is shown. Genotypes are Uruguay strain (Uru); Guadeloupe strain (Gua); H3, hybrid with Uruguay mother; and H4, hybrid with Guadeloupe mother.

Differences in expression between parental subspecies

At a 5% FDR, applying the conservative request of consistency across two commonly used approaches to detect statistically significant differences in expression, and requiring at least a $\log_2$ fold-change threshold of 0.5, we found that the testes exhibited proportionally more differentially expressed coding and lncRNA genes than accessory glands and ovaries between the parental subspecies (3-sample test for equality of proportions, $\chi^2=670.09$, d.f.=2, $P<2.2\times10^{-26}$; Table 2.2; Figure 2.2A). This pattern was consistent when the cut-off thresholds were increased to $\log_2$ fold-changes of 1 and 2 (Tables S5-S7, respectively). When focusing on the two directions of differential expression (*i.e.* genes overexpressed or underexpressed in Guadeloupe relative to Uruguay), we found evidence of a heterogenous association with tissue-type (3-sample test for equality of proportions, $\chi^2=6.911$, d.f.=2, $P=3.2\times10^{-2}$). This pattern is due to a significantly higher proportion of overexpressed genes in Guadeloupe accessory glands relative to Uruguay, compared to the testes and ovaries (Figure 2.2B). When considering the patterns of expression across all three tissue-types globally, we found only 46 (0.37%) of the differentially expressed genes with identical relationships between the parental subspecies. Most of the remaining differentially expressed genes, 6491 (52.74%), exhibit inconsistencies across tissues, in the directionality of expression, or both between the subspecies.

Table 2.2: Salient patterns of differential expression between the two parental subspecies.

	Gene Category		
Pattern	**Coding**	**Non-Coding ***	**All**
Testes	10,523	1,021 (980, 29, 4, 8)	11,544
Gua = Uru	8,846	639 (605, 25, 1, 8)	9,485 (82.16%)
Gua > Uru	839	235 (230, 2, 3, 0)	1,074 (9.30%)
Gua < Uru	838	147 (145, 2, 0, 0)	985 (8.53%)
Accessory Glands	9,030	616 (567, 34, 1, 14)	9,646
Gua = Uru	8,388	504 (461, 30, 1, 12)	8,892 (92.18%)
Gua > Uru	353	82 (78, 3, 0, 1)	435 (4.51%)
Gua < Uru	289	30 (28, 1, 0, 1)	319 (3.31%)
Ovaries	7,886	342 (333, 2, 0, 7)	8,228
Gua = Uru	7,312	252 (244, 1, 0, 7)	7,564 (91.93%)
Gua > Uru	290	60 (59, 1, 0, 0)	350 (4.25%)
Gua < Uru	284	30 (30, 0, 0, 0)	314 (3.82%)
All 3 samples #	11,022	1,285 (1,229, 34, 5, 17)	12,307
Consistent pattern	5,691	125 (124, 1, 0, 0)	5,816 (47.26%)
Gua = Uru	5,652	118 (118, 0, 0, 0)	5,770 (99.21%)
Gua > Uru	14	6 (5, 1, 0, 0)	20 (0.34%)
Gua < Uru	25	1 (1, 0, 0, 0)	26 (0.45%)
Inconsistent pattern ¦	5,331	1,160 (981, 32, 5, 17)	6,491(52.74%)

Direction of the differential expression between the two subspecies: > overexpression, < underexpression.

* In parenthesis the number of lncRNAs, rRNAs, tRNAs, and snoRNA, respectively.

Only genes expressed across the three types of biological samples.

¦ Genes that show differences in mRNA levels for at least one tissue in a given direction between the subspecies that are not observed in at least one other tissue.

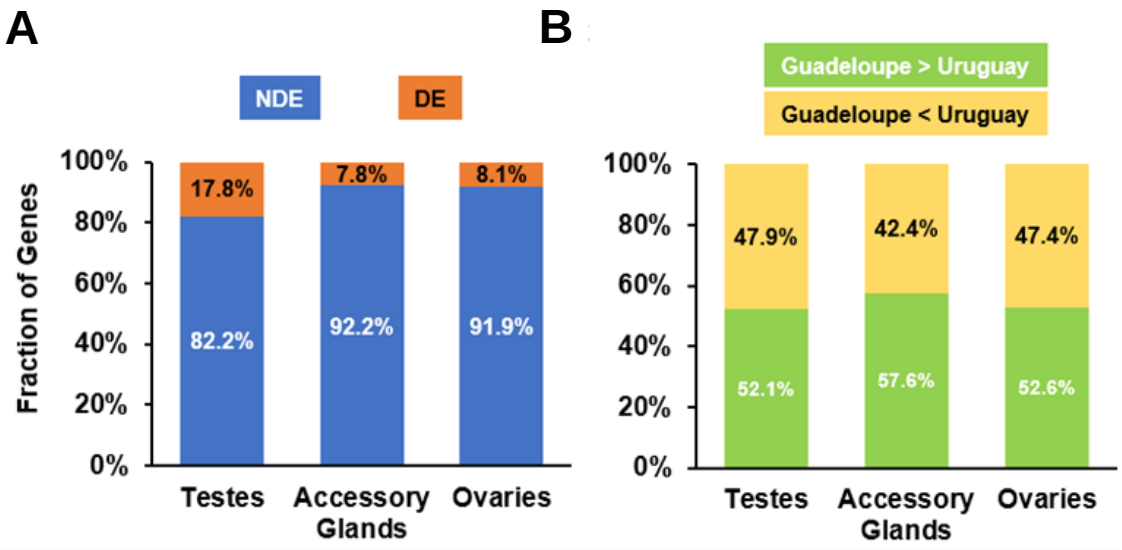


Figure 2.2: Relationship between tissue types and patterns of differential expression between the parental subspecies. (A) Frequency of non-differentially (NDE) and differentially (DE) expressed genes between the parental subspecies. Testes show significantly more differentially expressed than accessory glands and ovaries. (B) Frequency of differentially expressed genes relative to the species in which they exhibit overexpression. Accessory glands feature a more marked excess of overexpressed genes in Guadeloupe relative to Uruguay compared to testes and ovaries.

Tissue-specific genes in the parental subspecies

The three tissues assayed showed evidence of expressing 12,307 gene models, with similar numbers of gene models showing expression per genotype within tissues (Table S3). The testes, followed by the accessory glands, showed significantly more gene models expressed than ovaries. A pattern also observed when protein-coding and lncRNA genes are considered separately (one-way ANOVA, $P<0.001$ in all three contrasts; Table S4). Among the genes expressed in the tissues, we identified genes with tissue-specific expression in the parental subspecies. First, we obtained differentially expressed genes with at least a two-fold increase in expression in one tissue relative to the others. We then calculated a tissue-specificity score, τ, for these genes and only retained genes with τ

scores greater than 0.9. With these metrics, 2540 (22%), 636 (6.6%), and 269 (3.3%) of genes expressed were classified as testis, accessory gland, and ovary specific respectively in the Guadeloupe subspecies. The Uruguay subspecies had very similar proportions of testis, accessory gland, and ovary specific genes at 2508 (21.7%), 582 (6.0%), and 279 (3.4%) respectively. Between the two subspecies, 2270 (19.7%), 505 (5.2%), and 225 (2.7%) of genes were testis, accessory gland, and ovary specific respectively.

Previous studies on *Drosophila* showed a general trend of underrepresentation of male-biased genes and an overrepresentation of female-biased genes on the X-chromosomes (Parisi et al. 2003; Sturgill et al. 2007; Assis et al. 2012). Among the genes that showed tissue-specific expression in the parental subspecies, only 32% of testes genes and 19.2% of accessory gland genes were found on the X-chromosome compared to 48% of ovary specific genes being found on the X-chromosome. This low representation of X-chromosome genes in the male tissue compared to the ovary suggests that the observed trend of demasculinisation and feminisation of the *Drosophila* X-chromosome may also be occurring in the *D. willistoni* subspecies pair.

Patterns of hybrid expression

We examined the magnitude and patterns of expression in the sterile (H4) and fertile (H3) hybrids in relation to the parental subspecies. We categorized expression in the hybrids relative to the parentals as additive (*i.e.* hybrid gene expression falling within the ranges of the expression levels found in the parental subspecies) or transgressive (*i.e.* hybrid gene expression above or below the expression levels of the parental subspecies) (Table 2.3; Figure 2.3A). Within each tissue assayed, the fraction of differentially

expressed genes relative to the parental subspecies that are shared between the two hybrids instead of showing unique differential expression in either hybrid was significantly lower for both testes (18 shared vs. 349 unique) and accessory glands (4 shared vs. 116 unique) but similar for the ovaries (53 shared vs. 63 unique) (4.9% & 3.3% vs 45.7% respectively; two-tailed Fisher's exact test, $P=2.2\text{x}10^{-16}$) (Table 2.3; Figure 2.3B). This pattern was largely influenced by genes that showed additive rather than transgressive expression (two-tailed Fisher's exact test, 73 genes with shared additive expression and $P=1.99\text{x}10^{-14}$ vs. 2 with shared transgressive and $P=0.216$), a pattern that was consistent across all three tissues (testes:accessory glands:ovaries; additive: 17:4:52; transgressive: 1:0:1). These results indicate that the patterns of differential expression between the hybrids and the parental subspecies are dependent on tissue type.

Notably, genes displaying differential expression unique to only one of the hybrids were often biased toward an overrepresentation of differential expression in the sterile (H4) rather than fertile (H3) hybrid (Testes: 334 vs. 15; Accessory glands: 62 vs. 54; Ovaries: 52 vs. 11) (two-tailed Fisher's exact test, $P=2.2\text{x}10^{-16}$) (Table 2.3; Figure 2.3C). Moreover, the proportion of genes with additive expression were more abundant in the testes and accessory glands of sterile (H4) than fertile (H3) hybrids (Testes: 101 vs. 5; Accessory glands: 36 vs. 2) but had a more similar proportion between hybrids for the ovaries (22 vs. 11) (two-tailed Fisher's exact test, $P=5.87\text{x}10^{-5}$). In contrast, transgressive expression was more common in sterile (H4) than fertile (H3) hybrids for testes and ovaries (Testes: 233 vs. 10; and Ovaries: 30 vs. 0) but the pattern was reversed in the accessory glands (26 vs. 52) (two-tailed Fisher's exact test, $P=2.2\text{x}10^{-16}$). Overall, and

when considering both transgressive and additive differential expression jointly or separately among H4 and H3 hybrids relative to the parental subspecies, we found a prevalence of misexpression in the sterile (H4) relative to fertile (H3) hybrids (additive: 89.8% vs. 10.2%; transgressive: 82.3% vs. 17.7%; both: 84.8% vs. 15.2%).

As transgressive expression has been proposed to be particularly relevant in understanding hybrid sterility (Moehring et al. 2007; Catron and Noor 2008; Sundararajan and Civetta 2011; Gomes and Civetta 2015; Brill et al. 2016; Civetta 2016; Mack and Nachman 2017), we examined several aspects. First, we analysed the degree of commonality in the identity of genes showing transgressive expression across the three tissues assayed in both sterile and fertile hybrids. In sterile hybrids (H4), we found that only 9 (3.2%) of the 281 genes that showed transgressive expression do so in more than one tissue (Figure 2.4A). In fertile hybrids (H3), the pattern is similar with all 63 genes showing transgressive expression in only one particular tissue (Figure 2.4B). The comparison of the identity of genes with transgressive expression in H3 and H4 hybrids only showed one gene in common, GK14558, an orthologue of the *D. melanogaster* regulator of the Ras protein signal transduction pathway CG34393, which is misexpressed in the ovaries. These results indicate that first, transgressive misexpression is fundamentally tissue-dependent, a property observed in both hybrids, and second, that the virtual entirety of genes with transgressive expression in H3 and H4 hybrids are different.

Next, we analysed whether transgressively expressed genes were preferentially expressed in the tissue where they exhibit misexpression as genes with narrow expression (*i.e.* tissue-specific genes) are considered to be less pleiotropic and under weaker

selective constraints leading to higher rates of evolution (Mank et al. 2008; Meisel 2011; Assis et al. 2012). This makes genes with tissue-specific expression more susceptible to transgressive misexpression in the hybrids. We found that among the 224 transgressive genes uniquely misexpressed in the testes of H4 hybrids, only 31 (13.8%) were considered tissue-specific in Guadeloupe, 25 (11.2%) in Uruguay and 23 (10.3%) in both parental subspecies. None of the 10 transgressive genes unique to the testes of the H3 hybrids were considered tissue-specific in the parental subspecies. Of the 70 genes showing unique transgressive expression in the accessory glands of either H3 or H4 hybrids, we found 21 (30%) with tissue-specific expression in Guadeloupe, 25 (35.7%) in Uruguay, and 20 (28.6%) showing tissue-specific expression in both subspecies. For transgressive genes uniquely misexpressed in the ovaries of either hybrids, we found 8 (25.8%) genes with tissue-specific expression in Guadeloupe, 9 (29%) in Uruguay, and 8 (25.8%) in both parental species.

Overall, transgressive genes had a paucity for tissue-specificity, especially for genes misexpressed in the testes. This suggests that rapid rates of evolution do not fully explain the transgressive expression observed in the hybrids.

Table 2.3: Patterns of differential expression in hybrids relative to parental subspecies.

Tissue & Pattern	Category			Subtotal
	Unique to H3	Unique to H4	Shared	
Testes				
Additive	5 (2)	101 (23)	17 (4)	123 (29)
Transgression_over	10	202 (8)	1 (1)	213 (9)
Transgression_under	0	31 (3)	0	31 (3)
Subtotal	15 (2)	334 (31)	18 (5)	367 (38)
Accessory Glands				
Additive	2	36 (9)	4	42 (9)
Transgression_over	1	20 (1)	0	21 (1)
Transgression_under	51 (1)	6	0	57 (1)
Subtotal	54 (1)	62 (10)	4	120 (11)
Ovaries				
Additive	11 (2)	22 (4)	52 (8)	85 (14)
Transgression_over	0	15 (1)	1	16 (1)
Transgression_under	0	15	0	15
Subtotal	11 (2)	52 (5)	53 (8)	116 (15)

H3 (fertile male hybrid), Uruguay mother x Guadeloupe father; H4 (sterile male hybrid), Guadeloupe mother x Uruguay father. Female hybrids from both crosses are always fertile.

Differentially expressed lncRNAs are shown in parenthesis.

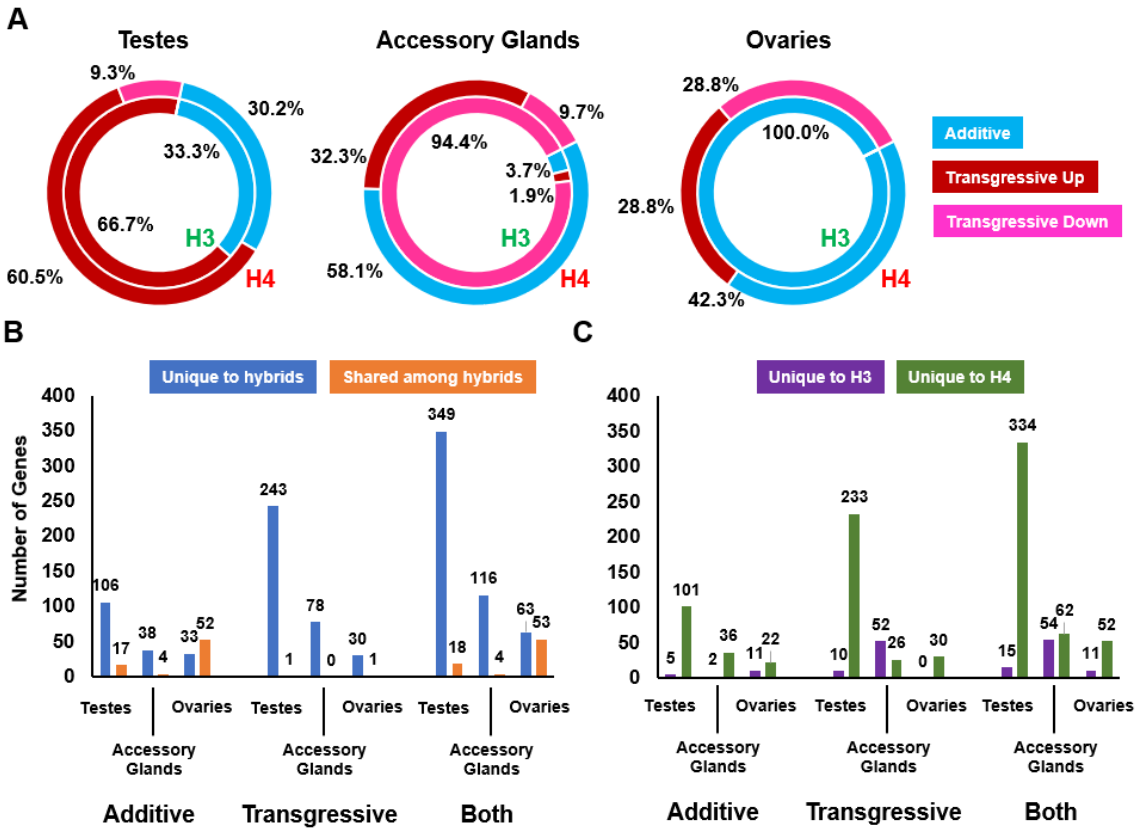


Figure 2.3: Patterns of differential expression in H3 and H4 hybrids relative to the parental subspecies of *D. willistoni*. (A) Pie charts showing the percentage of genes showing different patterns of differential expression across the three tissues assayed. Additive, when hybrid gene expression falls within the ranges of the expression levels of the parental subspecies; transgressive up and transgressive down, when hybrid gene expression is above or below the expression levels of the parental subspecies, respectively. (B) Bar graph showing the number of genes differentially expressed in the hybrids relative to the parental subspecies per tissue. Unique to hybrid, when the expression difference is only shown by one of the two hybrids; shared among hybrids, when the expression difference is shown by both hybrids. (C) Bar graph showing the break down of number of genes differentially expressed in one particular hybrid relative to the parental subspecies per tissue.

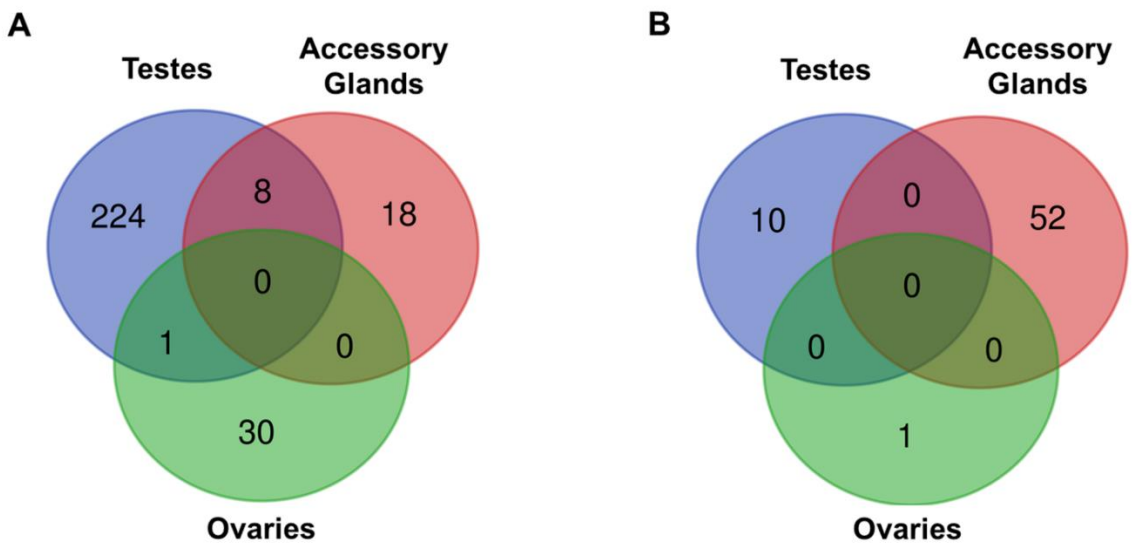


Figure 2.4: Venn diagram showing differentially expressed genes unique to each hybrid. (A) Shows transgressive genes in H4. (B) Shows transgressive genes in H3.

Identification of genome-wide regulatory incompatibilities in the hybrid background

Divergent regulatory elements present in the hybrid genome can cause transgressive expression. To determine the extent of such incompatibilities, we used fixed SNPs between the parental subspecies to identify allele-specific gene expression in the hybrids and infer the contributions of *cis*- and *trans*-divergence. In total, we identified nearly 48 million usable SNPs among the parental subspecies, the fertile, and sterile hybrids for all three tissues assayed. A summary of SNP counts per sample is presented in Table 2.4 and revealed an unexpected result. The H3 fertile hybrid is the F_1 progeny of Uruguay mothers and Guadeloupe fathers, since males obtain their X-chromosomes maternally, the expected abundance of allele specific SNPs in this hybrid should be $H3_{Uru} > H3_{Gua}$. The strikingly low abundance of Uruguay alleles in the H3 testes and accessory gland samples suggests the H3 hybrids may not be an F_1 progeny. Given the

uncertainty around the genomic background of H3 samples, they were discarded from further SNP analysis.

With the remaining parental and H4 sterile samples, we have identified 7,537 genes expressed in the ovaries with usable SNP information. Since determining the patterns of regulatory divergence in the hybrids requires the presence of both parental alleles, the analysis can only be performed on autosomal genes found in the testes and accessory glands, we identified 3,768 and 4,052 genes with usable SNPs in these tissues respectively. The SNP data were used to identify patterns of regulatory divergence for genes expressed in each of the three tissues. The results are summarised in Figure 2.5. We found that the majority of genes were conserved and showed no evidence of regulatory divergence (66.5%, 71.3% and 68.0% for the testes, accessory glands, and ovaries respectively). Of the remaining proportion of genes that showed evidence of regulatory divergence, we found that the gonads (*i.e.* testes and ovaries) were significantly driven by *cis*- instead of *trans*-regulatory divergence (Testes: 8.6% *cis*-only *vs.* 2.1% *trans*-only; $Z=12.544$; $P<0.00001$; Ovaries: 6.8% *cis*-only *vs.* 3.3% *trans*-only; $Z=9.75$; $P<0.00001$). On the other hand, genes in the accessory glands experienced higher divergence of *trans*-factors than *cis*-regulatory elements (5.6% *trans*-only *vs.* 4.4% *cis*-only; $Z=2.44$; $P=0.0147$).

Table 2.4: Total SNP counts for genes between the parental subspecies Guadeloupe (Gua) and Uruguay (Uru) and allele specific counts in the sterile (H4) and fertile (H3) hybrids.

Tissue	Gua	Uru	$H4_{Gua}$	$H4_{Uru}$	$H3_{Gua}$	$H3_{Uru}$
Testes	572,131	499,537	438,397	221,059	509,788	68,256
Accessory Glands	550,239	552,759	447,356	247,734	424,620	56,907
Ovaries	12,540,046	10,119,389	5,279,263	4,843,206	5,483,206	5,032,870

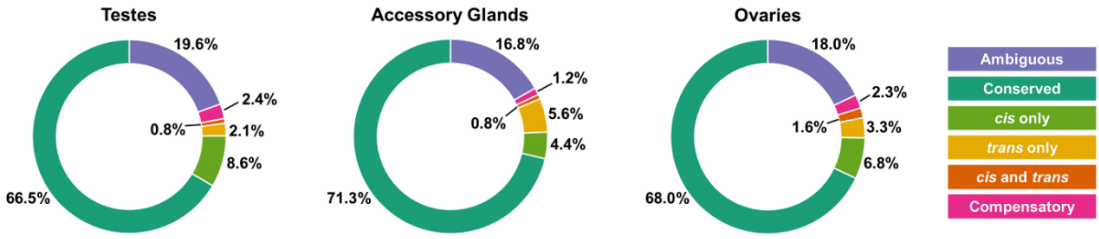


Figure 2.5: Types of regulatory divergence between the parental subspecies. Pie charts show the percentage of genes with the different types of regulatory divergence as inferred from the allele specific expression analysis and classified using significance patterns shown in table 2.1. Since males are hemizygous for the X chromosome, only autosomal genes were available for analysis in the testes and accessory glands samples. "*cis* only" refers to genes driven by *cis* regulatory elements, "*trans* only" for genes driven by *trans* factors. Genes driven by both *cis* and *trans* regulatory factors are termed "*cis* and *trans*". "Compensatory" refers to genes that show no significant expression differences between the parental subspecies despite having evidence of both *cis* and *trans* divergence. The term "Ambiguous" refers to cases where the expression patterns observed in the parental subspecies and the hybrid have no clear biological interpretation.

Functional clusters and interaction networks among transgressive genes

The presence of *cis*- and *trans*-regulatory incompatibilities within the hybrid background could lead to a cascade of transgressive expression. This predicts that transgressive genes expressed in the hybrid tissues will have clusters of functionally related proteins and the interaction of genes in a shared network or pathway. We tested

this prediction using default settings for STRING (v11.0; Szklarczyk et al. 2019) and by performing Gene Ontology functional annotations using g:Profiler (Raudvere et al. 2019).

The analysis of the 31 genes showing transgressive expression in the ovaries of H4 hybrids and the 28 genes with transgressive expression in the accessory glands of the sterile H4 male hybrids revealed small but significant (*i.e.* more interactions than randomly expected) protein-protein interactions (PPI) (Figure 2.6A; PPI enrichment $P=3.61x10^{-7}$ and Figure 2.6B; PPI enrichment $P=3.1x10^{-2}$, respectively). Transgressive genes in the ovaries had an overrepresentation of "Signal" genes based on UniProt keywords (FDR corrected $P=1.9x10^{-3}$) and a Gene Ontology: Molecular Function for peptidyl-dipeptidase activity (FDR corrected $P=1.341x10^{-2}$). No functional enrichment was found for transgressive genes in the accessory glands however, a KEGG pathway analysis showed an overrepresentation for genes belonging to the sphingolipid metabolism pathway (FDR corrected $P=9.8x10^{-3}$). Lastly, the 233 transgressive genes in the testes also showed a significant PPI network consisting of 58 nodes and 52 edges (Figure 2.7; PPI enrichment $P=3.38x10^{-12}$). A functional enrichment for 13 UniProt keywords (Table 2.5), 5 Gene Ontology: Molecular Functions (Table 2.6), and 12 Gene Ontology: Biological Processes (Table 2.7) were also identified among the transgressive genes in the testes of H4 hybrids.

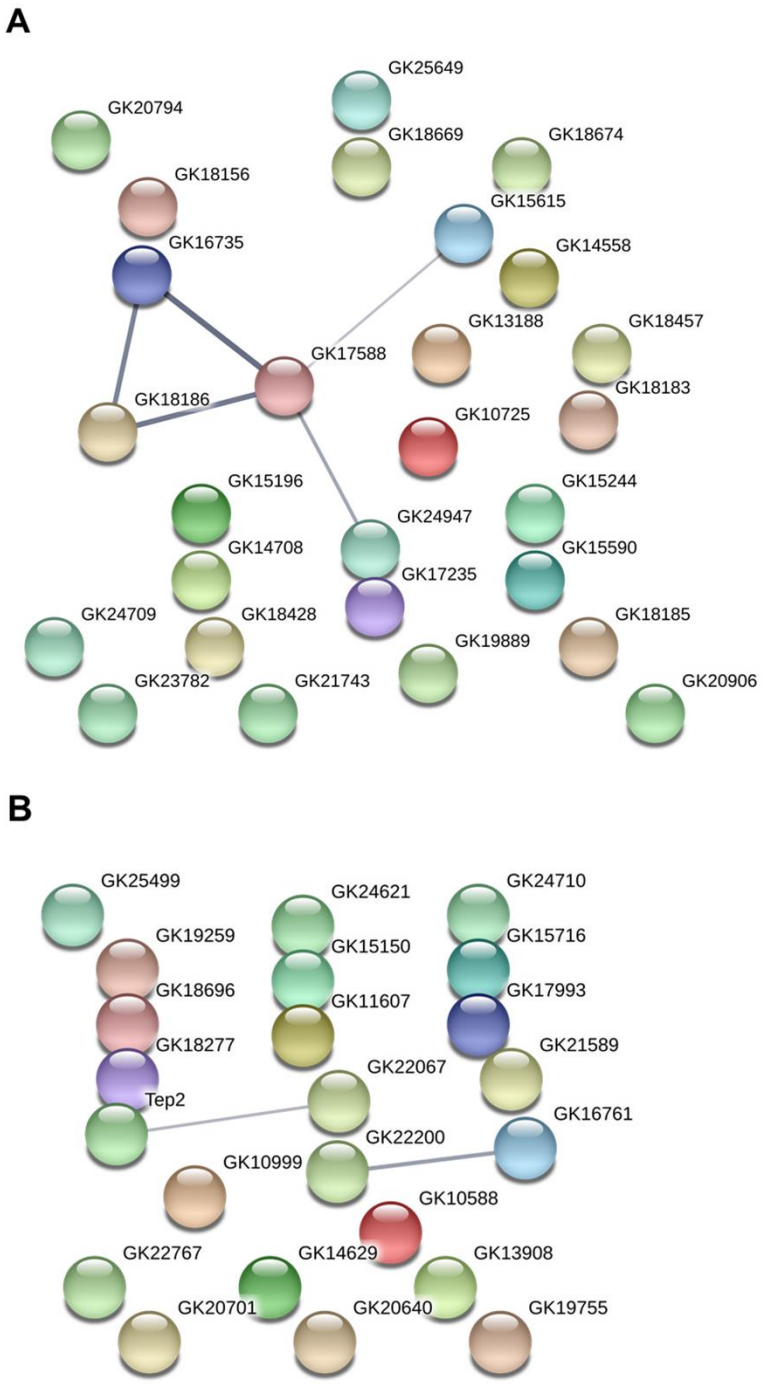


Figure 2.6: STRING PPI networks for transgressive genes expressed in the ovaries (A) and accessory glands (B) of H4 hybrids. Lines between nodes show interacting proteins with thickness representing the confidence of the interaction.

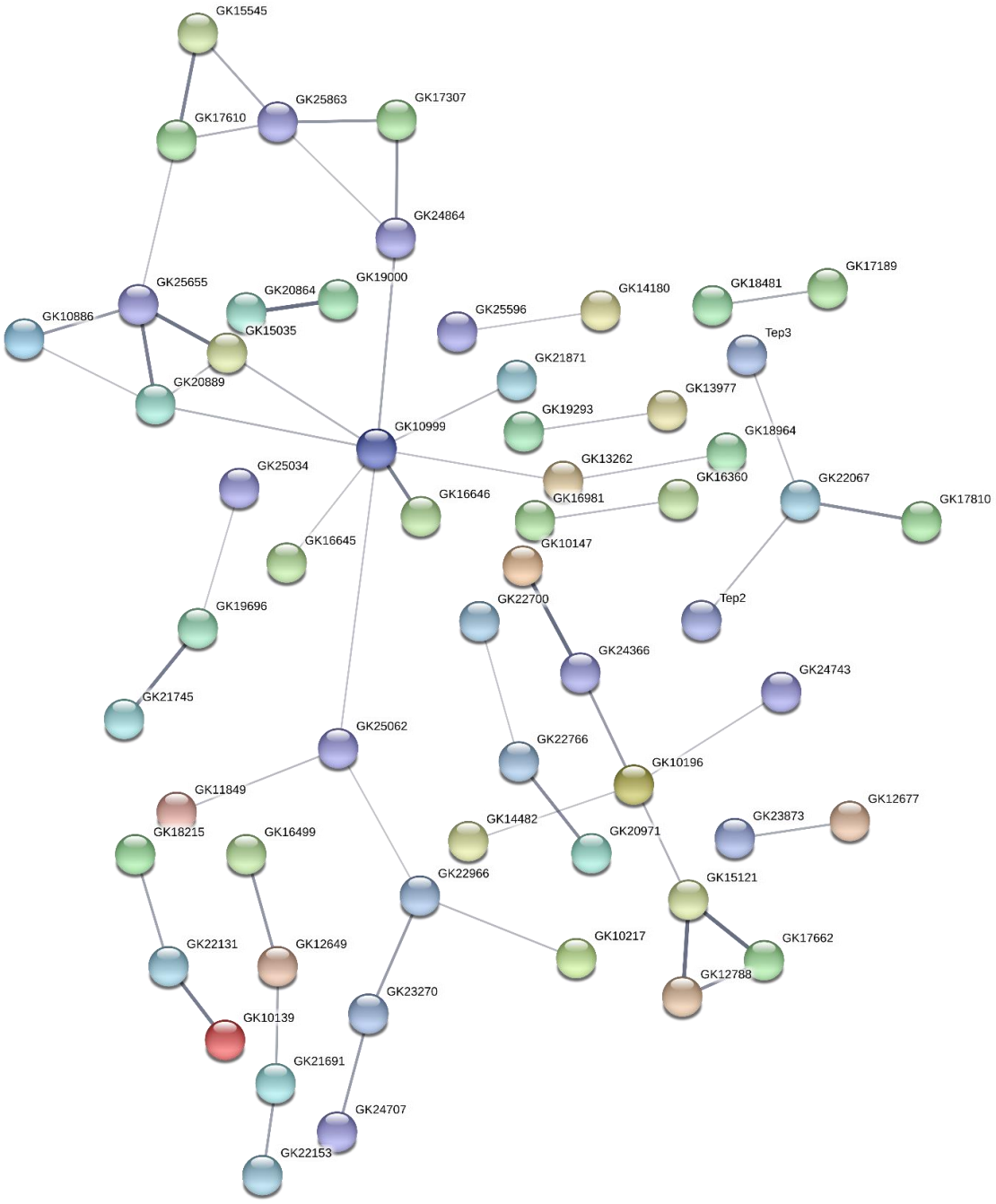


Figure 2.7: STRING PPI network for transgressive genes expressed in the testes of H4 hybrids. Only connected nodes are shown. Lines between nodes show interacting proteins with thickness representing the confidence of the interaction.

Table 2.5: Functional enrichment clusters based on UniProt keywords for genes showing transgressive expression in the testes of the H4 sterile male hybrids.

UniProt Keyword	**Count in Gene Set**	**FDR $P_{adj.}$**
Signal	69 of 2052	$9.84x10^{-11}$
Transmembrane helix	75 of 2576	$3.55x10^{-9}$
Glycosidase	7 of 74	$1.80x10^{-3}$
Actin-binding	5 of 36	$2.90x10^{-3}$
Repeat	14 of 336	$3.60x10^{-3}$
Disulfide bond	17 of 473	$3.90x10^{-3}$
Oxidoreductase	16 of 471	$9.10x10^{-3}$
Calcium/phospholipid-binding	2 of 2	$9.10x10^{-3}$
Annexin	2 of 2	$9.10x10^{-3}$
Integrin	2 of 5	$2.33x10^{-2}$
Cytoskeleton	4 of 44	$2.38x10^{-2}$
Lipid transport	2 of 6	$2.64x10^{-2}$
Cell adhesion	2 of 8	$3.89x10^{-2}$
Monooxygenase	5 of 88	$4.20x10^{-2}$
Laminin EGF-like domain	2 of 9	$4.20x10^{-2}$

Table 2.6: Overrepresented Gene Ontology: Molecular Functions for genes showing transgressive expression in the testes of the H4 sterile male hybrids as determined by g:Profiler.

Gene Ontology: Molecular Function	**Count in Gene Set**	**FDR $P_{adj.}$**
Imaginal disc growth factor receptor binding	4 of 5	$7.73x10^{-5}$
Chitinase activity	5 of 15	$6.77x10^{-4}$
Peptidase regulator activity	7 of 60	$1.26x10^{-2}$
Growth factor receptor binding	4 of 15	$1.85x10^{-2}$
Actin binding	8 of 88	$2.28x10^{-2}$

Table 2.7: Overrepresented Gene Ontology: Biological Processes for genes showing transgressive expression in the testes of the H4 sterile male hybrids as determined by g:Profiler.

Gene Ontology: Biological Processes	**Count in Gene Set**	**FDR $P_{adj.}$**
Apical junction assembly	9 of 34	$2.97x10^{-6}$
Septate junction assembly	8 of 30	$2.15x10^{-5}$
Cell-cell junction	9 of 43	$2.80x10^{-5}$
Tight junction organisation	8 of 31	$2.86x10^{-5}$
Tight junction assembly	8 of 31	$2.86x10^{-5}$
Cell-cell junction organisation	10 of 59	$4.00x10^{-5}$
Dorsal trunk growth, open tracheal system	4 of 9	$8.27x10^{-3}$
Cell adhesion	11 of 133	$1.31x10^{-2}$
Biological adhesion	11 of 133	$1.31x10^{-2}$
Wound healing	8 of 69	$1.76x10^{-2}$
Response to wounding	9 of 93	$2.32x10^{-2}$
Arp2/3 complex-mediated actin nucleation	4 of 13	$4.45x10^{-2}$

DISCUSSION

Here we performed a genome-wide expression analysis using RNA-sequences extracted from the testes, accessory glands, and ovaries of *D. w. willitsoni*, *D. w. winge*, their sterile F_1 male hybrid, and a fertile hybrid. Using tissue-specific transcriptomes, we identified how regulatory divergence between the parental subspecies could lead to gene misregulation and sterility in the hybrid. We found early signs of expression divergence between the parental subspecies where 17.8%, 7.8%, and 8.1% of genes expressed in the testes, accessory glands, and ovaries respectively were differentially expressed. The testes showed the highest proportion of differentially expressed genes between the subspecies, a trend not unexpected given that genes expressed in the testes and those involved in spermatogenesis are rapidly evolving among species of *Drosophila* (Coulthart and Singh 1988; Civetta and Singh 1995; Haerty et al. 2007; Harrison et al. 2015). Similar to the testes, genes expressed in the accessory glands are among the most rapidly evolving genes in *Drosophila* (Dorus et al. 2006). The involvement of these accessory gland proteins in male reproductive success (Ravi Ram and Wolfner 2007) causes natural selection to drive their rapid evolution (Swanson and Vacquier 2002). Given this, the relatively low proportion of differentially expressed genes in the accessory glands between the parental subspecies is somewhat surprising. However, not all genes that show expression in the accessory glands necessarily code for proteins involved in male reproduction and only around 200 seminal fluid proteins have been identified in *D. melanogaster* that are actively involved in male reproductive success (Findlay et al. 2008; Sepil et al. 2019).

Our analysis of tissue-specific genes showed that some genes considered tissue-specific in Guadeloupe did not show the same pattern in Uruguay. This suggests that a small degree of regulatory divergence between the subspecies at the tissue level might be

occurring. Among the three tissues, the testes showed the highest proportion of tissue-specific genes. Here we used the consensus of three different metrics to identify genes with tissue-specific expression. Although different methods used for the identification of tissue-specific genes could lead to different results due to differences in statistical testing and biological assumptions (Assis et al. 2012), the number of tissue-specific genes we identified in the *D. willistoni* subspecies pair followed the observed distributions seen in *D. melanogaster* and *D. pseudoobscura* (Chintapalli et al. 2007; Assis et al. 2012). The low proportion of accessory gland and ovary specific genes could be due to lower gene expression within these tissues. Alternatively, this could also suggest that genes expressed in both the accessory glands and ovaries may be more pleiotropic and involved in other non-tissue-specific functions. Some studies have shown that genes expressed in the seminal fluids produced by the accessory glands or within the female reproductive tract were also involved in immune functions (Samakovlis et al. 1991; Lung et al. 2001; Peng et al. 2005; Mueller et al. 2007).

Another interesting observation among the differentially expressed genes and those showing tissue-specific expression between the parental subspecies is the amount of long noncoding RNAs (lncRNAs). lncRNAs are involved with regulating gene expression at both the transcriptional and post-transcriptional levels. Not much is known about the role of lncRNAs in the context of speciation but studies on vertebrate evolution show that lncRNAs exhibit weaker functional constraint and rapid rates of turnover (reviewed in Kapusta and Feschotte 2014). An analysis of the *D. melanogaster* transcriptome revealed that 30% (or 563) of the identified lncRNAs had the highest expression in the testes with 125 of these lncRNAs only showing expression in the testes (Brown et al. 2014). Among the lncRNAs only expressed in the testes, individual knockouts of 32 lncRNAs led to sterility due to defects in spermatogenesis, these

lncRNAs have also been shown to undergo a more rapid evolution compared to protein-coding genes (Wen et al. 2016). The expression divergence we detected for lncRNAs expressed in the testes between the *D. willistoni* subspecies pair could indicate evidence of early divergence for the regulation of genes expressed in the testes of *D. willistoni*.

Overall, the proportion of differentially expressed genes between the *D. willistoni* subspecies pair is low compared to other distantly related species pairs of *Drosophila* (Table 2.8). A study on *D. melanogaster* and *D. simulans* which diverged ~2.5 mya (Cutter 2008) showed around 75% of differentially expressed genes between females of the species (Coolon et al. 2014). Between females of *D. melanogaster* and *D. sechellia*, which diverged 1.2 mya (Cutter 2008), 78% of genes were differentially expressed (McManus et al. 2010). 8% of genes were differentially expressed in the ovaries of *D. mojavensis* and *D. arizonae* (Lopez-Maestre et al. 2017) which diverged around 0.6-1 mya (Reed et al. 2007, 2008; Bono et al. 2009; Matzkin and Markow 2013). Coolon et al. (2014) found ~70% of genes were differentially expressed between females of *D. simulans* and *D. sechellia* which diverged around 0.25 mya (Garrigan et al. 2012). Lastly, a genome-wide analysis for RNA-sequences extracted from the male reproductive tract of *D. p. pseudoobscura* and *D. p. bogotana* which diverged around 0.23 mya (Schaeffer and Miller 1991; Wang et al. 1997) showed 14% of the genes were differentially expressed (Gomes and Civetta 2015). A caveat to both the Lopez-Maestre et al. (2017) and Gomes and Civetta (2015) studies is their use of 2 replicates for RNA-sequencing. This could lead to an underestimation of differentially expressed genes (Schurch et al. 2016). Overall, the low proportion of differentially expressed genes we found between the *D. willistoni* subspecies pair likely reflects their relatively recent divergence time and the degree of gene flow (Mardiros et al. 2016).

Despite the limited differential expression between the *D. willistoni* subspecies pair, the interaction of two divergent genomes during hybridisation could lead to hybrid dysfunction due to gene misregulation leading to transgressive expression. We found more genes with transgressive expression in the sterile (H4) hybrids relative to the fertile (H3) hybrids. Our SNP analysis showed that the H3 hybrids may not be the F_1 progeny of Uruguay and Guadeloupe, but instead might be a backcross progeny of the two subspecies. This limits the interactions between divergent regulatory elements in the H3 genomic background and could explain the low proportion of transgressive gene expression. However, fertile backcross progeny are still beneficial in this analysis as they can be used to differentiate between gene misregulation that might be associated with sterility from those that result from asymmetries in a hybrid genomic background (Michalak and Noor 2004; Ma et al. 2011; Brill et al. 2016; Alhazmi et al. 2019; Go et al. 2019).

Among the genes showing transgressive expression in the H4 hybrids, we found a propensity toward over-expression instead of under-expression. This is in stark contrast with other studies in *Drosophila* which showed a bias toward transgressive under-expression instead of over-expression in hybrids (Haerty and Singh 2006; Landry et al. 2007; Moehring et al. 2007; McManus et al. 2010; Llopart 2012; Coolon et al. 2014). Gomes and Civetta (2015) also found a higher proportion of transgressive over-expression instead of under-expression in sterile F_1 hybrids between the *D. pseudoobscura* subspecies pair and speculated that transgressive over-expression may be more common between sterile hybrids of *Drosophila* species that show a less severe sterility phenotype. Sterile hybrids between species of *Drosophila* that show an overrepresentation of transgressive under-expression were unable to produce individualised sperm (Kulathinal and Singh 1998; Moehring et al. 2006), while both

sterile hybrids from the *D. pseudoobscura* and *D. willistoni* subspecies pairs can produce individualised sperm, sterility in the case of the *D. willistoni* hybrids is due to the failure of transferring sperm into the female reproductive tract (Civetta and Gaudreau 2015; Gomes and Civetta 2014; Davis et al. 2020).

Hybrid male sterility in the *D. willistoni* subspecies pair is due to the misdevelopment of the testes forming a blockage at the basal end that prevents the transfer of sperm (Davis et al. 2020). Females in both directions of the cross are fertile and sterile males were shown to effectively transfer seminal fluids that triggered the appropriate female morphological response (Davis et al. 2020). This suggests that both the accessory glands and ovaries are experiencing little to no dysfunction. Interestingly, this is reflected by the proportions of transgressive gene expression in the H4 hybrids at the tissue level. Consistent with the sterility phenotype, the testes had the greatest proportion of transgressive gene expression. Furthermore, only 9 of the 281 transgressive genes in the sterile hybrid showed misexpression in more than one tissue suggesting that misregulation in the hybrids is tissue dependent. Given the tissue dependence of transgressive gene expression, we sought to determine whether tissue-specific genes were overrepresented among genes showing transgressive expression. Genes with narrow breadths of expression are likely functionally limited and less pleiotropic allowing them to evolve faster than genes with broader expression patterns (Duret and Mouchiroud 2000; Zhang and Li 2004; Liao et al. 2006; Haerty et al. 2007). Male-biased genes, especially those involved in spermatogenesis, experience faster rates of evolution than female- or non-biased genes (Coulthart and Singh 1988; Civetta and Singh 1995; Haerty et al. 2007; Harrison et al. 2015). This makes them more prone to misregulation in the hybrids. Surprisingly, only a small proportion of tissue-specific genes showed transgressive expression in our analysis suggesting that rapidly evolving genes are not the

main driving force for hybrid misregulation in the *D. willistoni* subspecies pair. Instead, the misregulation of genes with broader patterns of expression may be responsible for hybrid breakdown.

Using allele specific expression data, we provide a first-glance analysis of genome-wide regulatory divergence between the *D. willistoni* subspecies pair. Since the analysis required the expression of both parental alleles, we were only able to perform the analysis among autosomal genes for the H4 sterile male hybrids. We found a high proportion of genes that showed no evidence of regulatory divergence between the parental subspecies (*i.e.* conserved). Although our analysis only included autosomal genes for the H4 male hybrids, the analysis on the H4 female hybrids which included all genes with usable SNP information also showed an identical proportion of conserved regulatory elements between the subspecies. This is in contrast with other studies in *Drosophila* that showed a preponderance for *cis*-regulatory divergence or compensatory *cis*-*trans* mutations (Wittkopp et al. 2004; Wittkopp et al. 2008; Coolon et al. 2014; Brill et al. 2016). Gomes and Civetta (2015) also found a preponderance for *cis*-regulatory divergence for the closely related *D. pseudoobscura* subspecies pair suggesting that the recent divergence of the *D. willistoni* subspecies pair might not fully explain the high degree of conserved regulatory elements. The time of divergence between *D. w. willistoni* and *D. w. winge* is unknown and a haplotype analysis using a mitochondrial barcoding gene for this subspecies pair found a considerable degree of gene flow which may have reduced the amount of deleterious interactions (Mardiros et al. 2016) that were allowed to accumulate in the geographically separated *D. pseudoobscura* subspecies pair (Dobzhansky 1936; Dobzhansky et al. 1964).

Among the genes that showed evidence of regulatory divergence, we found a significantly higher proportion of *cis*- rather than *trans*-regulatory changes. Although

trans-factors have more genome-wide targets than *cis*-regulatory elements, changes in *cis* can cause a cascade of misexpression if they affect the expression of *trans*-factors. An analysis of the *D. melanogaster* and *Caenorhabditis elegans* genomes showed that genes encoding proteins with high degrees of regulatory complexity (e.g. transcription factors and signaling proteins) are flanked by large regions of non-coding DNA compared to other genes with more limited functions (Nelson et al. 2004). This suggests that genes encoding transcription factors or signaling proteins contain more enhancers making them more susceptible to acquiring *cis*-regulatory changes than other classes of genes. However, we found no evidence of *cis*-regulatory changes affecting the expression of known transcription factors among the genes showing transgressive expression in the hybrids. Alternatively, *cis* factors themselves can cause misregulation through interactions with other *cis* elements (Schoenfelder and Fraser 2019). *Cis*-regulatory elements are non-coding segments of DNA and include promoters and enhancers. Promoters are immediately upstream of the transcription start site and recruit transcription factors and RNA polymerase II to initiate transcription. Enhancers, on the other hand, activate or increase the expression of their target genes and can be located further upstream, downstream, or within introns. Enhancer-promoter interactions occur along with gene expression and there is evidence to support that the regulatory information to direct transcription is conveyed through enhancer-promoter interactions (Carter et al. 2002). Furthermore, enhancer-promoter interactions alone can induce transcription in the absence of transcription factors through forced chromatin looping (Deng et al. 2012). A study of the *D. melanogaster* genome showed that each enhancer on average interacted with multiple other enhancers and promoters and that such interactions are common in the highly compact *D. melanogaster* genome (Ghavi-Helm et

al. 2014). It is therefore possible for widespread gene misregulation to occur as a result of novel enhancer-promoter interactions found in a hybrid genome.

Gene misregulation can cause hybrid dysfunction if they lead to significant disruptions among interacting genes in a shared network. We focus on the testes since hybrid male sterility between the *D. willistoni* subspecies pair is due to improper testes development (Davis et al. 2020). Among the transgressive genes misexpressed in the testes, we found a significant protein-protein interaction (PPI) network of 58 genes (Figure 2.7), several overrepresented UniProt keywords and biological processes. However, a caveat to these overrepresentations is that small gene set sizes could lead to inflated significance scores. Nonetheless, the misregulation of genes with molecular functions suggested by the UniProt keywords and the biological processes determined by gene ontology could lead to disruptions in the development of the male reproductive tract. In *Drosophila*, development of the male reproductive tract depends on the recognition and fusion of two separate tissues, the genital disc and gonads (Rothenbusch-Fender et al. 2017). The gonads develop to form the testes while the genital disc forms the internal male reproductive organs (accessory glands, seminal vesicles, and ejaculatory bulb) and the external genitalia (Stern 1941; Greig and Akam 1995; Estrada et al. 2003). During pupation, myoblast cells start to accumulate around the developing seminal vesicle and form myotubes (Kuckwa et al. 2016). These myotubes then migrate toward the developing testes and form a muscle sheath that surrounds the developing testes (Kozopas et al. 1998; Kuckwa et al. 2016). Underneath this muscle sheath, the epithelia of the testes and seminal vesicles begin to fuse and form a continuous passage, the testes then begin to take on their spiral shape shortly after (Stern 1941). Overall, the development of the male reproductive tract requires proper coordination among several gene classes and disruption in the expression of these genes could lead to significant

defects in tissue development. Cytoskeletal components and actin proteins have been implicated in the cell migration process (Campellone and Welch 2010), while cell adhesion genes can help mediate fusion between neighbouring cells (Bulgakova et al. 2012) or help cells transition from an adhesive state to a migratory state (Lim and Thiery 2012). We found that genes (GK10886, GK13346, GK15121, GK17662, GK12788, GK22131, GK25655, and GK20889) with these molecular functions were misexpressed in the testes of the sterile male hybrids suggesting that the abnormal blockage that prevents the transfer of sperm stems from failures in early testes development. Given that hybrid breakdown likely occurs due to failures in early development, the levels of gene misexpression observed in the adult stage may not accurately reflect the degree of misregulation required for hybrid breakdown. Instead, focus should be directed toward the pupal stage for better characterisation of the genes involved in hybrid male sterility between the *D. willistoni* subspecies pair.

Table 2.8: Summary of genome-wide expression analyses between different species of *Drosophila* and the predominant type of regulatory divergence seen in their hybrids.

Species Pairs	Divergence	Expression Divergence	Hybrids	Approach	Conclusions	Citation
D. melanogaster and *D. simulans*	~2.5 mya	75% between females.	F_1 female hybrids with *D. melanogaster* mothers.	RNA-seq and allele-specific expression analysis.	Regulatory divergence between species is mostly in *cis*.	Coolon et al. 2014
D. melanogaster and *D. sechellia*	~1.2 mya	78% between females.	F_1 female hybrids with *D. melanogaster* mothers.	RNA-seq and allele-specific expression analysis.	*Trans*-regulatory divergence affected gene expression divergence the most.	McManus et al. 2010
D. mojavensis and *D. arizonae*	0.6-1 mya	8% between ovaries.	Reciprocal F_1 female hybrids.	RNA-seq and small RNA-seq.	Absence of piRNAs in hybrids leads to misregulation of transposable elements.	Lopez-Maestre et al. 2017
D. simulans and *D. sechellia*	0.25 mya	70% between females.	F_1 female hybrids with *D. simulans* mothers.	RNA-seq and allele-specific expression analysis.	Regulatory divergence between species is mostly in *cis*.	Coolon et al. 2014
D. p. pseudoobscura and *D. p. bogotana*	0.15-0.23 mya	14% between male reproductive tracts.	Reciprocal F_1 male hybrids.	RNA-seq and allele-specific expression analysis.	Regulatory incompatibilities in hybrids mostly driven by *cis*-only regulatory divergence.	Gomes and Civetta 2015
D. w. willistoni and *D. w. winge*	No formal estimates of divergence.	18% between testes, 8% between accessory glands and ovaries.	Reciprocal F_1 male and female hybrids.	RNA-seq and allele-specific expression analysis.	Regulatory elements between species are mostly conserved.	This chapter.

REFERENCES

Alhazmi, D., Fudyk, S. K., and Civetta, A. (2019). Testes proteases expression and hybrid male sterility between subspecies of *Drosophila pseudoobscura*. *G3 Genes, Genomes, Genet.* 9, 1065–1074. doi:10.1534/g3.119.300580.

Andrews, S. (2010). FastQC. *Babraham Bioinforma.* doi:citeulike-article-id:11583827.

Assis, R., Zhou, Q., and Bachtrog, D. (2012). Sex-biased transcriptome evolution in drosophila. *Genome Biol. Evol.* 4, 1189–1200. doi:10.1093/gbe/evs093.

Ayala, F.J. & Dobzhansky, T.H. (1974). A new subspecies of Drosophila pseudoobscura (Diptera: Drosophilidae). *Pan-Pac. Entomol.* 50: 211–219.

Ayala, F. J., and Tracey, M. L. (1973). Enzyme variability in the Drosophila willistoni group: VIII. Genetic differentiation and reproductive isolation between two subspecies. *J. Hered.* doi:10.1093/oxfordjournals.jhered.a108367.

Ayala, F. J., Tracey, M. L., Hedgecock, D., and Richmond, R. C. (1974). Genetic Differentiation During the Speciation Process in Drosophila. *Evolution*. doi:10.2307/2407283.

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., et al. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* doi:10.1186/gb4161.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bono, J. M., and Markow, T. A. (2009). Post-zygotic isolation in cactophilic Drosophila: Larval viability and adult life-history traits of D. mojavensis/D. arizonae hybrids. *J. Evol. Biol.* doi:10.1111/j.1420-9101.2009.01753.x.

Brill, E., Kang, L., Michalak, K., Michalak, P., and Price, D. K. (2016). Hybrid sterility and evolution in Hawaiian *Drosophila*: Differential gene and allele-specific expression analysis of backcross males. *Heredity (Edinb).* 117, 100–108. doi:10.1038/hdy.2016.31.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., et al. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*. doi:10.1038/nature12962.

Bulgakova, N. A., Klapholz, B., and Brown, N. H. (2012). Cell adhesion in Drosophila: versatility of cadherin and integrin complexes during development. *Curr. Opin. Cell Biol.* doi:10.1016/j.ceb.2012.07.006.

Campellone, K. G., and Welch, M. D. (2010). A nucleator arms race: Cellular control of actin assembly. *Nat. Rev. Mol. Cell Biol.* doi:10.1038/nrm2867.

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y. feng, and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nat. Genet.* 32, 623–626. doi:10.1038/ng1051.

Catron, D. J., and Noor, M. A. F. (2008). Gene expression disruptions of organism versus organ in Drosophila species hybirds. *PLoS One*. doi:10.1371/journal.pone.0003009.

Chen, Y., Lun, A. T. L., and Smyth, G. K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* 5, 1–48. doi:10.12688/F1000RESEARCH.8987.2.

Chintapalli, V. R., Wang, J., and Dow, J. A. T. (2007). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat. Genet.* 39, 715–720. doi:10.1038/ng2049.

Civetta, A. (2016). Misregulation of gene expression and sterility in interspecies hybrids: Causal links and alternative hypotheses. *J. Mol. Evol.* 82, 176–182. doi:10.1007/s00239-016-9734-z.

Civetta, A., and Gaudreau, C. (2015). Hybrid male sterility between Drosophila willistoni species is caused by male failure to transfer sperm during copulation. *BMC Evol. Biol.* 15, 1–8. doi:10.1186/s12862-015-0355-8.

Civetta, A., and Singh, R. S. (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. *J. Mol. Evol.* doi:10.1007/BF00173190.

Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in *Drosophila*. *Genome Res.* 24, 797–808. doi:10.1101/gr.163014.113.

Coulthart, M. B., and Singh, R. S. (1988). High level of divergence of male-reproductive-tract proteins, between Drosophila melanogaster and its sibling species, D. simulans. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a040484.

Cowles, C. R., Hirschhorn, J. N., Altshuler, D., and Lander, E. S. (2002). Detection of regulatory variation in mouse genes. *Nat. Genet.* 32, 432–437. doi:10.1038/ng992.

Coyne, J. A., and Orr, H. A. (1989). Patterns of Speciation in *Drosophila*. *Evolution* 43, 362. doi:10.2307/2409213.

Cutter, A. D. (2008). Divergence times in Caenorhabditis and Drosophila inferred from direct estimates of the neutral mutation rate. *Mol. Biol. Evol.* doi:10.1093/molbev/msn024.

Davis, H., Sosulski, N., and Civetta, A. (2020). Reproductive isolation caused by azoospermia in sterile male hybrids of Drosophila. *Ecol. Evol.* 10, 5922–5931. doi:10.1002/ece3.6329.

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., et al. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell* 149, 1233–1244. doi:10.1016/j.cell.2012.03.051.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Dobzhansky, T. H. (1936). Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics*, *21*(2), 113.

Dobzhansky T. (1937). Genetics and the origin of species. In Columbia biological series. New York: Columbia University Press

Dobzhansky, T., Anderson, W. W., Pavlovsky, O., Spassky, B., and Wills, C. J. (1964). Genetics of Natural Populations. XXXV. A Progress Report on Genetic Changes in Populations of Drosophila pseudoobscura in the American Southwest. *Evolution*. doi:10.2307/2406389.

Dorus, S., Busby, S. A., Gerike, U., Shabanowitz, J., Hunt, D. F., and Karr, T. L. (2006). Genomic and functional evolution of the Drosophila melanogaster sperm proteome. *Nat. Genet.* doi:10.1038/ng1915.

Duret, L., and Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol. Biol. Evol.* doi:10.1093/oxfordjournals.molbev.a026239.

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.

Findlay, G. D., Yi, X., MacCoss, M. J., and Swanson, W. J. (2008). Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. *PLoS Biol.* 6, 1417–1426. doi:10.1371/journal.pbio.0060178.

Garrigan, D., Kingan, S. B., Geneva, A. J., Andolfatto, P., Clark, A. G., Thornton, K. R., et al. (2012). Genome sequencing reveals complex speciation in the Drosophila simulans clade. *Genome Res.* doi:10.1101/gr.130922.111.

Ghavi-Helm, Y., Klein, F. A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., et al. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* 512, 96–100. doi:10.1038/nature13417.

Go, A., Alhazmi, D., and Civetta, A. (2019). Altered expression of cell adhesion genes and hybrid male sterility between subspecies of *Drosophila pseudoobscura*. *Genome* 62, 657–663. doi:10.1139/gen-2019-0066.

Gomes, S., and Civetta, A. (2014). Misregulation of spermatogenesis genes in *Drosophila* hybrids is lineage-specific and driven by the combined effects of sterility and fast male regulatory divergence. *J. Evol. Biol.* 27, 1775–1783. doi:10.1111/jeb.12428.

Gomes, S., and Civetta, A. (2015). Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci. Rep.* 5, 11976. doi:10.1038/srep11976.

Greig, S., and Akam, M. (1995). The role of homeotic genes in the specification of the Drosophila gonad. *Curr. Biol.* doi:10.1016/S0960-9822(95)00210-7.

Haerty, W., Jagadeeshan, S., Kulathinal, R. J., Wong, A., Ram, K. R., Sirot, L. K., et al. (2007). Evolution in the fast lane: Rapidly evolving sex-related genes in Drosophila. *Genetics* 177, 1321–1335. doi:10.1534/genetics.107.078865.

Haerty, W., and Singh, R. S. (2006). Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. *Mol. Biol. Evol.* 23, 1707–1714. doi:10.1093/molbev/msl033.

Harrison, P. W., Wright, A. E., Zimmer, F., Dean, R., Montgomery, S. H., Pointer, M. A., et al. (2015). Sexual selection drives evolution and rapid turnover of male gene expression. *Proc. Natl. Acad. Sci. U. S. A.* 112, 4393–4398. doi:10.1073/pnas.1501339112.

Kapusta, A., and Feschotte, C. (2014). Volatile evolution of long noncoding RNA repertoires: Mechanisms and biological implications. *Trends Genet.* doi:10.1016/j.tig.2014.08.004.

Kozopas, K. M., Samos, C. H., and Nusse, R. (1998). DWnt-2, a Drosophila Wnt gene required for the development of the male reproductive tract, specifies a sexually dimorphic cell fate. *Genes Dev.* 12, 1155–1165. doi:10.1101/gad.12.8.1155.

Kuckwa, J., Fritzen, K., Buttgereit, D., Rothenbusch-Fender, S., and Renkawitz-Pohl, R. (2016). A new level of plasticity: Drosophila smooth-like testes muscles compensate failure of myoblast fusion. *Dev.* doi:10.1242/dev.126730.

Kulathinal, R., and Singh, R. S. (1998). Cytological characterization of premeiotic versus postmeiotic defects producing hybrid male sterility among sibling species of the *Drosophila melanogaster* complex. *Evolution* 52, 1067. doi:10.2307/2411237.

Landry, C. R., Hartl, D. L., and Ranz, J. M. (2007). Genome clashes in hybrids: Insights from gene expression. *Heredity (Edinb).* 99, 483–493. doi:10.1038/sj.hdy.6801045.

Larracuente, A. M., Sackton, T. B., Greenberg, A. J., Wong, A., Singh, N. D., Sturgill, D., et al. (2008). Evolution of protein-coding genes in Drosophila. *Trends Genet.* 24, 114–123. doi:10.1016/j.tig.2007.12.001.

Liao, B. Y., Scott, N. M., and Zhang, J. (2006). Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. *Mol. Biol. Evol.* doi:10.1093/molbev/msl076.

Liao, Y., Smyth, G. K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656.

Lim, J., and Thiery, J. P. (2012). Epithelial-mesenchymal transitions: Insights from development. *Dev.* doi:10.1242/dev.071209.

Llopart, A. (2012). The rapid evolution of X-linked male-biased gene expression and the large-X effect in *Drosophila yakuba*, *D. santomea*, and their hybrids. *Mol. Biol. Evol.* 29, 3873–3886. doi:10.1093/molbev/mss190.

Lopez-Maestre, H., Carnelossi, E. A. G., Lacroix, V., Burlet, N., Mugat, B., Chambeyron, S., et al. (2017). Identification of misexpressed genetic elements in hybrids between Drosophila-related species. *Sci. Rep.* 7, 1–13. doi:10.1038/srep40618.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold-change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15, 1–21. doi:10.1186/s13059-014-0550-8.

Lung, O., Kuo, L., and Wolfner, M. F. (2001). Drosophila males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates. *J. Insect Physiol.* doi:10.1016/S0022-1910(00)00151-7.

Ma, D., and Michalak, P. (2011). Ephemeral association between gene CG5762 and hybrid male sterility in Drosophila sibling species. *J. Mol. Evol.* doi:10.1007/s00239-011-9466-z.

Mack, K. L., and Nachman, M. W. (2017). Gene Regulation and Speciation. *Trends Genet.* 33, 68–80. doi:10.1016/j.tig.2016.11.003.

Mank, J. E., Hultin-Rosenberg, L., Zwahlen, M., and Ellegren, H. (2008). Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. *Am. Nat.* 171, 35–43. doi:10.1086/523954.

Mardiros, X. B., Park, R., Clifton, B., Grewal, G., Khizar, A. K., Markow, T. A., et al. (2016). Postmating Reproductive isolation between strains of Drosophila willistoni. *Fly (Austin).* 10, 162–171. doi:10.1080/19336934.2016.1197448.

Matute, D. R., Butler, I. A., Turissini, D. A., and Coyne, J. A. (2010). A test of the snowball theory for the rate of evolution of hybrid incompatibilities. *Science (80-.).* 329, 1518–1521. doi:10.1126/science.1193440.

Matzkin, L. M., and Markow, T. A. (2013). "Transcriptional differentiation across the four subspecies of Drosophila Mojavensis," in *Speciation: Natural Processes, Genetics and Biodiversity*.

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi:10.1101/gr.102491.109.

Meisel, R. P. (2011). Towards a more nuanced understanding of the relationship between sex-biased gene expression and rates of protein-coding sequence evolution. *Mol. Biol. Evol.* 28, 1893–1900. doi:10.1093/molbev/msr010.

Michalak, P., and Noor, M. A. F. (2003). Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol. Biol. Evol.* 20, 1070–1076. doi:10.1093/molbev/msg119.

Moehring, A. J., Llopart, A., Elwyn, S., Coyne, J. A., and Mackay, T. F. C. (2006). The genetic basis of postzygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to hybrid male sterility. *Genetics* 173, 225–233. doi:10.1534/genetics.105.052985.

Moehring, A. J., Teeter, K. C., and Noor, M. A. F. (2007). Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Mol. Biol. Evol.* 24, 137–145. doi:10.1093/molbev/msl142.

Mueller, J. L., Page, J. L., and Wolfner, M. F. (2007). An ectopic expression screen reveals the protective and toxic effects of Drosophila seminal fluid proteins. *Genetics*. doi:10.1534/genetics.106.065318.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biol. Symp*.

Nelson, C. E., Hersh, B. M., and Carroll, S. B. (2004). The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* doi:10.1186/gb-2004-5-4-r25.

Orr, H. A. (1995). The population genetics of speciation: The evolution of hybrid incompatibilities. *Genetics*.

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics*.

Orr, H. A., and Turelli, M. (2001). The evolution of postzygotic isolation: Accumulating Dobzhansky-Muller incompatibilities. *Evolution* 55, 1085–1094. doi:10.1111/j.0014-3820.2001.tb00628.x.

Parisi, M., Nuttall, R., Naiman, D., Bouffard, G., Malley, J., Andrews, J., et al. (2003). Paucity of genes on the Drosophila X chromosome showing male-biased expression. *Science (80-.).* doi:10.1126/science.1079190.

Peng, J., Zipperlen, P., and Kubli, E. (2005). Drosophila sex-peptide stimulates female innate immune system after mating via the toll and Imd pathways. *Curr. Biol.* doi:10.1016/j.cub.2005.08.048.

Prakash, S. (1972). Origin of reproductive isolation in the absence of apparent genic differentiation in a geographic isolate of Drosophila pseudoobscura. *Genetics* 72, 143–155.

Ram, K. R., and Wolfner, M. F. (2007). Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. *Integr. Comp. Biol.* doi:10.1093/icb/icm046.

Ranz, J. M., Namgyal, K., Gibson, G., and Hartl, D. L. (2004). Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* 14, 373–379. doi:10.1101/gr.2019804.

Reed, L. K., LaFlamme, B. A., and Markow, T. A. (2008). Genetic architecture of hybrid male sterility in Drosophila: Analysis of intraspecies variation for interspecies isolation. *PLoS One*. doi:10.1371/journal.pone.0003076.

Reed, L. K., Nyboer, M., and Markow, T. A. (2007). Evolutionary relationships of Drosophila mojavensis geographic host races and their sister species Drosophila arizonae. *Mol. Ecol.* doi:10.1111/j.1365-294X.2006.02941.x.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11. doi:10.1186/gb-2010-11-3-r25.

Rothenbusch-Fender, S., Fritzen, K., Bischoff, M. C., Buttgereit, D., Oenel, S. F., and Renkawitz-Pohl, R. (2017). Myotube migration to cover and shape the testis of Drosophila depends on Heartless, Cadherin/Catenin, and myosin II. *Biol. Open* 6, 1876–1888. doi:10.1242/bio.025940.

Samakovlis, C., Kylsten, P., Kimbrell, D. A., Engstrom, A., and Hultmark, D. (1991). The Andropin gene and its product, a male-specific antibacterial peptide in Drosophila melanogaster. *EMBO J.* doi:10.1002/j.1460-2075.1991.tb07932.x.

Schaeffer, S. W., and Miller, E. L. (1991). Nucleotide sequence analysis of Adh genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U. S. A.* 88, 6097–6101. doi:10.1073/pnas.88.14.6097.

Schoenfelder, S., and Fraser, P. (2019). Long-range enhancer–promoter contacts in gene expression control. *Nat. Rev. Genet.* 20, 437–455. doi:10.1038/s41576-019-0128-0.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna* 22, 839–851. doi:10.1261/rna.053959.115.

Sepil, I., Hopkins, B. R., Dean, R., Thézénas, M. L., Charles, P. D., Konietzny, R., et al. (2019). Quantitative Proteomics Identification of Seminal Fluid Proteins in Male Drosophila melanogaster. *Mol. Cell. Proteomics* 18, S46–S58. doi:10.1074/mcp.RA118.000831.

Snook, R. R. (1998). Sperm Production and Sterility in Hybrids Between Two Subspecies of Drosophila pseudoobscura. *Evolution*. doi:10.2307/2410943.

Stern, C. (1941). The growth of testes in Drosophila. I. The relation between vas deferens and testis within various species. *J. Exp. Zool.* doi:10.1002/jez.1400870109.

Sturgill, D., Zhang, Y., Parisi, M., and Oliver, B. (2007). Demasculinization of X chromosomes in the Drosophila genus. *Nature*. doi:10.1038/nature06330.

Sundararajan, V., and Civetta, A. (2011). Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in drosophila sterile hybrids. *J. Mol. Evol.* doi:10.1007/s00239-010-9404-5.

Swanson, W. J., and Vacquier, V. D. (2002). The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* doi:10.1038/nrg733.

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

Wang, R. L., Wakeley, J., and Hey, J. (1997). Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. *Genetics* 147, 1091–1106.

Wen, K., Yang, L., Xiong, T., Di, C., Ma, D., Wu, M., et al. (2016). Critical roles of long noncoding RNAs in Drosophila Spermatogenesis. *Genome Res.* 26, 1233–1244. doi:10.1101/gr.199547.115.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2004). Evolutionary changes in cis and trans gene regulation. *Nature* 430, 85–88. doi:10.1038/nature02698.

Wittkopp, P. J., Haerum, B. K., and Clark, A. G. (2008). Regulatory changes underlying expression differences within and between *Drosophila* species. *Nat. Genet.* 40, 346–350. doi:10.1038/ng.77.

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., et al. (2005). Genome-wide midrange transcription profiles reveal expression level relationships in human tissue-specification. *Bioinformatics*. doi:10.1093/bioinformatics/bti042.

Zhang, L., and Li, W. H. (2004). Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. *Mol. Biol. Evol.* doi:10.1093/molbev/msh010.

Chapter 3: Selection and transgressive gene expression in hybrids between two closely related subspecies of *Drosophila*

Alwyn C. Go and Alberto Civetta

Department of Biology, The University of Winnipeg

This chapter was submitted on August 2020 to Frontiers in Genetics and is in press.

ABSTRACT

Genome-wide assays of expression between species and their hybrids have identified genes that become either over or under-expressed relative to the parental species (*i.e.* transgressive). Transgressive expression in hybrids is of interest because it highlights possible changes in gene regulation directly linked to hybrid dysfunction. Previous studies in *Drosophila* that used long-diverged species pairs with complete or nearly complete isolation (*i.e.* full sterility and partial inviability of hybrids) and high-levels of genome misregulation have found correlations between expression and coding sequence divergence. The work highlighted the possible effects of directional selection driving sequence divergence and transgressive expression. Whether the same is true for taxa at early stages of divergence that have only achieved partial isolation remains untested. Here, we reanalyze previously published genome expression data and available genome sequence reads from a pair of partially isolated subspecies of *Drosophila* to compare expression and sequence divergence. We find a significant correlation in rates of expression and sequence evolution, but no support for directional selection driving transgressive expression in hybrids. We find that most transgressive genes in hybrids show no differential expression between parental subspecies and used SNP data to explore the role of stabilizing selection through compensatory mutations. We also examine possible misregulation through cascade effects that could be driven by interacting gene networks or co-option of off-target *cis*-regulatory elements.

INTRODUCTION

Studies that have addressed the genetic basis of incompatibilities in hybrids between species, or diverging populations, have traditionally resorted to mapping loci and interactions between them (Coyne and Orr 1989; Masly and Presgraves 2007; Presgraves 2008; Cattani and Presgraves 2012; Dufresnes et al. 2016). This approach has been fruitful in that ultimately a few major protein coding genes have been identified (Ting et al. 1998; Masly et al. 2006; Phadnis and Orr 2009; Mihola et al. 2009), but in all cases the effect of these major genes requires interactions with other genetic factors. Major genes often show patterns of rapid evolution between divergent populations or species (Ting et al. 1998; Presgraves et al. 2003; Maheshwari and Barbash 2011) suggesting that, at least in part, changes in protein composition might exert effects on phenotype and function through alterations in patterns of expression of genes targeted by such proteins. Moreover, genome-wide surveys have provided evidence to support that many genes and complex systems of epistasis are linked to hybrid incompatibility phenotypes. (Morán and Fontdevila 2014; Turner and Harr 2014; Turner et al. 2014; Fontdevila 2016). While coevolution among interacting genes keeps function within populations and species, hybridization between divergent isolated populations and incipient species brings together incompatible interloci allele interactions resulting in a reduction in hybrid fitness (Dobzhansky 1937; Muller 1942; Orr 1996). The reduced fitness of hybrids serves as a postzygotic barrier among divergent taxa.

The role of divergence in the regulation of gene expression has been long acknowledged (King and Wilson 1975) but not until recently has genome-wide divergence in gene expression during speciation has been addressed. Recent reviews have

summarized how changes in gene expression could impact hybrid phenotypes (Civetta 2016; Mack and Nachman 2017). Using genome-wide approaches, questions have been addressed as to the proportion of genome-wide misregulation in hybrids, the relative contribution of *cis*- *vs*. *trans*-regulatory elements in gene misregulation, and the identity of misregulated genes that might contribute to hybrid fitness breakdown (Ranz et al. 2004; Haerty and Singh 2006; Renaut et al. 2009; Tirosh et al. 2009; McManus et al. 2010; Llopart 2012; Coolon et al. 2014; Gomes and Civetta 2015; Brill et al. 2016; Mack et al. 2016). Often, genome-wide assays of expression in hybrids reveal gene regulatory dysfunctions as patterns of transgressive gene expression (*i.e.*, expression beyond levels found in parental species). This can be a consequence of directional selection or drift causing changes at *cis*- and *trans*-regulatory elements that drive divergence in expression between taxa and transgressive expression in hybrids. Previous studies have found positive correlations between protein coding evolution and gene expression divergence between species of *Drosophila* (Castillo-Davis et al. 2004; Nuzhdin et al. 2004; Lemos et al. 2005; Artieri et al. 2007). Moreover, the finding of a similar significant positive correlation between nonsynonymous (d_N) and nonsynonymous/synonymous (d_N/d_S) divergence and gene expression differences between hybrids and parental species has been used to suggest sequence divergence driving regulatory incompatibilities and to highlight the potential effects of directional selection in gene expression during speciation (Artieri et al. 2007; Hunt et al. 2013). However, the species pairs used were typically long-diverged with hybrids exhibiting complete or nearly complete isolation and high-levels of genome misregulation (Ranz et al. 2004; Haerty and Singh 2006; Artieri et al. 2007; McManus et al. 2010; Coolon et al. 2014). The use of divergent populations

within species of copepods have found no significant relationship between hybrid transgressive expression and estimates of sequence divergence and the authors offered an alternative physiological explanation for the detected pattern (Barreto et al. 2015, 2018).

There are in fact alternative explanations that could explain the lack of relationship between sequence and expression divergence. Mutations within taxa can work to compensate the effect of deleterious mutations on expression (*i.e.* stabilizing selection). The possibility that *cis–trans* mutations may cause compensation within species but lead to transgressive expression in hybrids is supported by studies that report abundant *cis–trans* epistasis (Mackay 2014; Mackay and Moore 2014; He et al. 2016; Vonesch et al. 2016). However, the strength of selection for a secondary compensatory mutation might be small (Bourguet 1999). It is also possible for transgressive expression in hybrids to arise as a response to hybrid dysfunction within gene interacting networks or metabolic pathways. While this could work to ameliorate fitness problems in hybrids, it could also exacerbate hybrid dysfunction. This might be particularly the case for fitness breakdown between diverging populations (Barreto et al. 2015, 2018). Finally, we speculate that newly arising mutations in *trans* regulatory elements that result from divergence between taxa or compensatory mutations within, could co-opt pre-existing *cis*-regulatory elements among multiple genes thereby causing widespread misregulation.

Here, we used a pair of geographically separated subspecies of *D. pseudoobscura*, *D. p. pseudoobscura* and *D. p. bogotana*, that have diverged for at least 0.15 Myr (Schaeffer and Miller 1991; Wang et al. 1997) and whose hybrids exhibit unidirectional male sterility where only male hybrids produced by *D. p. bogotana* females are sterile. We reanalyze previously published transcriptomics data (Gomes and Civetta 2015) using

a newer *D. p. pseudoobscura* genome release (r3.04) and updated mapping and expression analysis tools to explore relationships between genome expression and gene coding sequence divergence. Our report identifies no relationship between sequence divergence and transgressive expression in hybrids suggesting a need for broader examinations of transgressive expression between recently diverged populations and species across taxa. We find that most transgressive genes in hybrids are not differentially expressed between subspecies. We explore explanations for transgressive expression other than incompatibilities in regulation arising from rapid divergence between subspecies, such as compensatory mutations, gene-interaction networks, and the co-option of multiple *cis*-regulatory elements by *trans*-regulatory elements. While we find some support for these alternative hypotheses, we acknowledge that they do not fully explain transgressive expression in hybrids, we discuss some caveats and offer other possible explanations in the hope that they will trigger further inquiry. Ultimately, full comprehension of transgressive expression in hybrids will require combining information on genome expression and sequencing with the identification of interactomes and a proper characterization of mechanism of *trans* effects on characterized *cis*-regulatory targets.

MATERIALS AND METHODS

RNA-sequence data

Raw RNA sequence data used in this analysis were from a genome-wide transcriptomics study of the *Drosophila pseudoobscura* subspecies pair and their reciprocal hybrids by Gomes and Civetta (2015). Briefly, RNA was extracted from the whole male reproductive tract. Biological replicates were obtained for the parental subspecies and their reciprocal F_1 hybrids with each replicate containing 30-40 male reproductive tracts. cDNA libraries were prepared using the Illumina TruSeq Stranded mRNA sample preparation kit and multiplexed on a single lane of an Illumina HiSeq2000 platform with 100 bp paired-end sequencing. A quality check was performed on the raw reads using FastQC (Andrews 2010). Read processing and adapter trimming were performed with Trimmomatic (Bolger et al. 2014) and reads with a Phred score below 30 and a final length of 50 bp were excluded.

Mapping and differential expression analysis

We mapped processed reads to the latest release (r3.04) of the *D. p. pseudoobscura* reference genome (http://flybase.org/) using STAR, chosen for its reliability (Dobin et al. 2013; Baruzzo et al. 2017) over the previously used TopHat approach (Gomes and Civetta 2015). Read counting was performed at the gene level using featureCounts (Liao et al. 2014) with the reversely stranded (-s 2) and fragment counting (-p) parameters and the latest version of the *D. p. pseudoobscura* annotation serving as a guide.

Pairwise differential expression across all groups was performed using both DESeq2 (Love et al. 2014) and edgeR (Robinson et al. 2009). In the analysis using edgeR, genes with less than 1 count per million (CPM) in at least one group were excluded from further analysis and the per gene counts for each sample were normalised using the TMM method (Robinson and Oshlack 2010). The default settings were used to obtain normalised counts from the DESeq2 analysis. The consensus list of differentially expressed genes from both tools were used for all downstream analyses. Differentially expressed genes among the hybrids were identified as transgressive if their expression were significantly above or below the range found in the parental subspecies. Further, $\log_2$ fold-changes (lfc) thresholds of 0.5 and 1 were applied to increase our statistical yield of true positives (Schurch et al. 2016). All tools for the analysis were ran on Galaxy (http://usegalaxy.org).

Coding sequence and expression divergence

Rates of coding sequence divergence between *D. p. bogotana* and *D. p. pseudoobscura* were estimated for differentially expressed genes between the parental subspecies and for transgressive genes in fertile and sterile F_1 hybrids. Since the RNA-seq data provided only partial sequences from each gene analyzed, we retrieved raw DNA sequence reads from the sequence read archives (SRA) under the accession number SRX091468 (*D. p. bogotana*). The *D. p. bogotana* raw sequence reads were aligned to all gene regions from the r.3.04 *D. p. pseudoobscura* reference genome (http://flybase.org/) using BWA (Li 2010) ran on Galaxy (http://usegalaxy.org/) under default settings except for the maximum number of gap extensions which was set to 4. The ‘extract consensus

from assembly' workflow in UGene (Okonechnikov et al. 2012) was then used to extract the *D. p. bogotana* gene regions and these were aligned to the longest available transcript for *D. p. pseudoobscura* from FlyBase (http://flybase.org/) using MAFFT (Katoh 2013). The alignments were modified using Gblocks v.0.91b (Castresana 2000) with default settings except for the block parameters which allowed gap positions with half within the final blocks – this removes unaligned introns from the *D. p. bogotana* gene region while preserving possible indels. Alignments from Gblocks were inspected to ensure that the coding sequences were intact open reading frames and were a multiple of three.

Rates of synonymous (d_S) and nonsynonymous (d_N) nucleotide substitutions were estimated using the SeqinR package (Charif and Lobry 2007) loaded on RStudio version 1.1.463. Non-parametric Spearman rank sum correlation coefficients were calculated to test the relationship between coding sequence divergence (d_N, d_S, and d_N/d_S) and expression difference. For the parental subspecies, expression differences were calculated as the absolute difference of $[\log_2(\bar{x}_{D.\ p.\ pseudoobscura}) - \log_2(\bar{x}_{D.\ p.\ bogotana})]$. For the transgressive genes, expression differences were calculated for each hybrid relative to each parental subspecies as the absolute difference of $[\log_2(\bar{x}_{\text{Fert or Ster}}) - \log_2(\bar{x}_{D.\ p.\ pseudoobscura\ or\ D.\ p.\ bogotana})]$. The lower absolute difference value was kept as a measure of minimum transgressive expression (Barreto et al. 2015).

Allele specific expression

To determine the role of *cis* and/or *trans* changes to transgressive gene expression in the hybrids, we identified fixed species-specific single nucleotide polymorphisms (SNPs) and their relative allele expression in the hybrids. SNPs between

the parental subspecies were identified from their mapped reads using Naïve variant caller followed by processing with the Variant annotator (Blankenberg et al. 2014). SNPs were considered fixed in each parental subspecies if each parent had a single different allele and at least 3 supporting reads. Allele specific expression in the hybrids was measured by first assigning their RNA-seq reads to a parent of origin based on the identity of the allele at fixed SNP positions in each parent. Reads with fixed SNPs mapping to a single gene were summed and any gene with less than 20 mapped reads from both parental subspecies combined were discarded from further analysis (McManus et al. 2010; Gomes and Civetta 2015). SNP counts for each gene were then adjusted to account for differences in sequencing depth between samples. Samples with zero SNP counts were given a value of 1 to allow for statistical testing. To detect significant differences between the ratio of parental SNP counts to counts of each parental allele in the sterile and fertile hybrids respectively, the Fisher's exact test was used (McManus et al. 2010; Gomes and Civetta 2015). Transgressive genes that showed differential expression between the parental subspecies were classified as driven by *cis-trans* divergence if the Fisher's exact test was significant and *cis* regulatory divergence when the Fisher's exact test was not significant (McManus et al. 2010). For transgressive genes that were not differentially expressed between the parental subspecies, a significant result for the Fisher's exact test indicated evidence for compensatory *cis* and *trans* mutations (McManus et al. 2010) while a non-significant result suggested a conservation in regulatory interactions and classified as non-compensatory.

Interactions and sequence similarity

Interactions among proteins were predicted using STRING (v11.0; Szklarczyk et al. 2019). Gene-Ontology and UniProt keyword enrichments were assessed from outputs using STRING and DAVID (v6.8; Huang et al. 2009a, b). We used the extended gene regions (which includes 2kb 5' and 3') for genes that showed transgressive expression driven by *trans* regulatory elements (*i.e. cis-trans* divergent or compensatory) to perform a BLASTn against a database containing all transgressive genes and against another database with all *D. p. pseudoobscura* extended gene regions within the genome to identify similarities between upstream regions for plus/plus matches or between the upstream and downstream regions for plus/minus matches. We retained only hits that were lower than 1×10^{-14} and unique among transgressive sequences and not shared with other genes in the genome. Retained hits had E-values lower than 8×10^{-15}, with nucleotide alignments of at least 173 base pairs and identities higher than 64%.

RESULTS

The re-analysis of our previously published data (Gomes and Civetta 2015) by mapping reads onto a newer released genome assembly and using more recently developed analytical pipelines found similar results in terms of lack of bias in mapping, low proportion of differentially expressed genes between subspecies, and significant excess of transgressive expression in sterile relative to fertile hybrids (Supplementary material).

Transgressive gene expression in hybrids does not correlate with accelerated rates of evolution as expected under a scenario of divergent selection between subspecies.

Under the assumption that regulatory evolution and structural protein evolution are under similar selective pressures, a correlation is expected between expression difference and nucleotide sequence evolution. Of the 819 differentially expressed genes between the parental subspecies, 604 (73.7%) were protein coding genes with the remaining 215 (26.3%) being non-coding RNAs or coding genes without full coding sequences available for both subspecies. The percentage of differentially expressed protein coding genes between subspecies increases significantly when a less stringent lfc threshold of 0.5 was applied (82.7%; Z= 5.51, $P< 0.001$) (Figure S1). We found a significant correlation for expression differences between subspecies and nonsynonymous (d_N) sequence divergence (N= 604; Spearman's $\rho= 0.091$, $P= 0.026$) but not between differences in expression and synonymous substitutions (d_S) (Spearman's $\rho=$ 0.-0.046, $P= 0.261$). The d_N/d_S ratio was also positively correlated with expression differences ($\rho= 0.108$, $P= 0.011$). Using the less stringent lfc threshold of 0.5, d_N, d_S, and

d_N/d_S were all significantly correlated with gene expression divergence between subspecies (N= 1,801; $\rho = 0.121$, $P= 2.39x10^{-7}$; $\rho= 0.065$, $P= 0.005$; and $\rho=0.096$, $P= 8.4x10^{-5}$, respectively) (Figure 3.1A). These results are overall in agreement with previous findings in *Drosophila* and other organisms confirming that protein sequence and expression divergence are influenced by similar selective processes (Nuzhdin et al. 2004; Castillo-Davis et al. 2004; Khaitovich et al 2005; Artieri et al. 2007; Ortiz-Barrientos et al. 2007).

Given that protein coding sequence differentiation serves as a good predictor of expression divergence, some studies have explored correlations between rates of protein divergence with expression of misregulated genes in hybrids. Misregulated genes with transgressive expression in hybrids are of interest in speciation as they associate with hybrid disrupted phenotypes (Moehring et al. 2007; Catron and Noor 2008; Sundararajan and Civetta 2011; Gomes and Civetta 2015; Brill et al. 2016; Civetta 2016). Significant positive correlations are suggestive of either directional selection or relaxation of selective constraints fueling regulatory incompatibilities (Artieri et al. 2007; Hunt et al. 2013; Barreto et al. 2015). Of the 44 transgressive genes in the hybrids, 35 had available sequence data for the estimation of coding sequence divergence. The analysis showed no significant correlations between sequence divergence and expression difference (N=35; d_N, $\rho= 0.078$, $P= 0.655$; d_S, $\rho= 0.242$, $P= 0.161$; d_N/d_S, $\rho= -0.112$ $P= 0.547$). This result holds when a less stringent lfc threshold of 0.5 was used, with 223 of the 262 transgressive genes having sequence data available for analysis (N= 223 d_N, $\rho = -0.078$, $P= 0.245$; d_S, $\rho= 0.081$, $P= 0.230$; d_N/d_S, $\rho=-0.092$, $P= 0.208$) (Figure 3.1B).

Alternative explanations for transgressive expression in hybrids: Compensatory mutations, interaction networks, and transcriptional drive by sequence similarity among targets

One possibility for a lack of correlation between transgressive expression in hybrids and sequence divergence is that transgressive expression might be a consequence of occasional

deleterious mutations that are followed by compensatory DNA changes to overcome detrimental effects on gene expression (*i.e*. a side effect of stabilizing selection between divergent taxa) (Figure 3.2A – Gene 1). Our data shows that 32 out of 44 (72.72%) transgressive genes in the hybrids were not differentially expressed between parental subspecies. The low number of transgressive genes is likely a consequence of our stringent use of a two-fold-change (lfc= 1) in expression threshold to maximize our statistical yield of true positives. Given the low sample size, we decided to continue using a less stringent lfc threshold of 0.5 and found, as with the more stringent threshold, a large proportion of transgressive genes without differential expression between parental subspecies (79%, 207/262). If genes without differential expression between subspecies are under stabilizing selection favouring compensatory mutations to buffer deleterious mutations and restore expression to similar levels among parental subspecies, we expect their rate of sequence divergence to be lower than those of genes experiencing divergence in regulation, and thus expression, between subspecies. Our data shows no significantly lower rates of change (d_N and d_N/d_S) for genes with transgressive expression in hybrids and no differential expression between parentals (Mann-Whitney FDR corrected *P*-values) (Table 3.1).

We used informative SNPs to identify genes with transgressive expression in hybrids driven by compensatory mutations or *cis-trans* divergence (Figure 3.2A – Gene 1 and Figure 3.2B – Gene 3). Twenty five percent of the transgressive genes (65/262) had non-informative SNPs to allow us to classify parent of origin for the alleles found in the hybrids. Of the remaining 197 transgressive genes, we found that for 65% of them, transgressive expression could be explained by compensatory mutations (97 genes) or *cis-trans* divergence (31 genes) (Figure 3.2A&B – Genes 1 and 3). The remaining being cases in which the transgressive gene shows similar ratios of subspecies allele expression in parents and hybrids. Of these, 62 were classified as non-compensatory and 7 as having experienced *cis* divergence (Figure 3.2A&B – Genes 2 and 4).

We explored whether transgressive expression in hybrids for genes that do not show evidence of compensatory or *cis-trans* mutations could be a cascade triggered by interactions in a shared gene network and/or pathway (Bader et al. 2015; Barreto et al. 2015). This will predict clusters of interacting and functionally related proteins to be misregulated in the hybrids. We detected a protein-protein interaction (PPI) network of 90 genes (34% of the 262 transgressive genes) (Figure 3.3) with a significant (*i.e.* more interactions than randomly expected) PPI enrichment ($P = 4.29\times10^{-2}$). We found no evidence of known functional enrichment in the network, but a significant overrepresentation of "Signal" genes based on UniProt keywords (FDR corrected $P = 1.25\times10^{-7}$). The PPI analysis was still significant for the subset of transgressive genes in the sterile hybrids (PPI enrichment $P = 1.06\times10^{-2}$, 79 nodes) but not for fertile hybrids (PPI enrichment $P = 0.106$, 4 nodes). Twenty-two genes in the network were *cis* or non-compensatory, thus their misregulation could be driven by interactions with other

misregulated genes in the network (Figure 3.3). We found no significant PPI for transgressive genes differentially expressed between subspecies (P= 0.597). Finally, we also explored whether transgressive expression in hybrids could be a consequence of transcriptional drive caused by *trans* mutations affecting multiple genes with *cis* sequence similarity (Figure 3.2A – Red arrows). We found 46 genes (18% - 46/262) with possible evidence of co-option by newly evolved *trans* mutations. Of these genes, 15 were classified as compensatory, 10 had *cis*-*trans* divergence, 9 were non-compensatory, and 12 had non-informative SNPs for classification (Table S8).

DISCUSSION

Genome-wide, our results are in agreement with previous reports of correlated evolution between sequence and expression divergence (Castillo-Davis et al. 2004; Nuzdhin et al. 2004; Khaitovich et al. 2005; Lemos et al. 2005; Artieri et al. 2007; Hunt et al. 2013; Whittle et al. 2014; Barreto et al. 2015), but provide no support for positive selection or relaxation of selective constraints as drivers of change causing misregulation and transgressive expression in hybrids. Genes with no differential expression between subspecies and transgressive expression in hybrids did not show overall evidence of lower sequence divergence than transgressive genes with differential expression between subspecies. This result is unexpected under a scenario of compensation favouring mutations that restore divergence in gene expression between parental subspecies (*i.e*. stabilizing selection). We used SNPs to tease apart regulatory divergence among transgressive genes in hybrids. Transgressive expression results from divergence in *cis* and *trans* regulatory elements, leading to differential expression between parental species as well as hybrids. Alternatively, such changes can be buffered by compensatory mutations within lineages to restore levels of expression to similar levels between species but cause misexpression in hybrids (Landry et al. 2005, McManus et al. 2010, Mack and Nachman 2017). Studies of divergence in gene expression between species provides support for changes in transcript levels being often deleterious, with large mutational effects, and equilibrium levels of genetic variation maintained by stabilizing selection (Rifkin et al. 2003; Lemos et al. 2005; Hodgins-Davis et al. 2015). Our study shows that the majority (79%) of transgressive genes in hybrids between *D. p. pseudoobscura* and *D. p. bogotana* were not differentially expressed between the subspecies, and the SNP

analysis supports a good proportion of transgressive expression caused by compensatory changes (49%) during early stages of species divergence, with another (16%) caused by *cis-trans* divergence.

A caveat to our results is that informative SNPs are limited between closely related subspecies. Thus 25% of transgressive genes could not be analyzed this way. Moreover, for any gene, not all reads have informative SNPs imposing some analytical limitations. While this might lead to an underestimation, our result of 49% compensatory evolution for a pair of very closely related subspecies of *Drosophila* is expected when compared to estimates of 73% compensatory evolution for hybrids between more distantly related species of *D. simulans* and *D. sechellia* (Coolon et al. 2014) and 67% for yeast (Wang et al. 2015). The proportion of compensatory mutations within lineage (49%) is larger than *cis-trans* divergence between lineages (16%) and suggests that hybrids between closely related taxa might be more vulnerable to a breakdown of coadaptations within species than misregulation caused by divergent evolution.

We explored possible alternative explanations for a large proportion of transgressive genes which could not be explained by *cis-trans* compensation or divergent *cis-trans* evolution. We found that genes with transgressive expression in hybrids that experienced divergence in regulation between subspecies produced proteins that did not show enrichment for interactions. On the other hand, transgressive genes with no evidence of divergence between subspecies were enriched for protein interactions. This result suggests that in some cases misregulation and transgressive expression could be a cascade effect driven by networks of interacting proteins and that such domino effect could work to exacerbate initial incompatibilities in hybrids between early stage

diverging lineages. The role of gene-network effects is expected under the Bateson-Dobzhansky-Muller model of speciation (Turner et al. 2014) and while there has been some support for gene-networks buffering allelic variation among yeast strains (Bader et al. 2015) its importance in speciation is largely unexplored. Finally, we entertained the idea that newly arising *trans* mutations in either divergent or compensatory cases could possibly generate a cascade effect of misregulation of targets that might have not experienced *cis*-regulatory mutations between divergent taxa (Figure 3.2A – Red arrows). We explored the idea of "transcriptional drive by sequence similarity among targets" by seeking sequence similarity within proximal (2,000bp) putative *cis*-regulatory elements between transgressive genes showing evidence of *cis*-*trans* divergence or compensation and those showing no evidence of such sequence divergence. Our analysis showed some support for this idea with 18% of genes being possibly co-opted. However, only 9 genes classified as non-compensatory appear as possible targets. One important limitation is that we only addressed sequence similarities between nearby upstream sequence regions of compensatory or *cis*-*trans* transgressive genes and upstream sequence regions of other transgressive genes, leaving unexplored the possibility that misregulation could be exerted by more distant *cis*-regulatory elements.

Table 3.1: Average evolutionary rates (± SD) for differentially expressed genes between parental subspecies that do not show transgressive expression in hybrids ($(P1 \neq P2)_{NT}$), transgressive genes that show differential expression between subspecies ($(P1 \neq P2)_{T}$), and transgressive genes that do not show differential expression between subspecies ($(P_1 = P_2)_{T}$). FDR corrected Mann-Whitney tests show no significant differences between rates of non-synonymous substitutions (d_N), synonymous substitutions (d_S), and the d_N/d_S ratio across all three comparisons.

	Non-transgressive	Transgressive	
	$(P_1 \neq P_2)_{NT}$	$(P_1 \neq P_2)_{T}$	$(P_1 = P_2)_{T}$
N	1763	49	174
d_N	5.022×10^{-3} ($\pm 1.74 \times 10^{-2}$)	4.461×10^{-3} ($\pm 6.02 \times 10^{-3}$)	4.060×10^{-3} ($\pm 5.90 \times 10^{-3}$)
d_S	2.290×10^{-2} ($\pm 2.82 \times 10^{-2}$)	2.086×10^{-2} ($\pm 1.83 \times 10^{-2}$)	1.890×10^{-2} ($\pm 1.60 \times 10^{-2}$)
d_N/d_S	2.513×10^{-1} ($\pm 3.81 \times 10^{-1}$)	2.171×10^{-1} ($\pm 2.34 \times 10^{-1}$)	2.389×10^{-1} ($\pm 3.19 \times 10^{-1}$)

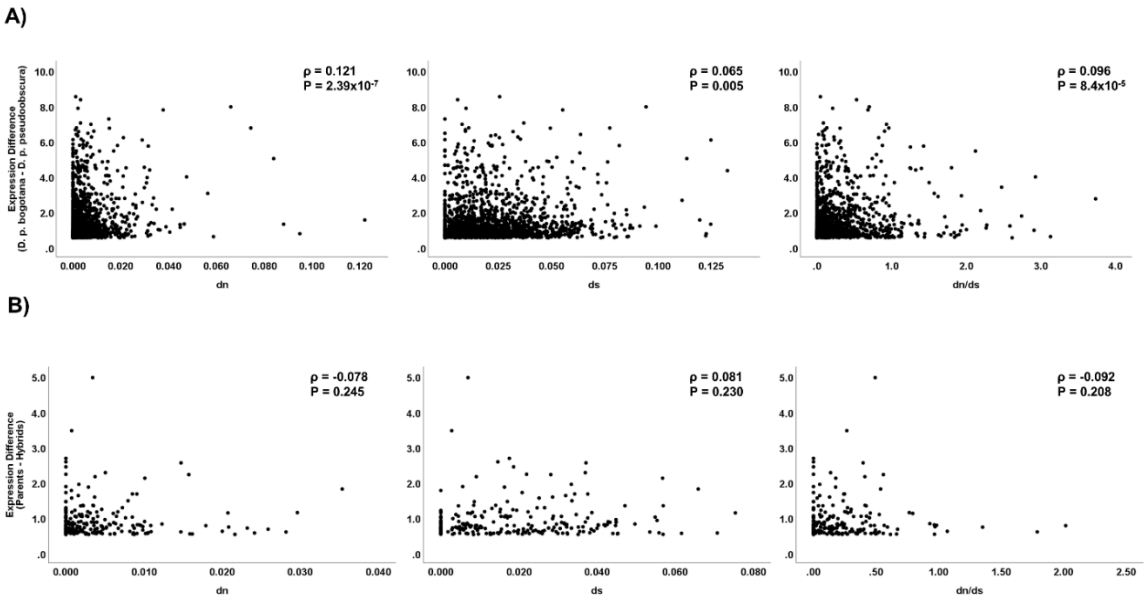


Figure 3.1: Correlation analysis between expression and coding sequence divergence. Spearman's rank-sum coefficient and *P*-values are displayed in each frame. (A) Analysis on differentially expressed genes between the parental subspecies. (B) Analysis on genes showing transgressive expression in hybrids.

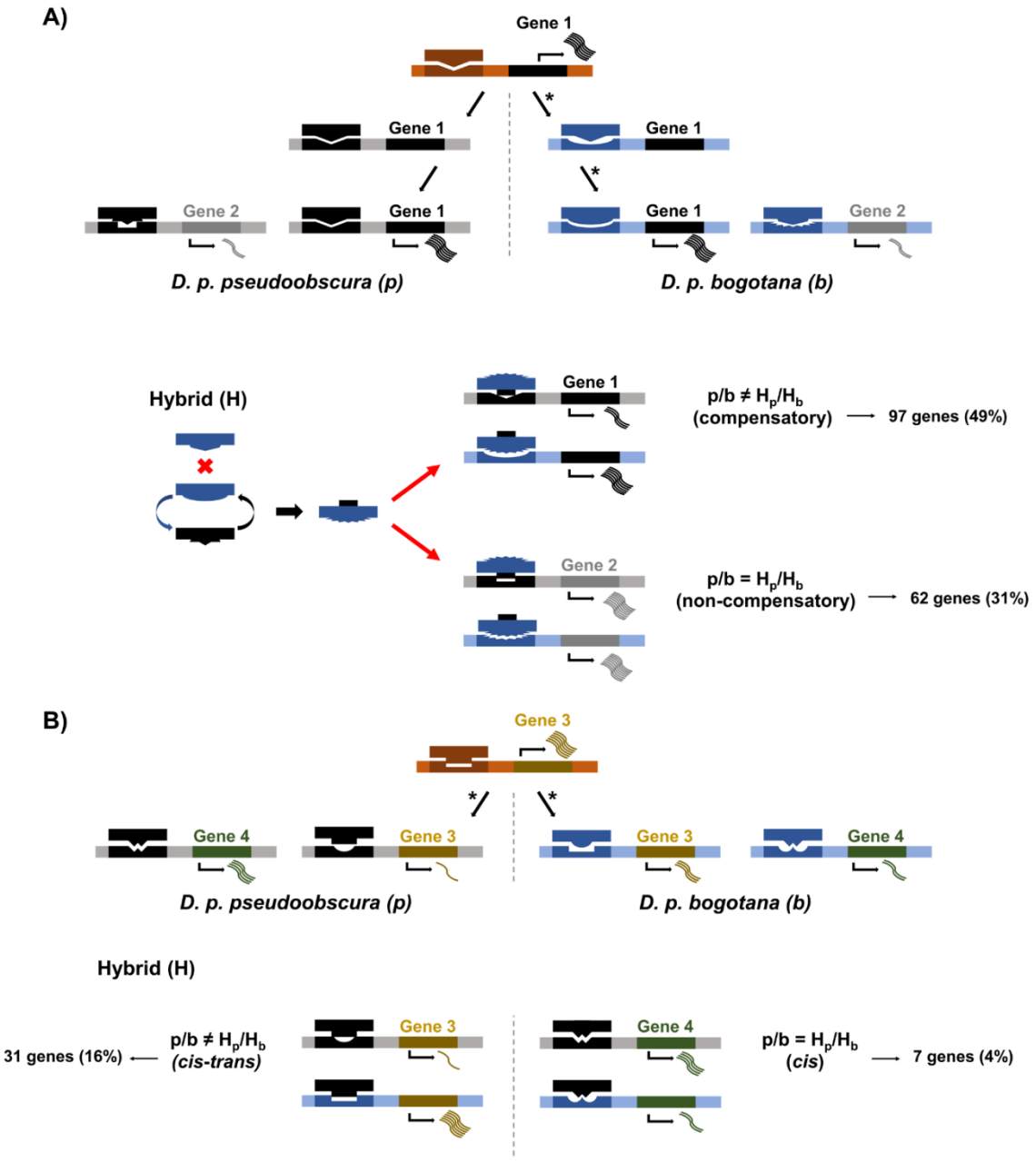

A)
Gene 1
Gene 1
Gene 1
Gene 2
Gene 1
Gene 1
Gene 1
Gene 2
D. p. pseudoobscura (p)
D. p. bogotana (b)
Hybrid (H)
Gene 1
p/b ≠ Hp/Hb
(compensatory)
97 genes (49%)
Gene 2
p/b = Hp/Hb
(non-compensatory)
62 genes (31%)
B)
Gene 3
Gene 4
Gene 3
Gene 3
Gene 4
D. p. pseudoobscura (p)
D. p. bogotana (b)
Hybrid (H)
31 genes (16%)
p/b ≠ Hp/Hb
(cis-trans)
Gene 3
Gene 4
p/b = Hp/Hb
(cis)
7 genes (4%)

Figure 3.2: Scenarios of regulatory divergence for *cis*- and *trans*-regulatory divergence. (A) Gene 1 shows compensatory *cis* and *trans* mutations wherein *D. p. bogotana* experiences an initial mutation in *cis* followed by a mutation in *trans* restoring gene expression to similar levels between parental subspecies. Gene 2 shows similar levels of expression in parental subspecies. In the hybrid background, the *D. p. bogotana trans* factor for gene 1 interacts with the *D. p. pseudoobscura trans* factor for gene 2 leading to a conformation change. This new *trans* factor complex can now bind optimally to the *cis* region of genes 1 and 2 (red lines) resulting in transgressive expression (*i.e.* expression above parental levels). The allelic ratio of gene 2 in the hybrid is equal and the gene is classified as non-compensatory through SNP analysis. (B) Gene 3 shows divergence in *cis* in one subspecies and *trans* in the other subspecies. This leads to sub-optimal binding in both subspecies and differential expression. The regulatory incompatibilities persist within the hybrid background leading to unequal allelic ratios. Gene 3 is classified as *cis*-*trans* divergent by SNP analysis. Gene 4 shows a situation of *cis*-only divergence between the parental subspecies. Regulatory incompatibilities would occur in *D. p. bogotana* but not *D. p. pseudoobscura* resulting in differential expression between the subspecies. Similar interactions for this gene would occur in the hybrid resulting in equal allelic ratio. Gene 4 is classified as *cis*-only by SNP analysis.

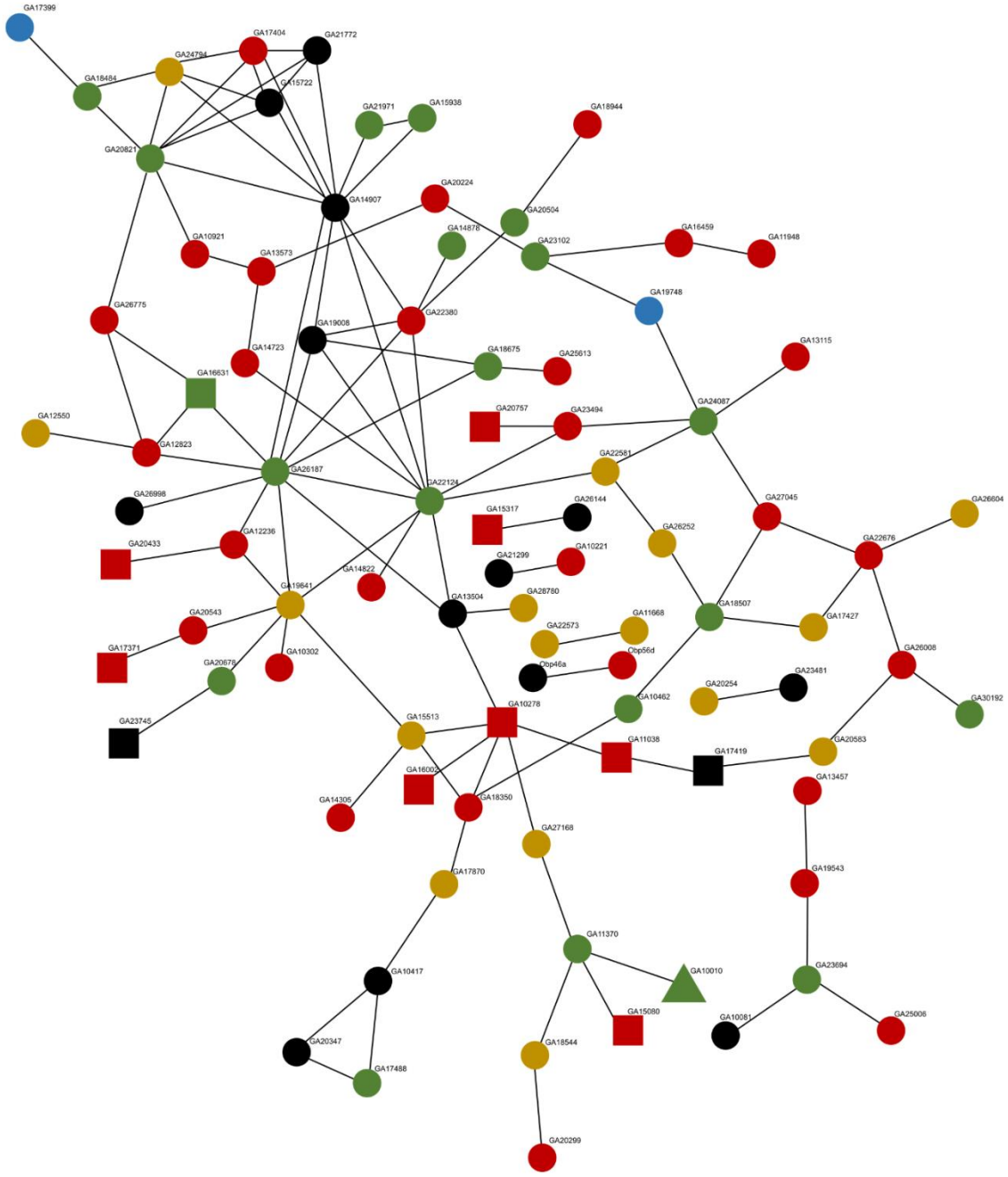


Figure 3.3: STRING protein-protein interaction network for all transgressive genes in hybrids. Circles represent transgressive genes that are unique to the sterile hybrids (78), squares are genes unique to the fertile hybrid (11), and the triangle represents a gene that shows transgressive expression in both fertile and sterile hybrids. Non-compensatory genes (20) are coloured green, red represents compensatory genes (37), yellow for genes with *cis-trans* divergence (16), blue for *cis*-only genes (2), and black represents genes with no informative SNPs (15).

REFERENCES

Andrews, S. (2010). FastQC. *Babraham Bioinforma.* doi:citeulike-article-id:11583827.

Artieri, C. G., Haerty, W., and Singh, R. S. (2007). Association between levels of coding sequence divergence and gene misregulation in *Drosophila* male hybrids. *J. Mol. Evol.* 65, 697–704. doi:10.1007/s00239-007-9048-2.

Bader, D. M., Wilkening, S., Lin, G., Tekkedil, M. M., Dietrich, K., Steinmetz, L. M., et al. (2015). Negative feedback buffers effects of regulatory variants. *Mol. Syst. Biol.* 11, 785. doi:10.15252/msb.20145844.

Barreto, F. S., Pereira, R. J., and Burton, R. S. (2015). Hybrid dysfunction and physiological compensation in gene expression. *Mol. Biol. Evol.* 32, 613–622. doi:10.1093/molbev/msu321.

Barreto, F. S., Watson, E. T., Lima, T. G., Willett, C. S., Edmands, S., Li, W., et al. (2018). Genomic signatures of mitonuclear coevolution across populations of *Tigriopus californicus*. *Nat. Ecol. Evol.* 2, 1250–1257. doi:10.1038/s41559-018-0588-1.

Baruzzo, G., Hayer, K. E., Kim, E. J., DI Camillo, B., Fitzgerald, G. A., and Grant, G. R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nat. Methods* 14, 135–139. doi:10.1038/nmeth.4106.

Blankenberg, D., Von Kuster, G., Bouvier, E., Baker, D., Afgan, E., Stoler, N., et al. (2014). Dissemination of scientific software with Galaxy ToolShed. *Genome Biol.* doi:10.1186/gb4161.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 30, 2114–2120. doi:10.1093/bioinformatics/btu170.

Bourguet, D. (1999). The evolution of dominance. *Heredity (Edinb).* doi:10.1038/sj.hdy.6885600.

Brill, E., Kang, L., Michalak, K., Michalak, P., and Price, D. K. (2016). Hybrid sterility and evolution in Hawaiian *Drosophila*: Differential gene and allele-specific expression analysis of backcross males. *Heredity (Edinb).* 117, 100–108. doi:10.1038/hdy.2016.31.

Castillo-Davis, C. I., Hartl, D. L., and Achaz, G. (2004). cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* 14, 1530–1536. doi:10.1101/gr.2662504.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552. doi:10.1093/oxfordjournals.molbev.a026334.

Catron, D. J., and Noor, M. A. F. (2008). Gene expression disruptions of organism versus organ in Drosophila species hybirds. *PLoS One*. doi:10.1371/journal.pone.0003009.

Charif, D., and Lobry, J. R. (2007). “SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis,” in, 207–232. doi:10.1007/978-3-540-35306-5_10.

Civetta, A. (2016). Misregulation of gene expression and sterility in interspecies hybrids: Causal links and alternative hypotheses. *J. Mol. Evol.* 82, 176–182. doi:10.1007/s00239-016-9734-z.

Coolon, J. D., McManus, C. J., Stevenson, K. R., Graveley, B. R., and Wittkopp, P. J. (2014). Tempo and mode of regulatory evolution in *Drosophila*. *Genome Res.* 24, 797–808. doi:10.1101/gr.163014.113.

Coyne, J. A., and Orr, H. A. (1989). Patterns of Speciation in *Drosophila*. *Evolution* 43, 362. doi:10.2307/2409213.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., et al. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. doi:10.1093/bioinformatics/bts635.

Dobzhansky T. (1937). Genetics and the origin of species. In Columbia biological series. New York: Columbia University Press.

Dufresnes, C., Majtyka, T., Baird, S. J. E., Gerchen, J. F., Borzee, A., Savary, R., et al. (2016). Empirical evidence for large X-effects in animals with undifferentiated sex chromosomes. *Sci. Rep.* 6. doi:10.1038/srep21029.

Fontdevila, A. (2016). Hybrid incompatibility in *D rosophila*:An updated genetic and evolutionary analysis. *eLS*, 1–16. doi:10.1002/9780470015902.a0020896.pub2.

Gomes, S., and Civetta, A. (2015). Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci. Rep.* 5, 11976. doi:10.1038/srep11976.

Haerty, W., and Singh, R. S. (2006). Gene regulation divergence is a major contributor to the evolution of Dobzhansky-Muller incompatibilities between species of *Drosophila*. *Mol. Biol. Evol.* 23, 1707–1714. doi:10.1093/molbev/msl033.

He, X., Zhou, S., St. Armour, G. E., Mackay, T. F. C., and Anholt, R. R. H. (2016). Epistatic partners of neurogenic genes modulate Drosophila olfactory behavior. *Genes, Brain Behav.* doi:10.1111/gbb.12279.

Hodgins-Davis, A., Rice, D. P., Townsend, J. P., and Novembre, J. (2015). Gene expression evolves under a house-of-cards model of stabilizing selection. *Mol. Biol. Evol.* 32, 2130–2140. doi:10.1093/molbev/msv094.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* 37, 1–13. doi:10.1093/nar/gkn923.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211.

Hunt, B. G., Ometto, L., Keller, L., and Goodisman, M. A. D. (2013). Evolution at two levels in fire ants: The relationship between patterns of gene expression and protein sequence evolution. *Mol. Biol. Evol.* 30, 263–271. doi:10.1093/molbev/mss234.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi:10.1093/molbev/mst010.

Khaitovich, P., Hellmann, I., Enard, W., Nowick, K., Leinweber, M., Franz, H., et al. (2005). Evolution: Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science (80-.).* 309, 1850–1854. doi:10.1126/science.1108296.

King, M. C., and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science (80-.).* 188, 107–116. doi:10.1126/science.1090005.

Landry, C. R., Wittkopp, P. J., Taubes, C. H., Ranz, J. M., Clark, A. G., and Hartl, D. L. (2005). Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. *Genetics* 171, 1813–1822. doi:10.1534/genetics.105.047449.

Lemos, B., Meiklejohn, C. D., Cáceres, M., and Hartl, D. L. (2005). Rates of divergence in gene expression profiles of primates, mice, and flies: Stabilizing selection and variability among functional categories. *Evolution* 59, 126–137. doi:10.1111/j.0014-3820.2005.tb00900.x.

Liao, Y., Smyth, G. K., and Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930. doi:10.1093/bioinformatics/btt656.

Llopart, A. (2012). The rapid evolution of X-linked male-biased gene expression and the large-X effect in *Drosophila yakuba, D. santomea,* and their hybrids. *Mol. Biol. Evol.* 29, 3873–3886. doi:10.1093/molbev/mss190.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold-change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15. doi:10.1186/s13059-014-0550-8.

Mack, K. L., Campbell, P., and Nachman, M. W. (2016). Gene regulation and speciation in house mice. *Genome Res.* 26, 451–461. doi:10.1101/gr.195743.115.

Mack, K. L., and Nachman, M. W. (2017). Gene regulation and speciation. *Trends Genet.* 33, 68–80. doi:10.1016/j.tig.2016.11.003.

Mackay, T. F. C. (2014). Epistasis and quantitative traits: Using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* doi:10.1038/nrg3627.

Mackay, T. F. C., and Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Med.* doi:10.1186/gm561.

Maheshwari, S., and Barbash, D. A. (2011). The genetics of hybrid incompatibilities. *Annu. Rev. Genet.* 45, 331–355. doi:10.1146/annurev-genet-110410-132514.

Masly, J. P., Jones, C. D., Noor, M. A. F., Locke, J., and Orr, H. A. (2006). Gene transposition as a cause of hybrid sterility in *Drosophila*. *Science (80-.).* 313, 1448–1450. doi:10.1126/science.1128721.

Masly, J. P., and Presgraves, D. C. (2007). High-resolution genome-wide dissection of the two rules of speciation in *Drosophila*. *PLoS Biol.* 5, 1890–1898. masdoi:10.1371/journal.pbio.0050243.

McManus, C. J., Coolon, J. D., Duff, M. O., Eipper-Mains, J., Graveley, B. R., and Wittkopp, P. J. (2010). Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Res.* 20, 816–825. doi:10.1101/gr.102491.109.

Mihola, O., Trachtulec, Z., Vlcek, C., Schimenti, J. C., and Forejt, J. (2009). A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science (80-.).* 323, 373–375. doi:10.1126/science.1163601.

Moehring, A. J., Teeter, K. C., and Noor, M. A. F. (2007). Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Mol. Biol. Evol.* 24, 137–145. doi:10.1093/molbev/msl142.

Morán, T., and Fontdevila, A. (2014). Genome-wide dissection of hybrid sterility in drosophila confirms a polygenic threshold architecture. *J. Hered.* 105, 381–396. doi:10.1093/jhered/esu003.

Muller, H. J. (1942). Isolating mechanisms, evolution and temperature. *Biol. Symp.*

Nuzhdin, S. V., Wayne, M. L., Harmon, K. L., and McIntyre, L. M. (2004). Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. *Mol. Biol. Evol.* 21, 1308–1317. doi:10.1093/molbev/msh128.

Okonechnikov, K., Golosova, O., Fursov, M., Varlamov, A., Vaskin, Y., Efremov, I., et al. (2012). Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* 28, 1166–1167. doi:10.1093/bioinformatics/bts091.

Orr, H. A. (1996). Dobzhansky, Bateson, and the genetics of speciation. *Genetics*.

Ortíz-Barrientos, D., Counterman, B. A., and Noor, M. A. F. (2007). Gene expression divergence and the origin of hybrid dysfunctions. *Genetica* 129, 71–81. doi:10.1007/s10709-006-0034-1.

Phadnis, N., and Allen Orr, H. (2009). A single gene causes both male sterility and segre. *Science* 323, 376–379. doi:10.1126/science.1163934.A.

Presgraves, D. C. (2008). Sex chromosomes and speciation in *Drosophila*. *Trends Genet.* 24, 336–343. doi:10.1016/j.tig.2008.04.007.

Presgraves, D. C., Balagopalan, L., Abmayr, S. M., and Orr, H. A. (2003). Adaptive evolution drives divergence of a hybrid inviability gene between two species of *Drosophila*. *Nature* 423, 715–719. doi:10.1038/nature01679.

Ranz, J. M., Namgyal, K., Gibson, G., and Hartl, D. L. (2004). Anomalies in the expression profile of interspecific hybrids of *Drosophila melanogaster* and *Drosophila simulans*. *Genome Res.* 14, 373–379. doi:10.1101/gr.2019804.

Renaut, S., Nolte, A. W., and Bernatchez, L. (2009). Gene expression divergence and hybrid misexpression between lake whitefish species pairs (Coregonus spp. Salmonidae). *Mol. Biol. Evol.* doi:10.1093/molbev/msp017.

Rifkin, S. A., Kim, J., and White, K. P. (2003). Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat. Genet.* 33, 138–144. doi:10.1038/ng1086.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140. doi:10.1093/bioinformatics/btp616.

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11. doi:10.1186/gb-2010-11-3-r25.

Schaeffer, S. W., and Miller, E. L. (1991). Nucleotide sequence analysis of Adh genes estimates the time of geographic isolation of the Bogota population of *Drosophila pseudoobscura*. *Proc. Natl. Acad. Sci. U. S. A.* 88, 6097–6101. doi:10.1073/pnas.88.14.6097.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., et al. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *Rna* 22, 839–851. doi:10.1261/rna.053959.115.

Sundararajan, V., and Civetta, A. (2011). Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in drosophila sterile hybrids. *J. Mol. Evol.* doi:10.1007/s00239-010-9404-5.

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

Ting, C. T., Tsaur, S. C., Wu, M. L., and Wu, C. I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science (80-.).* 282, 1501–1504. doi:10.1126/science.282.5393.1501.

Tiroshauth, I., Reikhav, S., Levy, A. A., and Barkai, N. (2009). A yeast hybrid provides insight into the evolution of gene expression regulation. *Science (80-.).* 324, 659–662. doi:10.1126/science.1169766.

Turner, L. M., and Harr, B. (2014). Genome-wide mapping in a house mouse hybrid zone reveals hybrid sterility loci and Dobzhansky-Muller interactions. *Elife* 3, 1–25. doi:10.7554/eLife.02504.

Victoria Cattani, M., and Presgraves, D. C. (2012). Incompatibility between X chromosome factor and pericentric heterochromatic region causes lethality in hybrids between drosophila melanogaster and its sibling species. *Genetics* 191, 549–559. doi:10.1534/genetics.112.139683.

Vonesch, S. C., Lamparter, D., Mackay, T. F. C., Bergmann, S., and Hafen, E. (2016). Genome-Wide Analysis Reveals Novel Regulators of Growth in Drosophila melanogaster. *PLoS Genet.* doi:10.1371/journal.pgen.1005616.

Wang, R. L., Wakeley, J., and Hey, J. (1997). Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. *Genetics* 147, 1091–1106.

Wang, Z., Sun, X., Zhao, Y., Guo, X., Jiang, H., Li, H., et al. (2015). Evolution of gene regulation during transcription and translation. *Genome Biol. Evol.* doi:10.1093/gbe/evv059.

Whittle, C. A., Sun, Y., and Johannesson, H. (2014). Dynamics of transcriptome evolution in the model eukaryote *Neurospora*. *J. Evol. Biol.* 27, 1125–1135. doi:10.1111/jeb.12386.

Chapter 4: General Discussion

Studies on the genetic basis of hybrid male sterility have established that a single gene alone is not enough to cause hybrid male sterility. Although single genes with major contributions have been identified, their effects often require the cooperation of other genes. For example, *OdsH*, the first hybrid male sterility gene identified in *Drosophila* that contributes to hybrid male sterility between *D. simulans* and *D. mauritiana* (Ting et al. 1998), does not cause hybrid male sterility in of itself but requires the interaction of genes found on the Y-chromosome and 4th autosome (Bayes and Malik 2009; Phadnis and Malik 2013). Similarly, *Ovd*, a major hybrid male sterility gene between *D. p. pseudoobscura* and *D. p. bogotana* (Phadnis and Orr 2009), requires the interaction of genetic targets found on the 2nd and 3rd autosomes for the manifestation of its sterility effect (Phadnis 2011). This highlights the importance of complex interactions between multiple genes from different loci in the establishment of reproductive isolation.

The breakdown and possible novel interactions between *cis*- and *trans*-regulatory elements in a hybrid background may disrupt gene interaction networks and cause hybrid male sterility. In both the *D. pseudoobscura* and *D. willistoni* subspecies pairs analysed in this thesis, gene interaction networks among transgressive genes misexpressed in their sterile F_1 hybrids have been identified. Among the transgressive genes belonging to the *D. pseudoobscura* F_1 sterile hybrid gene interaction network (Figure 3.3), four genes (GA10010, GA10921, GA17404, and GA18484) are potential targets of the major sterility gene *Ovd* (Appendix: Targets of *Ovderdrive* poster). GA10921 and GA17404 both encode proteins with cell adhesion domains while GA18484 encodes a protein with both a cell adhesion and protease domain (Alhazmi et al. 2019; Go et al. 2019). Protease and cell adhesion were previously found to be two of the largest gene ontologies whose

changes in expression were linked to hybrid male sterility in the *D. pseudoobscura* subspecies pair (Gomes and Civetta 2015). Follow up gene expression assays on some candidate genes, which includes GA10921, GA17404, and GA18484, using fertile backcross progeny and an introgression progeny in which the *Ovd* allele was swapped to produce sterile and fertile male progeny genotypically similar to F_1 sterile hybrids have confirmed GA10921 as a strong candidate for one of the interacting partners of *Ovd* linked to sterility (Alhazmi et al. 2019; Go et al. 2019). Although not a direct target of *Ovd*, GA17404 and GA18484 may still contribute to hybrid male sterility through gene interaction networks (Figure 3.3).

Among the network of transgressive genes expressed in the testes associated with sterility in the *D. willistoni* subspecies pair (Figure 2.7), cell adhesion genes were also overrepresented. Furthermore, three of these cell adhesion genes, GK11667, GK20889, and GK21871, were orthologues of genes (GA17404, GA18484, and GA20821, respectively) found in the network of transgressive genes associated with sterility in the *D. pseudoobscura* subspecies pair. The role of cell adhesion genes in the onset hybrid male sterility in *Drosophila* has not been characterised beyond the observation of Gomes and Civetta (2015). In general, aside from genes broadly associated with spermatogenesis (Michalak and Noor 2003), no recurrent class of genes have been consistently linked with the onset of hybrid male sterility. The representation of cell adhesion genes among transgressive genes linked to sterility, especially some with orthologous pairs, between two different subspecies pairs of *Drosophila* whose sterile hybrid males show different phenotypes for sterility potentially highlights a role for cell adhesion genes in the onset of interspecies hybrid male sterility and speciation.

The commonality I have found for a possible role of cell adhesion genes in hybrid male sterility is surprising as *Drosophila willistoni* and *D. pseudoobscura* diverged from each other approximately 50 million years ago (Median= 54 mya; CI= 35-70 mya. Source: http://www.timetree.org/). More striking is that genome-wide studies that have aimed to identify genes involved in hybrid male sterility in mammals have also found cell adhesion genes as one of the classes of genes that contribute to hybrid male sterility. In mice hybrids between *Mus musculus musculus* and *M. m. domesticus* (two subspecies in the early stages of speciation and whose hybrid males exhibit unidirectional sterility), quantitative trait locus mapping has identified a region on the X-chromosome with a strong association to hybrid male sterility. Functional annotation of the genes found in this region of the X-chromosome found a cluster of 25 cell adhesion genes (Turner et al. 2014). A genome-wide association study on Savannah and Bengal interspecific hybrid cat breeds, have identified 8 autosomal genes linked to hybrid male sterility (Davis et al. 2015). One of these genes, *CADM1*, encodes a cell adhesion molecule. The identification of a cell adhesion gene responsible for sterility in hybrids between Savannah and Bengal cat breeds is particularly interesting given that the sterility phenotype between these breeds is somewhat reminiscent of the sterility phenotype seen in sterile male hybrids from both the *D. willistoni* and *D. pseudoobscura* subspecies pairs. Like the F_1 male sterile hybrids between the closely related subspecies of *D. willistoni* analyzed in chapter 2, F_1 hybrids from both cat breeds suffer from azoospermia and severe degeneration of the seminiferous tubules of the testes (Davis et al. 2015; Davis et al. 2020). Later generation backcross hybrids among Savanah and Bengal cats display a phenotype more similar to that of hybrids between the *D. pseudoobscura* subspecies, with defects in meiosis and low amounts of sperm with high proportions of abnormalities (Gomes and Civetta 2014; Davis et al. 2015).

Overall, these studies suggest a previously unexplored role for cell adhesion genes in the manifestation of hybrid male sterility and the onset of speciation. Future studies on species pairs in the early stages of speciation across a wide range of organisms are required to establish whether or not the misregulation of cell adhesion genes is indeed linked with testes development and spermatogenesis defects seen in sterile male interspecies hybrids.

Policy Implications

The policy implications of my work would come from its potential translational aspects. In this thesis, I used RNA-sequencing at the genome-wide level to identify genes linked to sterility in *Drosophila*. Since much of the *Drosophila* genome share orthologues with the human genome (Rubin et al. 2000), the sterility related genes identified from my project may have implications in our understanding of human sterility as well.

Approximately 10-15% of couples experience challenges when trying to conceive. In about half of these cases, infertility can be attributed to the male (Moore and Reijo-Pera 2000). Among these infertile men, 12% have untreatable conditions such as Klinefelter's syndrome or testicular atrophy, 13% have a potentially treatable condition like genital tract obstructions while the remaining 75% suffer from low sperm counts and/or low sperm motility (Baker et al. 1986).

The sterility phenotype in the *D. pseudoobscura* subspecies pair is similar to the 75% of infertility issues in men. The sterile F_1 hybrids between these subspecies pair are capable of producing mature although non-motile sperm (Snook 1998; Gomes and Civetta 2014). Among the transgressive genes associated with sterility, three genes with

known human orthologues with functions related to sperm development were found. GA14907 and GA20504 have human orthologues (LAP3 and ANPEP, respectively) whose functions are similar in *Drosophila* in that they both affect sperm function (Agarwal et al. 2015; Laurinyecz et al. 2019). GA10278, has a human orthologue HMGCR which is involved in the migration of primordial germ cells in the testes during the early stages of sperm development (Van Doren et al. 1998). The fact that these three genes are found in the network of transgressive genes in hybrids of *D. pseudoobscura* (Figure 3.3) suggests a potential sterility pathway that may aid in our understanding of human infertility. Identifying a pathway that leads to the production of non-motile sperm may also help further the development of male-directed oral contraceptives as an alternative to the female birth-control pill that often comes with side-effects due to hormonal imbalances (Liao and Dollin 2012).

Beyond male fertility the results of my work may also have policy implications in the field of pest control. Traditionally, broad-spectrum insecticides like neonicotinoids have been used to control insect populations in an agricultural setting. In the recent years, neonicotinoid use has gained scrutiny over its potential accumulation in the environment and impact on non-target organisms. For example, neonicotinoid use has been implicated in the decline of bee populations known as colony collapse disorder (Whitehorn et al. 2012) and exposure to neonicotinoids and its metabolites have been associated with reduced growth and impaired immune function in other species (Thompson et al. 2020). These issues prompted a renewed interest in alternative species-specific methods of pest control. One alternative method is the sterile insect technique (SIT). SIT is a non-insecticidal method that relies on the release of sterile males who mate with wild females and prevent offspring. However, the classical method of sterilising males uses radiation

which may lower their ability to compete with wild males for females (Guerfali et al. 2011). An alternative method to radiation sterilisation has been developed in mosquitoes and relies on the use of RNAi to silence male reproductive genes (Whyard et al. 2015).

The results of my work provide potential gene targets for SIT applications. Of the three genes likely associated with sperm development mentioned above, GA14907 and GA10278 have orthologues in *Anopheles gambiae* and *Aedes aegypti*, two mosquito species that commonly act as disease vectors for malaria and dengue fever respectively. The misexpression of these genes in sterile hybrids of the *D. pseudoobscura* subspecies pair leads to the production of non-motile sperm and no other apparent reductions in fitness. This makes the orthologues of GA14907 and GA10278 ideal candidate genes for SIT since their altered expression will only cause male sterility without affecting the ability to compete for females.

Overall, the implications of my work are not only beneficial in furthering our understanding of speciation, but my identification of gene interaction networks linked to sterility may also have applications in understanding male sterility and in providing potential gene targets for genetic based techniques of pest control. The findings of this thesis add to our growing body of knowledge needed in making informed decisions and policies.

REFERENCES

Agarwal, A., Sharma, R., Durairajanayagam, D., Ayaz, A., Cui, Z., Willard, B., et al. (2015). Major protein alterations in spermatozoa from infertile men with unilateral varicocele. *Reprod. Biol. Endocrinol.* 13, 1–22. doi:10.1186/s12958-015-0007-2.

Alhazmi, D., Fudyk, S. K., and Civetta, A. (2019). Testes proteases expression and hybrid male sterility between subspecies of *Drosophila pseudoobscura*. *G3 Genes, Genomes, Genet.* 9, 1065–1074. doi:10.1534/g3.119.300580.

Baker, H.W. (1986). Relative incidence of etiological disorders in male infertility.

Bayes, J. J., and Malik, H. S. (2009). Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. *Science (80-.).* 326, 1538–1541. doi:10.1126/science.1181756.

Go, A., Alhazmi, D., and Civetta, A. (2019). Altered expression of cell adhesion genes and hybrid male sterility between subspecies of *Drosophila pseudoobscura*. *Genome* 62, 657–663. doi:10.1139/gen-2019-0066.

Gomes, S., and Civetta, A. (2014). Misregulation of spermatogenesis genes in *Drosophila* hybrids is lineage-specific and driven by the combined effects of sterility and fast male regulatory divergence. *J. Evol. Biol.* 27, 1775–1783. doi:10.1111/jeb.12428.

Gomes, S., and Civetta, A. (2015). Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Sci. Rep.* 5, 11976. doi:10.1038/srep11976.

Guerfali, M. M. S., Parker, A., Fadhl, S., Hemdane, H., Raies, A., and Chevrier, C. (2011). Fitness and reproductive potential of irradiated mass-reared mediterranean fruit fly males ceratitis capitata (Diptera: Tephritidae): Lowering radiation doses. *Florida Entomol.* 94, 1042–1050. doi:10.1653/024.094.0443.

Hodgins-Davis, A., Rice, D. P., Townsend, J. P., and Novembre, J. (2015). Gene expression evolves under a house-of-cards model of stabilizing selection. *Mol. Biol. Evol.* 32, 2130–2140. doi:10.1093/molbev/msv094.

Laurinyecz, B., Vedelek, V., Kovács, A. L., Szilasi, K., Lipinszki, Z., Slezák, C., et al. (2019). Sperm-Leucylaminopeptidases are required for male fertility as structural components of mitochondrial paracrystalline material in Drosophila melanogaster sperm. *PLoS Genet.* 15, 1–24. doi:10.1371/journal.pgen.1007987.

Liao, P. V., and Dollin, J. (2012). Half a century of the oral contraceptive pill: historical review and view to the future. *Can. Fam. Physician* 58, 757–760.

Michalak, P., and Noor, M. A. F. (2003). Genome-wide patterns of expression in Drosophila pure species and hybrid males. *Mol. Biol. Evol.* 20, 1070–1076. doi:10.1093/molbev/msg119.

Moore, F. L., and Reijo-Pera, R. A. (2000). Male sperm motility dictated by mother's mtDNA. *Am. J. Hum. Genet.* 67, 543–548. doi:10.1086/303061.

Phadnis, N. (2011). Genetic architecture of male sterility and segregation distortion in *Drosophila pseudoobscura* bogota-USA hybrids. *Genetics* 189, 1001–1009. doi:10.1534/genetics.111.132324.

Phadnis, N., and Malik, H. S. (2013). "The molecular and evolutionary basis of hybrid sterility: From Odysseus to Overdrive," in *Speciation: Natural Processes, Genetics and Biodiversity*.

Phadnis, N., and Orr, H. A. (2009). A single gene causes both male sterility and segregation distortion in *Drosophila* hybrids. *Science (80-.).* 323, 376–379. doi:10.1126/science.1163934.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor Miklos, G. L., Nelson, C. R., Hariharan, I. K., et al. (2000). Comparative genomics of the eukaryotes. *Science (80-.).* doi:10.1126/science.287.5461.2204.

Snook, R. R. (1998). Sperm Production and Sterility in Hybrids Between Two Subspecies of Drosophila pseudoobscura. *Evolution*. doi:10.2307/2410943.

Thompson, D. A., Lehmler, H. J., Kolpin, D. W., Hladik, M. L., Vargo, J. D., Schilling, K. E., et al. (2020). A critical review on the potential impacts of neonicotinoid insecticide use: Current knowledge of environmental fate, toxicity, and implications for human health. *Environ. Sci. Process. Impacts* 22, 1315–1346. doi:10.1039/c9em00586b.

Ting, C. T., Tsaur, S. C., Wu, M. L., and Wu, C. I. (1998). A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science (80-.).* 282, 1501–1504. doi:10.1126/science.282.5393.1501.

Turner, L. M., White, M. A., Tautz, D., and Payseur, B. A. (2014). Genomic Networks of Hybrid Sterility. *PLoS Genet.* 10, 18–22. doi:10.1371/journal.pgen.1004162.

Whitehorn, P. R., O'Connor, S., Wackers, F. L., and Goulson, D. (2012). Neonicotinoid pesticide reduces bumble bee colony growth and queen production. *Science (80-.).* 336, 351–352. doi:10.1126/science.1215025.

Whyard, S., Erdelyan, C. N. G., Partridge, A. L., Singh, A. D., Beebe, N. W., and Capina, R. (2015). Silencing the buzz: A new approach to population suppression of mosquitoes by feeding larvae double-stranded RNAs. *Parasites and Vectors* 8, 1–11. doi:10.1186/s13071-015-0716-6.

APPENDIX

Supplementary Text S1

The lower proportion of uniquely mapped sequences in the accessory gland and testis compared to ovary samples results from a higher presence of rRNA. FastQC reports for the raw data in these samples showed two jagged peaks in the "Per sequence GC content" compared to one smooth curve in the reads from the ovaries (Figure S1A). This is indicative of incomplete ribo-depletion. To confirm this extent, we performed BLASTn homology searches for the top five overrepresented sequences according to the FastQC reports of the indicated samples finding highly significant hits at 1E-15 to ribosomal DNA sequences from other *Drosophila* species such as *D. virilis* and *D. subobscura*. Considering multi-mapping sequencing reads (STAR: MAPQ value for unique mappers was set to the maximum of 255 and the number of alignments to include in the output was increased to 100; featureCounts: multi-mapping reads were included in the counts) confirmed the higher abundance rRNA genes compared to unique mapped reads in the samples of accessory glands and testes but not in ovaries (Figure S1B). It is not apparent at this time whether these differences are reflective of a technical bias during the library construction of the samples for different differences, a biological difference in the relative amount of rRNA across tissues, or both.

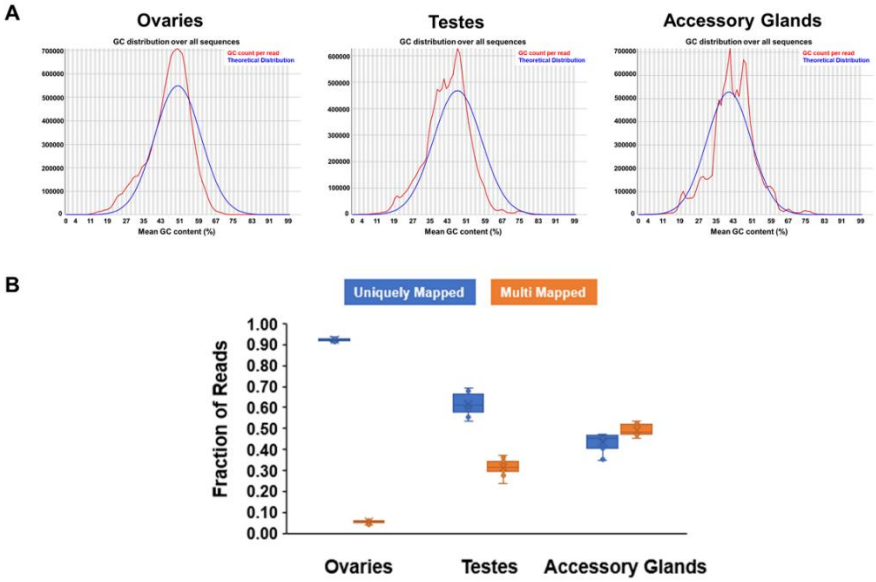


Figure S1. Mapping trends for sequence mapping across different tissue libraries. A) Example of the "Per sequence GC content" profile as part of the FastQC report for the raw reads of a Guadeloupe sample from each of the three tissues (ovaries, testes, and accessory glands). Jagged peaks are seen in the curves for both testes and accessory glands samples while one smooth peek is seen for the ovaries samples. This general trend is observed across all other RNA-seq datasets for testes and accessory glands generated in this study. B) Fraction of sequencing reads between the counts obtained from the default settings of STAR and the setting that allowed multi-mapping reads to be included and counted.

Table S1. Library sequenced in chapter 2.

Tissue	Sample *	Number of Sequencing Reads †			
		Total	Uniquely Mapped	Multi-Mapped	Unmapped
Ovaries	Gua1	13,262,062	12,369,796	663,131	218,824
	Gua2	15,189,890	14,122,507	820,318	235,443
	Gua3	13,404,121	12,226,748	919,365	243,955
	H31	12,673,546	11,733,158	677,699	263,609
	H32	12,405,196	11,405,645	737,567	261,749
	H33	12,435,718	11,561,403	663,458	210,163
	H41	10,882,317	10,060,481	587,349	233,968
	H42	12,595,168	11,718,609	686,620	190,187
	H43	12,169,043	11,384,190	479,247	305,442
	Uru1	11,629,406	10,712,994	660,462	225,846
	Uru2	12,245,110	11,195,557	716,567	330,066
	Uru3	11,787,776	10,856,053	714,256	218,073
Testes	Gua1	11,435,347	7,842,724	2,746,969	846,215
	Gua2	14,384,785	7,680,629	4,331,565	2,372,051
	Gua3	12,283,756	7,198,005	4,028,898	1,056,403
	H31	11,625,014	6,943,100	3,982,956	698,663
	H32	13,702,980	7,867,941	4,958,747	875,620
	H33	12,528,083	8,521,339	3,685,225	321,971
	H41	13,060,365	8,065,931	4,442,605	551,147
	H42	12,813,981	8,863,975	3,536,810	413,891
	H43	13,105,681	7,974,057	4,235,769	896,428
	Uru1	12,166,360	7,579,445	3,650,269	935,593
	Uru2	12,567,210	6,985,788	4,695,604	885,988
	Uru3	11,503,927	7,186,932	3,565,688	751,206
Accessory Glands	Gua1	12,182,182	5,628,737	5,779,480	771,132
	Gua2	10,872,905	3,803,831	5,676,509	1,392,819
	Gua3	12,227,786	5,535,298	5,929,958	763,013
	H31	11,134,815	3,956,607	5,944,994	1,232,624
	H32	12,230,539	4,958,503	6,436,023	835,345
	H33	13,066,538	5,902,311	6,315,226	849,324
	H41	12,183,723	5,022,714	6,184,759	975,916
	H42	12,360,302	5,774,254	5,939,611	646,443
	H43	11,233,767	5,083,313	5,681,418	468,448
	Uru1	13,402,346	6,323,392	6,068,351	1,010,536
	Uru2	12,838,171	6,045,045	6,009,212	784,412
	Uru3	11,662,816	5,478,497	5,537,227	647,286

* Genotype replicate number. *D. willistoni g*enotypes: Gua, Guadeloupe; Uru, Uruguay; H3 (fertile male hybrid), Uruguay mother x Guadalupe father; H4 (sterile male hybrid), Guadeloupe mother x Uruguay father

† Illumina 100 nt paired end reads.

Table S2. Percentage of uniquely mapped for each tissue and genotype.

Genotype *	Ovaries	Accessory Glands	Testes
Guadalupe	92.49 ± 0.90	42.15 ± 5.08	60.19 ± 6.30
Uruguay	91.88 ± 0.32	47.08 ± 0.09	60.12 ± 3.20
H3	92.50 ± 0.42	40.41 ± 3.94	61.72 ± 4.55
H4	93.01 ± 0.45	44.40 ± 2.33	63.92 ± 3.73
Global	92.47 ± 0.73	43.51 ± 4.42	61.49 ± 5.07

Average ± SD. Uniquely mapped reads are those that mapped to a single site in the reference genome.
* H3 (fertile male hybrid), Uruguay female × Guadeloupe male; H4 (sterile male hybrid), Guadeloupe female × Uruguay male.

Table S3. Genes found expressed across the parental subspecies and their hybrids

	Gene Type		
Sample & Genotypes *	**Coding**	**Non-Coding †**	**All**
Testes	10,523	1,021 (980, 29, 4, 8)	11,544
Gua	10,198	954 (915, 28, 4, 7)	11,152
Uru	10,139	850 (813, 29, 0, 8)	10,989
H3	10,336	958 (924, 26, 0, 8)	11,294
H4	10,168	922 (890, 20, 4, 8)	11,090
Accessory Glands	9,030	616 (567, 34, 1, 14)	9,646
Gua	8,804	571 (524, 34, 1, 12)	9,375
Uru	8,379	523 (478, 31, 1, 13)	8,902
H3	8,452	554 (507, 34, 1, 12)	9,006
H4	8,848	577 (529, 33, 1, 14)	9,425
Ovaries	7,886	342 (333, 2, 0, 7)	8,228
Gua	7,599	294 (289, 1, 0, 4)	7,893
Uru	7,667	289 (283, 0, 0, 6)	7,956
H3	7,707	317 (309, 1, 0, 7)	8,024
H4	7,793	317 (310, 1, 0, 6)	8,110
≥1 Sample	11,022	1,285 (1,229, 34, 5, 17)	12,307
Gua	10,706	1,202 (1,151, 34, 5, 14)	11,910
Uru	10,678	1,091 (1,043, 31, 1, 16)	11,769
H3	10,844	1,234 (1,182, 34, 1, 17)	12,078
H4	10,718	1,185 (1,131, 33, 5, 16)	11,903

***** *D. willistoni g*enotypes: Gua, Guadeloupe; Uru, Uruguay; H3 (fertile male hybrid), Uruguay mother x Guadalupe father; H4 (sterile male hybrid), Guadeloupe mother x Uruguay father.

† In parenthesis the number of lncRNAs, rRNAs, tRNAs, and snoRNA, respectively.

Table S4. Differences in number of expressed gene models across the tissues surveyed

Gene Type	F	P^*
Coding	272.1	8.94×10^{-9}
lncRNA	325.9	4.02×10^{-9}
Both	318.8	4.43×10^{-9}

* One-way ANOVA.

Table S5. Salient patterns of differential expression between subspecies with a 2-fold-change threshold.

	Gene Category		
Pattern	**Coding**	**Non-Coding ***	**All**
Testes	10,523	1,021 (980, 29, 4, 8)	11,544
Gua = Uru	9,974	787 (751, 27, 1, 8)	10,761 (93.22%)
Gua > Uru	282	161 (157, 1, 3, 0)	443 (3.84%)
Gua < Uru	267	73 (72, 1, 0, 0)	340 (2.95%)
Accessory Glands	9,030	616 (567, 34, 1, 14)	9,646
Gua = Uru	8,643	525 (482, 30, 1, 12)	9,168 (95.04%)
Gua > Uru	237	66 (62, 3, 0, 1)	303 (3.14%)
Gua < Uru	150	25 (23, 1, 0 ,1)	175 (1.81%)
Ovaries	7,886	342 (333, 2, 0, 7)	8,228
Gua = Uru	7,703	293 (285, 1, 0, 7)	7,996 (97.18%)
Gua > Uru	92	31 (30, 1, 0, 0)	123 (1.49%)
Gua < Uru	91	18 (18, 0, 0, 0)	109 (1.32%)
All 3 samples #	11,022	1,285 (1,229, 34, 5, 17)	12,307
Consistent pattern	6,679	149 (148, 1, 0, 0)	6,828
Gua = Uru	6,669	146 (146, 0, 0, 0)	6,815 (99.81%)
Gua > Uru	3	3 (2, 1, 0, 0)	6 (0.09%)
Gua < Uru	7	0	7 (0.10%)
Inconsistent pattern ¦	4,343	1,136 (1,081, 33, 5, 17)	5,479

Direction of the differential expression between the two subspecies: > overexpression, < underexpression.

* In parenthesis the number of lncRNAs, rRNAs, tRNAs, and snoRNA, respectively.

Only genes expressed across the three types of biological samples.

¦ Genes that show differences in mRNA levels for at least one tissue in a given direction between the subspecies that are not observed in at least one other tissue.

Table S6. Salient patterns of differential expression between subspecies with a 4-fold-change threshold

	Gene Category		
Pattern	**Coding**	**Non-Coding ***	**All**
Testes	10,523	1,021 (980, 29, 4, 8)	11,544
Gua = Uru	10,418	965 (924, 29, 4, 8)	11,383
Gua > Uru	44	32 (32, 0, 0, 0)	76
Gua < Uru	61	24 (24, 0, 0, 0)	85
Accessory Glands	9,030	616 (567, 34, 1, 14)	9,646
Gua = Uru	8,934	579 (533, 31, 1, 14)	9,513
Gua > Uru	58	28 (26, 2, 0, 0)	86
Gua < Uru	38	9 (8, 1, 0, 0)	47
Ovaries	7,886	342 (333, 2, 0, 7)	8,228
Gua = Uru	7,848	328 (319, 2, 0, 7)	8,176
Gua > Uru	16	7 (7, 0, 0, 0)	23
Gua < Uru	22	7 (7, 0, 0, 0)	29
All 3 samples #	11,022	1,285 (1,229, 34, 5, 17)	12,307
Consistent pattern	7,060	169 (168, 0, 0, 1)	7,229
Gua = Uru	7,059	169 (168, 0, 0, 1)	7,228
Gua > Uru	1	0	1
Gua < Uru	0	0	0
Inconsistent pattern ¦	3,962	1,116 (1,061, 34, 5, 16)	5,078

Direction of the differential expression between the two subspecies: > overexpression, < underexpression.

* In parenthesis the number of lncRNAs, rRNAs, tRNAs, and snoRNA, respectively.

Only genes expressed across the three types of biological samples.

¦ Genes that show differences in mRNA levels for at least one tissue in a given direction between the subspecies that are not observed in at least one other tissue.

Table S7. Three-sample test for equality of proportions for expression patterns between the parental subspecies

	Comparisons Across Tissues	
Criteria	**DE* / All Expressed Genes**	**(Gua > Uru) / DE**
5% FDR + DESeq2 & edgeR + 0.5 FC	χ^2=670.09, d.f.=2, $P<2.2\times10^{-26}$	χ^2=6.91, d.f.=2, $P=3.2\times10^{-2}$
5% FDR + DESeq2 & edgeR + 2 FC	χ^2=157.52, d.f.=2, $P<2.2\times10^{-26}$	χ^2=8.88, d.f.=2, $P=1.2\times10^{-2}$
5% FDR + DESeq2 & edgeR + 4 FC	χ^2=28.302, d.f.=2, $P<7.1\times10^{-7}$	χ^2=10.81, d.f.=2, $P=4.5\times10^{-3}$

* DE, differentially expressed: Gua > Uru and Gua > Uru.

Supplementary Results Chapter 3

Limited differential expression between subspecies and few genes with transgressive expression in hybrids

Approximately 290 million reads were mapped to the *D. p. pseudoobscura* reference genome. We checked for potential mapping bias against *D. p. bogotana* but found that *D. p. bogotana* and *D. p. pseudoobscura* had very similar percentages of total RNA sequence reads mapped (83.54% and 82.11% respectively). Sterile and fertile hybrids also had similar proportions of mapped reads to the reference genome (85.09% and 82.95% respectively). Moreover, out of a total of 16,726 genes annotated, we did not find an overrepresentation of genes with higher average expression in *D. p. pseudoobscura* than *D. p. bogotana* (8,944 and 8,621 respectively), which would have been expected if there was biased mapping.

Under the more stringent lfc threshold of 1, only 819 genes were differentially expressed between the parental subspecies (4.9% of annotated genes), with equal proportions of genes with higher expression in one species or the other (398 in *D. p. bogotana vs.* 421 in *D. p. pseudoobscura*). A limited number of genes showed transgressive expression in hybrids (44) with a significantly higher proportion in the sterile F_1 hybrid males (39) than fertile F_1 hybrid males (4) ($Z = 7.5$; $P<0.00001$). One gene showed transgressive expression in both sterile and fertile hybrids. Using the less stringent threshold of lfc 0.5, the number of differentially expressed genes between the parental subspecies increases to 2,179 (13.03% of total annotated genes; Figure S1). The proportion of genes with higher expression in one species than the other remains similar with 1,103 genes with higher expression in *D. p. bogotana vs.* 1,076 in the *D. p. pseudoobscura*. The trend of a few genes showing transgressive expression in the hybrids and a higher proportion of transgressive genes in the sterile hybrids relative to the fertile hybrids remains the same with a lfc threshold of 0.5. Of the 262 transgressive genes between the hybrids, a significant proportion belonged to the sterile F_1 hybrids (240) ($Z = 18.4$, $P<0.00001$), while only 18 genes showed transgressive expression in the fertile hybrids. Four genes had transgressive expression in both hybrids.

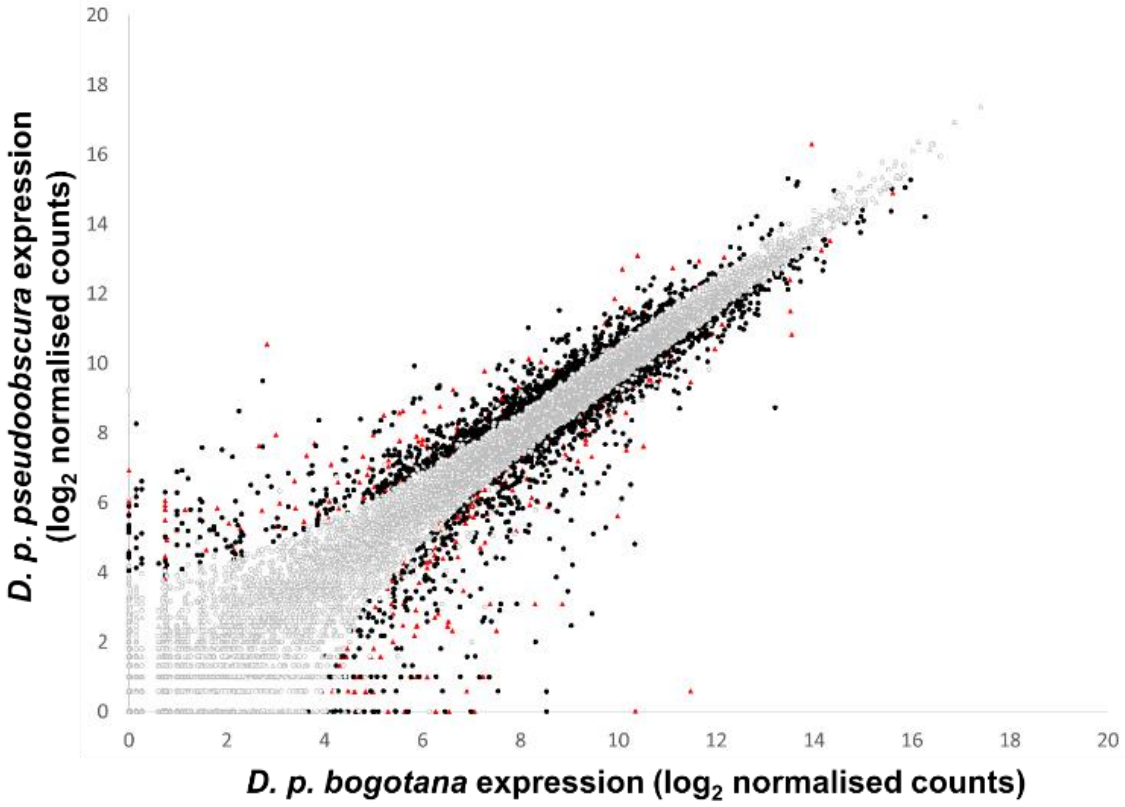


Figure S1. Differential gene expression between the two parental subspecies. DESeq2 normalised counts were used as a measure of gene expression. The expression of 16,726 genes were measured. Genes that are differentially expressed between subspecies are shaded black and red while those that do not show differential expression are grey. Circles denote protein coding genes and non-protein coding genes are represented by triangles. Under the less stringent $\log_2$-fold-change threshold of 0.5, 2,179 (303 non-coding) genes were differentially expressed between subspecies.

Table S8. Significant BLASTn results using compensatory and cis-trans transgressive genes as queries against D. p. pseudoobscura extended gene regions. For matches, black FBgn= compensatory genes, black underlined= non-compensatory; blue= *cis-trans*; and grey= NA. GWH= Genome-wide hits.

Query	*Matches*	*E-values*	*% Identity*	*Best GWH*
FBgn0070459	FBgn0271996	0	100	Unique
FBgn0072286	FBgn0077417; FBgn0079358 FBgn0248786; FBgn0271117 FBgn0271521; FBgn0272358 FBgn0273410	9e-20; 6e-29 7e-28; 3e-20 2e-34; 7e-28 1e-23	98; 88 86; 94 91; 86 88	Unique
FBgn0078355	FBgn0071944; FBgn0078943 FBgn0079731; FBgn0080577 FBgn0245144	1e-23; 4e-49 5e-17; 2e-41 2e-41	73; 77 72; 80 78	Unique
FBgn0078546	FBgn0074132; FBgn0247626	2e-30; 4e-58	75; 85	Unique
FBgn0078680	FBgn0071944; FBgn0075542 FBgn0078943; FBgn0080577 FBgn0081015; FBgn0244760 FBgn0245144	5e-33; 2e-51 6e-64; 3e-43 3e-30; 5e-33 4e-47	77; 79 85; 84 73; 74 82	Unique
FBgn0079637	FBgn0071944; FBgn0245851 FBgn0075000; FBgn0077499 FBgn0078943; FBgn0080577 FBgn0244760;	7e-16; 6e-17 9e-27; 2e-15 8e-15; 8e-15 4e-25; 1e-18	72; 74 64; 71 70; 71 75; 70	FBgn0080782 (6.00E-13)
FBgn0079731	FBgn0071944	1e-18; 1e-18	75; 74	FBgn0272290 (4.00E-08)
FBgn0245605	FBgn0071718; FBgn0077417 FBgn0247977	4e-75; 3e-26 3e-96	78; 97 74	Unique
FBgn0248096	FBgn0071718	9e-20; 0	81; 100	Unique
FBgn0250421	FBgn0079358; FBgn0248786 FBgn0271521; FBgn0273410	2e-26; 6e-33 1e-35; 6e-33 2e-21	87; 76 92; 76 87	FBgn0271418 (6.00E-04)
FBgn0262055	FBgn0271812	0	100	Unique
FBgn0271245	FBgn0246515; FBgn0250150 FBgn0250421	2e-34; 6e-34 2e-21	82; 80 81	FBgn0081107 (2.00E-05)
FBgn0271910	FBgn0272900	2e-160	100	Unique
FBgn0272358	FBgn0079358; FBgn0248786 FBgn0271521	4e-23; 0 4e-24	90; 100 85	Unique

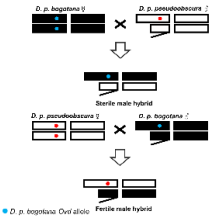
Targets of *Overdrive* poster

Identifying gene regulatory interactions associated with hybrid male sterility in *Drosophila pseudoobscura*

Alwyn Go and Alberto Civetta
Department of Biology, University of Winnipeg, Winnipeg, MB, Canada

BACKGROUND

Incompatibilities between divergent genomes lead to reproductive barriers that promote speciation. The *Drosophila pseudoobscura* subspecies pair is representative of the earliest stages of speciation and is a useful system in understanding the genetics leading to it. They exhibit unidirectional hybrid male sterility:

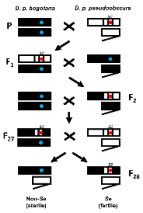


Hybrid male sterility is due to X-autosomal incompatibilities. *Overdrive* (*Ovd*) has a major contribution and is found within the X chromosome, this gene is predicted to have a DNA binding domain making it a possible transcription factor[1].

Objective: To identify the set of genes whose expression is regulated by the state of the *Ovd* alleles.

METHODS

To identify the interacting partners of *Ovd*, we took advantage of the fact that *Ovd* is tightly linked to the *sepia* (*se*) eye colour mutation in *D. p. pseudoobscura*[1]. This allowed us to replace the sterile *D. p. bogotana* allele with the fertile *D. p. pseudoobscura* one using the introgression design below:



Triplicate samples of RNA were extracted from the testes of the parental subspecies, F_1 sterile male hybrids, *se* males and non-*se* males. Differential expression analysis was performed using DESeq2 and edgeR. Results were filtered using a log fold change threshold of 1 and significance (FDR corrected $p < 0.05$).

The overlap of all differentially expressed genes from these comparisons are targets of *Ovd* related to sterility.

RESULTS

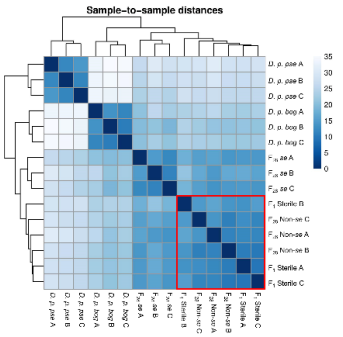


Figure 1: DESeq2 sample clustering based on normalised data. The replicates from the parental subspecies as well as *se* males cluster closely with each other while replicates from the sterile samples (F_1 sterile male hybrids and non-*se* males) form one big cluster.

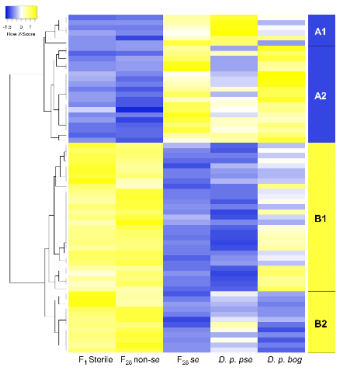


Figure 3: Expression correlation of the 66 targets of *Ovd* across all samples. The two major clusters (A and B) are based on under- and over- expression of the sterile samples relative to the fertile samples. Two sub-clusters (1 and 2) based on differential expression between the parental subspecies and F_{28} *se* are within the main clusters. The proportion of sterility targets of *Ovd* in each sub-cluster are: A1 (17%), A2 (16%), B1 (10%), and B2 (25%).

Table 1: Number of differentially expressed genes for each of the pairwise comparisons from both edgeR and DESeq2 along with the number of consensus genes from the two tools.

Pairwise comparison	edgeR	DESeq2	Consensus
D. p. bogotana - *D. p. pseudoobscura*	1720	1098	1096
D. p. bogotana - F_1 sterile male hybrids	703	357	357
D. p. bogotana - F_{28} *se* male hybrids	544	278	277
D. p. bogotana - F_{28} non-*se* male hybrids	685	375	374
D. p. pseudoobscura - F_1 sterile male hybrids	1115	546	546
D. p. pseudoobscura - F_{28} *se* male hybrids	935	452	452
D. p. pseudoobscura - F_{28} non-*se* male hybrids	1142	572	572
F_1 sterile male hybrids - F_{28} *se* male hybrids	125	58	58
F_1 sterile male hybrids - F_{28} non-*se* male hybrids	1	2	1
F_{28} *se* male hybrids - F_{28} non-*se* male hybrids	110	66	66

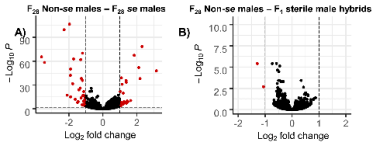


Figure 2: Volcano plot based on DESeq2 results for the comparisons between A) non-*se* vs *se* males. This comparison shows all the genes whose expression is regulated by the state of the *Ovd* alleles. These are potential *Ovd* targets. B) non-*se* vs F_1 sterile male hybrids: shows that after 28 generations of backcrosses the non-*se* males are now nearly identical to the F_1 sterile male hybrids.

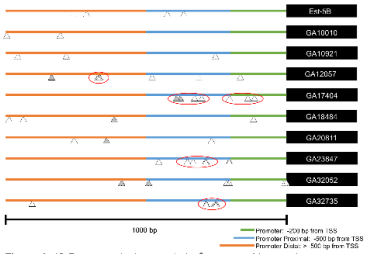


Figure 4: 43 *D. p. pseudoobscura* strains[2] were used to examine polymorphisms and substitutions 1000 bp upstream of the transcription start site of the sterility *Ovd* targets. Δ represent the relative positions of the fixed substitutions between *D. p. bogotana* and *D. p. pseudoobscura*. Grey Δ show *D. p. bogotana* allele frequency of less than 5% while those with dotted lines reflect the scarcity of *D. p. pseudoobscura* sequences (n<10). Red circles show clusters of fixed substitution as putative alternative binding sites for *Ovd* alleles.

CONCLUSIONS

- The non-*se* males are roughly equivalent to the F_1 sterile male hybrids. 66 differentially expressed genes between *se* and non-*se* males are potential targets of *Ovd*. 10 of these genes are sterility targets of *Ovd*.
- The 66 targets of *Ovd* group into 4 networks of correlated expression with significant enrichment for genes involved in *proteolysis, regulation of proteolysis, cell adhesion, cuticle development*, and *mitotic cell cycle checkpoint* based on gene ontology.
- The 10 sterility targets of Ovd will be validated by qPCR and clusters of fixed nucleotide changes upstream of these genes are possible Ovd binding sites which can be tested using ChIP-PCR.

REFERENCES

1. Phadnis, N., & Orr, H. A. (2009). A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *science*, 323(5912), 376-379.
2. Fuller, Z. L., Haynes, G. D., Zhu, D., Batterton, M., Chao, H., Dugan, S., ... & Ongeri, F. (2014). Evidence for stabilizing selection on codon usage in chromosomal rearrangements of Drosophila pseudoobscura. *G3: Genes, Genomes, Genetics*, 4(12), 2433-2449.