NATURAL VARIATION AND DIVERGENCE OF REPETITIVE DNA IN *DROSOPHILA*

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Heng Chin Kevin Wei

February 2016

NATURAL VARIATION AND DIVERGENCE OF REPETITIVE DNA IN *DROSOPHILA*

Heng Chin Kevin Wei, Ph. D.

Cornell University 2016

Most eukaryotic genomes harbor substantial amounts of transposable elements and satellite DNA, collectively known as repetitive sequences. They are characterized as genomic parasites as they increase copy number often at the expanse of host fitness. Because activities of these elements can cause genomic instability, they are usually highly regulated and maintained in the repressive chromatin environment known as heterochromatin. Yet, there are many examples where repetitive DNA adopts crucial cellular functions, including recruiting centromeric proteins and extending telomeres. Additionally, they are often evolving at high rates such that closely related species have vastly different types and abundances of repetitive sequences. However, the evolutionary dynamics of repetitive sequences and mechanisms driving their rapid turnover remain elusive. Here, I present several methods and studies that shed light on to the evolution of repetitive sequences.

In the first study, I present k-Seek, a computational method to identify and quantify simple sequence repeats from whole genome sequences. Using this method, we characterized variation in satellite DNA in *Drosophila melanogaster* natural populations. In the second study, I applied k-Seek on nine *Drosophila* species, to determine the divergence pattern of satellite DNA. I find that the rapid turnover of satllites is not universal but specific to only some fly lineages. In the third study, I investigated one mechanism, known as meiotic drive, by which repeats can

evolve rapidly. I present a novel method to identify meiotic drivers utilizing pooled sequencing. This method was used to test whether different abundances of the telomeric retrotransposon HeT-A, and by extension telomere length, can bias chromosome transmission. In the last study, I examined one striking consequence of the rapid divergence of repetitive DNA – hybrid incompatibility. Using RNA-Seq, I examined, transcriptome-wide, the mis-regulation in hybrids associated with the lethality caused by the *Hybrid Male Rescue* gene which regulates the expression of repetitive sequences in *D. melanogaster* and *D. simulans*.

BIOGRAPHICAL SKETCH

Heng Chin (Kevin) Wei was born in Taipei, Taiwan. At the age of eleven, he, along with his family, immigrated to Vancouver, Canada in 1997. Growing up, Kevin wanted to become a doctor, but little did he know that this would lead him down a very different road. Kevin attended McGill University majoring in Biology and Cognitive Science. As any good pre-med student, he was keen on arguing for grades. In one of his attempts, he instead was invited to join the lab as a research assistant in the lab of Prof. Ehab Abouheif to study the evolution and development of polyphenism in ants. While this was at first a great opportunity to extend his resume, critical for medical school applications, he became fascinated by the scientific approach and the field of evolution. After this initial experience, he spent several summers at different laboratories with Undergraduate Summer Research Awards from the National Science and Research Council of Canada, further developing his interest for research.

After graduating from McGill in 2009, Kevin spent a year in the lab of Chau-Ti Ting at the National Taiwan University in Taiwan to study the process of speciation in Drosophila. There, he became aware of concepts in Population Genetics and Molecular Evolution and learned of the works of Andy Clark, Dan Barbash, and Chip Aquadro. With the desire to further study Evolution, Kevin decided to attend the graduate program of Genetic, Genomics, and Development at Cornell University in 2010. Kevin joined the labs of Andy Clark and Dan Barbash where he focused on the evolution of repetitive DNA in Drosophila.

ACKNOWLEDGEMENTS

I thank my two advisors, Dan Barbash and Andy Clark, for supporting and encouraging me throughout my Ph.D. They provided me lots of freedom and independence to explore not just topics included in this dissertation, but many additional projects. While not all of them worked out, they were all enlightening endeavors that taught me new biology and experimental methods. Moreover, Dan and Andy have both been incredibly insightful and encouraging in guiding me. It is with their tremendous mentorship that I've had such a satisfying and enjoyable experience.

Much gratitude also goes to Paul Soloway, the third member of my thesis committee. On the one hand, I must somewhat apologize to him for roping him into a heavily evolution- and bioinformatics-oriented project, things that are barely tangential to his core interest. On the other hand, I need to thank him as he has been incredibly resourceful throughout my Ph.D.

The completion of several of my studies would not have been possible without the help of several people. Amanada Manfredo has made numerous Illumina libraries for me at incredible speed, as well as explained to me the intricacies and biochemistry of library preparation. Jen Grenier was my go-to person when I have questions about sequence data that were generated before I joined the Clark lab. Julien Aroyles was the person who showed me the basics of analyzing short-read sequences when I began my foray into bioinformatics. Shuqing Ji has been incredibly helpful with all of my wet work. I thank her for putting up with my inane questions and frequent requests. I also like to thank John Lis for letting me work with radioactive material in his lab, which was instrumental in validating k-Seek. In addition, Satyaki Prasad taught me nearly all the fly dissections I have done, including the not-so-pipe-resembling larval brain which were used to generate mitotic spreads.

While many would characterize life in Ithaca as mundane, I actually enjoyed my time here thoroughly. This is, in no small part, thanks to the great friends I have. Thanks to Fabiana for being an understanding and supportive friend, with whom I've devoured many buckets of popcorn. Also, gratitude goes to Ted Chen for the multiple hours we spent on dropping MMR together. Fellow non-Mac using Barbashers, Mike McGurk, Satyaki Prasad, and Dan Zinshteyn (not anymore) were all wonderful individuals with whom I shared many laughs, bad puns, sarcastic witticisms, inappropriate jokes, and philosophical discussions that went nowhere.

Most of my family have spent extended amount of time in Ithaca with me for different reasons. I thank them all for their tremendous support and help throughout my time here, including taking care of my cats while I am away. Most importantly, I thank them for putting up with my temperament and more-than-I-would-like-to-admit frustrating aloofness.

Lastly I would like to thank my partner Lizzie Shih. She has been extremely patient with me during the ups and downs of my Ph.D. While most of our time was spent painfully in a long distance relationship, we pulled through, and I will be forever indebted to her for dropping everything she had in Taiwan to take care of me and our cats in Ithaca.

TABLE OF CONTENTS

Biographical Sketch ……..……...……v

Acknowledgements ……...……vi

Table of Contents ……...……..viii

1 INTRODUCTION

1.1 The C-value paradox and genomic parasites ……....1

1.2 Deleterious and rapidly evolving junk ……3

1.3 Junk with function and influence ……6

1.4 Pieces of the difficult puzzle ……7

2 CORRELATED VARIATION AND POPULATION DIFFERENTIATION IN SATELLITE DNA ABUNDANCE AMONG LINES OF *DROSOPHILA MELANOGASTER*

2.1 Introduction ……10

2.2 Results ……13

2.2.1 Identification and quantification of tandem repeats

2.2.2 Identification of known and novel kmers

2.2.3 Population structure

2.2.4 Concerted evolution of kmer abundance

2.2.5 Interspersion of kmer blocks drives correlation

2.3 Discussion ……..33

2.3.1 Potential causes of population variation

2.3.2 Concerted change of *k*mer abundances

2.4 Materials and Methods ……...38

3 RAPID ACCUMULATIOIN OF SATELLITE DNA IS LINEAGE SPECIFIC IN *DROSOPHILA*

3.1 Introduction ……42

3.2 Results ……44

3.2.1 Identification of satellite DNA in *Drosophila* species with k-Seek

3.2.2 Satellites on the Y chromosome

3.2.3 Gains and losses of satellites along *Drosophila* phylogeny

3.3 Discussion ……..57

3.3.1 Satellite, transposable elements, and genome size divergence

3.3.2 Population genetic considerations of the lineage-specific satellite accumulation

3.3.3 Accumulation of satellites on the Y chromosome

3.4 Materials and Methods ……...60

4 MEIOTIC TRANSMISSION FIDELITY IS ROBUST TO EXTREME TELOMERE-LENGTH DIFFERENCES IN *DROSOPHILA*

4.1 Introduction ……63

4.2 Results ……66

4.2.1 Drastic telomere length variation

4.2.2 Loci and phenotypes associated with telomere length

4.2.3 Assessing non-Mendelian segregation using pooled sequencing

4.2.4 Telomere length does not bias segregation

4.3 Discussion ……..83

4.3.1 *HeT-A* variation and and truncations

4.3.2 Massive telomere length variation in Drosophila compared to other species

4.3.3 Identification of biased segregation using whole-genome sequencing

4.3.4 Potential genomic elements causing meiotic drive

4.4 Materials and Methods ……...90

5 LIMITED GENE MISREGULATION IS EXACERBATED BY ALLELE-SPECIFIC UP-REGULATION IN LETHAL HYBRIDS BETWEEN *DROSOPHILA MELANOGASTER* AND *D. SIMULANS*

5.1 Introduction ……96

5.2 Results ……99

5.2.1 *Hmr* has different effects on *D. melanogaster* and *hybrid* genomes

5.2.2 The *D. melanogaster* alleles are more sensitive to the presence of *Hmr*

5.2.3 Variable growth rates account for a large portion of the differential expression in hybrids

5.2.4 Candidate *Hmr* targets are highly repressed in *D. simulans*

5.2.5 Hybrid-specific expression differences are limited

5.3 Discussion ……117

5.3.1 Expression profile in hybrids

5.3.2 X-chromosome misregulation

5.3.3 Discrepancies and similarities with microarray data

5.3.4 *Hmr* function in hybrids

5.3.5 The role of *Hmr* in allele-specific regulation in hybrids

5.3.6 Divergence of *Hmr* regulation and function

5.4 Materials and Methods …….125

References ……...130

CHAPTER 1

INTRODUCTION

1.1 The C-value paradox and genomic parasites

Evolutionary Genomics is the study of the history and evolutionary forces that shape the DNA content of organisms in its components and totality. This field flourished in recent years in large part due to the advent of high-throughput technologies which made rapid and cost-efficient sequencing of genomes possible. Prior to these methods that have come to define the field, one of the first attempts to study the genome as a whole is estimation of genome size. By measuring the mass of DNA from isolated nuclei, studies calculated the amount of DNA per haploid genome, a measurement that became known as the C-value (Swift, 1950).

When researchers began characterizing C-values across different organisms, an unexpected pattern emerged. At the time, there was a strong suspicion that more "complex" organisms such as ourselves are going to have more intricate and by extension larger genomes. As revealed by C-value estimates across large number of species, genome size of organisms correlates very poorly with the presumed complexity and estimated gene counts (Cavalier-Smith, 1978; Mirsky and Ris, 1951). Moreover, it is poorly conserved and can vary at several orders of magnitude between closely related taxa (Gregory, 2015; Holm-Hansen, 1969; Sparrow et al., 1972). One of the best and most popular examples of this is the C-value of the onion, which is estimated to be over four times more than that of humans (Gurushidze et al., 2012). This unexpected pattern has come to be known as the C-value paradox (Thomas, 1971).

Part of the solution to the puzzle came from the discovery of non-coding DNA, i.e. sequences that do not encode proteins. Such sequences include introns (Vinogradov, 1999), regulatory elements, genes encoding non-coding RNAs, and intergenic sequences. Many of these non-coding DNA are essential and functional features of the genome as they play various roles in genome, cell, and developmental regulation. However, these elements constitute a fraction of the genome, leaving the rest of the genetic content unaccounted for. The full solution came with the discovery of highly repetitive DNA sequences (Gregory, 2001; Hancock, 2002). These sequences are categorized into two types: satellite DNA (Britten and Kohne, 1968) and transposable elements (McClintock, 1950) both which are commonly found in eukaryotic genomes at substantial copy numbers.

The prevalence of repetitive sequences in eukaryotic genomes is due to their ability to increase in copy number. Transposable elements are protein coding elements with the sole purpose of self-replicating in the genome. They are usually divided into two classes: Retrotransposons and DNA transposons. The former replicates via transcription of the full length element followed by reverse-transcription and insertion into the genome, often described as a "copy and paste". The latter excise itself out of its original place in the genome and insert into a new location – "cut and paste". Satellite DNAs are composed of simple noncoding sequences that are tandemly repeating for long stretches. Because of their repeating nature, satellite DNAs are prone to homology mediated crossing-over with unequal exchange with either homologous alleles or satellite blocks elsewhere in the genome. The result of the unequal exchange is a copy with increased size and one with decreased size.

Due to the mechanism by which they expand in copy number, the presence of repetitive elements in the genome is highly deleterious. The activity of transposable elements can often

lead to disruption of gene function as new elements are inserted into coding or regulatory. Moreover, large scale mobilization of transposable elements can cause numerous double strand breaks throughout the genome. Satellite DNAs can induce intra- and inter-chromosomal recombination leading to large scale chromosomal rearrangements that are highly deleterious. As all of these effects negatively disrupt the integrity of the genome, these elements can be extremely detrimental to the organism. As such, they are often characterized as genomic parasites and selfish DNA (Doolittle and Sapienza, 1980; Orgel and Crick, 1980).

1.2 Deleterious and rapidly evolving junk

The preponderance of deleterious DNA elements in eukaryotic genome is therefore fascinating. As a means to subdue their detrimental effects, repetitive DNAs are sequestered in a highly repressive and recombination free regions of the genome, known as heterochromatin. These regions are predominantly around the centromere and telomere with highly compact chromatin that remains condensed throughout cell cycle. The tight packaging of histones is the result of several proteins that deposits and recognizes repressive histone modifications which are primarily di- and tri-methylation of the 9^{th} lysine residue on histone 3 (Grewal and Moazed, 2003; Grewal and Rice, 2004). This methylation mark is deposited by the protein Su(var)3-9 and recognized by Heterochromatin Protein 1a (HP1a) (Lachner et al., 2001; Rea et al., 2000). HP1a in turn recruits Su(var)3-9 which methylates histones in neighboring nucleosome. This feedforward system allows for the spreading of heterochromatin across large stretches of DNA thereby regulating the vast amounts of repetitive sequences in the genome (Howe et al., 1995; Weiler and Wakimoto, 1995).

The ability for genomes to combat the activity of repetitive DNA neutralizes their deleterious effect. Once repressed, they contribute little to cellular and genomic processes and are thus given the label "junk DNA" along with other nonfunctional DNA elements like pseudogenes (Comings, 1972). This term was formalized by Susumu Ohno, who demonstrated that the number of genes in a genome cannot exceed a maximum which is determined by the mutation rate (Ohno, 1972). If the gene number is greater than this upper-bound, the mutational load on the genome would be too large resulting in inevitable fitness decline. He argued that junk DNA reduces the chance for mutations to be disruptive, or "import[ant] for doing nothing" (Ohno, 1972). This view is no longer popular and it is more commonly accepted that junk DNAs, at least with respect to repetitive sequences, exist in genome at a selection mutation balance where their abundance is the product of their rate of amplification, the genome's rate of loss, and their fitness effect if any (Charlesworth et al., 1994; Palazzo and Gregory, 2014). The deleterious effect of the vast quantities of repeats is likely so weak that selection is inefficient at removing them.

While the selection mutation balance model of repetitive sequence may account for the preponderance of repeats in the genome, it is unsatisfactory in explaining the vast differences observed across taxa. It is unlikely that the expansion of genome size via large-scale transposable element insertions will have no deleterious consequences. In fact, mis-regulations of repeats are usually detrimental, causing lethality and/or sterility. Moreover, even the failure to regulate one type or family of transposable element is sufficient to cause sterility. This was observed in *Drosophila melanogaster*, when females from strains naïve to the DNA transposon P-element were mated to males with P-elements, producing sterile progenies. This incompatibility between naïve and exposed genomes, known as hybrid dysgenesis, results from the inability of the

maternally deposited material to silence the P-element inherited from the paternal genome (Boussy et al., 1988; Kidwell, 1983). The severity of the phenotype underscores the deleterious nature of mis-regulation of repetitive sequences and raises the question of how repeats could emerge in the first place, let alone, expand in genomes.

The case of P-element invasion presents a fascinating case that provides insight into this question. P-element invaded the *D. melanogaster* genome in the past 50 years through horizontal transfer from *D. willistoni* (Daniels and Strausbaugh, 1986; Daniels et al., 1990). The fact that it had spread so rapidly indicates that the severity of the dysgenesis phenotype is not representative of the fitness of the wild population when first exposed. One likely possibility is that low amount of P-element in the genome at the initial exposure may only have minor fitness consequences due to limited transposition rates. The genome then developed the necessary defense suppressing the activity of the P-element. In support of this view, theoretical consideration of the population dynamics of invasions has indicated successful invasions are likely of elements that are regulated and have intermediate transposition rates (Barrón et al., 2014; Charlesworth and Langley, 1989; Le Rouzic and Capy, 2005).

As satellite DNAs are not coding elements with ability to self-replicate, they are unlikely spread via horizontal transfer. Therefore, gains of satellite DNA are the result of mutational events that are vertically inherited to the next generation. Rolling circle replication (Okumura et al., 1987) and polymerase slippage (Levinson and Gutman, 1987) are thought to be mechanisms that create small blocks of tandem repeats which could be further amplified through multiple rounds of unequal crossing over (Charlesworth et al., 1994). While the precise mechanism remains unclear, comparison between closely related species have revealed that satellite DNA likely has high rates of change. For example, between closely related Drosophila species, the

satellite DNA content can vastly differ (Lohe and Brutlag, 1987; Lohe and Roberts, 2000) as will be further discussed in Chapter 3. The fitness consequences associated with the intermediate changes, however, is unknown.

1.3 Junk with function and influence

It is worth noting that the C-value paradox is not paradoxical at all, but rather an unexpected observation. As such, it has been often relabeled as the C-value enigma (Gregory, 2001). However, repetitive sequences do present a true paradox or a contradiction: they are fast changing and deleterious, yet they can have important cellular function. For transposable elements, there are documented cases where they provide regulatory sequences for adaptive purposes. In one example, insertion of the DNA transposon mPING into a regulatory region conferred stress-resistance (Naito et al., 2009). Perhaps more surprisingly, transposable elements can also take on essential cellular functions. As Chapter 4 will describe in further detail, three transposable elements in Drosophila have been domesticated to maintain telomere length thereby replacing the necessity of telomerase (Abad et al., 2004; Biessmann et al., 1992; Pardue et al., 2005). In the case of satellites, they are commonly the primary sequences that compose centromeres. Through yet unknown mechanisms, centromeric satellites recruit centromeric histone variants like CENP-A in humans (Chueh et al., 2005; Shelby et al., 1997) and Cid in flies (Blower and Karpen, 2001; Malik and Henikoff, 2001).

Satellite's association with centromeres has prompted geneticist to speculate whether its rapid turnover is driven by mechanisms related to chromosomal segregation (Malik, 2009; Malik and Henikoff, 2001; Zwick et al., 1999). One mechanism of interest is meiotic drive (discussed further in Chapter 4), where a locus acquires the ability to distort meiotic segregation frequency

in females. The inheritance of a meiotic driver will be non-Mendelian and its population frequency will quickly fix (Novitski, 1951; Sandler et al., 1959). Subsequent emergence of a new stronger driver will lead to the replacement of the previous driver in the population, and so on. Although this mechanism is often invoked to explain for the rapid evolution of satellites (Malik, 2009; Malik and Henikoff, 2001), evidences for clear meiotic drive is sparse (Fishman and Willis, 2005).

More recently, repetitive sequence have received a lot of interest for their ability to modulate gene expression genome-wide. In several studies, identical genomes with different repetitive contents on the Y chromosome have been shown to cause significant differences in gene expression and phenotypes, such as immune response (Lemos et al., 2008, 2010; Paredes and Maggert, 2009). The difference is thought to result from repetitive sequences acting as a sink titrating away the limited amount of repressive proteins from the rest of the genome (Weiler and Wakimoto, 1995). Therefore, different amounts of repetitive content will cause differential distribution of repressive proteins and gene regulation in trans. The resulting phenotypic difference may have fitness consequence subjecting repetitive content to selection.

1.4 Pieces of the difficult puzzle

As outlined, repetitive DNA has several conflicting qualities: it is deleterious while but numerous in the genome; it provides little function in most cases but essential in some; it requires proper regulation but is constantly changing. To understand the evolution of both satellite DNA and transposable elements, several questions need to be better answered:

1. What is the deleterious impact of satellite DNA when properly and improperly regulated?
2. What are the population dynamic of satellite DNA expansion and contraction?

3. Can repetitive sequences act as meiotic drivers?
4. How does the rest of the genome evolve in relation with repetitive DNAs.

Our inability to answer these questions is primarily due to technical failures. Historically, repetitive DNA has been difficult to analyze due low sequence complexity and high redundancy. Early identifications of satellites were accomplished using cesium density gradient separation which separated DNA based on G-C content (Peacock et al., 1974). However, such methods can only identify highly abundant satellites due to limits in resolution. Repetitive DNAs are also recalcitrant to most genetic manipulations as heterochromatin is genetically inert. As mentioned at the beginning, the field of Evolutionary Genomics has benefited tremendously from the advent of high-throughput technologies. However, the new technologies have not been particularly useful for studying repetitive sequences. The most popular and flexible high-throughput sequencing method is Illumina short-read sequencing, which generates hundreds of millions of small 50-150bp sequences, or reads. Much like pieces of a scrambled puzzle, the reads are then assembled together using various computational methods to reconstitute the sequences of the genome. The repetitive nature of repeats, unfortunately, excludes the possibility of assembly and unique alignments. While this issue plagues both transposable elements as well, it is particularly problematic for satellite DNA as they have very low sequence complexity.

In this dissertation, I will present studies that attempt to address these challenges and questions pertaining to the evolution repetitive sequence. The investigations will focus exclusively on *Drosophila melanogaster* (and related species, as it provides a tractable genome size, wealth of population genetic data, and powerful genetic tools. In Chapter 2, I will present a novel approach that identifies and quantifies satellite DNA from short-read whole genome sequences. This method will be applied to investigate the population dynamic of satellite DNA in

natural populations of Drosophila melanogaster. In Chapter 3, I will discuss further application of this method to exhaustively identify satellite DNA across different Drosophila species in order identify the rate of gains and losses of satellite DNA. Chapter 4 will layout an investigation of telomeric retrotransposons as meiotic drivers. In Chapter 5, I will discuss the function of a protein that has evolved in response to the rapid evolution of repetitive sequences. Finally, I will conclude and speculate on future directions that will further our understanding of repeat biology and evolution.

CHAPTER 2

CORRELATED VARIATION AND POPULATION DIFFERENTIATION IN SATELLITE DNA ABUNDANCE AMONG LINES OF *DROSOPHILA MELANOGASTER*

2.1 Introduction

Heterochromatin occupies a substantial portion of most eukaryotic genomes and contains vast quantities of tandemly-repeating, non-coding DNA elements known as satellite DNA. These sequences, along with transposable elements, are often described as selfish elements or genomic parasites, as they can increase their copy numbers irrespective of host fitness (Doolittle and Sapienza, 1980; Orgel and Crick, 1980). Indeed, they can be highly deleterious for the host genome; for example, ectopic recombination between homologous satellite repeats can lead to devastating chromosomal rearrangements (Bzymek and Lovett, 2001; Peng and Karpen, 2007). Consequently, these elements are mostly sequestered in repressive chromatin environments around the centromeres and telomeres where there is minimal recombination and transcriptional activity. Yet, paradoxically, repetitive sequences are also crucial components of euchromatic genomes, as they recruit the centromeric histone H3 variant to form centromeres in many species (Malik and Henikoff, 2001; Shelby et al., 1997), thereby affecting the fidelity of chromosome segregation (Henikoff et al., 2001; Karpen et al., 1996).

Adding to the perplexity, satellite DNA turns over at remarkably high rates between species (Kamm et al., 1995; Lohe and Roberts, 2000). In *Drosophila melanogaster*, satellite DNA is estimated to occupy over 20% of the genome. With the exception of the 359-bp (Lohe and Brutlag, 1986), Rsp (Wu et al., 1988), and dodeca (Abad et al., 1992) satellites, most known

satellites are tandem repeats of simple sequences (≤ 10 bp); the most abundant include AAGAG (aka GAGA-satellite), AACATAGAAT (aka 2L3L), and AATAT (Lohe and Brutlag, 1986; Lohe et al., 1993; Peacock et al., 1974). In comparison, the genome of its sister species *D. simulans*, from which *D. melanogaster* diverged ~2.5 mya, is estimated to have only 5% satellite DNA, more than 10-fold less AAGAG, and little to no AACATAGAAT (Lohe and Brutlag, 1987). For further contrast, nearly 50% of the *D. virilis* and less than 0.5% of *D. erecta* genomes are satellite DNA (Gall et al., 1971; Lohe and Brutlag, 1987). These drastic differences illustrate that the quantities and composition of satellite DNA are highly labile. Strikingly, such rapid changes in genomes have been implicated in post-zygotic isolation of species in the form of hybrid incompatibility in several species of flies (Bayes and Malik, 2009; Ferree and Barbash, 2009; Satyaki et al., 2014), demonstrating the critical role satellite DNA has on the evolution of genomes and species.

The expansions and contractions of satellite sequences are thought to result from a combination of molecular events such as unequal crossing over (Smith, 1976), rolling circle replication (Okumura et al., 1987), and polymerase slippage (Levinson and Gutman, 1987). Early population genetic studies assumed that satellite DNA has no function and that small changes in copy number are neutral, although total abundance may be under constraint due to the potential burden on metabolism, nuclear volume, and DNA replication (Charlesworth et al., 1994). Under such assumptions, early simulation studies demonstrated that unequal crossing over, drift, and reduced recombination are sufficient to generate long stretches of satellite DNA from random sequences (Smith, 1976; Stephan, 1986). Nevertheless, selection also appears to play an important role in shaping satellite DNA. For example, Stephan and Cho (1993) showed that selection is important in determining the length and heterogeneity of satellites, suggesting

that the drastic inter-specific differences may not be neutral (Stephan and Cho, 1994). Since then, multiple authors have emphasized the importance of both genetic drift and natural selection in the evolution of repetitive DNA (Hartl, 2000; Petrov, 2001). Furthermore, recent studies have shown that repetitive sequences can have remarkable effects on the rest of the genome. For example, natural variation in the Y chromosome, which is nearly entirely heterochromatic, can modulate differential gene expression and cause variable phenotypes including differences in immune response (Lemos et al., 2008, 2010). These results reveal that changes in repetitive sequences can have fitness consequences on which selection will act. Meiotic-drive models have also been proposed, in which centromeric satellites that bias the rate of transmission in female meiosis will quickly fix in the population (Henikoff et al., 2001; Malik, 2009), providing an additional mechanism for rapid turnover of satellite DNA.

However, technical challenges have hindered research on heterochromatin. Because heterochromatic regions do not recombine, common genetic manipulations are mostly ineffective. Repetitive sequences, particularly low complexity satellite DNA, present severe challenges for making sequence assemblies and unique alignments (Hoskins et al., 2007). A handful of techniques have been applied to study heterochromatin. High-density cesium-chloride gradient centrifugation has been instrumental in identifying major satellite blocks of different buoyancy, but fails to isolate less abundant repeats (Lohe and Brutlag, 1986). Hybridization approaches can only label known repeats and are often difficult to quantify precisely. More recently, flow cytometry has been used to indirectly estimate heterochromatic content, but it cannot distinguish the different types of satellites contributing to the observed total (Bosco et al., 2007).

To address these shortfalls, we developed a novel computational method, named k-Seek, that exhaustively identifies and quantifies short tandemly repeating sequences from whole-

genome sequences. We applied this method to 84 inbred *D. melanogaster* lines derived from natural populations, and characterized the natural variation in satellite DNA. This allowed us to answer three questions: 1) what are the abundances of all simple tandem repeat sequences in *D. melanogaster*; 2) how are their quantities changing within species and populations; and 3) how do they change with respect to each other?

2.2 Results

2.2.1 Identification and quantification of tandem repeats

We developed and validated a software package (k-Seek) that identifies and quantifies tandem repeats of 2- to 10mers from short-read-based whole-genome sequences (Figure 2.1A). In short, each raw read is first broken into small fragments of equal lengths. Identical fragments are then clustered. Whereas complex sequences are expected to yield clusters with very few members, short repetitive fragments will form a large cluster. Once the kmer is identified, the number of repeats from the read is then tallied based on a word-search procedure. To capture tandem counts, only kmers that are either immediately preceded or followed by the same kmer are scored. Additionally, we exclude tandem repeats that span less than 50 bp to avoid microsatellites, most of which are less than 30 bp in *D. melanogaster* (Fondon et al., 2012), and to guard against ascertainment bias for small kmers (2-4mers) as they are easier to identify from short stretches of DNA than larger kmers. Counts are summed across all reads and divided by the average read depth of the uniquely mapped autosomal genome, allowing us to estimate the abundance of every identified kmer in the genome Benchmarking with simulated reads reveals that k-Seek is highly accurate at identifying tandem repeats from 100 bp reads, and the counts are robust against point mutations and indels (Figures 2.1B).

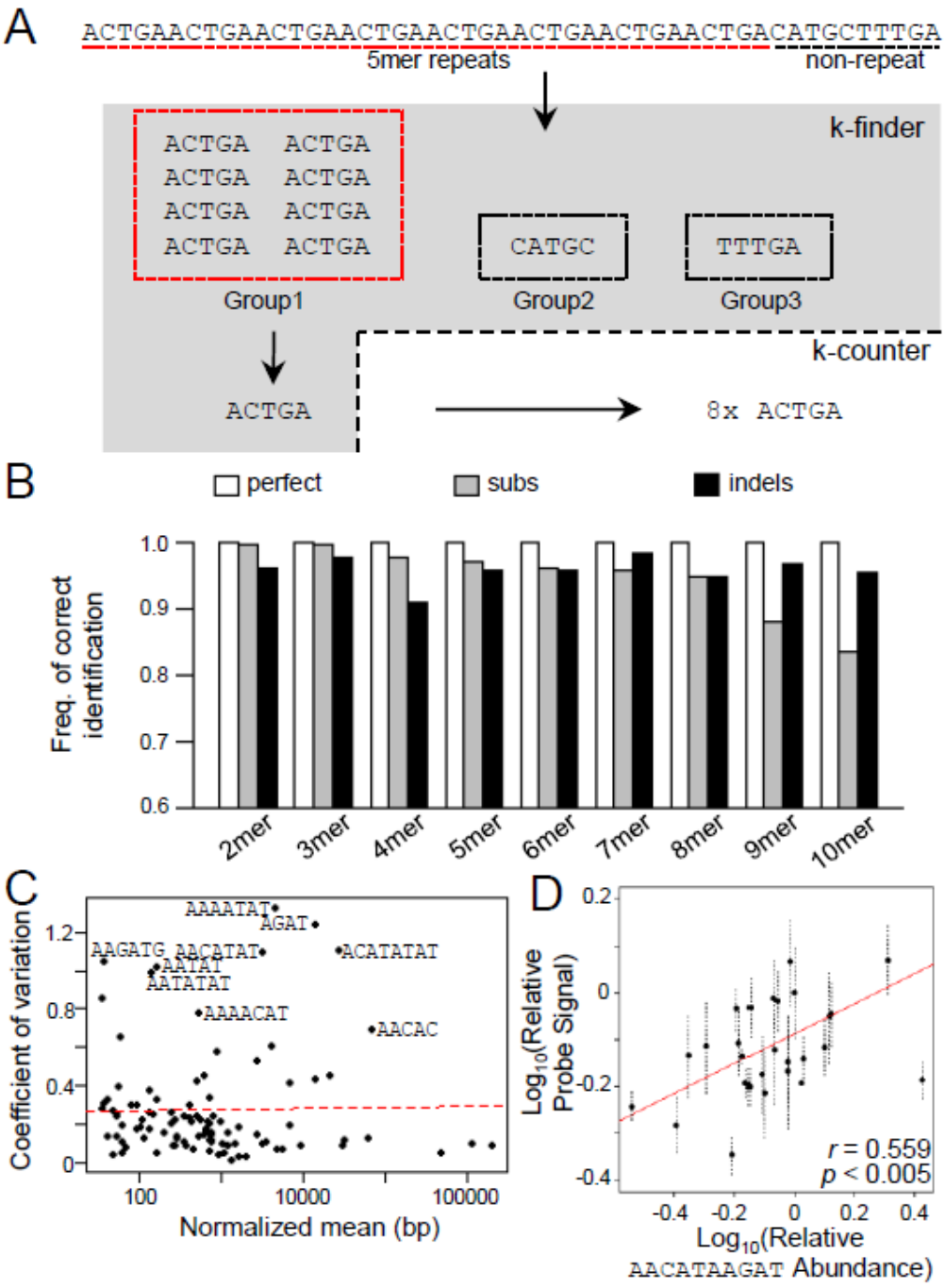

A
ACTGAACTGAACTGAACTGAACTGAACTGAACTGAACTGACATGCTTTGA
5mer repeats
non-repeat
k-finder
ACTGA ACTGA
ACTGA ACTGA
ACTGA ACTGA
ACTGA ACTGA
Group1
CATGC
Group2
TTTGA
Group3
k-counter
ACTGA
8x ACTGA
B
perfect
subs
indels
Freq. of correct identification
1.0
0.9
0.8
0.7
0.6
2mer
3mer
4mer
5mer
6mer
7mer
8mer
9mer
10mer
C
Coefficient of variation
1.2
0.8
0.4
0
AAAATAT
AGAT
AAGATG
AACATAT
ACATATAT
AATAT
AATATAT
AAAACAT
AACAC
100
10000
100000
Normalized mean (bp)
D
Log10(Relative Probe Signal)
0.2
0
-0.2
-0.4
r = 0.559
p < 0.005
-0.4
-0.2
0
0.2
0.4
Log10(Relative AACATAAGAT Abundance)

Figure 2.1. The k-Seek package identifies and quantifies tandem kmers. (A) k-finder identifies kmers de novo by fragmenting short reads and grouping them. k-counter then quantifies the number of tandem occurrences. (B) k-Seek applied to simulated 100-bp reads containing tandem arrays of kmers of different lengths. Simulated tandem arrays contained either perfect repeats, up to four substitutions (subs), or indels. Frequency of correct identification is plotted. (C) Variability of kmers between three independent library preparations is plotted against the kmer abundance. Some of the highly variable kmers are labeled. Dotted line depicts the line of best fit. (D) Dot blot with DNA from 27 lines hybridized with a probe targeting AACATAAGAT. Signal intensity is plotted against abundance inferred by k-Seek, both relative to a reference line, with regression line plotted in red. Error bars are SEs calculated from three replicates of each sample in the dot blot.

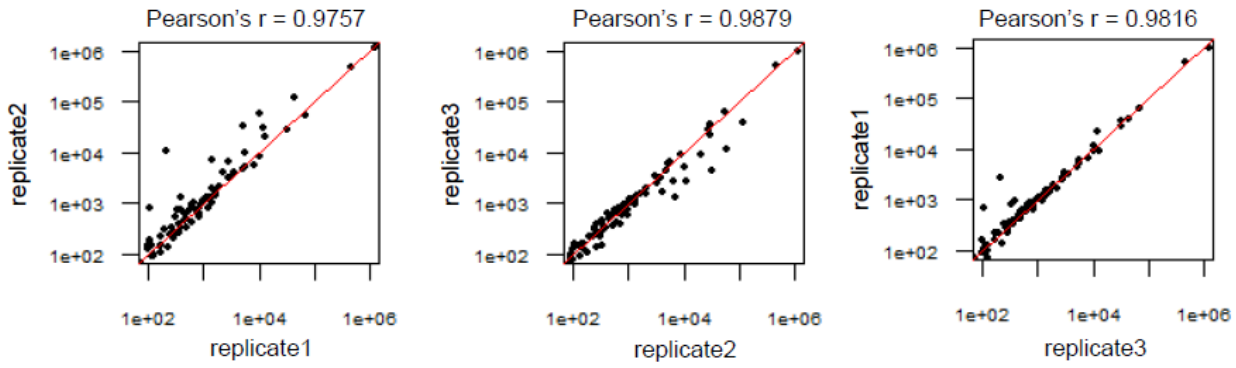


Figure 2.2. Pairwise correlation of ZW155 replicates. DNA extracted from ZW155 was used to generate three independent libraries and sequenced. The replicates were plotted against each other with respect to their kmer content.

To determine the reproducibility of k-Seek across library preparations, we applied it to three DNA libraries independently generated from line ZW155. The kmer quantities are highly correlated (Pearson's *r* ranges from 0.976 to 0.988, Figure 2.2). Furthermore, the variability is not influenced by the abundance of the repeat (Figure 2.1C). Nevertheless, several kmers have an elevated degree of variation. To independently assess the accuracy of k-Seek, we quantified the abundance of the 10mer AACATAGAAT by measuring the radioactivity of [32]P-labeled probes hybridized to blotted DNA from 27 lines. We found significant correlation between the methods (Figure 2.1D, Pearson's $r = 0.559$, $p < 0.005$). We also attempted to quantify the 5mer AAGAG using the same approach. However, this probe was problematic, and we were unable to obtain consistent results across replicates and experiments.

2.2.2 Identification of known and novel kmers

We applied k-Seek to a collection of 84 inbred *D. melanogaster* lines sampled from Beijing, Ithaca, Netherlands, Tasmania, and Zimbabwe, known as the Global Diversity Lines. Although there are 73,001 possible 2-10mers, we only identified 72 distinct kmers with population median abundance of more than 100 bp per 1x depth across all lines (Table 2.1). This list includes all previously identified kmers (Figure 2.3). As expected, AAGAG and AACATAGAAT have the largest quantities, as they are two of the most abundant satellites known in *D. melanogaster*. Curiously, we detected AATAT at substantially lower abundance than expected. This is likely due to under-amplification of sequences depleted of CGs during the PCR stage of library preparation (Aird et al., 2011). The most abundant kmer lengths were 5mers and 10mers, while only a single 4mer was found. Most repeats are A-rich with at least two A's in tandem. Most but not all (58/72) follow the $(RRN)_m(RN)_n$ formula where R represents a purine

Table 2.1. Counts for all possible kmers and identified kmers.

	Possible kmers	Median > 100bp	Top 100	All
2mer	4	3	3	3
3mer	10	5	6	9
4mer	33	1	1	18
5mer	102	17	21	59
6mer	350	12	14	113
7mer	1170	7	13	115
8mer	4140	4	4	77
9mer	14560	6	8	72
10mer	52632	17	30	202
total	73001	72	100	668

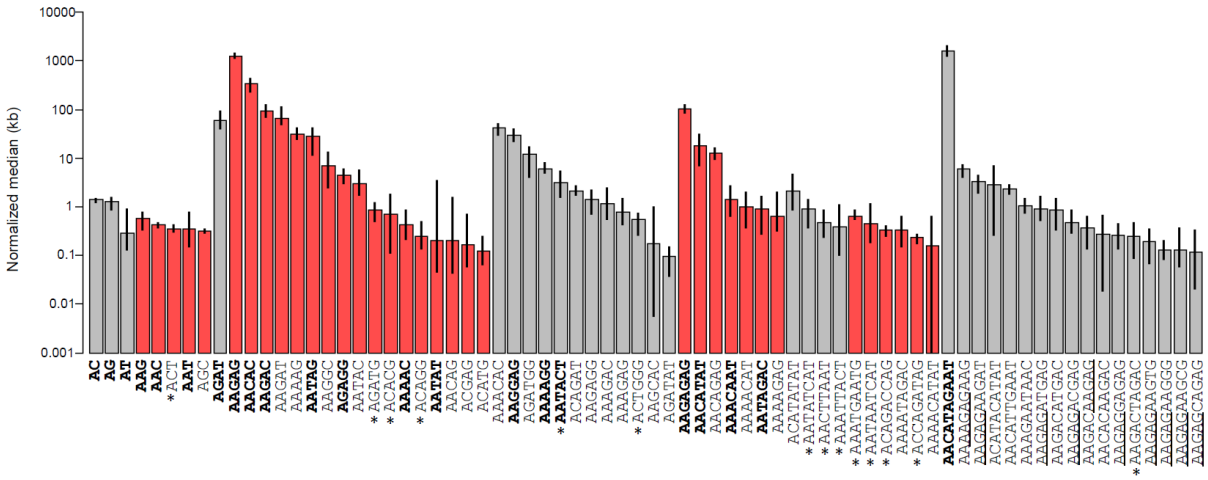


Figure 2.3. kmer abundance. Medians of the top kmers across all strains are plotted in log10 scale. Error bars represent the first and third quartiles. Gray and red bars are even- and odd-number kmers, respectively. Previously characterized kmers are labeled in bold. kmers not following (RRN)m(RN)n are labeled by an asterisk. AAGAG-containing 10mers are in red.

and N represents any nucleotide, thought to be canonical for *D. melanogaster* satellites (Lohe and Brutlag, 1986).

Among these 72 kmers, 50 were previously unknown. Most are 5mers and 10mers ranging from normalized mean abundance of 105 bp (AAGAGCAGAG) to 66,564 bp (AAGAT) across lines. 11 of the 16 new 10mers contain AAGAG, suggesting that they originated from a mutation in one copy of AAGAG, followed by amplification of it and its non-mutated neighbor to generate a new 10mer. Additionally, we find four 8mers and six 9mers, lengths that had not previously been identified. Notably, most of the 8-9mers (7/10) do not follow the $(RRN)_m(RN)_n$ formula and may therefore represent a qualitatively distinct group of satellite sequences.

2.2.3 Population structure

Across all Global Diversity lines, the average total kmer count is 4.03 Mb per 1x read depth. Strikingly, the lowest and highest lines differ by 2.50-fold (equating to 4.29 Mb difference), indicating high intraspecific variability (Figure 2.4). AAGAG and AACATAGAAT, the two most abundant kmers, comprise 74% of the total kmer counts on average but can be as high as 88% and as low as 57%, further revealing marked differences in the repeat composition among the lines.

The phylogenetic relationships among the lines were inferred from the genome-wide SNP calls, and that analysis largely recapitulates the expected demographic history of *D. melanogaster* (Grenier et al. 2014), with an African origin and a relatively recent global spreading along human trade routes (for review see (Stephan and Li, 2006)). The simple expectation is that kmer abundance will also reflect the same population structure. However, hierarchical clustering of kmer abundances failed to differentiate the lines into their respective

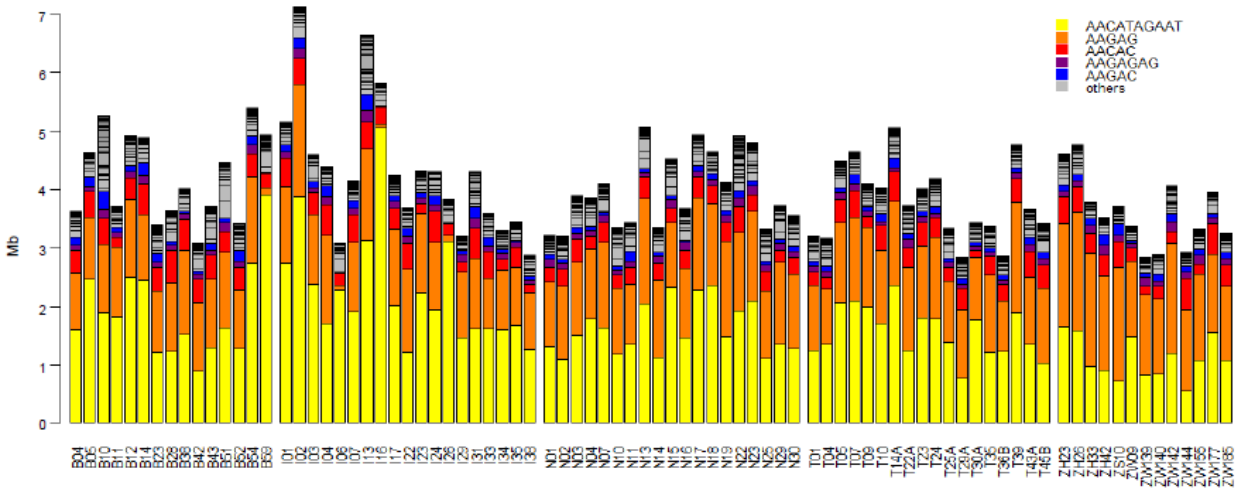


Figure 2.4. Distribution of all kmers across lines. For each line in Global Diversity Lines, the kmer abundance is plotted in the cumulative barplot. The 5 kmers contributing the most to the total are differentiated by colors, and the rest are in gray.

populations (Figure 2.5). To further investigate, we applied principal components analysis on the top 100 kmers (Figure 2.6A). The Zimbabwe lines fall into a diffuse cluster with minimal overlap with other populations, as expected. The Netherlands lines broadly cluster with the Tasmanian lines, consistent with the introduction of *D. melanogaster* to Australia by European settlers (David and Capy, 1988). Surprisingly though, the Beijing lines largely overlap with the Ithaca lines, even though North American populations are thought to be of European origin (Begun and Aquadro, 1993; David and Capy, 1988) and distinct from Asian populations established shortly after the initial out-of-Africa migration (Schlötterer et al., 2006). These discrepancies suggest that satellite DNA abundance is subject to a distinct evolutionary history from the rest of the genome.

To infer the population differences for each kmer, we applied *Rst* statistics, assuming a step-wise mutational model (Figure 2.6B) (Hardy et al., 2003; Slatkin, 1995). Of the top 100 kmers, the majority (n = 54) display very little population differentiation, many (n = 46) have *Rst* of >0.1, showing appreciable population differentiation, and some (n = 7) have *Rst* of >0.4, revealing high population differences. For example, the AT 2mer has a startlingly high *Rst* of 0.554, which appears to be due to elevated levels in the Netherlands and Tasmania populations (Figure 2.6C). AAGAG and AACATAGAAT, the most abundant kmers, have moderate levels of differentiation, with the Zimbabwe population having highest and lowest abundance, respectively (Figures 2.6D and E). Because different kmers have distinct patterns of population differentiation, we conclude that they experience different evolutionary dynamics.

Among the most differentiated kmers, two 10mers (AACATATAAT and AAAATAGAAT) are surprisingly found only in the Netherlands, Tasmania and Zimbabwe populations, while being completely absent in Beijing and Ithaca populations (Figure 2.6F).

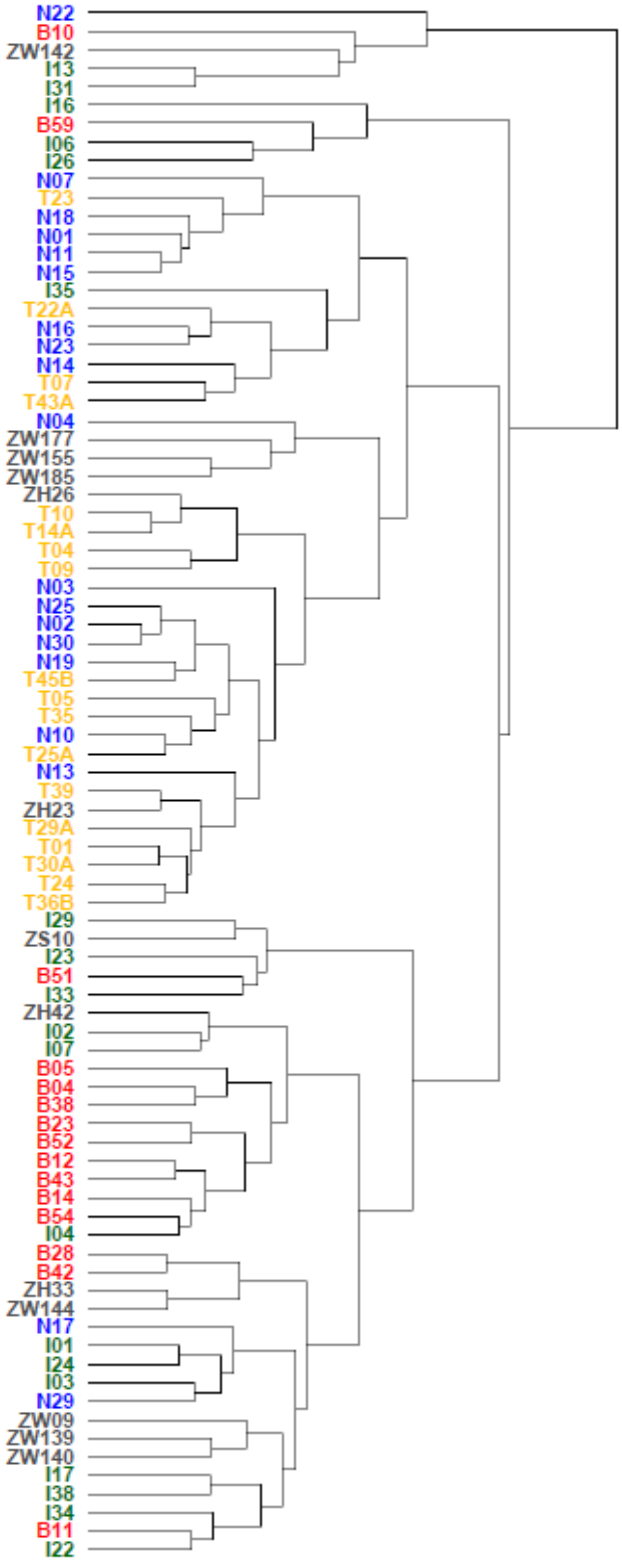

N22
B10
ZW142
I13
I31
I16
B59
I06
I26
N07
T23
N18
N01
N11
N15
I35
T22A
N16
N23
N14
T07
T43A
N04
ZW177
ZW155
ZW185
ZH26
T10
T14A
T04
T09
N03
N25
N02
N30
N19
T45B
T05
T35
N10
T25A
N13
T39
ZH23
T29A
T01
T30A
T24
T36B
I29
ZS10
I23
B51
I33
ZH42
I02
I07
B05
B04
B38
B23
B52
B12
B43
B14
B54
I04
B28
B42
ZH33
ZW144
N17
I01
I24
I03
N29
ZW09
ZW139
ZW140
I17
I38
I34
B11
I22

Figure 2.5. Hierarchical clustering of the Global Diversity Lines based on kmer quantities. Lines are clustered based on euclidean distance. Euclidean distances between lines were calculated based on log10 kmer quantities relative to the mean. The top 100 kmers were used. The different populations are color coded: Beijing (red), Ithaca (green), Netherlands (blue), Tasmania (yellow), and Zimbabwe (gray).

Within Tasmania and Netherlands populations, there is high variation even among individual lines, which suggests that these 10mers have high turnover rates. To confirm the presence/absence polymorphism we designed FISH probes targeting AACATATAAT and observed fluorescent foci from mitotic chromosomes of Netherland and Tasmania but not Beijing lines (Figures 2.6G and 2.7). The foci are autosomal and appear to be centromeric as they are located near the primary constriction and do not overlap with the predominantly pericentric AAGAG foci. This finding provides the first report of population-specific satellite DNA in Drosophila and further underscores the high rate of satellite DNA turnover.

2.2.4 Concerted evolution of kmer abundance

Interestingly, the two population-specific 10mers are highly positively correlated across lines (Figure 2.6F; Pearson's $r = 0.993$, $p < 2.2 \times 10^{-16}$), suggesting that they undergo coordinated changes in copy number. To comprehensively identify kmers that are evolving in a concerted fashion, we generated a pairwise correlation matrix for the top 100 kmers, and clustered those that are highly correlated (Figure 2.8A). This was accomplished using Modulated Modularity Clustering which rearranges rows and columns of the correlation matrix to identify clusters of variables with maximal pairwise correlations among all cluster members (in this case, kmers) without pre-determined knowledge or an arbitrary decision on number of clusters (Stone and Ayroles, 2009). Overall, we find 9 major clusters of correlated kmers. The number of kmers within each cluster ranges from 2 to 21. As expected the two population-specific 10mers are found within the same cluster; in addition, this cluster contains the AATAT 5mer. Clustering appears to be driven in part by sequence similarity; several clusters are either AT-rich, AG-rich, or AC-rich. For example, the AG-rich cluster contains AAGAG, as well as related sequences

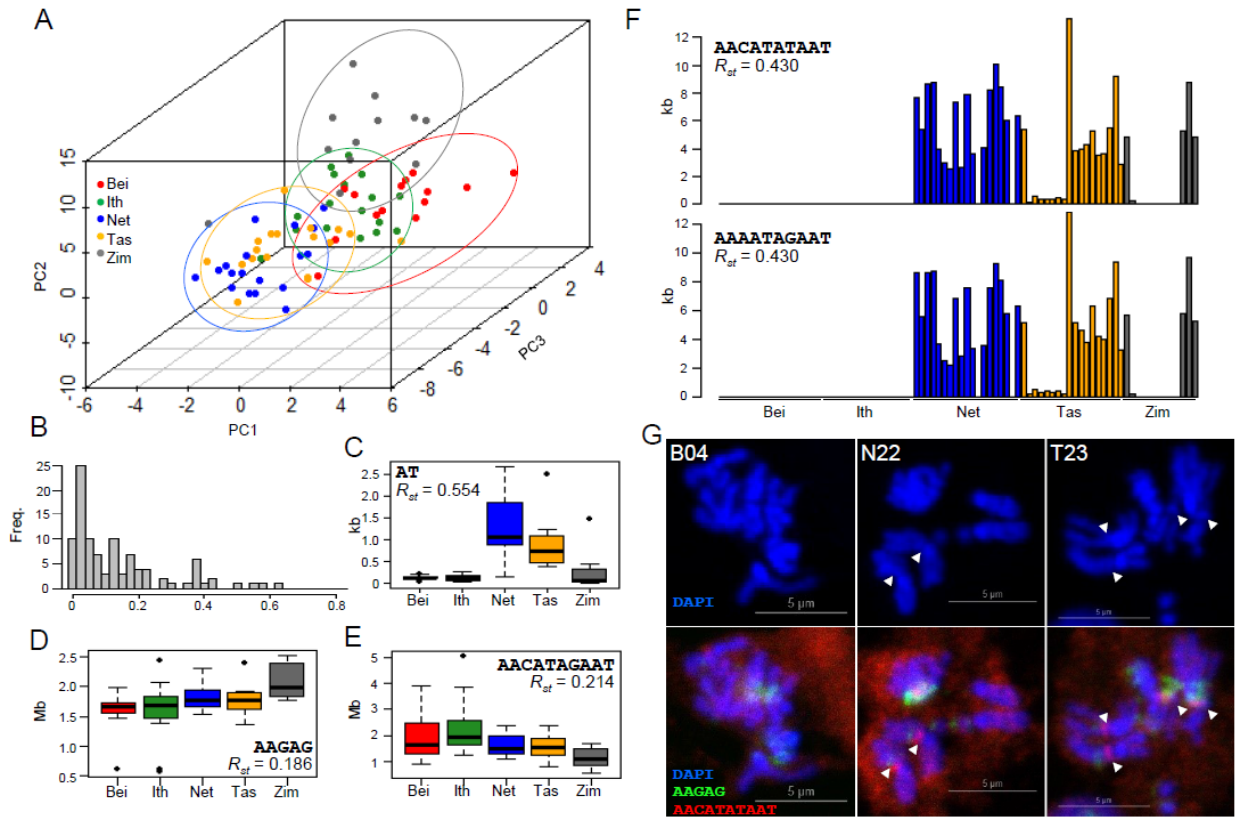


Figure 2.6. Population structure of kmers. (A) Lines are plotted based on the first three principal components derived from the top 100 kmers. Lines from the same populations are circled with the respective colors. (B) Distribution of population differentiation index RST. (C–E) Distribution of abundance of selected kmers in the five populations. (F) Abundance of AACATATAAT and AAAATAGAAT across lines. (G) FISH applied to mitotic chromosomes of lines from Beijing and Tasmania. Probes for AAGAG are labeled green, AACATATAAT red, and DAPI blue. Arrowheads indicate red foci.

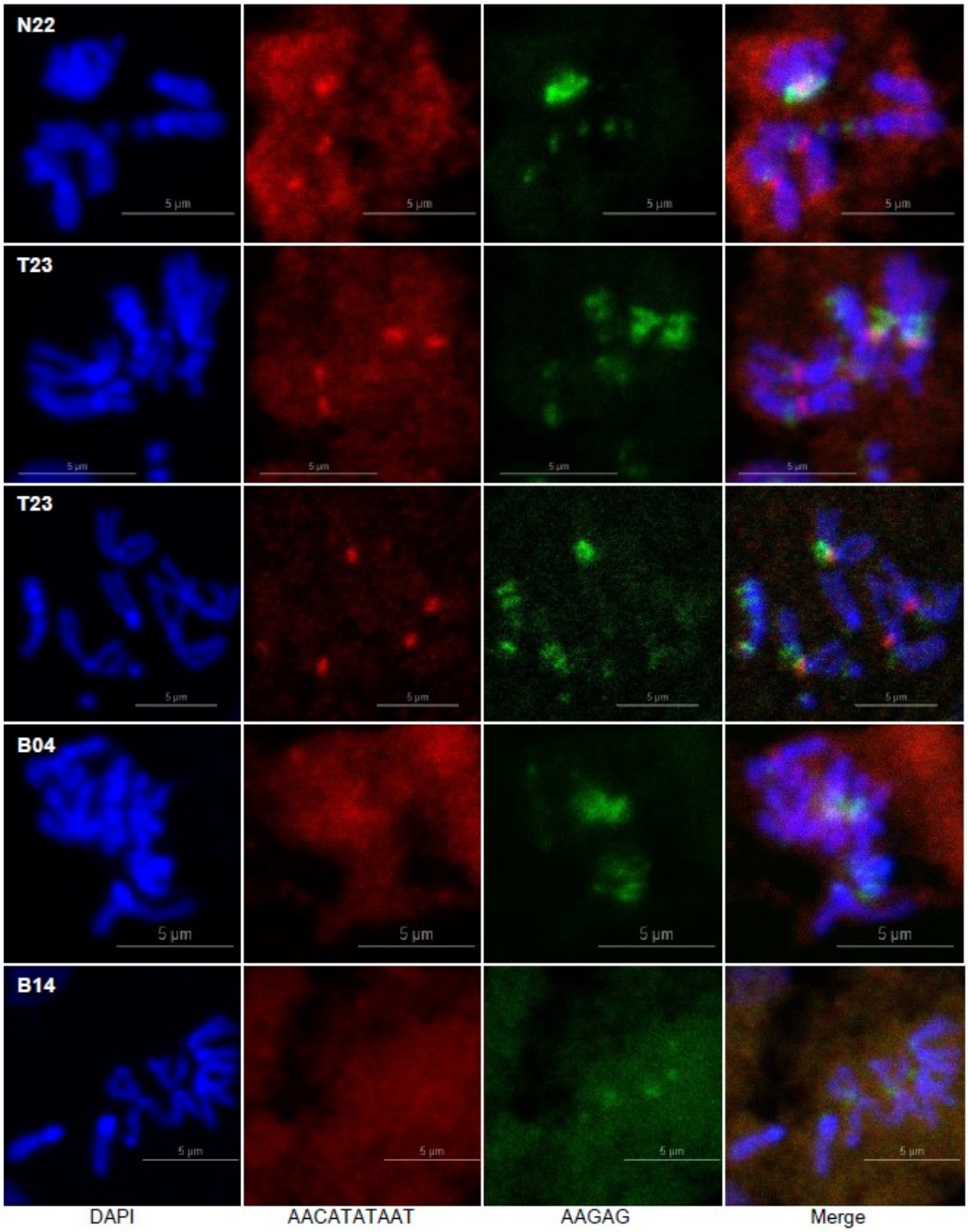


Figure 2.7. Fluorescent in situ hybridization for population specific kmer. Probes targeting 5mer AAGAG and population-specific 10mer AACATATAAT are green and red, respectively. Mitotic chromosomes are labeled by DAPI in blue.

AAGAGAG and AAGAGAGAG (Figure 2.8.D). However, we note that many highly related kmers fall into separate clusters, for example, AACATAGAAT (the most abundant 10mer) and AACATATAAT (one of the two population-specific kmers), even though they only differ by one nucleotide. Surprisingly, we also observe relatively weak but significant negative correlations among a small number of kmers (Figures 2.8A and Figure 2.9). Notably, the AG-rich kmers are negatively correlated with the AT-rich kmers; not only are the two respective clusters anti-correlated (Figure 2.8A, arrowhead), the 10mer AAGAGCAGAG that is grouped within the AT-rich cluster is also negatively correlated with all other AT-rich kmers (Figure 2.8B). These negative correlations suggest that different kmers can have antagonistic relationships such that expansion of one comes at the expense of another.

2.2.5 Interspersion of kmer blocks drives correlation

One possible cause of the observed positive correlations is that correlated kmers represent physically linked and interspersed satellite blocks. Therefore, deletions or duplications of these repetitive blocks will decrease and increase both kmers in concert. To test this possibility, we identified, across all lines, paired-end reads where kmers are found in both mate-pairs and determined the frequency of their occurrences relative to the abundance of the identified kmers (Figure 2.10A). As expected, almost every kmer is most frequently paired with itself, reflecting that many of them comprise sizeable and homogenous blocks. However, kmers found within positively correlated clusters tend to be found in mate pairs more frequently than those outside (Figure 2.10B-D), consistent with our hypothesis. For example, kmers within the AG-rich cluster are highly interspersed with one another. This is further supported by a significant and positive

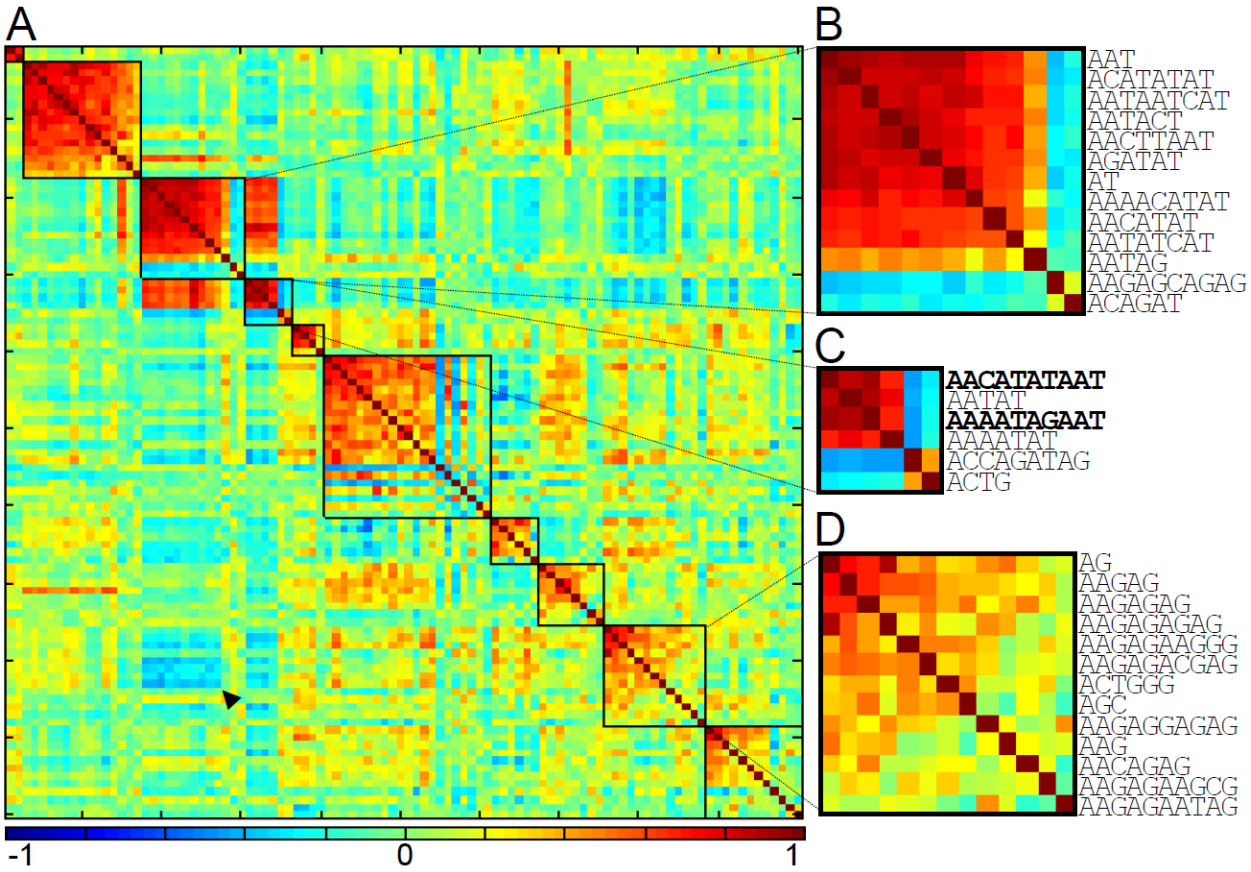


Figure 2.8. Correlation structure of kmer variation. (A) Pair-wise correlation matrix is reorganized such that kmers with correlated change across lines are clustered into groups demarcated by boxes. Colors represent strength of Spearman's correlation. Arrowhead indicates clusters that are negatively correlated. (B–D) Magnified clusters with AT-rich, the two population-specific (bold), and AG-rich kmers, respectively.

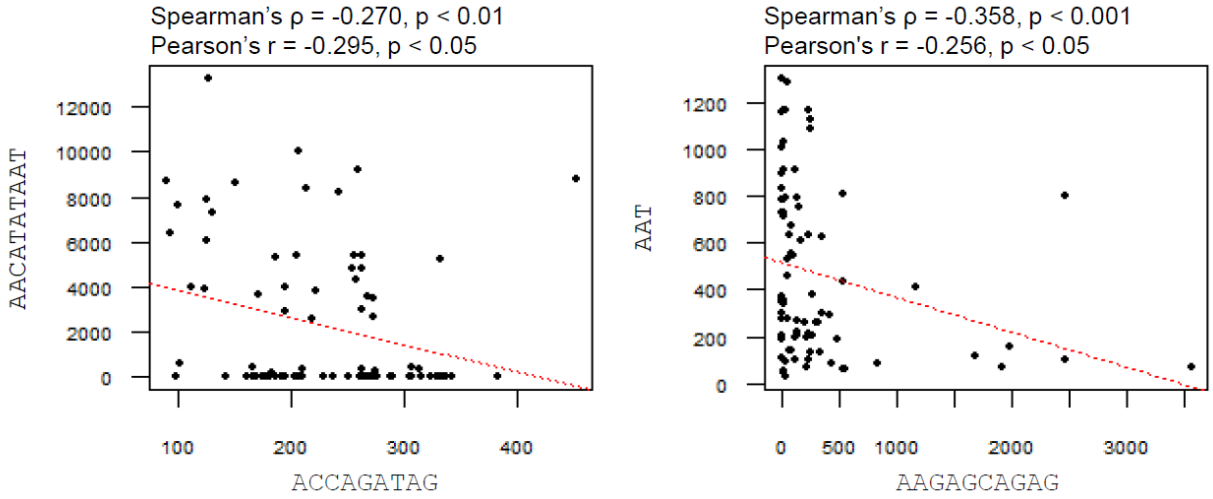


Figure 2.9. Anti-correlated kmers. Spearman's ρ is calculated with all samples. Pearson's r is calculated after removing samples that have kmer quantities lower than 1% of the average. Red line depicts regression line using all samples.

association between the correlation values and the interspersion frequency of the kmers (Figure 2.10E).

The two population-specific 10mers identified are also highly interspersed (Figure 2.10C). Interestingly, they are most frequently paired with each other, such that mate pairs containing the same 10mer are rarely found. This result suggests that the two 10mers are interspersed with each other in small blocks that are roughly the length of the insert size, which is ~450bp (Figure 2.10A). This is also true for the AAC 3mer, and the AAAATAACAT 10mer, suggesting that they also exist in small interdigitated blocks.

We note that there are many instances where interspersed kmers are not correlated. This is unsurprising since interspersion itself is insufficient to drive correlated change if the blocks do not experience duplication and/or deletion. Additionally, for kmers that are found interspersed with many other kmers, presumably in separate blocks, independent indels in different blocks will result in local concerted change in abundance, but their abundances aggregated across the genome will likely be uncorrelated. Of further interest are correlated kmers that are not interspersed, such as the two population-specific 10mers and the 5mer AATAT, as they indicate additional mechanisms underlying the concerted change. However, it is difficult to distinguish these from interspersion that we fail to capture due to low coverage or under-representation. Furthermore, any junction between satellite blocks that is gapped by complex sequences, such as transposable elements, will also likely be missed.

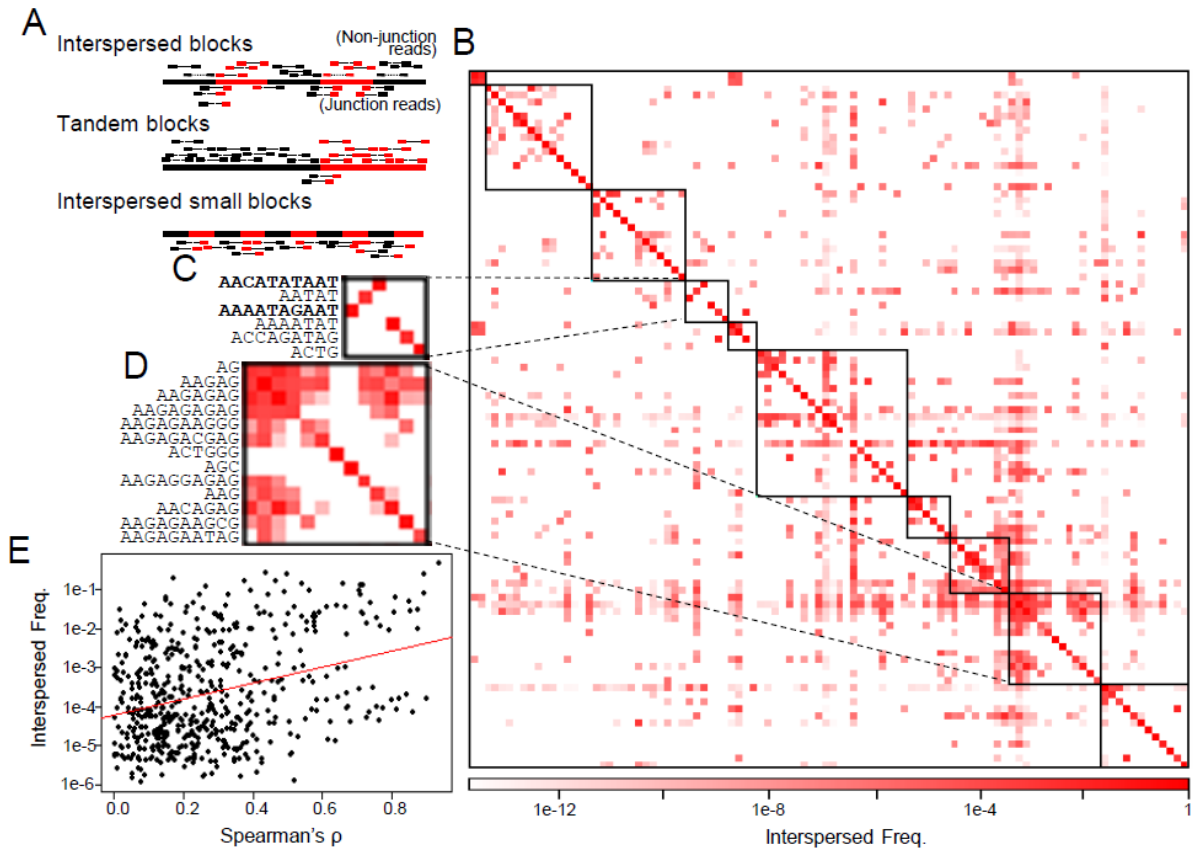


Figure 2.10. Interspersed kmer blocks. (A) Paired-end reads are used to infer interspersion. Interspersed kmers will have many mate pairs spanning the junctions (Left). Large kmer blocks are in tandem and will have few mate pairs containing different kmers (Middle). Interspersion of small blocks (~450 bp) will yield only mate-pair gapping junctions (Right). (B) Interspersion frequency matrix for kmers organized as in Figure 4A. (C and D) Magnified cluster containing the two population-specific 10mers (in bold) and AG-rich cluster, respectively. (E) The correlation strength for each kmer pair (from Figure 4A) is plotted against the interspersed frequency (from B). Self-correlation and interspersion (cells across the diagonal) and pairs with no interspersion are excluded. Regression line is plotted in red (Pearson's r = 0.325, P = 9.77 x 10−15).

2.3 Discussion

Many important questions in heterochromatin biology are now accessible using our software pipeline (k-Seek) to identify and quantify short tandemly-repeated satellite sequences from short-read whole-genome shotgun sequences. Previously, identification of satellite sequences was mostly accomplished through labor-intensive methods that have low sensitivity (Lohe and Brutlag, 1987). As a result, the current catalog of satellite DNA contains exclusively kmers that are present in large quantities. Our method is accurate at identifying tandem kmers and discovered many previously unknown kmers of low to medium abundance. We expected that PCR would be major source of bias during library preparation as the polymerase under-amplifies AT-rich sequences, and we indeed found lower abundance of AATAT compared to previous characterizations (Lohe and Brutlag, 1986). Nevertheless, using three replicate libraries made from a single sample, we found such bias to be consistent across the libraries, allowing us to characterize population variation of individual kmers.

2.3.1 Potential causes of population variation

By applying k-Seek to the Drosophila Global Diversity lines, we characterized natural variation in heterochromatin repeat structure. The mean satellite abundance in a population is expected to approximate an equilibrium determined by the mutation rate, the degree of selective constraint, potentially positive selection, and population size (Charlesworth et al., 1994; Stephan, 1986). For many kmers, the difference between populations is small, and the low *Rst* suggests a high rate of migration or turnover. Nevertheless, we identified multiple kmers with appreciable to high population differentiation. Notably, the pattern of inter-population differentiation is also variable among repeats, revealing that some kmers evolve relatively independently of others. The

process driving the population differences could be either neutral drift or natural selection. According to the out-of-Africa model, the Zimbabwe population is expected to have the highest level of genetic variation, provided that the differences are nearly neutral. While this is, as expected, true for all kmers considered together (Figure 2.11), we found many exceptions which may be revealing of their modes of evolution. For example, the Netherlands population not only has a significantly higher abundance of the AT 2mer compared to Zimbabwe, but the between-line variability is also substantially greater. These differences may be indicative of a relaxation of constraint within the Netherlands population, allowing for labile expansion and contraction. In contrast, the differentiation pattern of AAGAG shows significant reduction in the non-African populations, potentially reflecting an increase in the level of selective constraint after the out-of-Africa migration, or reduced variability due to the out-of-Africa bottleneck as is seen for most of the genome (Pool et al., 2012).

The incongruity between the population structure inferred from kmer abundance and from demographic history and SNPs is intriguing. This is reminiscent of the well documented phenomenon of the homogenization of multi-copy gene families and tandemly arrayed genes such as rDNA, a process that has been called “molecular drive” (Dover, 1982). Resulting from sequence exchanges via gene conversion, paralogs that predate the species split may display a high degree of within-species sequence homogeneity. Depending on the stochastic or potentially biased dynamics of the process that results in homogenization of the repeated arrays, phylogenetic relationships between species and by extension populations may not be preserved in these sequences.

Alternatively, the discrepancy may be due to incomplete lineage sorting (Pollard et al., 2006) of some repeats. In one possible scenario, individuals without the correlated AT-rich

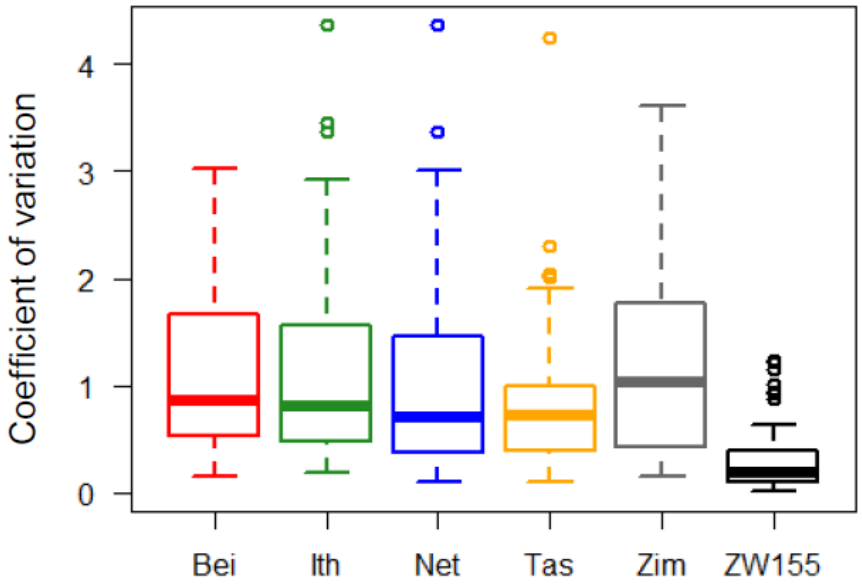


Figure 2.11. Distribution of coefficient of variation. The coefficient of variation for the top 100 kmers are plotted. The Zimbabwe population has the highest average coefficient of variation ($p = 0.0244$, Kruskal Wallis test.)

repeats segregated at low frequency in the European population and were subsequently introduced by chance only to North American but not to Australia. Additionally, we cannot rule out the possibility that the similarity between Beijing and Ithaca lines is due to selection from a common environmental pressure.

Meiotic drive and segregation distortion present an additional explanation of strong population-specific patterns. In these scenarios chromosomes with a particular kmer abundance or composition can have a segregation advantage in some populations. If these repetitive sequences have pleiotropic deleterious effects, fixation of a suppressor can quickly purge these sequences from populations. Our discovery of the two population-specific 10mers (Figure 3G) is particularly striking and suggestive of the rapid evolution predicted by most models of meiotic drive. Notably, these lines provide genetic material for direct empirical tests of possible segregation differences (now underway).

2.3.2 Concerted change of *k*mer abundances

Our characterization of the kmers across different lines enabled investigation of the dynamics of their changes in abundance. We identified several groups of repeats that are highly correlated, revealing that different satellites undergo concerted evolution in abundance. The fact that kmers within a correlated group have sequence similarity is suggestive of the potential underlying mechanism. Several DNA binding proteins have been identified to target satellite DNA (for review see Csink & Henikoff 1998). GAGA-factor is a transcription factor responsible for key developmental regulation (Biggin and Tjian, 1988), heat-shock response, and chromatin remodeling (Kerrigan et al., 1991; Lu et al., 1993; Tsukiyama et al., 1994) that localizes to AAGAG and AAGAGAG satellites during mitosis (Platero et al., 1998; Raff et al., 1994). Here,

we have shown that these two kmers are correlated along with other AG-rich kmers, thus raising the possibility that the concerted change is driven by binding to a common protein. In one scenario, an increase in GAGA-factor more effectively packages AG-rich repeats into heterochromatin, thereby raising host tolerance to the repeats. Subsequently, both repeats will increase in number. Conversely, a decrease in the protein level may result in sub-optimal regulation of repeats and reduction in organismal fitness; therefore selection will favor individuals with less of the targeted repeats, resulting in correlated contraction. Similarly, the concerted change between AT-rich kmers may reflect proteins that recognize AT-rich satellites, including ORC2, an essential component of the origin recognition complex that initiates DNA replication (Pak et al., 1997), and D1, an essential protein implicated in chromatin remodeling (Rodriguez Alfageme et al., 1980). Additionally, the PROD protein (proliferation disrupter) binds to the AACATAGAAT 10mer on mitotic chromosomes (Török et al., 1997). Our results suggest that it may also bind to other satellites as several kmers are correlated with this 10mer.

However, the observed concerted changes are not necessarily driven by selection. We demonstrate that the structure of kmers can also account for some of the observed correlated patterns. For satellite blocks that are interspersed with each other, a block duplication or deletion of the region will increase or decrease the kmers together. Indeed, kmers that are highly interspersed tend to have higher correlation in abundance. Notably, the two population-specific 10mers are highly interspersed in small tandem blocks, which explains their striking degree of correlation. Furthermore, AAGAG and AAGAGAG are also moderately interspersed, suggesting that GAGA-factor binding may not be the only mechanism driving their correlation. We therefore conclude that the complex architecture of satellite DNA is likely the result of both neutral and selected mutational changes.

Surprisingly, we also identified negative correlations, albeit weak ones, both between individual kmers and between groups of correlated kmers, most notably between the AT- and AG-rich clusters. This is suggestive of antagonistic relationships such that some kmers increase at the expense of others. We speculate that this reflects an optimal load of satellite DNA that the genome can tolerate. Therefore, the deleterious effect of an increase in one satellite DNA group can be alleviated by a decrease in a different group. This load may be determined by the fitness benefits of maintaining an optimal genome size (Charlesworth et al., 1994), but we find this unlikely given the variability of genome sizes among as well as within species (Ellis et al., 2014) and that total kmer quantities differ greatly between lines. Alternatively, the load may be chromosome-specific, as satellite DNAs often have chromosome-specific distributions. An optimal load of satellite DNA may ensure faithful chromosomal transmission or prevent deleterious rearrangements, as lengthy satellite blocks may be more prone to unequal crossing over or ectopic recombination. Regardless of the specific molecular and evolutionary mechanism, the observed antagonistic relationships intimate the curious possibility that satellite DNAs are not only at odds with the host genome, but also with each other.

2.4 Materials and methods

2.4.1 Drosophila lines and sequence reads

The 84 lines of Drosophila melanogaster used in this study were sib-mated from isofemales lines from Beijing, Ithaca, Netherlands, Tasmania, and Zimbabwe (Greenberg et al., 2011). Details of the whole-genome shotgun sequencing using the Illumina platform can be found in Grenier et al. 2015.

2.4.2 Processing and normalizing tandem kmer counts

kmer identification is described in Figure 1A and supplementary materials. To normalize kmer counts between lines, we divided all counts by the average read-depth at autosomal regions and then multiplied the counts by the kmer length, to obtain number of nucleotides per 1x depth. The average read-depth was obtained by mapping all sequences with BWA on standard settings to the *D. melanogaster* reference r.546 (Flybase), followed by sorting with Samtools (Li et al., 2009). We then used Picard tools to compute the distribution of read-depth and averaged across the autosomes. We note that very few reads map to the Y chromosome (genomic sequences were from females), indicating very little contribution from any sperm present in the females' reproductive tracts.

2.4.3 Simulation of reads with tandem repeats

For each kmer length (k=2 to 10), we generated 600,000 100 bp reads. Each read contained a random number of tandem occurrences for a randomly generated kmer. One-third of the reads contained perfect tandem repeats, one-third contained 1 to 4 point mutations in the tandem repeats, and one-third contained an indel of varying size within the tandem repeats. k-Seek was applied to the simulated reads and correct identification for tandem repeats greater than 50 bp was recorded. 200,000 100-bp reads were also generated containing random sequences. No kmers were identified from these.

2.4.4 Quantifying satellites with dot blots

50 pmol of AACATAAGATAACATAAGATAACATAAGAT (Sigma) was radiolabeled with [γ-32]ATP using T4 polynucleotide kinase (NEB), followed by clean-up with Micro Bio-Spin

P30 Column (Bio-Rad) to remove unincorporated labels. Probe was denatured at 95°C for 10 min and immediately put on ice. We extracted DNA of 50 females from 27 lines with the DNeasy kit (Qiagen). Using a Bio-Dot Microfiltration Apparatus (Bio-Rad), we loaded 100 ng of each sample in 0.4 M NaOH and 10 mM EDTA in triplicate onto a Zeta-Probe GT membrane (Bio-Rad) following the manufacturer's instructions, in addition to a 3-fold serial dilution of DNA from line B10. The placement of samples was randomized. After drying in an oven at 80°C for 30 min, the membrane was incubated in 25 ml of hybridization buffer (0.5M sodium phosphate, 7% SDS) with 100 µl of denatured salmon sperm DNA (10 mg/ml) for 30 min at 60°C in a rotating oven. The buffer was then replaced with 25 ml of fresh hybridization buffer containing the denatured probe and incubated overnight. The membrane was washed at 68°C twice with 50 ml 1X SSC and 0.1% SDS followed by two washes with 0.1X SSC and 0.1% SDS. The membrane was wrapped with plastic wrap, placed into a phosphoimager for 48 hrs and scanned with a Typhoon 9400. Signal intensity was processed in ImageJ with background subtraction (Schneider et al., 2012). The intensity of each sample was calculated according to the standard curve constructed from the dilution series.

2.4.5 In situ hybridization

Brains from wandering third instar larvae were dissected and washed in 0.7% NaCl, transferred to 0.5% sodium citrate for 10 min, followed by fixation in 20 ul of 50% acetic acid and 4% paraformaldehyde for 2 min on a siliconized coverslip. The samples were then squashed onto a glass slide and flash frozen in liquid nitrogen. Slides were then immersed in 100% EtOH for 10 min and air-dried in room temp for 2-3 days with the coverslip removed. Hybridization procedure was conducted as in (Dernburg, 2011). 250 ng of

AAGAGAAGAGAAGAGAAGAGAAGAG-Cy3 and AAAATAGAATAAAATAGAATAAAATAGAAT-Cy5 probes (ordered from Sigma) were used for probe mixture. The samples were imaged on Zeiss confocal microscope, and images were processed on Zen software.

2.4.6 kmer correlation and interspersion matrix

We applied the publicly available software Modulated Modularity Clustering (Stone and Ayroles, 2009) on the normalized counts of the top 100 repeats to generate the clustered correlation matrix. Using custom Perl scripts on .sep outputs from k_counter.pl, we identified the number of mate pairs where both reads contain kmers and tallied across all lines to obtain n_{ij}, the number of mate pairs containing kmer i and kmer j. Interspersed frequency is calculated as: $n_{ij}/\sqrt{n_i n_j}$ where and n_i and n_j are number of pairs where at least one of the reads contains kmer i and kmer j, respectively.

CHAPTER 3

RAPID ACCUMULATIOIN OF SATELLITE DNA IS LINEAGE SPECIFIC IN *DROSOPHILA*

3.1 Introduction

The Drosophila genus has long been a model for evolutionary studies. Given its well-defined phylogeny with a plethora of species that have richly characterized behavioral, anatomical, developmental, and genetic differences, the genus and species within it represent an unprecedented resource to investigate and understand the evolution of genomes and phenotypes. Unfortunately, satellite DNAs have received relatively little attention, even though they constitute large portions of eukaryotic genomes. Efforts to define satellite DNA have primarily focused on *Drosophila melanogaster* and closely related species, including *D. simulans* and *D. erecta* (Lohe and Brutlag, 1986, 1987; Lohe and Roberts, 2000). Even though the rapid rate at which satellites change in abundance was readily apparent from these three species, the dearth of knowledge regarding satellite composition in distant species has stymied our understanding of satellite evolution. The release of the 12 Drosophila species genomes (Clark et al., 2007) has shed little light on these low complexity sequences as they are severely under-represented in the genome assemblies. This is primarily due to the fact that plasmids, fosmids, and BAC libraries used for shotgun sequencing of the 12 species tend to be biased against maintenance of highly repetitive DNA. With recent advances in sequencing technologies and our development of the k-Seek pipeline, exhaustive and high-throughput characterization of satellite DNAs is now possible (Wei et al., 2014).

Changes in satellite DNA are inextricably linked with the evolution of several genomic features. As repetitive sequences which include both satellites and transposable elements often represent the bulk of eukaryotic genomes, the divergence of genome size is predominantly driven by changes in repeat content (Gregory, 2001). The rapid rate of repeat turnover has led to the observation that genome size correlates poorly with organismal complexity. This discrepancy has led to the famous "C-value paradox" (Cavalier-Smith, 1978) and the popularity of the term "junk DNA" to describe the repetitive sequences that have no functional relevance in the genome (Ohno, 1972; Palazzo and Gregory, 2014). However, the evolutionary mechanisms driving the changes in repetitive sequence, and by extension genome-size, remain elusive. For example, it remains unclear whether changes in genome size can be adaptive and under selection. Better characterization of repeat turnover will, therefore, provide a richer understanding of genome size evolution.

The evolution of satellite DNA is also closely coupled with the evolution of sex chromosomes. The X and Y chromosomes typically originate from a homologous pair of autosomes. Differentiation of the pair begins when one chromosome acquires a male sex-determining factor and recombination around the locus stops (Charlesworth, 1991). While the details about the intermediate stages are murky, the differentiation process is thought to have a predictable trajectory, where the proto-Y chromosome loses gene content and accumulates repetitive sequences, ultimately creating a degenerate Y chromosome with little to no genes and high amounts of repetitive sequences, and potentially even the loss of the chromosome altogether (Charlesworth and Charlesworth, 2000; Koerich et al., 2008; Vicoso and Bachtrog, 2015). Therefore, the Y chromosome provides an excellent opportunity to understand the expansion and growth of satellites that are likely to be neutral.

Here, we characterize the satellite DNA content in 9 Drosophila species by applying k-Seek on whole genome sequences of females and males. We find that high satellite content is not universal to all species, but only specific lineages. These lineages have high rates of gains of both Y-linked and autosomal/X-linked satellites. On the other hand, many lineages have very little repeat content, inconsistent with a model where repeats are simply accumulating neutrally. Surprisingly, very few loss events were detected indicating that the ancestral state may have contained very few satellites. Overall, this study illustrates that the evolution and divergence of satellites is highly dynamic and requires appreciation finer than that implied by "junk DNA".

3.2 Results

3.2.1 Identification of satellite DNA in *Drosophila* species with k-Seek

Illumina libraries were prepared separately from males and females of 9 species: *D. melanogaster, D. simulans, D. sechellia, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D mojavensis, and D. virilis* (Figure 3.1). The libraries were prepared using PCR-free library kits in order to minimize under-representation of AT-rich sequences. We applied k-Seek on the whole genome sequences of the samples to identify and quantify simple satellites. The copy number of each kmer was then multiplied by length of the kmer and normalized by the average autosomal depth of each sample. The list of kmers was trimmed to include only those that are over 10 kb in at least one sample. Across the species (Figure 3.2A), *D. virilis* has, by far, the most satellites at 8.0 Mb, consistent with previous reports (Bosco et al., 2007). This is nearly three times as much as the second highest species, *D. sechellia*, at 2.8 Mb. On the other side of the extremes, *D. erecta*, *D. persimilis*, and *D. pseudoobscura* have the lowest amounts with less than 300 kb. For the species within the *melanogaster* complex, the males all have higher

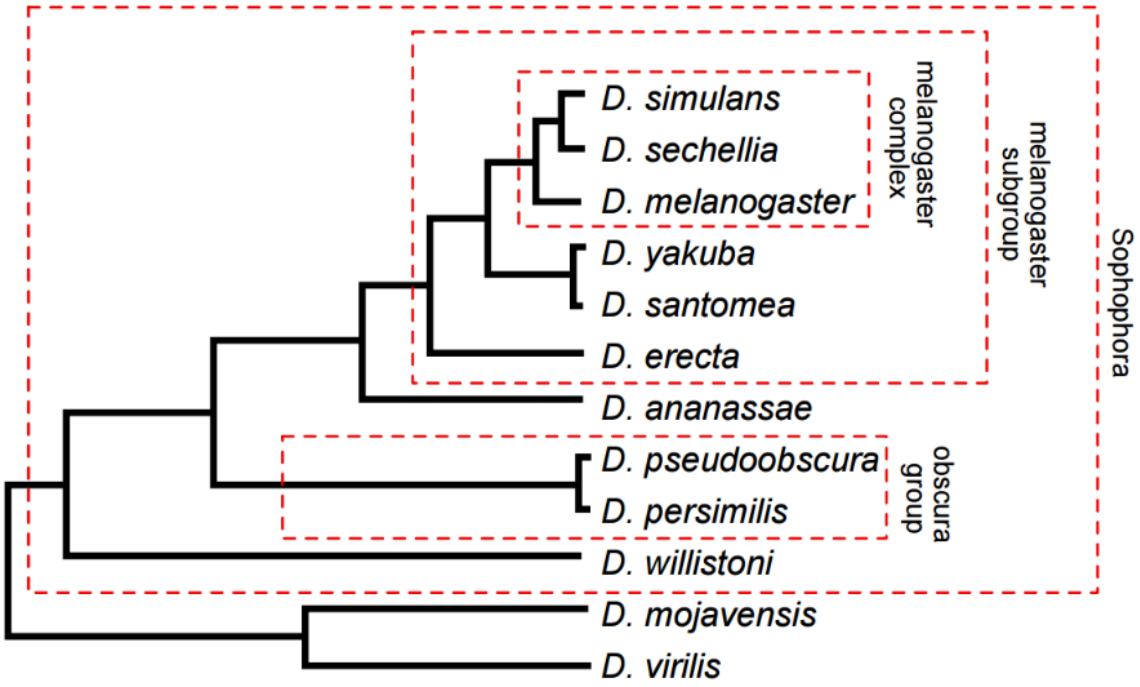


Figure 3.1. Phylogeny of *Drosophila* species sequenced. Commonly referenced lineages are boxed in dotted lines and labeled.

abundance of satellites than females, consistent with the fact that the Y chromosome has accumulated large amounts of repetitive sequences. This, however, does not appear to be a general rule, as females have slightly larger satellite amounts in *D. ananasse, D. mojavensis, and D. virilis*, likely caused by X-linked satellites which will be at twice the amount in females compared to males.

Only 4 satellites are present at greater than 1 kb across all species, the most abundant of which is the mononucleotide repeat of A, found on average of 300 kb per species (Figure 3.3B). The A mononucleotide repeat is also the largest satellite in the three low abundance species, as well as in *D. mojavensis* (Table 3.1). The rest of the satellites are either species-specific or found in fewer than 3 species, revealing that very few repeats are shared between species, further underscoring the speed at which they turnover. Consistent with previous reports, the most abundant repeats in *D. melanogaster* and *D. virilis* are the 5mer AAGAG and 7mer AAACTAC, respectively. The latter accounts for, astonishingly, 80% of the satellite quantities in *D. virilis* females. *D. ananassae* has the 6mer AAGGTC as the most abundant kmer. While the vast majority of kmers identified are smaller than 10 bps, the most abundant repeat in *D. simulans* and *D. sechellia* is the 15mer AACAGAACATGTTCG. Interestingly, it is composed of the 5mer GAACA twice and the palindrome TGTTC when rearranged to a different phase (GAACA-GAACATGTTC), suggesting that it may have originated from duplications of the AACAG 5mer. Other than two consecutive As, there appears to be little commonality between the largest repeats suggesting that repeat length and composition have little bearing on the potential for massive expansion.

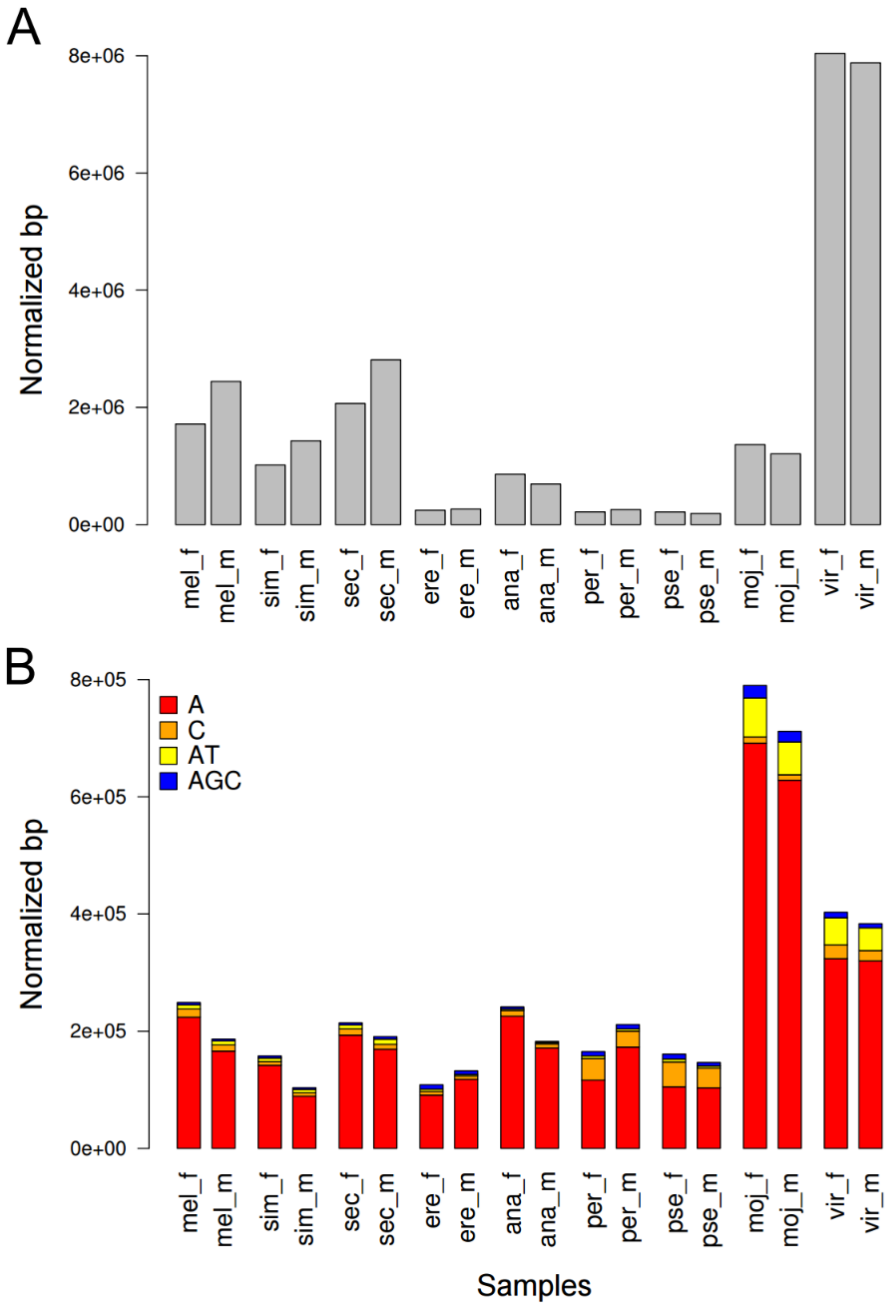


Figure 3.2. Satellite contents across samples. A. The total satellite content for each sample is plotted. B. The amounts of satellites shared across all species are plotted.

Table 3.1. Most abundant satellite among species

Species	No. of kmers		Most abundant kmer	Quantity (kb)	% total satellites
	Females	Males			
D. melanogaster	28	35	AAGAG	388.47	22.6%
D. simulans	18	23	AACAGAACATGTTCG	439.02	43.2%
D. sechellia	20	27	AACAGAACATGTTCG	102.41	49.5%
D. erecta	12	12	A	90.96	37.2%
D. ananassae	13	12	AAGGTC	326.09	38.0%
D. persimilis	8	8	A	116.44	53.4%
D. pseudoobscura	8	8	A	105.09	48.6%
D. mojavensis	21	23	A	691.44	50.6%
D. virilis	25	33	AAACTAC	6442.84	80.1%

It has been suggested that 5mers and multiples of it are particularly conducive for forming satellites due to the ~10 bp rotational periodicity of DNA wrapping around the nucleosome (Lohe and Brutlag, 1986). Adding support to this idea, euchromatic nucleosome-bound DNA shows an enrichment of AA dinucleotides with 10 bp periodicity at positions facing the histones (Langley et al., 2014; Mavrich et al., 2008; Segal et al., 2006). This is thought to provide an intrinsic curve to the double helix that promotes binding affinity to the nucleosome. However, the high abundance of 5mers and 10mers appears to be specific to the *melanogaster* complex, rather than a general pattern, as across all the species, 5mers and 10mers are no more numerous than other kmers (Figure 3.3A). To further test whether satellite DNA is over-represented for periodic AA dinucleotides, for each kmer we generated a homogeneous satellite block that is 150 bp, roughly the length of DNA wrapping around the histone core of one nucleosome. The AA dinucleotide frequency across the 150 bp is then determined across all generated kmer blocks. For kmers found in the *melanogaster* complex, the AA dinucleotide frequency indeed shows peaks with a 10 bp periodicity (Figure 3.4B). However, the kmers in *D. virilis* fail to show such a pattern, indicating that the expansion and emergence of satellites do not require an optimal nucleosome binding.

3.2.3 Satellites on the Y chromosome

To identify satellites on the Y chromosome, we compared the kmer quantities between males and females (Figure 3.4), reasoning that male-biased repeats must be at least partially Y-linked. Consistent with the highly heterochromatic and degenerate state of their Y, the species in the *melanogaster* complex all have multiple Y-enriched repeats. Several of them are AT-rich and shared among the three species, arguing that they are likely on the Y of the last common ancestor.

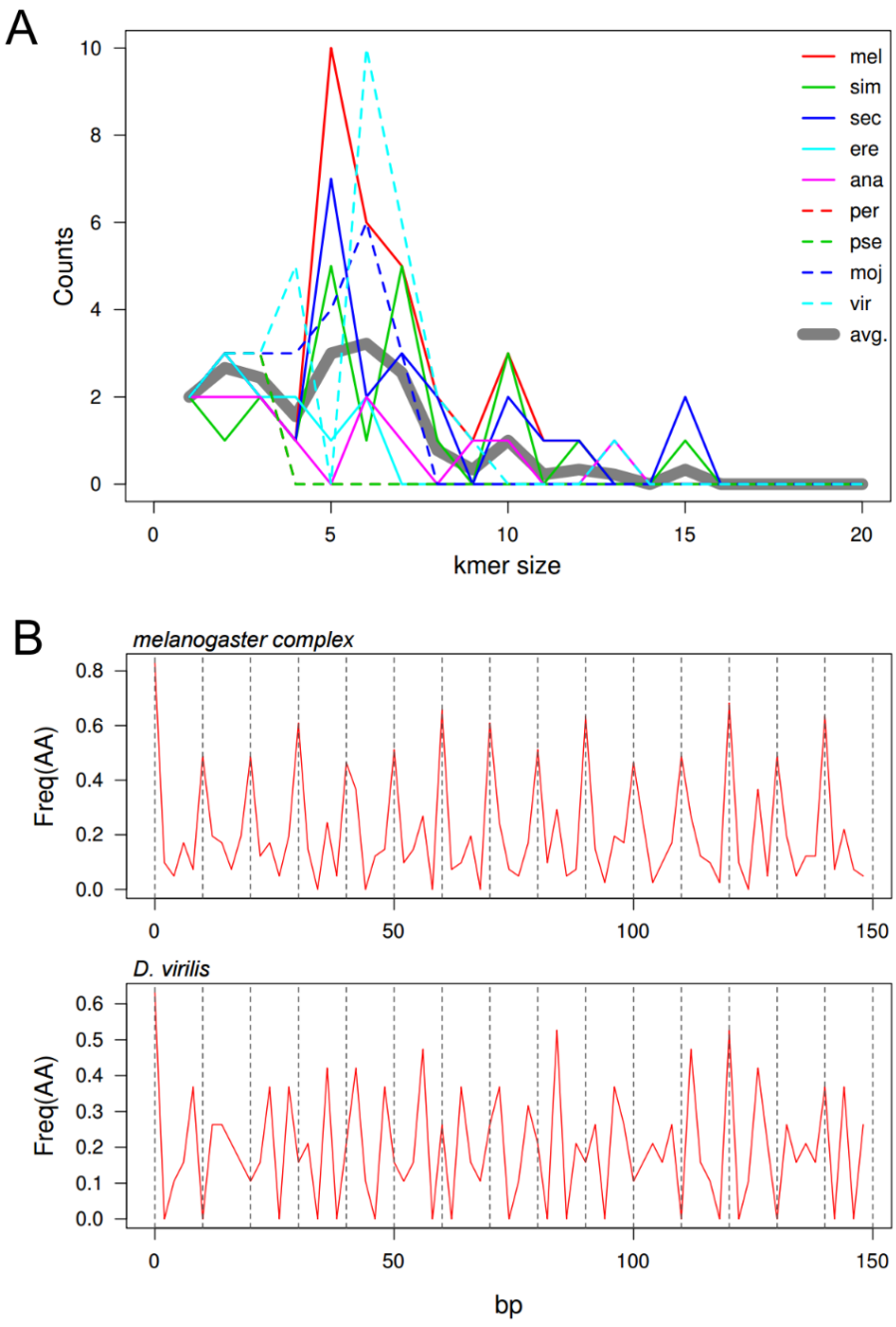


Figure 3.3. Satellite length distribution and AA dinucleotide phasing . A. The distribution of kmer lengths is plotted for each species. The average across all species is in gray. B. The AA dinucleotide frequency of all kmers within the labeled lineages is plotted across 150bp (the length of DNA wrapping around a single nucleosome. 10bp intervals are labeled by dotted gray lines.

Notably, half of the *D. melanogaster* Y-enriched repeats have essentially no presence in females and thus are specific to the Y. These Y-specific kmers are found only in *D. melanogaster* and not the other two species in the complex, suggestive of recent origins. Curiously, *D. erecta* and *D. ananassae*, both of which share an orthologous Y as those species in the *melanogaster* complex, have only one Y-linked satellite. This raises the possibility that the accumulation of satellites is specific to the *melanogaster* complex and these two species either have less degenerate Y chromosomes or are less tolerant of high repeat content.

Differential accumulation of satellites is also seen between *D. mojavensis* and *D. virilis*. The former has only 3 distinct repeat classes on the Y chromosome while the latter has 11. Interestingly, one of the three Y-linked satellites in *D. mojavensis* is AAGAG which is the most abundant, though not male-biased, satellite in *D. melanogaster*, suggesting independent gains of this satellite. The Y-linked satellites on *D. virilis* are all specific to the lineage, and, similar to *D. melanogaster*, a large subset of them are exclusively on the Y.

The two species in the *pseudoobscura* group completely lack male-biased repeats. Unlike the other species, the *obscura* group experienced a translocation of the ancestral Y onto an autosome less than 18 mya and the current Y had arisen de novo (Carvalho and Clark, 2005). Therefore the absence of satellites likely reflects the recent acquisition such that the neo-Y has not had enough time for substantial degeneration and accumulation of repeats.

3.2.4 Gains and losses of satellites along *Drosophila* phylogeny

Based on parsimony, gains and losses of satellites were determined across the Drosophila phylogeny (Figure 3.5). Interestingly, there are only 6 unambiguous losses compared to the 62 gains, suggesting that the ancestral state is relatively free of satellites. The gains of satellites are

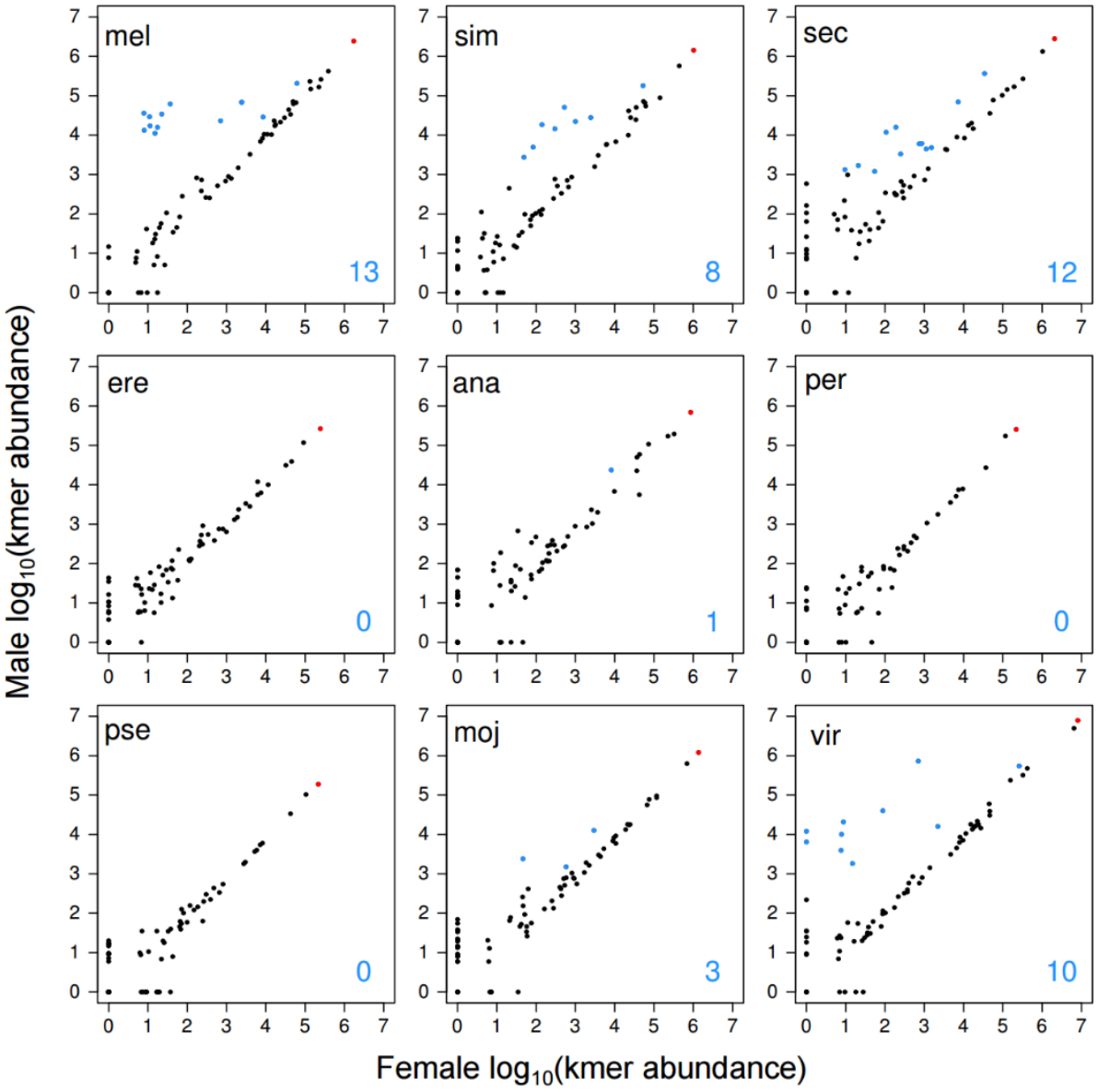


Figure 3.4. Y-biased satellites. For each species, kmer abundances in the female sample are plotted against the male sample. kmers that are greater than 10kb and more than 2-fold higher in males are categorized as y-biased (blue). The number of Y-biased satellites are indicated on the bottom right of each panel. The total satellite content is labeled red.

Table 3.2. kmers on the Y chromosomes

Species	Y-linked kmers	No.	
		Y-biased	Y-specific
D. melanogaster	AATAT, AAGAC, AATATAT, AAGAGG, **AATAC**, **AAAGAC**, **AATAGAC**, **AAGACATGAC**, AATAG, AGATG, **AAAAC**, **AAGACTAGAC**, **AAACAAT**, **AAAATAGAC**	14	7
D. simulans	AATAT, AAACAAC, AAGAGAATAG, AAT, AATATAT, AACAATC, AGATATAT, **AAACAAT**, **AAGAGAG**	9	2
D. sechellia	AATAT, AAT, AATAATAT, AAAAC, ACAGAT, AG, AC, AATATT, AGATATAT, **AAAAT**, **AATGAC**, **ACAGCAT**	12	3
D. erecta	AAATAT	1	0
D. ananassae	AAAGGT	1	0
D. persimilis	na	0	0
D. pseudoobscura	na	0	0
D. mojavensis	AAAAT, **AAAAC**, AAGAG	3	1
D. virilis	ACAG, AAACTAT, AC, **AAAC**, **AACTATT**, ACACAT, **AACAATCC**, **AATAATAG**, **ACAGACAGG**, **ACAGACAGACAGG**, **AAACAC**	11	7

kmers in bold are Y-specific

predominantly at the terminal branches, with *D. virilis* and the species in the *melanogaster* complex accounting for the vast majority. *D. melanogaster* has the fastest rate of acquisition with 22 new satellites gained within the ~2.5 million years after its split from *D. simulans*. Furthermore, the entire *melanogaster* complex has markedly faster rates of satellite acquisition than any other lineage. This is in stark contrast with other lineages in the *Sophophora* subgenus which have relatively few gains. The rate of gain is slow in *D. mojavensis*, but moderate in *D. virilis* given the branch length. However, because of the sparsity of species outside the *Sophophora* subgenus, it is difficult to determine the timing of the 17 gains in *D. virilis*, and they may be very recent and rapid events.

Given the paucity of shared kmers, we were surprised to see six satellites found in distant lineages consistent with independent gains (Figure 3.6). Two of them are di- and tri-nucleotide satellites which may have independently emerged through stochastic polymerase slippage followed by amplification via unequal crossing-over in the different lineages. The rest are longer kmers and, therefore, unlikely to have originated from polymerase error.

To determine whether satellite compositions across species can be used to infer the species relationship, we used hierarchical clustering to group the different species and sexes (Figure 3.6). As expected, the males and females of the same species are almost always clustered together. Although the three species within the *melanogaster* complex cluster as expected, the clustering of additional species fails to recapitulate the known phylogenetic relationships. *D. virilis* and *D mojavensis* cluster together as expected, however, their satellite composition appear to be more similar than *D. melanogaster* and *D simulans*, even though the latter pair diverged much more recently. Similarity, here, is determined by Euclidean distance which assumes a linear, clock-like change of repeats abundance. The high degree of similarity between *D.*

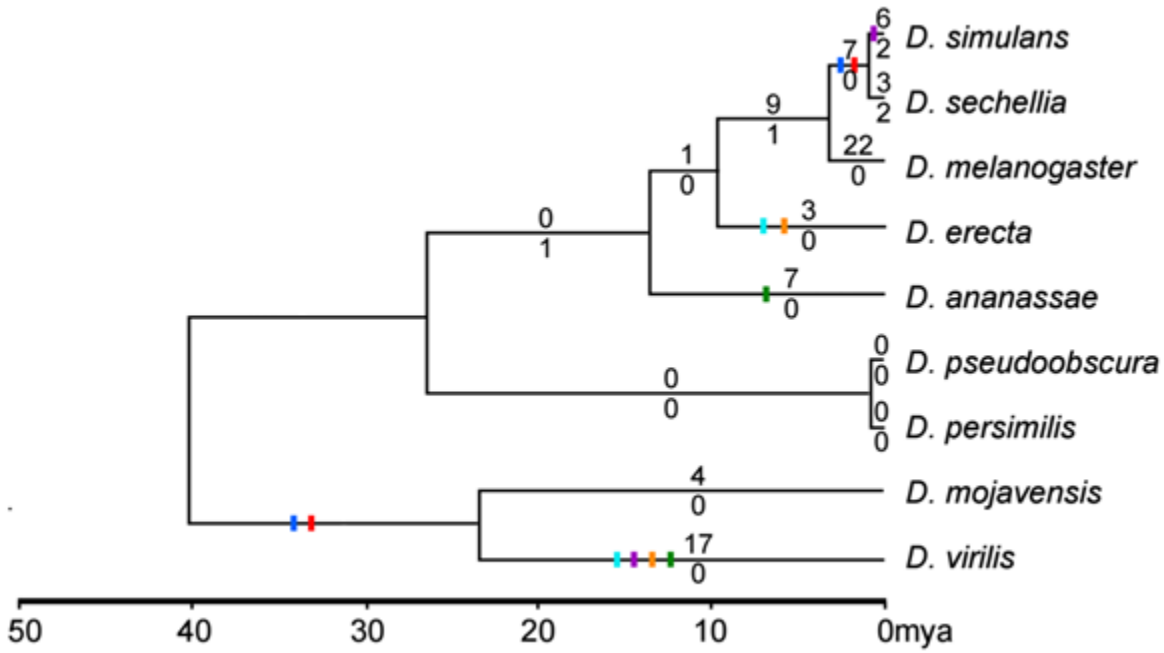


Figure 3.5. Gains and losses of satellites across *Drosophila* phylogeny. The number of unambiguous gains and losses, as determined by parsimony, are labeled above and below the branches, respectively. Independent gains are labeled with slashes of the same color on different branches.

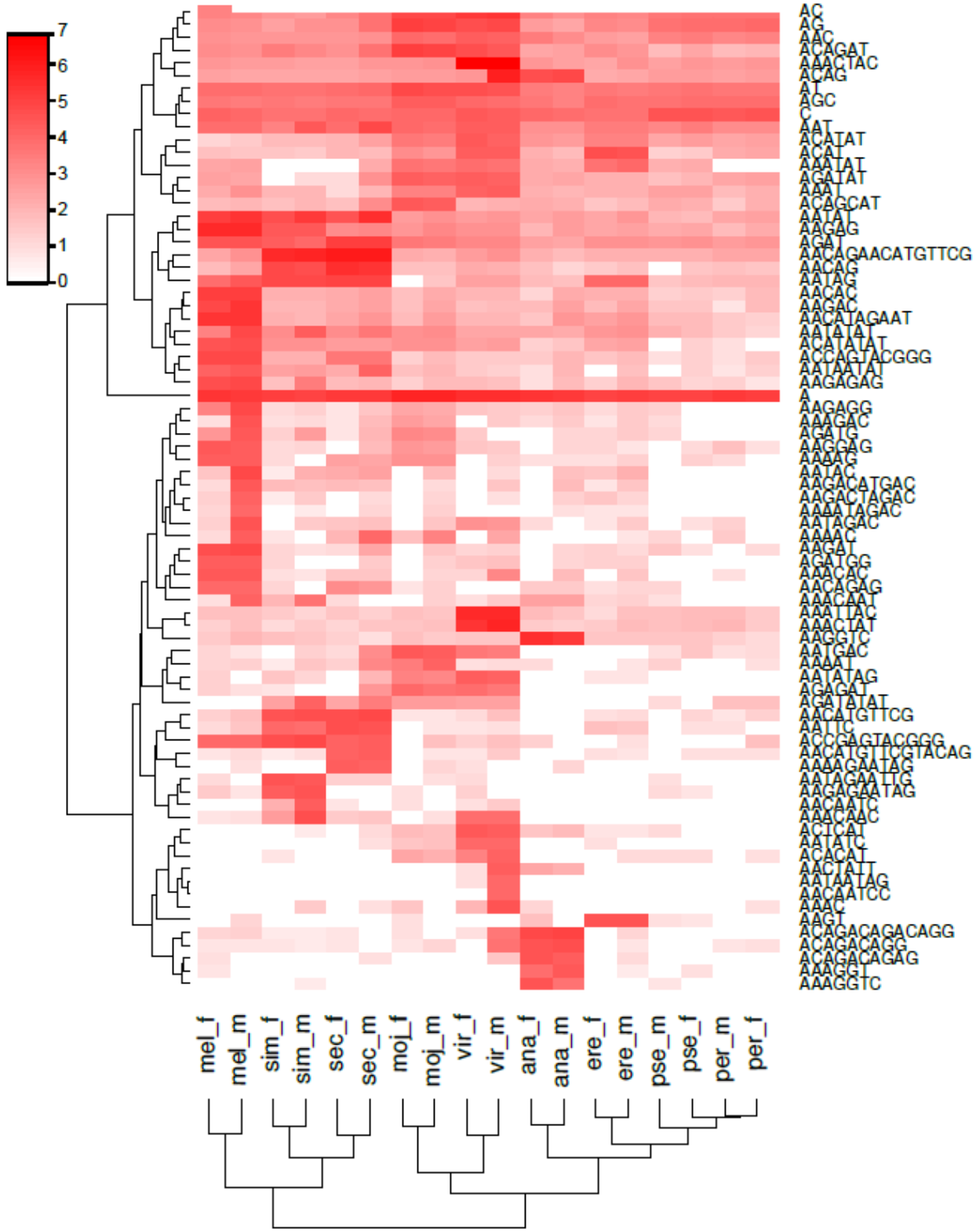


Figure 3.6. Hierarchical clustering of species. Clustering was based on Euclidean distance of the log-transformed kmer abundance. Color represents the log10 kmer abundance.

ananassae, *D erecta*, *D. pseudoobscura*, and *D. persimilis* is partly driven by the low abundance and absence of large number of repeats. After further trimming the list of satellites to only those greater than 1 kb in at least one of these 4 species, the species cluster as expected. As these low abundance repeats are likely to be loss or gained sporadically, an Euclidean distance metric may be inappropriate.

3.3 Discussion

3.3.1 Satellite, transposable elements, and genome size divergence

By applying k-seek on WGS of several *Drosophila* species, we exhaustively quantified simple sequence repeats. Bosco et al., 2007 used flow cytometry to estimate both genome size and the heterochromatic content across *Drosophila* species taking advantage of the fact that heterochromatic sequences are under-replicated in polyploid follicle cells. While the total satellite content from our study is highly correlated with their results for both genome size (Pearson's r = 0.968, p-value = 9.906e-07) and heterochromatic content (Pearson's r = 0.986, p-value = 9.906e-07), several species show glaring differences. For example, the genome of *D. virilis* is estimated to be twice as large as that of *D. melanogaster*, but the total satellite content is nearly 4-fold more in our estimation. These discrepancies are likely due to additional repetitive content such as transposable elements and complex satellites like the 359-bp satellite. Other genomic features and evolutionary events may further exacerbate the difference: for example, the genome of *D. ananassae* has integrated the genome of the prokaryotic symbiont *Wolbachia* (Choi et al., 2015; Dunning Hotopp et al., 2007).

While it is easy to envision the direct contribution of satellite content to genome size, the moderate correlation between them may also reflect some degree of independence of factors that

impact each. For example, the satellite content in some species may represent only a tiny fraction of the genome content as is the case for species like *D. erecta*, and therefore contribute little to the absolute genome size. On the other hand, a possible causal mechanism driving the positive correlation may be a genome-wide defense against both transposable elements and satellite DNA. In flies, defense against activities of transposable elements is predominantly accomplished by the piRNA machinery, where short RNAs of 26-31bp are used to recognized TE transcripts for degradation and transcriptional silencing. While the relationship between piRNAs and satellite DNAs is unclear, the fact that some satellites are expressed as mRNA suggest the possibility of piRNA targeting (Rošić et al., 2014; Usakin et al., 2007). Many components of the piRNA pathway are rapidly evolving, which is thought to be the result of the arms-race with TEs that evolve to escape repression (Castillo et al., 2011; Lee and Langley, 2010; Simkin et al., 2013). An efficient TE defense system can therefore produce a small genome as has been suggested for *D. simulans* which has fewer transposable elements, and, as shown here, less satellite DNA content, as compared to *D. melanogaster*.

3.3.2 Population genetic considerations of the lineage-specific satellite accumulation

While satellite DNA is thought to turnover rapidly, our data indicates that rapid change may be specific to certain *Drosophila* lineages. The highest rate is within the *melanogaster* complex, where large numbers of satellite repeats were gained within a period ~10 million years. This starkly contrasts with other lineages such as the obscura group which appears to have little to no gain over ~26 million years. These results suggest that the emergence and expansion of satellites are not purely stochastic processes that have no fitness impact, i.e. neutral. If the gains are completely neutral, different lineages are expected to have accumulated

comparable levels of satellites. However, the accumulated satellites may be weakly deleterious, in which case the species or lineages with larger effective populations are expected to accumulate less as selection is more efficacious. This may account for the differences between the species within the *melanogaster* complex: *D. simulans,* which has a larger effective population size (Hu et al., 2013; Nolte and Schlötterer, 2008) than *D. melanogaster,* has less satellite content. Similarly, *D. sechellia,* which has a small effective population size (Legrand et al., 2009), has more satellites than *D. simulans,* from which it recently diverged. However, there are few estimates for the effective population size of the remaining species, making this model difficult to test beyond the *melanogaster* complex.

Alternatively, the lineage-specific accumulation of satellites may have precipitated after specific changes in regulation of repeats. Csink and Henikoff 1998 suggested that the gains of AG-rich repeats in *D. melanogaster* were possible because the genome evolved a mechanism to ameliorate their deleterious effect. They argued that the transcription factor GAGA-factor which targets AG-rich promoters acquired the novel binding to AG-rich repeats. Beyond this case, the coupling of satellite expansion and emergence of sequence-specific binding proteins may be a general mechanism, as several proteins in D. *melanogaster* target satellites sequence-specifically. This includes Origin of replication complex 2 which targets the AT-rich repeats (Pak et al., 1997) and proliferation disruptor which targets the 10mer 2L3L (Török et al., 1997). Both of these satellites have expanded in *D. melanogaster*. Similarly, the highly abundant repeats found in *D. simulans* (AACAGAACATGTTCG) and *D. virilis* (AAACTAC) may be under the regulation of proteins that evolved binding specificity. In contrast to the arms-race model, this model posits that satellites are deleterious if improperly regulated, and the emergence of sequence-specific binding regulation reduces the constraint on satellite abundance. Therefore, expansions are not

bouts of activity resulting from escape of repression, but rather permissive growth due to relaxed constraint.

3.3.3 Accumulation of satellites on the Y chromosome

The trajectory of Y chromosome degeneration is thought to be stereotypical; repeats will inevitably accumulate as genes become pseudogenized. With the exception of the replacement of the Y chromosome along the obscura lineage, drosophila species are thought to share the same Y which originated from a supernumerary a B chromosome (Carvalho, 2002). The absence of recombination in Drosophila males is thought to expedite the process of Y degeneration as deleterious mutations on the Y are necessarily linked to the rest of the chromosome, and the entire chromosome is subject to Muller's ratchet (Charlesworth and Charlesworth, 2000). It is, therefore, to our surprise that not all the Y chromosomes accumulated high satellite content, as exemplified by *D. erecta* and *D. ananassae*. One possibility is that the accumulated satellites were lost via chromosome fragmentation or intra-chromosomal exchange between satellites through gene conversion. Alternatively, it may simply reflect the fact that these species have low satellite content overall. It is also a possibility that these Y chromosomes are nonetheless accumulating large amounts of transposable elements and the lack of satellites may be due to a genomic environment that is particularly intolerant to satellite DNA.

3.4 Materials and methods

3.4.1 DNA extraction, library preparation and sequencing

10 males and 10 females were collected for each species. DNA was extracted using the DNeasy kit (Qiagen). Libraries were made using the Illumina Truseq PCR-free kit with 24 barcodes. The

samples were sequenced first on one lane of HiSeq 2500 at 100 bp single-end, followed by a second single lane run at 100 bp paired end.

3.4.2 Sequence analysis

We used FastQC to check the quality of the samples. The most recent version of the reference genome for each species was downloaded from Flybase, with the exception of the *D. simulans* reference which was released by P. Andolfatto. We aligned the reads using BWA on standard settings, followed by samtools and picards tool to infer the average autosomal read-depth. For references where the contigs are not labeled by chromosomes, we inferred the autosome by comparing the read-depth of the contigs between males and females, removing the contigs that are fewer than 5Mbs or are at half the read depth in males.

3.4.3 kmer identification and quantification

We updated kseek to identify kmers of up to 20 bps. We applied k-Seek on all samples and summed the samples between the sequencing runs. The kmer counts were then normalized by the kmer length and the average autosomal read-depth. kmers with less than 10 kb in at least one species were removed.

3.4.4 dinucleotide frequency of satellites

For each kmer, a homogeneous satellite block of 150 bp was computationally created (e.g. AAGAG*30). The dinucleotide frequency across the 150 bp was then determined by scanning across the 150bp of all the satellites in 2 bp windows.

3.4.5 Inferring satellite gains and losses

For satellites shared between more species, the branches in which they were gained or lost were determined manually using a parsimony approach. As gains and losses have equal weight, cases where they are equally probable are removed.

CHAPTER 4

MEIOTIC TRANSMISSION FIDELITY IS ROBUST TO EXTREME TELOMERE-LENGTH DIFFERENCES IN *DROSOPHILA*[1]

4.1 Introduction

Eukaryotic linear chromosomes end in specialized structures called telomeres in order to solve the challenge of fully replicating chromosome ends, known as the end replication problem. Most eukaryotes use the enzyme telomerase to add RNA-templated short repeats to chromosome ends to prevent terminal erosion into functional genes (Blackburn *et al.*, 2006). But despite telomeres being essential for all linear chromosomes, several lineages, including *Drosophila,* have lost telomerase-mediated telomere formation (Mason *et al.*, 2015). *Drosophila* instead maintain their telomeres using three specialized retrotransposons, *HeT-A*, *TART* and *TAHRE* (Mason et al., 2008; Pardue and DeBaryshe, 2011), collectively known as the HTT elements. Telomeres are thus maintained through activation and reverse-transcription, leading to incorporation of new elements to the chromosome ends. As active copies only transpose to chromosome ends, this is a prime example of the domestication of transposons (TE) to take on an essential function. However, broken *Drosophila* chromosomes can be stably maintained without telomeric TEs, indicating that the HTT array is not strictly necessary and that there is no sequence-specificity to telomeres (Biessmann *et al.*, 1990; Rong, 2008).

[1] This work was in collaboration with James M. Mason and Hemakumar M. Reddy at the National Institute of Environmental Health Science, and Shuqing Ji, Chandramouli Rathnam, Jimin Lee, and Deanna Lin at Cornell. HMR and SJ assayed telomere length using qPCR across DGRP lines. SJ, CR, JL, and DL collected embryos and prepared the library for sequencing. KHW analyzed all data, developed the computation framework for the method and wrote the manuscript.

The evolutionary forces driving the transition between distinct forms of telomere formation remain unclear. Organisms such as the silkworm *Bombyx mori* and the beetle *Tribolium castaneum*, which have both telomere-specific short repeats and TEs (Fujiwara *et al.*, 2005; Osanai-Futahashi and Fujiwara, 2011), may resemble intermediate stages of this transition. However, recent phylogenetic analyses of insect telomeres suggest that the Dipteran lineage leading to Drosophila had neither the canonical telomerase nor telomeric TEs (Mason *et al.*, 2015). Regardless of the previous state, the transition may have initiated with TE insertions at the telomere that drifted to fixation. The increase in telomere length would then have rendered the canonical telomerase-mediated system or alternative mechanisms unnecessary, subsequently these mechanisms would have been lost due to neutral accumulation of mutations. Alternatively, the replacement may have been non-neutral. Beyond protecting chromosome ends, telomeres play essential roles in meiosis such as in homologous pairing, separation of chromosomes at anaphase, and centromere assembly (Klutstein *et al.*, 2015; Rockmill and Roeder, 1998). These roles make telomeres a potential hotspot for selfish elements that can influence meiotic segregation frequencies, specifically in oogenesis, where only one of four meiotic products will be included into the pronucleus and become the oocyte, while the rest become polar bodies (Zwick *et al.*, 1999). If a selfish element can manipulate oogenesis to increase its rate of transmission, it will cause meiotic drive and quickly fix in the population. Such a scenario may mediate transitions in telomere sequence type, whereby insertions of selfish TEs create or modify a telomere that is then more frequently included in the egg pronucleus. Unlike the neutral model, the emergence of telomeric meiotic drivers will result in the rapid replacement of telomeric sequences at every chromosomal end.

Telomere-mediated meiotic drive may also explain high levels of variability in telomere-associated sequences. The telomeric TEs among *Drosophila* species vary widely in sequence and structure (Piñeyro *et al.*, 2011; Villasante *et al.*, 2007). Given that telomeric TEs are multi-copy, the fact that their sequences recapitulate the species tree suggests that they undergo concerted and rapid turnover. This can be mediated through gene-conversion homogenizing different variants, or replacement by new copies that are positively selected. Other sequence classes also point to the evolutionary lability of telomere regions. Many species contain sub-telomeric repeat classes, which vary widely in sequence and arrangement (Anderson *et al.*, 2008; Mason and Villasante, 2014; Mefford and Trask, 2002). Sub-telomeric genes show presence/absence polymorphisms, suggesting high rates of telomeric deletions in populations (Kern and Begun, 2008; Walter *et al.*, 1995). In addition, high rates of divergence have been observed in the subtelomeric euchromatic sequences between *D. melanogaster* and its sister species *D. simulans* (Anderson *et al.*, 2008).

If meiotic drive underlies the rapid evolution of telomeres and related sequences, genes suppressing their effects would evolve under positive selection (Thomson and Feldman, 1976). This is because meiotic drivers can be deleterious either through pleiotropy or linkage with a deleterious locus. Several proteins that regulate telomeric chromatin state and HTT element expression show elevated rates of evolution. These include the telomere capping proteins Hoap and HipHop, multiple genes in the piRNA pathway, and TE repressors like *Lhr* and *Hmr* (Blumenstiel, 2011; Gao *et al.*, 2010; Lee and Langley, 2010; Raffa *et al.*, 2011; Satyaki *et al.*, 2014; Schmid and Tautz, 1997). Interestingly, mutant alleles of several of these genes produce long telomeres (Khurana *et al.*, 2010; Satyaki *et al.*, 2014; Shpiz and Kalmykova, 2012), raising

the possibility of genomic conflict between telomere length and host fitness, another hallmark of meiotic drive.

Telomere length is determined by rates of transposition, deletion, gene conversion, unequal crossing over, and number of cell divisions per generation. In flies, the first indication that telomere length is under genetic control was from a wild strain containing a mutation for the *Telomere elongation* (*Tel*) gene (Siriaco *et al.*, 2002). Although it was mapped to chromosome 3R, the identity of the *Tel* gene remains unknown. Nevertheless, this raises the question — how variable is *Drosophila* telomere length? Here, we examine a natural population of *D. melanogaster* for telomere repeats to determine telomere-length variation, in particular focusing on the most abundant repeat, *HeT-A*. We report the discovery of unprecedented variation in telomere length and test whether extreme telomere length variants cause meiotic drive using whole-genome sequencing of large pools of embryos.

4.2 Results

4.2.1 Drastic telomere length variation

To evaluate the extent to which telomere length varies in wild populations, we quantified relative *HeT-A* abundance in 182 lines from the *Drosophila* Genetic Reference Panel using qPCR. Since *HeT-A* elements are frequently 5' truncated (George *et al.*, 2006), we used primers flanking the 5' region of the *gag* coding sequence in order to capture mostly full-length copies (Figure 4.1A). We observed an enormous 288-fold range of *HeT-A* abundance between the highest and lowest lines (NC703 and NC852, respectively). Notably, the top 3 lines (NC161, NC882, and NC703) have higher *HeT-A* abundance than even the GIII line, which harbors the

Tel mutation and has accumulated long telomeres for over a decade, indicating that extremely long telomeres occur in natural populations, albeit at low frequency.

We used fluorescent in situ hybridization (FISH) on polytene chromosomes to confirm that the high abundance of *HeT-A* corresponds to long telomeres. We found that the high abundance line NC882 has elongated telomeres marked by *HeT-A* signal, while the low abundance NC332 has markedly less *HeT-A* signal at chromosome ends (Figure 4.1B). Furthermore, when we crossed the two lines to each other we clearly observed disparate telomere lengths at the tip of a polytene chromosome. We conclude that *HeT-A* abundance strongly correlates with telomere length, and henceforth refer to high and low abundance lines as having long and short telomeres, respectively.

The *HeT-A* quantities distribute along a logarithmic scale, arguing that the variation among lines is not simply due to new attachments or deletions, which instead predicts linear changes in abundance. Interestingly, the distribution of the log-transformed quantities is not normal (Shapiro-Wilk test $p < 0.001$, Figure 4.1C), and better fits a logistic distribution (Kolmogorov–Smirnov test, $p = 0.6278$, see Materials and Methods), indicating that more individuals fall at the extremes of the distribution than would be expected based on a normal distribution. As such distributions often characterize self-limiting processes like cell growth, this suggests that telomere length is under some form of constraint.

To further investigate the biological processes underlying the length differences, we quantified abundance at two additional regions of *HeT-A*, the 3' CDS and the promoter, in a subset of the lines. As amplification efficiency may differ between primer pairs, it is not possible to directly compare quantities between the three regions. We therefore calculated the respective *HeT-A* abundances relative to the short-telomere line NC852 (Figure 4.1D). In the three long

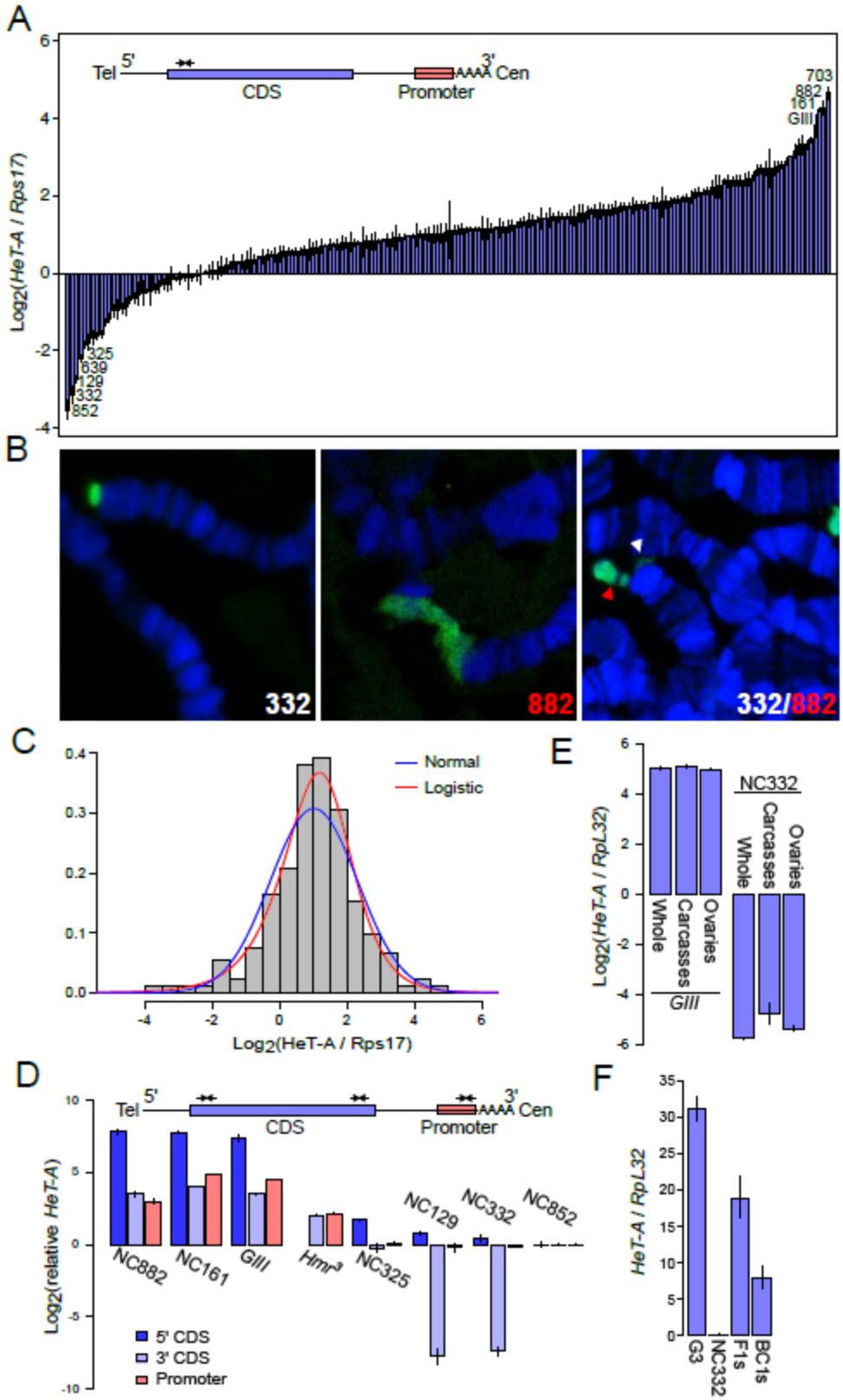

A
Tel
5'
CDS
Promoter
3'
AAAA
Cen
703
882
161
GIII
325
639
129
332
852
Log2(HeT-A / Rps17)
B
332
882
332/882
C
Normal
Logistic
Log2(HeT-A / Rps17)
D
Tel
5'
CDS
Promoter
3'
AAAA
Cen
Log2(relative HeT-A)
NC882
NC161
GIII
Hmr3
NC325
NC129
NC332
NC852
5' CDS
3' CDS
Promoter
E
Log2(HeT-A / RpL32)
Whole
Carcasses
Ovaries
GIII
NC332
Whole
Carcasses
Ovaries
F
HeT-A / RpL32
G3
NC332
F1s
BC1s

Figure 4.1. Natural variation in telomere length as assayed by *Het-A* quantities. (A) *Het-A* copy number was measured using qPCR of the DGRP lines as well as the long-telomere strain GIII (Mackay *et al.*, 2012; Siriaco *et al.*, 2002). The primer pair (convergent arrows) targeting the 5' region is depicted on the schematic of one full length HeT-A transcript where the *gag* CDS is colored in blue, the promoter in red, and UTRs are dotted lines. The most extreme lines at each end of the distribution are labeled. Error bars represent the relative error estimated from triplicates. (B) Polytene chromosome spreads of NC332, NC882, and a F1 heterozygote are probed with *Het-A* probe (green). DNA is labeled by DAPI (blue). In the heterozygote, the long allele is labeled by a red arrowhead, and the short allele by an white arrowhead. (C) The distribution of *HeT-A* quantities from (A) is fitted with either a normal or a logistic distribution. (D) Relative *HeT-A* quantities were determined at three regions as indicated by schematic in a subset of the DGRP and mutant lines. Quantities are plotted relative to NC852 so that the three regions can be compared. Hmr^{3} was assayed at only two of the regions. (E) *HeT-A* quantities were measured in whole females, ovaries, and carcasses with ovaries removed in a long and a short line, using the 3' CDS primer pair. (F) *HeT-A* quantities were measured in F1s heterozygous for long and short telomeres and backcross embryos of the F1s to the short parent using the 3' CDS primer pair. Note that the Y-axis here is in linear scale.

lines assayed, the relative abundance of the 3' CDS and promoter are between 8-29 fold higher than NC852, but these differences are drastically lower than for the 5' CDS (between 162-224 fold). Because *HeT-A* is highly prone to 5' truncations, these discrepant ratios are most likely due to a high number of 3' fragments in NC852 rather than a high number of 5' fragments in the long lines. Interestingly, in the two short lines NC129 and NC332, the 3' CDS is over 200-fold lower relative to NC852, even though the 5' CDS and promoter share similar quantities with NC852, revealing an excess of *HeT-A* 3' fragments that contain no CDS. This suggests that different lines may have different biases regarding the site of truncation. The excessive 5' truncations in the short lines may reflect multiple cycles of terminal erosion during replication, followed by new attachments at the end, since the 5' end of *HeT-A* is oriented toward the chromosome end. Alternatively, it could also be due to high rates of incomplete reverse transcription, such that only the 3' portion of the element is added, or gene conversion involving extended arrays of 3' ends.

The enormous range observed indicates that telomere length is highly labile. To test the stability of extremely long telomeres, we first compared *HeT-A* abundance in ovaries and carcasses of the extremely long GIII line. We found similar quantities in both, indicating that long telomeres are stable within individuals (Figure 4.1E). We then tested whether the length is stably transmitted across generations. We crossed GIII to the short NC332 line, reasoning that the discrepant chromosomes ends may be particularly unstable in a heterozygous background (Figure 4.1F). *HeT-A* abundance in these F1s was approximately intermediate to those of the parents. We then backcrossed the F1 females to the NC332 line creating a pool of BC1s and found that *HeT-A* abundance in the BC1s is also intermediate relative to the parents. These results indicate that telomere length is stably inherited across at least two generations.

4.2.2 Loci and phenotypes associated with telomere length

In order to identify genetic loci associated with telomere length, we performed a genome-wide association study (GWAS) on the log-transformed qPCR quantities. We note that unlike typical GWAS studies the "phenotype" is actually a genomic feature. A total of 34 genetic markers (32 SNPs and 2 indels) were significantly associated (see Materials and Methods) with telomere length (Figure 4.2A, Table 4.1). All loci but three are non-coding: 16 are found within introns, 2 are in UTRs, 4 are near genes, and 9 are intergenic (Figure 4.2B). The two nonsynonymous variants are missense SNPs in the genes *Hemolectin* (*Hml*) and *viking* (*vkg*). The dearth of coding variants suggests that telomere length may be controlled largely by regulatory changes, but none of the identified genes have been previously implicated in telomere or even heterochromatin regulation, and gene enrichment analysis on this small set of genes yielded no notable processes. Interestingly though, many of the associated loci are within binding sites of well characterized transcription factors that play essential roles in development. This includes key embryonic patterning genes like *Caudal*, *distalless* and *hunchback*.

We sought to find additional phenotypes and genomic features that are associated with telomere length. Based on a report of associations between telomere length and life history traits (Walter *et al.*, 2007), we tested whether telomere length correlates with fecundity or longevity but found no significant association. Interestingly though, we found that telomere length shows a weak but significant correlation with genome size (Pearson's $r = 0.1556$, $p = 0.03859$) (Ellis *et al.*, 2014). This is unlikely a direct effect of telomere length on genome size, since the longest line has at most 3-5 Mbps of telomeric sequence (see Discussion), less than 3% of the total

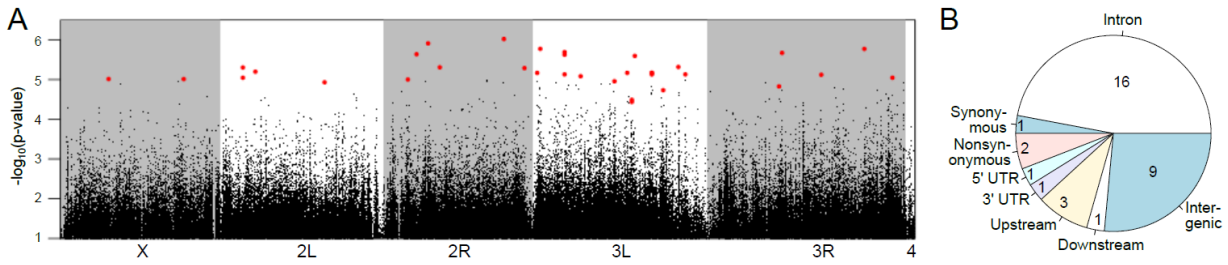


Figure 4.2. Genome wide association study of telomere-length variation. (A) A Manhattan plot of the p-values of variants used in the GWAS. Significant variants are labeled in red. (B) The significant SNPs are categorized by their genomic classes.

Table 3.1. Genes with loci significantly associated with telomere length

Chr	Pos	Gene	Site Class	TF binding
X	17359611	*CG43658*	intron	
2L	5014194	*vkg*	nonsynonymous	cad, chinmo
2R	3487145	*CG1358*	intron	bcd, dl, gt, Med, twi, HSA
	4707584	*sns*	intron	
	6336554	*G-oalpha47A*	intron	
		Cyp49a1	downstream	
	7976135	*jeb*	3' UTR	
	17001610	*king-tubby*	intron	
	19890974	*Unc-89*	intron	cad, chinmo, HSA
3L	538999	*klar*	intron	dl, hb, Trl
	968196	*Glut1*	intron	
	4387745	*slow*	intron	
	4387750	*slow*	intron	
	4387788	*slow*	intron	
	13207205	*tRNA:CR32127*	upstream	twi, cad
	13852596	*Hml*	synonymous	cad, sens
		CG8745	downstream	cad, sens
	13852625	*Hml*	nonsynonymous	cad, sens
		CG8745	downstream	cad, sens
	14271015	*fz*	intron	da, dl, gt, twi, HSA, Kr, sens
	16666956	*CG13032*	upstream	cis-regulatory modules (CRMs)
		zetaCOP	upstream	CRMs
	16666957	*CG13032*	upstream	CRMs
		zetaCOP	upstream	CRMs
	16666963	*CG13032*	upstream	CRMs
		zetaCOP	upstream	CRMs
	21386410	*rgn*	intron	
3R	10031156	*DopR*	intron	
	10457816	*CG7886*	intron	
	22034454	*CR44320*	intron	
	25982149	*CG11498*	5' UTR	twi, cad, chinmo

Site classes and annotations are based on GWAS output from the DGRP website.

genome size. Instead, we suggest that this correlation reflects polymorphisms that modulate the expansion and contraction of multiple repetitive sequences including the telomeric repeats.

4.2.3 Assessing non-Mendelian segregation using pooled sequencing

To test whether drastic telomere length differences cause biased meiotic segregation, we devised a scheme to sample genotype frequencies from whole-genome sequencing of large pools of progeny. We generated heterozygous F1 females by crossing lines with long and short telomeres (P1 and P2, respectively) and then backcrossed the F1s to the P2 parental line (Figure 4.3A). Given Mendelian segregation, half of the offspring are expected to be heterozygous and the other half homozygous at any given site. When sampling a large population of individuals the expected ratio of the P1:P2 alleles is therefore 1:3. To minimize the effects of viability and/or developmental differences among individuals of different genotypes, we collected large numbers of early embryos within a short time window (3-4 hrs after egg laying). After pooling all embryos for each of the crosses, DNA was extracted and sequenced with Illumina to ~30x depth per cross (Table 4.2, see Materials and Methods).

To infer allele frequency, we determined heterozygous sites genome-wide in two ways. First, we used GATK to infer SNPs in the pooled sequence reads from which we identified heterozygous SNPs. Second, we called SNPs in the parental lines from available collections (DGRP lines) or our own sequencing (*Hmr*^{3} and GIII), and then identified sites such that the parents were homozygous for different nucleotides so that all offspring are expected to be heterozygous. We used only the set of SNPs found in both methods, resulting in ~400,000 heterozygous sites genome-wide for each cross (Table 4.2). Across these sites, the average frequency of the P1 allele is significantly higher than expected (Table 4.2). To determine

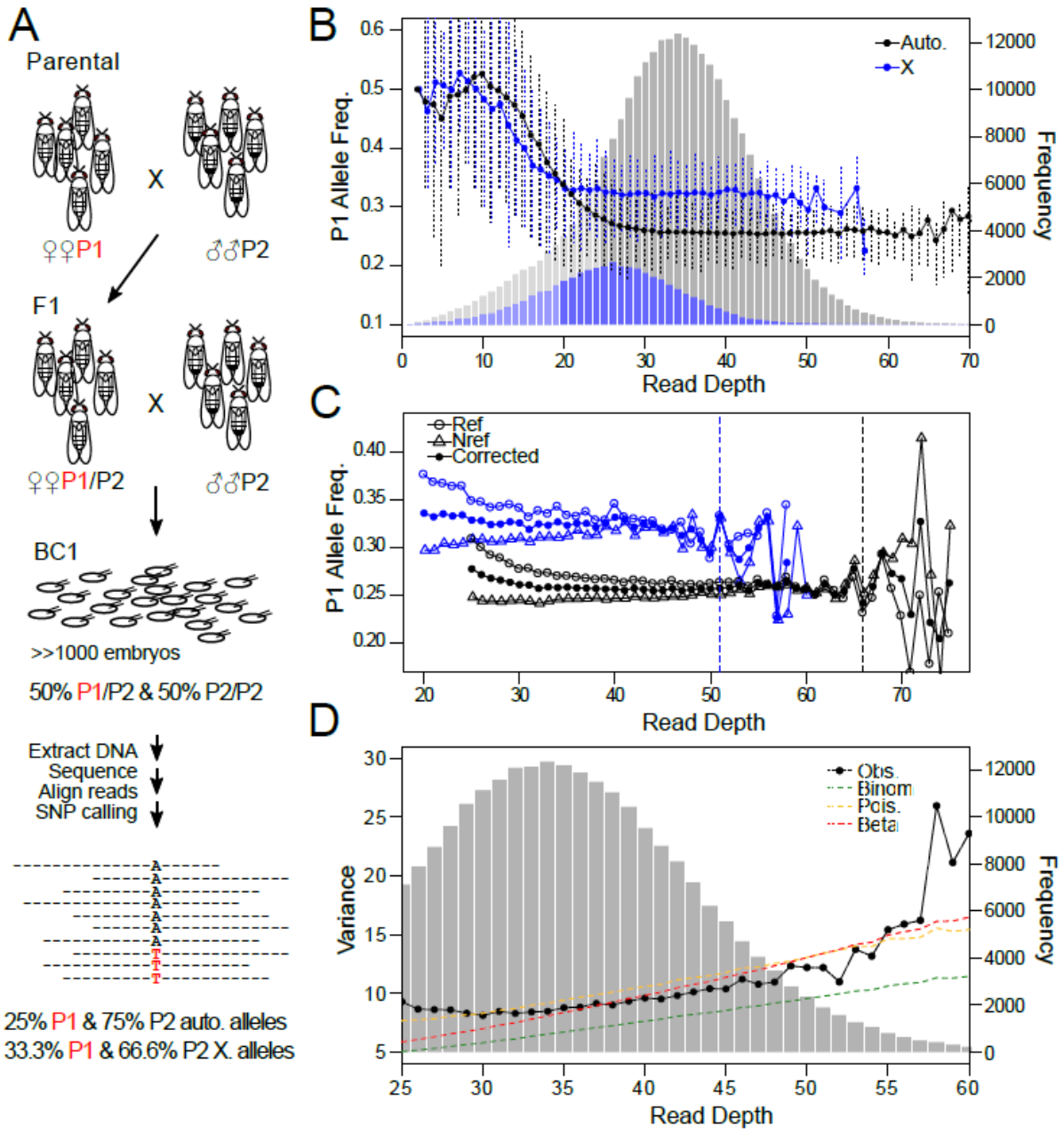

A
Parental
X
♀♀P1
♂♂P2
F1
X
♀♀P1/P2
♂♂P2
BC1
>>1000 embryos
50% P1/P2 & 50% P2/P2
Extract DNA
Sequence
Align reads
SNP calling
25% P1 & 75% P2 auto. alleles
33.3% P1 & 66.6% P2 X. alleles
B
P1 Allele Freq.
Frequency
Read Depth
Auto.
X
C
P1 Allele Freq.
Ref
Nref
Corrected
Read Depth
D
Variance
Frequency
Read Depth
Obs.
Binom
Pois.
Beta

Figure 4.3. Experimental strategy and statistical considerations for assessing allele frequency by pooled sequencing. (A) Strategy to measure distortion of Mendelian segregation using whole-genome sequencing of pooled embryos. Females (P1) and males (P2) of different telomere lengths are mated to generate F1s with disparate telomere lengths. The F1 females are backcrossed to P2 and a large number of 3-4 hr old BC1 embryos collected for sequencing. Heterozygous SNP sites are identified to infer segregation frequency. Expectation under Mendelian segregation is shown at bottom. (B-D) Analysis of SNPs from the NC882 and NC129 cross. (B) Average frequencies of the P1 allele across heterozygous sites at different read depths are plotted for the Autosomal (black line) and X-linked SNPs (blue line). Error-bars delineate the 0.25 and 0.75 quantiles. Underneath is the distribution of read depth at autosomal and X-linked SNP sites in gray and blue histograms, respectively. Bins with lighter shades are sites removed from allele frequency estimation in downstream analyses. (C) The average allele frequency is plotted when P1 is the reference allele (open circles) and when it is the alternative allele (triangles). The allele frequency after correction for reference is plotted in closed circles. Autosomal and X-linked sites are distinguished by black and blue, respectively. Dotted lines mark the cutoff for reference allele correction. (D) Variance estimates of the allele counts are plotted for each read depth (black). Variance is also plotted for each read depth, assuming either a Binomial (green), Poisson (yellow), or Beta-Binomial process (red).

Table 4.2. Summary of pooled sequencing

P1	P2	No. of embryos	No. of reads (million)	No. of het sites	Autosomes		X		No. of het sites after filtering
					Avg. depth	Avg. P1 freq.	Avg. depth	Avg. P1 freq.	
GIII	NC332	~3600	76.09	430,564	31.03	0.305*	23.91	0.360*	326,682
NC882	NC129	~2000	89.04	418,278	33.13	0.280*	25.95	0.341*	343,235
Hmr^3	NC129	~1800	64.15	339,204	24.81	0.270*	19.6	0.367*	182,183
NC332	**NC882**	~8000	88.4	440,510	33.37	0.270*	25.96	0.348*	362,844

* p.value < 1e-6 when compared to expected value of 0.25 for autosomes and 0.33 for X.
Lines with long telomeres are in bold.

whether this elevation is artefactual, we looked at the P1 frequency across sites of different read depth and found that low-depth sites have markedly higher P1 frequency (Figure 4.3B). We attribute this to the asymmetric ascertainment bias involved in calling heterozygous sites where the expected allele frequency is 0.25. In a binomial process, a 25% success rate is likely to yield trials where the P1 allele is not sampled, particularly when the number of draws is low (meaning here, at low read depth). This, therefore, causes a tendency to miss heterozygous sites with low P1 allele frequency introducing a bias that depends on read depth. To minimize this bias, we removed X-linked and autosomal sites with read depth less than 20 and 25, respectively, reducing the number of heterozygous sites to ~300,000 (Table 4.2). We note that the *Hmr*^{3} x NC129 cross had the largest reduction in heterozygous sites as it had shallower sequencing compared to the other crosses.

Another technical artifact that can influence inference of allele frequencies is ascertainment bias for the reference allele. As reference alleles often exist in sizeable haplotype blocks, this bias may result in local deviations from the expected frequencies. We therefore differentiated sites where the P1 allele is the reference from those where P2 allele is the reference. Indeed, the P1 frequency of the former is significantly elevated (Figure 4.3C). The difference between the frequencies of reference and non-reference alleles decreases as read depth increases. To correct for this, we estimated the extent of bias for each read depth, and increased the non-reference alleles accordingly (see Materials and methods).

While the sampling of alleles might at first be expected to be binomial, there are PCR steps and other features of the experiment that would be expected to inflate the sample-to-sample variance (Plagnol *et al.*, 2012). In fact, the variance of allele counts is notably greater than binomial (Figure 3D). After attempting to fit distributions of read counts that accommodate this

over-dispersion (e.g. beta-binomial), we settled on the Poisson to model the variance of P1 and P2 read counts. These data have the property that the variance in allele proportions at both low and high read depths is elevated. This is in part because the left and right tails of the read depth distribution capture structural variants in the genome such as deletions and duplications, respectively, which alters both the expected read depth and the allele frequency. We therefore removed sites with greater than twice the average read-depth.

4.2.4 Telomere length does not bias segregation

Instead of using frequencies of individual heterozygous sites, we aggregated heterozygous sites within 200 kb windows, reasoning that neighboring sites should have similar patterns of distortion as recombination will rarely break the genetic linkage at this scale. In 3 of the 4 crosses, allele frequencies mostly conform to the Mendelian expectation across all chromosomes (Figure 4A). The cross between *Hmr*^{3} and NC129 yielded the most uneven allele frequencies, with multiple sharp spikes and dips. We attribute these deviations to a combination of low embryo counts, low read depth causing high sampling variance, and a high degree of heterozygosity and structural variants in the *Hmr*^{3} line. These signals are unlikely to be drive loci because they fail to display the expected frequency elevation at linked sites. As our experiments are mapping across one generation of recombination, we expect that a driver will produce a broad peak of distortion. For a driver that is not contained within an inversion, as recombination breaks the linkage between it and flanking sites, distortion is expected to attenuate with genetic distance, vanishing by 50 cM. Drivers within inversions are expected to show a broader peak, with the shape of the signal further dependent on whether or not recombination is fully suppressed within the inversion.

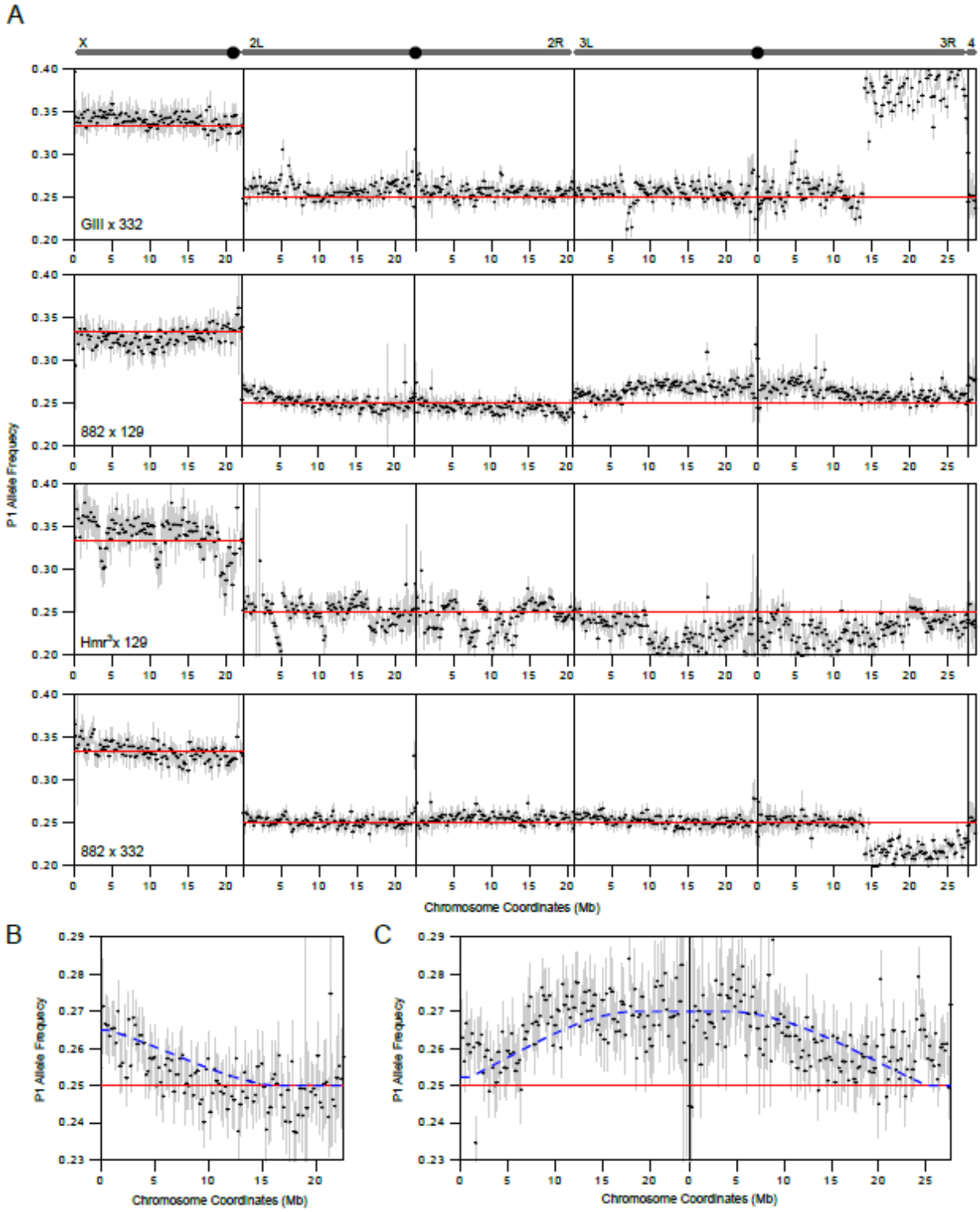

A
X
2L
2R
3L
3R
4
GIII x 332
882 x 129
Hmr³x 129
882 x 332
P1 Allele Frequency
Chromosome Coordinates (Mb)
B
C
P1 Allele Frequency
Chromosome Coordinates (Mb)

Figure 4.4. Allele frequency estimates across chromosomes. (A) Schematics of chromosomes are shown above with centromeres labeled as black circles, except for the 4th chromosome at the far right. For each of the crosses, the frequencies of allele counts at heterozygous sites are averaged in 200kb windows and plotted across all chromosomes. Red horizontal lines mark the Mendelian expectation. Error bars represent 99% confidence based on Poisson distribution of the aggregated P1 count in the window. (B-C) Magnified views of chromosomes 2L and 3 from the NC882 X NC129 cross. The blue dotted line indicates the expected decline in signal of distortion for a telomeric (B) or centromeric (C) drive locus, based on genome-wide recombination rate estimates.

On chromosome 3R of the GIII x NC332 and NC332 x NC882 crosses, we detected distortion spanning over half of the chromosome, starting at around 15 Mb; the allele frequency increased to an average of 0.384 and dropped to 0.217, respectively. The absence of attenuation of the distortion signal suggests that the deviations are not due to drive. As NC332 was used in both crosses and the direction of distortion was opposite in the two crosses, we reasoned that the non-Mendelian ratios are due to a large structural variation or polymorphism in the NC332 line. Using PCR across indel polymorphisms, we found that NC332 is heterozygous at three sites in the distal half of chromosome 3R. We note that this heterozygosity appears to be specific to our stock as it is not present in the published NC332 WGS. Therefore the apparent deviation is caused by unexpected heterozygosity in the parental line rather than non-Mendelian segregation.

Interestingly, we noticed slight elevations of the P1 frequency in the NC882 x NC129 cross at two loci, both of which show a gradual decay. The first is at the telomere of 2L, where the elevated P1 allele frequency attenuates toward the centromere (Figure 4B), and the second is at the centromere of chromosome 3, where the elevated allele frequency decreases on both arms distally (Figure 4C). To determine whether the observed decay of the signals is consistent with recombination breaking the linkage between the drive loci and distal loci, we simulated the decay of the signals based on published estimates of recombination rates across the chromosome arms (Fiston-Lavier *et al.*, 2010), and also determined the magnitude of deviation that best fits the recombination rates. We estimate that the distortion on 2L increases the allele frequency to 0.2605, equating to a 4.2% increase in transmission frequency. The observed allele frequency decays faster than the simulated decay, where the former drops to the Mendelian expectation at around 10 Mb while the latter is at ~16 Mb. One possible explanation is that this discrepancy reflects variation in recombination rate among lines. The centromere distortion on chromosome 3

elevates the allele frequency to 0.2709 equating to a 8.4% increase, and the observed decay fits the simulated decay strikingly well. Because NC882 shows no effect when crossed to NC332, we suspect that the observed deviation is caused by putative drive allele(s) in NC129. Unfortunately, we had little power to test for drive caused by NC129 in the Hmr^3 x NC129 cross, as the data are too inconsistent for reasons discussed above. Because only one chromosome end out of all the crosses between long and short telomere lines displayed deviation consistent with meiotic drive, we conclude that extreme telomere length differences are insufficient to cause biased segregation.

4.3 Discussion

4.3.1 *HeT-A* variation and and truncations

We assessed natural variation in telomere length across the DGRP lines by quantifying relative *HeT-A* copy number using qPCR. As most *HeT-A* elements are truncated at the 5' end, quantification of the 5' ends likely correspond to full length elements. Surprisingly, we observed an enormous 288-fold range between the longest and shortest lines. The abundance is distributed on a logarithmic scale indicating that telomere length increases non-linearly. This is inconsistent with a simple model where *HeT-A* additions occur at a constant rate, and suggests instead that high abundance arrays are more likely to have larger gains. The observed distribution can result from attachment events, if the rate of addition increases as more elements accumulate increasing the copy number of active elements. Alternatively, unequal crossing-over can increase and decrease large blocks of *HeT-A* arrays creating a large range of sizes. In addition, the probability of crossing-over is expected to increase as telomeres become longer.

On the other end of the extreme are the short lines, where we identified an excess of 3' fragments. This may be the result of high rates of incomplete reverse transcription and subsequent transposition of truncated elements. Because of the location of the promoter in the 3' UTR of the adjacent element, 5' truncated elements can still be expressed and subsequently reverse transcribed. Our observation that short lines have excesses of truncations at different regions suggests that specific truncations may be highly active. However, sequence analyses of newly transposed elements found only *HeT-A* copies with full length CDS, arguing that incorporation of large truncations is uncommon (Biessmann *et al.*, 1993; Biessmann *et al.*, 1994). Therefore we suggest an alternative scenario where rates of telomeric deletion are higher in short lines. Chromosome shortening due to the end replication problem is estimated to delete 70-75 bp per generation, which is offset by *HeT-A* incorporation at a rate of ~1 copy per 100 generations (Biessmann *et al.*, 1992). Variation in these rates can produce telomeres of different lengths as well as many 5' truncations after multiple cycles of telomeric deletions and additions. This model also accounts for the observed variation in the region of truncation. For example, in NC852 the rate of loss may be low such that a new addition occurs before terminal deletions reach the 3' of the CDS, resulting in elements that are mostly truncations proximal to the 5' end. In contrast, NC332 and NC882 may have higher rates of deletion, producing mostly elements with more distal truncations. However, the extent to which terminal erosion rate per generation can vary is unclear. Since the rate of loss per replication cycle is determined by the length of the terminal RNA primer which is relatively constant, the observed variation may be dependent on the number of cell cycles per generation. One possible source are the germline stem cells which needs to be renewed though mitosis. In addition to small end erosions, large fragmentation that creates terminal deletions has also been found to occur frequently in *Drosophila* (Golubovsky *et*

al., 2001; Levis, 1989). Altogether, we suspect that multiple mechanisms contribute to the observed distribution of *HeT-A* abundance and truncation patterns.

4.3.2 Massive telomere length variation in *Drosophila* compared to other species

Determining a precise quantitative estimate of telomere length in Drosophila is more challenging than in other eukaryotes because its telomeric repeat sequences are much more complex. Assuming that the lowest line has the minimum of 1 full-length *HeT-A*, the mean across the DGRP lines is 34 copies, and the *y w* line used as reference has 25 copies. These numbers are similar to that of the genome reference line, y^1 ; cn^1 bw^1 sp^1 which was independently estimated to have approximately 29 and 7 full-length copies of *HeT-A* and *TART*, respectively (George *et al.*, 2006), which amounts to 365 kb of *HeT-A* and *TART* sequences in total and 45.6 kb per chromosome end. Based on these numbers, the longest lines we assayed would have over 3.6 Mb of telomeric sequence equating to an average of ~450 kb per chromosome end. The shortest line would have only 12 kb in total and 1.5 kb per chromosome end, but given the large number of truncations, the actual sizes of the telomeres in short lines are likely larger. Because we only sampled *HeT-A*, we cannot rule out the possibility that *TART* and *Tahre* might compensate for low *HeT-A* numbers. However, this is unlikely, as surveys of *HeT-A* and *TART* quantities across several stocks, including GIII, found that increases in *Het-A* abundance are accompanied by increases in *TART* abundance (Siriaco *et al.*, 2002).

The range of estimated sizes that we detected is in stark contrast with organisms that utilize the canonical telomerase mechanism of telomere protection. Telomere lengths have been estimated to be between 5-10 kb in humans, ~1 kb to 9.3 kb in *Arabidopsis thaliana*, 132 to 2400 bp in *C. elegans*, and only <100 to 500 bp in yeast (Fulcher *et al.*, 2014; Liti *et al.*, 2009;

Thompson *et al.*, 2013; Vasa-Nicotera *et al.*, 2005). Furthermore, even interspecific differences such as that between mouse species pale in comparison (Zhu *et al.*, 1998). Notably, in all these organisms the variation is normally distributed rather than an over-dispersed log-normal as in *Drosophila*. These differences are consistent with the idea that the range and distribution found in flies are due to its retrotransposon-based mechanism of telomere evolution. We suspect that the evolution of this mechanism allows telomeres to be more labile in flies. Interestingly, this lability may also present challenges for proper regulation, thereby driving the rapid evolution of multiple components of the capping complex. More stable telomeres, in contrast, may lead to regulation that is highly conserved, as exemplified by the CST and shelterin complexes in yeast and humans, respectively (Raffa *et al.*, 2011).

4.3.3 Identification of biased segregation using whole-genome sequencing

Most known cases of non-Mendelian chromosome segregation involve loci of large effect (Fishman and Saunders, 2008; Larracuente and Presgraves, 2012). The difficulty in genotyping large numbers of individuals to minimize sampling error has limited our ability to identify weak drivers. However, weak drivers are likely to be more prevalent in nature, as they are expected to take longer to fix in the population and may be less prone to having pleiotropic deleterious consequences. Our strategy of genome sequencing pools of individuals has several advantages that allowed us to identify candidate loci of weak drivers. First, by assaying allele frequencies from pools, we are able to sample a large number of individuals, thereby minimizing sampling error and increasing sensitivity to detect distortion. Second, by using whole-genome sequencing, we can scan for drive loci genome-wide as opposed to targeted genotyping which requires a priori knowledge of the driver. Third, the large number of informative heterozygous sites

genome-wide provides information for error and bias correction. Lastly, the contiguous windows of allele frequencies across chromosomes allowed us to detect and fully visualize the attenuation of distortion as predicted from estimates of recombination rate.

A similar strategy has been applied on sperm pools of hybrid male mice to test for segregation distortion (Corbett-Detig *et al.*, 2015). Direct sampling of gametes as opposed to F1s has the major advantage of eliminating viability effects. This strategy may be applicable on sperm isolates in flies (Dorus *et al.*, 2006), but would not work for *Drosophila* oocytes as they maintain their polar bodies until after fertilization. Instead, generating haploid embryos by eliminating the paternal genome may allow for direct sampling of female meiotic segregation (Langley *et al.*, 2011). However, the efficacy will depend on how uniformly the haploid embryos arrest and the number of gynogenetic haploid escapers.

We stress that the expectation of gradual signal attenuation due to recombination plays a pivotal role in qualitative assessment of false positives. Significant deviation from the expected Mendelian ratio can be caused by structural variants like indels and duplications or by unaccounted residual heterozygosity in the parental lines. Both of these will produce alterations to the expected allele frequency, but the resulting distortion will be local, and, importantly, will have sharp edges rather than attenuating gradually.

However, this experiment does not allow us to distinguish distortion caused by meiotic drive from that caused by viability effects arising from polymorphism between the parental lines. In our scheme, the progeny pool contains heterozygotes and homozygotes. Genotypic differences may result in differential viability, embryo size, and/or growth rate, all of which will produce unequal representation of the two genotypes and, consequently, P1 allele frequency that deviates from 0.25. In addition, the heterozygous genotype may experience heterosis or incompatibilities

which will increase or decrease the P1 allele frequencies, respectively. As with a meiotic drive effect, these deviations will attenuate as recombination separates the causal locus from distal loci. The X-chromosome may be particularly susceptible to fitness effects since recessive deleterious alleles will be exposed in hemizygous males. This may explain the slight deviation in the middle of the X (at ~8 Mb) in the NC882 x NC129 cross. To minimize such effects, we selected early embryos that are closely timed (3-4 hr), reasoning that viability differences are likely to be more pronounced later in development. Nevertheless, we cannot eliminate this possibility altogether. We note, though, that the two regions identified are close to the telomere and centromere, consistent with proposed models of centromeric and telomeric meiotic drive.

There are several ways to test the possibility of viability effects. First, instead of backcrossing to P2, the F1 females can be backcrossed to P1 males. If the elevation we observed here is due to heterosis, P1 allele frequency will be lower than expected in this cross. However, if the distortion is due to drivers on the P1 allele, the P1 allele frequency will be higher. Second, F1 sons, instead of daughters, can be mated to P2 females, producing offspring that are genetically identical to those in our crosses. Because true meiotic drive occurs only in oogenesis, the distortion signal should be absent. If, however, the distortion persists, biased female meiosis can be ruled out as the underlying mechanism. The definitive test of meiotic drive versus viability effects would be to genotype and determine the viability of individual embryos.

4.3.4 Potential genomic elements causing meiotic drive

By generating F1s heterozygous for long and short telomeres, we tested the hypothesis that telomere length biases chromosome segregation in meiosis. However, among 4 crosses tested, only one chromosome end in one cross showed allele frequency distortion. Thus, we conclude

that telomere length differences, at least as assayed by *HeT-A* quantities, are insufficient to cause meiotic drive.

However, to our surprise we identified two candidate loci in the cross between NC882 and NC129, at the telomere of 2L and the centromere of chromosome 3. Since our results show that *HeT-A* abundance does not correlate with distortion, other sequences must be considered as candidates for this potential drive locus. Because *HeT-A* and *TART* abundances are highly correlated (George *et al.*, 2006), we suspect that *TART* is unlikely to be the driver. A better candidate may be the telomere-associated sequences (TAS), multiple kilobase arrays of repeats adjacent to the HTT-arrays that are defined as heterochromatic based on their ability to repress expression of inserted transgenes (Karpen and Spradling, 1992). There are several known TAS variants. For example, many lines including the reference strain lack the TAS on XL, and many lab stocks lack either the 2L or 3L TAS (Mason *et al.*, 2004). Moreover, the TAS repeats also display high rates of sequence turnover between species. The rapid evolution and presence/absence polymorphisms are consistent with a meiotic drive model. Interestingly, biased chromosome transmission has been observed in Arabidopsis with artificial terminal truncations (Teo *et al.*, 2011). Although meiotic drive has not yet been specifically tested in this case, the telomeres are intact, which suggests that subtelomeric sequences are important for Mendelian transmission.

The distortion signal on chromosome 3 spans the centromere. The potential for centromeres to accumulate meiotic drivers has been long speculated (for review see Malik, 2009). One clear example comes from a centromere-linked locus that causes non-Mendelian segregation in *Mimulus* hybrids (Fishman and Saunders, 2008). Chmátal et al. (2014) showed that Robertsonian fusions of mouse chromosomes create metacentric centromeres that segregate

into the pronucleus more frequently than the telocentric counterparts. This is mediated by increased attachment of kinetochore proteins at the "stronger" centromere. Differential amounts of repetitive DNA in flies may similarly create centromeres of different strengths. One of the most abundant satellites in *D. melanogaster,* AACATAAGAT, is a candidate for the effect we observed, as it is located in the pericentromeric regions of 2L and 3L. We also recently characterized two population-specific satellites, mapping near the centromeres of chromosomes 2 and 3; interestingly they are population-specific with a global distribution that is inconsistent with the neutral expectation (Wei *et al.*, 2014). The potential effects of these satellites on meiotic segregation can be tested using the method presented here.

4.4 Materials and methods

4.4.1 Estimating *HeT-A* quantities with qPCR

For quantification of the DGRP lines at the 5' CDS, around ten adult flies of mixed sex from each DGRP line were collected from stock vials and flash frozen in liquid nitrogen. Flies were homogenized by using beads and purification steps as per the manufacturer's protocol using Agencourt DNAdvance Genomic DNA Isolation Kit (Beckman Coulter). DNA isolation steps were handled by Biomek 4000 Liquid Handling System (Beckman Coulter robotic system). After purification using columns, DNA was eluted in 50 ul sterile water, concentration was estimated by using a NanoDrop 2000 (Thermo Scientific) and diluted to a concentration of 10 ng/µl using sterile water. 5′TTGTCTTCTCCTCCGTCCACC3′ (forward) and 5′GAGCTGAGATTTTTCTCTATGCTACTG3′ (reverse) were used for qPCR. The quantifications were normalized to quantities of RpS17 amplified by the primers 5′AAGCGCATCTGCGAGGAG3′ (forward) and 5′CCTCCTCCTGCAACTTGATG3′ (reverse).

Real-time PCR was run using ABI Prism 7900 HT Sequence detection system (Applied Biosystems). Each DNA sample was run in triplicate to estimate average Ct values. Mean and standard deviation values of the three replicate reactions were used to estimate the telomere length of each line.

For quantification of the 3' CDS and promoter regions, we collected ~10 1-3 day-old females. DNA was extracted from carcasses with ovaries removed using Puregene Core Kit, and concentration was measured by NanoDrop. Primer sequences for the two regions of *HeT-A* and for RpL32 (rp49) used for normalization were from (Klenov *et al.*, 2007).

4.4.2 Fluorescent *in situ* hybridization on polytene chromosomes

Third instar larvae were dissected in 0.7% NaCl. The salivary glands were separated from the brain and imaginal disks and fixed in freshly made 1.84% paraformaldehyde/45% glacial acetic acid for five min on siliconized coverslips. A frosted glass slide was then applied onto the coverslip and gentle pressure applied to dissociate the cells. The slides were then submerged in liquid nitrogen for at least 10 min with the coverslips then quickly removed with a blade and the slides washed and dehydrated with 95% EtOH. Prior to hybridization the slides were treated with 2X SSC at 70℃ for 30 min, followed by dehydration with 95% EtOH at room temperature for 10 min and air-drying for 5 min. The slides were then submerged in 0.07N NaOH for 3 min, followed by dehydration and air-drying again. Remaining steps of hybridization and washing were as in reference (Larracuente and Ferree, 2015). For the *HeT-A* probe, a 105 bp HeT-A fragment was amplified using the primer pair CGCAAAGACATCTGGAGGACTACC/ TGCCGACCTGCTTGGTATTG and cloned. The vector was used as a template to generate

probes using the PCR Dig Probe Synthesis Kit (Roche). Imaging was carried out with a Zeiss Confocal Microscope and images processed using Zeiss Zen software.

4.4.3 *Drosophila* stocks and crosses, embryo collection, and sequencing

The long-telomere stocks GIII and *Hmr*^{3} are described in (Siriaco *et al.*, 2002) and (Satyaki *et al.*, 2014), respectively. Identity of DGRP lines used for telomere drive crosses was tested using a subset of RFLPs described in (Mackay *et al.*, 2012). All crosses were performed at 25°. F1 females were generated by setting 3 vials each with ~25 virgin females and ~25 males, and flipping vials every 1-2 days for approximately 1 week. We aimed to have approximately 400 males and 400 virgin females aged 3-7 days as parents for each biological replicate to generate BC1 embryos. All vials containing F1 females were kept for several days and monitored for larvae to ensure that they contained only virgins.

Crosses were set with approximately 200 parents in a vial for one or two days, then transferred to an egg-collection cup containing a grape-juice/agar plate supplemented with yeast paste, and kept overnight in the dark. The next day, fresh plates were changed every hour and then aged for 3 hours after collecting. The approximate number of embryos was recorded, embryos rinsed in dH_20 and then dechorionated in 50% bleach for 2 minutes. Embryos were examined under a microscope and bleaching continued as necessary. Any larvae and late stage embryos were removed and embryos rinsed with dH_20. Embryos were then suspended in PBT and transferred to 1.5 ml eppendorf tubes, excess PBT removed, and vials frozen in liquid nitrogen and stored at -80° C. We used the DNAeasy Kit to extract DNA from embryo pools. Libraries were generated using the TruSeq Kit (Illumina) and sequenced on 2 lanes of HiSeq 2500 Rapid Run (100bp single-end reads).

4.4.4 Genome-wide association studies

HeT-A quantities were log-transformed and averaged across replicates. The quantities were uploaded to the DGRP website for GWAS (Huang *et al.*, 2014). The output provides two p-values per genomic locus: one based on a simple regression, and the other a mixed model. Because of the low number of significant hits, we selected sites that have a nominal p-value $< 10^{-5}$ in either one of the two models.

4.4.5 Identification of heterozygous sites

Sequences were aligned to *D. melanogaster* reference (r5.46) using BWA on standard settings. Using Samtools, we sorted, removed PCR duplicates, and merged all samples into one file. We applied the GATK package following the recommended practices (see https://www.broadinstitute.org/gatk/), up to variant calling with HaplotypeCaller. All subsequent steps were carried out by custom Perl scripts. To determine expected heterozygous sites from the parental lines, we identified single nucleotide homozygous sites that differ between the parents. To determine heterozygous sites from the crosses, we filtered for single nucleotide heterozygous sites with genotype quality > 20 (PHRED scale). This set of sites was then polarized using the parental set. Only heterozygous sites found in both sets are used for allele frequency quantification.

4.4.6 Allele frequency quantification and reference bias correction

Read counts for each of the two alleles at heterozygous sites are inferred from the AD and DP fields in the vcf files. For each read depth bin i, we determined the average P1 allele frequency

when it is the reference ($FreqR_i$) and when it is the non-reference allele ($FreqN_i$). The observed allele frequencies can be modeled as: $FreqR_i = \frac{p_i n_i}{p_i n_i + q_i}$ and $FreqN_i = \frac{p_i}{p_i + q n_i}$, where n is the factor to which the reference allele is elevated from the non-reference allele. p and q are the P1 and P2 counts in the absence of bias, respectively. Based on the two equations, it can be determined that: $n_i = \sqrt{\frac{FreqR_i(1-FreqN_i)}{FreqN_i(1-FreqR_i)}}$. The observed P1 and P2 counts are divided by n to determine p and q, when they are the respective reference alleles. This correction was applied to sites with read depth less than twice the average read depth. Across multiple sites within a 200 kb, the counts of alleles from the same parent are summed.

4.4.7 Over-dispersion modeling

For each read depth bin, n, binomial and Poisson distributions were fitted with a mean of np_n, where p is the average P1 frequency at the read depth. To fit the beta-binomial distribution, we used the R package vgam to infer the mean, np_n, and shape parameter, ρ_n, for each read depth. Beta-binomial distributions were then generated for each read depth with mean of np_n and $\bar{\rho}$, the average of $\rho_n$across read depths.

Recombination rate estimates and distortion signal decay

The formulae for genetic distance measured in cM as a function of physical distance measured in Mb for each chromosomes were taken from the *Drosophila melanogaster* recombination rate calculator [FistonLavier 2010]. The genetic distance (G) from the distortion loci was calculated in 200 kb windows, with windows >50 cM set to 50 cM. For each window (k) away from the distortion loci on the P1 chromosome, the proportion (p) of oocytes carrying the P1 allele can be

calculated as: $p_k = D(1 - G_k/100) + (1 - D)(G_k/100)$, where D is the proportion of P1 alleles at the distortion locus in the oocyte pool, calculated as twice the P1 allele frequency at the distortion loci on autosomes and 3/2 for loci on the X. The P1 allele frequency at the driver was determined using a least-squares approach, where multiple decay curves were generated with incremental changes (0.0001) in P1 allele frequencies; the P1 allele frequency/curve of best fit was selected. The left and right sides of the addition correspond to the proportions of gametes where the P1 alleles are linked with the distortion locus (i.e. no recombination happened in between) and the P1 alleles are not linked with the distortion locus (i.e. recombination happened in between), respectively. The expected P1 allele frequency is then $\frac{1}{2}p_k$ for autosomal sites and $\frac{2}{3}p_k$ for X-linked sites.

PCR genotyping

We identified 20-45 bp indel polymorphisms from the VCF files that distinguish NC332 from GIII across chromosome 3R. We designed two PCR primer pairs flanking indels outside the distortion region: TTTCCGTGTTTTGTTTCTCATCG (forward) /TGTTGTTCTTGTTGTTGTTGTCA (reverse) at 3R 7,215,009 and TGATGTTGATGAGCGCACAG/AAATGCTGTCACACGCTTTG at 3R 6,883,792. We also designed three pairs flanking indels inside the distortion region: CCAGGTGGGTACTCAATAGATTT/CTGTTGGAAATGGAGGTGAGAA at 3R 22,400,780, GGCTCTGGGCCATGTCAATA/GTGCGTGTTGGCCTGTTAAT at 3R 23,929,097, and CTGGGGAGTAGCACGTTTCC/GATGTGGATGTGGCTGTGGA at 24,917,176.

CHAPTER 5

LIMITED GENE MISREGULATION IS EXACERBATED BY ALLELE-SPECIFIC UP-REGULATION IN LETHAL HYBRIDS BETWEEN *DROSOPHILA MELANOGASTER* AND *D. SIMULANS*

5.1 Introduction

Interspecific hybrids have a wide range of fitness values and display diverse phenotypes. In plants, hybridization can be important in adaptation and speciation (Rieseberg and Willis 2007; Soltis and Soltis 2009). Animal hybrids, however, often display deleterious incompatibility phenotypes such as sterility and lethality (reviewed in Maheshwari and Barbash 2011). Hybrids with *Drosophila melanogaster* have been intensively studied due their ease of manipulation and suite of incompatibilities (Barbash 2010). F1 daughters of *D. melanogaster* females crossed to *D. simulans* males are fully viable at lower temperatures, progressively lethal at higher temperatures, and are completely sterile. The F1 sons are fully lethal, dying as larvae. Strikingly, all of these incompatibility phenotypes can be at least partially suppressed by mutations in the X-linked *D. melanogaster* gene *Hybrid male rescue* (*Hmr*).

Hmr encodes a putative DNA or chromatin binding protein that localizes to heterochromatin and represses expression of selfish DNA repeats including transposable elements and satellite DNAs (Thomae et al. 2013; Satyaki et al. 2014). Population genetic studies demonstrate that *Hmr* has diverged under positive selection in both *D. melanogaster* and *D. simulans* (Barbash et al. 2004). Antagonistic interactions with selfish elements are therefore a potential cause of the rapid evolution of *Hmr*. However, it is unclear how this heterochromatic

function in pure species relates to hybrid lethal activity. In fact, whether *Hmr* function is the same in parental species and their F1 hybrids remains a long-standing question in hybrid incompatibility studies (Maheshwari and Barbash 2011).

Gene expression analyses have been applied to *D. melanogaster*/*D. simulans* and other hybrids with distinct objectives in mind. One is to gain a better understanding of the physiological and developmental differences between hybrids and parental species. For example, *D. melanogaster*/*D. simulans* female hybrids show a reduced expression of female-biased genes, consistent with their lacking functional ovaries (Ranz et al. 2004). Other studies have found down-regulation of male-biased genes in sterile hybrid sons (Michalak and Noor 2003; Haerty and Singh 2006; Moehring et al. 2007; Llopart 2012). A second goal has been to identify genes that significantly differ from parental levels, in order to find candidate genes that have undergone regulatory divergence (Ranz et al. 2004; Landry et al. 2005; Lai et al. 2006; Renaut et al. 2009) . The ability to quantitate allele-specific expression in hybrids further allows one to distinguish the relative roles of *cis*- and *trans*-regulatory evolution (Wittkopp et al. 2004; Graze et al. 2009; McManus et al. 2010).

In addition, several gene expression analyses have attempted to identify specific regulatory pathways whose misexpression causes or contributes to hybrid incompatibility. Hybrid male sterility has been extensively studied, and has revealed that sterility can often occur with only few genes misregulated, especially for hybrids of closely related species in the *D. simulans* species complex (Michalak and Noor 2003; Haerty and Singh 2006; Moehring et al. 2007). Hybrid lethality is less well-characterized by genome-wide approaches. To address the extent of misregulation specifically associated with hybrid lethality, a microarray study compared gene expression in lethal and viable *D. melanogaster/D. simulans* hybrids, and found

only a handful of genes different between the two (Barbash and Lorigan 2007). However, because hybrids can have gross physiological defects, rarely can these studies disentangle direct misregulation from effects that are downstream consequences of the incompatibility phenotype.

Here we analyze RNA-Seq data in *D. melanogaster*/*D. simulans* hybrid males in order to contrast the role of *Hmr* in pure-species and hybrids, to determine the expression differences in hybrids that are associated with *Hmr*-induced lethality, and to characterize gene expression differences that are inherent to hybrids irrespective of lethality. First, we identify genes differentially expressed in *Hmr* mutants relative to wildtype *D. melanogaster* male larvae (*mel-Hmr*$^{-}$ and *mel-Hmr*$^{+}$, respectively) and genes differentially expressed between viable *Hmr*$^{-}$ and lethal *Hmr*$^{+}$ hybrid larvae (*hyb-Hmr*$^{-}$ and *hyb-Hmr*$^{+}$, respectively). Comparing these gene lists allows us to determine the functional overlap of *Hmr* in the *D. melanogaster* and hybrid backgrounds. Second, we distinguish the expression of the *D. melanogaster* and *D. simulans* alleles in hybrids, in order to determine allele-specific differences. Third, we evaluate the extent to which developmental differences contribute to the observed expression differences between hybrids. Fourth, we identify regulatory differences in putative *Hmr* targets between the species. Finally, we integrate the hybrid analyses with studies of gene expression divergence between *D. melanogaster* and *D. simulans*. We find that 1) *Hmr* does not maintain pure-species function in hybrids, 2) viable (*Hmr*$^{-}$) hybrids are associated with an excess of genes where only the *D. melanogaster* allele is down-regulated, and 3) that after accounting for development and tissue differences and intraspecific divergence, the number of genes in hybrids that are expressed outside of the parental range is surprisingly low.

5.2 Results

5.2.1 *Hmr* has different effects on *D. melanogaster* and *hybrid* genomes

To identify genes regulated by *Hmr* in *D. melanogaster*, we compared the transcript level between *mel-Hmr*$^{-}$ and *mel-Hmr*$^{+}$ male larvae. While very few genes (n = 40) are significantly different, they show a strong bias with 30 up-regulated and only 10 down-regulated in *mel Hmr*$^{-}$ (Figure 5.1A). The extent of up-regulation ranges from 1.53-fold to 8.10-fold. In this set of 40 candidate *Hmr* targets, genes located in pericentric heterochromatin are significantly over-represented when compared to the whole genome (9/40 = 22.5%; FET p = 6.40e-4). With the exception of one, all of these heterochromatic genes are up-regulated in *mel-Hmr*$^{-}$. These results differ from analyses in ovaries, where heterochromatic genes are slightly down-regulated in *Hmr* mutants (Satyaki et al. 2014). It is unclear whether this is because of sex-specific differences and/or differences between germline and somatic gene regulation. We note though that both studies agree on the relatively weak effect *Hmr* has on protein-coding genes, in contrast to its major effects on transposable element expression.

A strikingly different pattern is observed when we compared *hyb-Hmr*$^{+}$ to *hyb-Hmr*$^{-}$ male larvae. Not only are many more genes differentially regulated (n = 622) between the hybrids than between pure-species, the opposite bias is observed: significantly fewer genes are up-regulated than down-regulated in *hyb-Hmr*$^{-}$ (210 up-regulated versus 412 down-regulated, p = 8.383e-09 FET; Figure 5.1B). Furthermore only 23 (3.70%) of the differentially regulated genes are heterochromatic, which is not significantly different than the genome-wide proportion of heterochromatic genes (FET, p = 0.1143). Barbash and Lorigan (2007) used microarrays to compare expression of the same two hybrid genotypes, but at an earlier developmental time. They identified a substantially smaller set of 91 genes differentially regulated between the

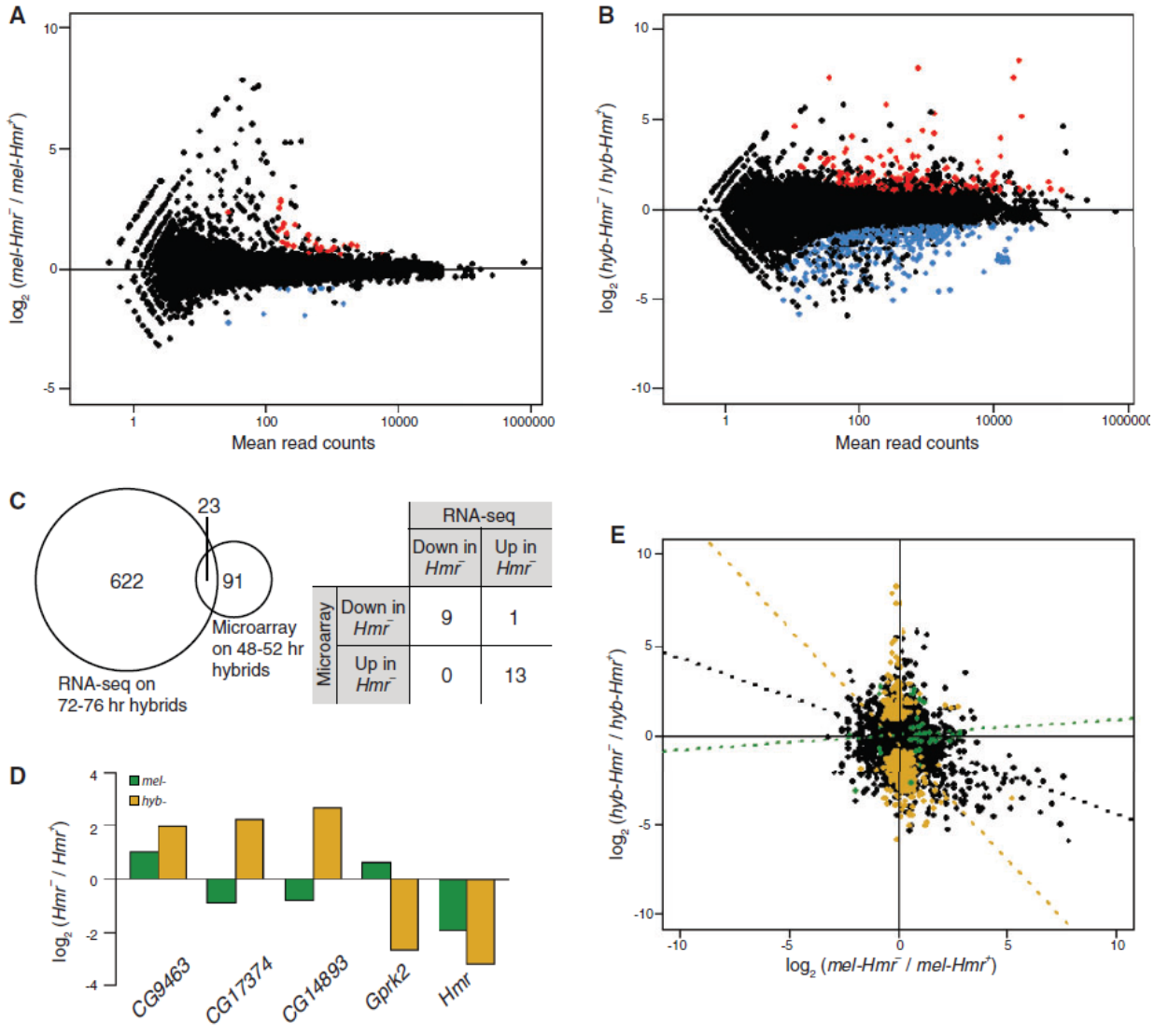

A
log2 (mel-Hmr⁻ / mel-Hmr⁺)
Mean read counts
B
log2 (hyb-Hmr⁻ / hyb-Hmr⁺)
Mean read counts
C
622
23
91
Microarray on 48-52 hr hybrids
RNA-seq on 72-76 hr hybrids
RNA-seq
Down in Hmr⁻
Up in Hmr⁻
Microarray
Down in Hmr⁻
9
1
Up in Hmr⁻
0
13
D
mel-
hyb-
log2 (Hmr⁻ / Hmr⁺)
CG9463
CG17374
CG14893
Gprk2
Hmr
E
log2 (hyb-Hmr⁻ / hyb-Hmr⁺)
log2 (mel-Hmr⁻ / mel-Hmr⁺)

Figure 5.1. Differential expression between *Hmr*$^{+}$ and *Hmr*$^{-}$ in *D. melanogaster* and hybrid male larvae. (A) Each dot represents the $\log_2$ of the fold-difference in expression of each gene in *mel-Hmr*$^{-}$ relative to *mel-Hmr*$^{+}$, plotted against the expression level calculated as mean number of reads. Red indicates genes significantly up-regulated, blue significantly down-regulated in *mel-Hmr*$^{-}$. (B) Same analysis as in part A, for *hyb-Hmr*$^{-}$ relative to *hyb-Hmr*$^{+}$. (C) Left, overlap of differentially regulated genes between *hyb-Hmr*$^{-}$ and *hyb-Hmr*$^{+}$ identified in this RNA-Seq study and in a previous microarray study (Barbash and Lorigan 2007). Right, comparison of the direction of change for the 23 overlapping genes. (D) Comparison of the direction of change for the 5 genes that are significant in both Figure 1A and 1B. Green, *mel-Hmr*$^{-}$ relative to *mel-Hmr*$^{+}$; yellow, *hyb-Hmr*$^{-}$ relative to *hyb-Hmr*$^{+}$. (E) Correlations between fold-differences of the *mel* (fig. 1A) and *hyb* (fig. 1B) comparisons for the whole transcriptome (black, Pearson's r = -0.257), for genes significantly mis-regulated in *mel-Hmr*$^{-}$ from fig. 1A (green, r = 0.079), and for genes significantly different between hybrids from fig. 1B (yellow, r = -0.310).

hybrids. Although only 23 are shared between the datasets, the microarray and our experiments agree on the direction of expression difference, with the exception of one gene (Figure 5.1C). Gene ontology terms identified here include chitin and cuticle-related processes (Table 5.1).

Expression of X-linked genes is especially susceptible to misregulation in hybrid males as a result of hemizygosity and potential breakdown of dosage compensation (Laurie 1997; Orr 1997). If directly related to hybrid lethality, then one might expect over-representation of X-linked genes in comparisons between hybrid genotypes. We observed no enrichment of X-linked gene mis-expression, however, consistent with the previous microarray study (Barbash and Lorigan 2007). In fact, of the genes down-regulated in *hyb-Hmr*$^{-}$, the number of X-linked genes is significantly underrepresented (n = 24, FET, $p < 0.0001$).

Hybrid incompatibility may arise either through *Hmr* regulating the same targets in hybrids that it normally regulates within *D. melanogaster*, or alternatively, by regulating a distinct set of genes in hybrids. To distinguish between the two possibilities, we looked for genes that are differentially regulated in both *mel-Hmr*$^{-}$ and *hyb-Hmr*$^{-}$, when compared to their respective *Hmr*$^{+}$ samples. Strikingly, only 4 genes (other than *Hmr*) overlap and only one of them (*CG9463*) shows the same direction of change (Figure 5.1D). Moreover, the fold-changes between the hybrid and *D. melanogaster* comparisons are not correlated, either for the differentially regulated genes or genome-wide (Figure 5.1E). *Hmr* therefore has essentially non-overlapping effects on the hybrid versus *D. melanogaster* genomes, strongly suggesting that *Hmr* has neomorphic function in the hybrid genetic background.

5.2.2 The *D. melanogaster* alleles are more sensitive to the presence of *Hmr*

Because *Hmr* has rapidly evolved in the two lineages (Barbash et al. 2004), *Hmr*, its co-factors, and its targets may have co-evolved separately in *D. melanogaster* and *D. simulans*.

Table 5.1. Significant gene ontology classes of differentially expressed genes between hybrids

Biological Processes	No.	*Genes*
chitin metabolic process, amino sugar metabolic process, glucosamine-containing compound metabolic process, aminoglycan metabolic process	19	*Mur2B, CG8192, obst-G, CG10154, CG10140, obst-F, CG7298, CG6996, CG14608, Cht5, CG14880, CG7714, CG7715, CG6403, CG13643, Muc26B, Muc96D, CG32284, obst-H*
cellular anion homeostasis, cellular acyl-CoA homeostasis, anion homeostasis	5	*CG8814,CG8629,CG15829,CG5804,CG8628*
Cellular Components		
extracellular region	62	*Argk, Gld, Lsp1alpha, Lsp2, y, Drs, PebIII, pnut, cher, Ag5r, Lcp65Ag1, Mur2B, TotA, prc, CG9400, CG5177, spz3, CG8192, CG5550, yellow-d2, Ilp2, obst-G, CG10154, CG10140, Ilp8, obst-F, CG7298, CG6996, Spn77Bc, CG11131, CG14608, beat-Vb, Cht5, CG14880, CG7714, CG7715, burs, Ccap, CG13618, CG11852, TwdlP, TwdlO, TwdlN, CG6403, CG14258, Obp99a, Obp99d, Obp99b, capa, CG13643, Muc26B, Peritrophin-15a, yellow-e, TotC, Muc96D, CG32284, QC, TwdlX, Twdlalpha, NLaz, obst-H, Cpr47Eg*
contractile fiber	8	*Actn, Mp20, Prm, Mlp84B, Pglym78, Scgdelta, CG14207, Mhc*
contractile fiber part, myofibril	7	*Actn, Prm, Mlp84B, Pglym78, Scgdelta, CG14207, Mhc,*
extracellular matrix	11	*Mur2B, prc, CG5550, Muc68Ca, TwdlP, TwdlO, TwdlN, Muc26B, Muc96D, TwdlX, Twdlalpha*
sarcomere	6	*Actn, Prm, Mlp84B, Pglym78, CG14207, Mhc*

Molecular Processes		
structural constituent of chitin-based cuticle,structural constituent of cuticle	32	*Lcp4, Edg91, Ccp84Ag, Ccp84Ab, Ccp84Aa, Lcp65Ag2, Lcp65Ag1, Lcp65Af, Lcp65Ac, Acp65Aa, Cpr49Ag, Cpr49Ah, Cpr50Cb, Cpr56F, Cpr62Bb, Cpr62Bc, Cpr64Ad, Cpr65Ec, Cpr67Fa2, Cpr73D, Cpr76Bb, Cpr78Cb, TwdlP, TwdlO, TwdlN, Cpr100A, Cpr65Aw, TwdlX, Twdlalpha, Cpr31A, Cpr47Eg, Lcp65Ag3,*
chitin binding, carbohydrate derivative binding	20	*Mur2B, CG8192, obst-G, CG10154, CG10140, obst-F, CG7298, CG6996, CG14608, Cht5, CG14880, CG7714, CG7715, CG6403, CG13643, Muc26B, Peritrophin-15a, Muc96D, CG32284, obst-H,*
diazepam binding	5	*CG8814, CG8629, CG15829, CG5804, CG8628,*
structural molecule activity	48	*Act87E, Lcp4, Prm, Edg91, Ccp84Ag, Ccp84Ab, Ccp84Aa, Mlp84B, Lcp65Ag2, Lcp65Ag1, Lcp65Af, Lcp65Ac, Acp65Aa, Mur2B, Scgdelta, CG1368, Cpr49Ag, Cpr49Ah, Cpr50Cb, Cpr56F, Cpr62Bb, Cpr62Bc, Cpr64Ad, Cpr65Ec, Cpr67Fa2, Muc68Ca, obst-G, CG10154, Cpr73D, Cpr76Bb, obst-F, CG7298, CG6996, Cpr78Cb, RpL24-like, TwdlP, TwdlO, TwdlN, Cpr100A, Muc26B, Muc96D, Cpr65Aw, TwdlX, Twdlalpha, Cpr31A, Cpr47Eg, Lcp65Ag3, Mhc,*
structural constituent of chitin-based larval cuticle	10	*Lcp4, Ccp84Ag, Ccp84Ab, Ccp84Aa, Lcp65Ag2, Lcp65Ag1, Lcp65Af, Lcp65Ac, Cpr56F, Cpr64Ad,*
fatty-acyl-CoA binding	5	*CG8814, CG8629, CG15829, CG5804, CG8628,*

Therefore, the observed hybrid differences may reflect differential regulation of the *D. melanogaster* and *D. simulans* alleles. As hybrid males contain only *Hmr* from *D. melanogaster* since it is X-linked, one might expect the presence or absence of *Hmr* to affect the regulation of *D. melanogaster* autosomal alleles (A_{mel}) more than *D. simulans* autosomal alleles (A_{sim}) in hybrids. To test this possibility, we determined the species-of-origin for each read, allowing us to distinguish between the expression level of A_{mel} and A_{sim} at all sites with orthology in the genome assemblies of the two species (see Material and Methods). We then compared the allele-specific expression of autosomal genes and looked for those where only one of the two alleles is differentially regulated between *hyb-Hmr*$^{+}$ and *hyb-Hmr*$^{-}$ (Figure 5.2A and B). For the majority of genes different between hybrids, both alleles are affected. Strikingly, however, of the genes down-regulated in *hyb-Hmr*$^{-}$, A_{mel}-specific regulation is in significant excess; 177 genes have A_{mel}-specific down-regulation, while only 76 have A_{sim}-specific down-regulation ($p = 0.0012$, FET; Figure 2B). In contrast, of genes up-regulated in *hyb-Hmr*$^{-}$, similar numbers show up-regulation of only A_{mel} or only A_{sim} ($n = 58$ and 46 respectively; $p = 0.487$). The specificity of the effect on A_{mel} strongly suggests that allele-specific effects are not caused by lethality or developmental differences in hybrids.

5.2.3 Variable growth rates account for a large portion of the differential expression in hybrids

Both lethal and viable hybrid males develop more slowly than the pure species, but lethal hybrids have an even slower growth rate than viable hybrids (Bolkan et al. 2007). Pure-species are at 3rd larval instar at 72-76 hours after egg laying (AEL), but the developmental progress of hybrids is less clear. To assess the developmental stage of the hybrids and how much

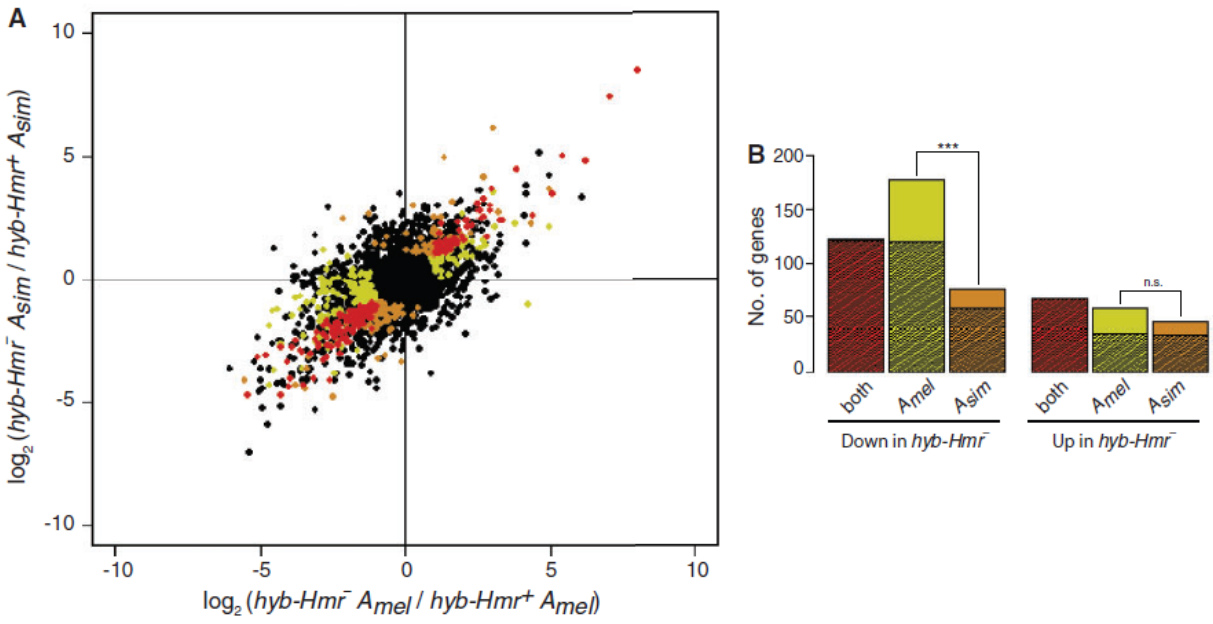
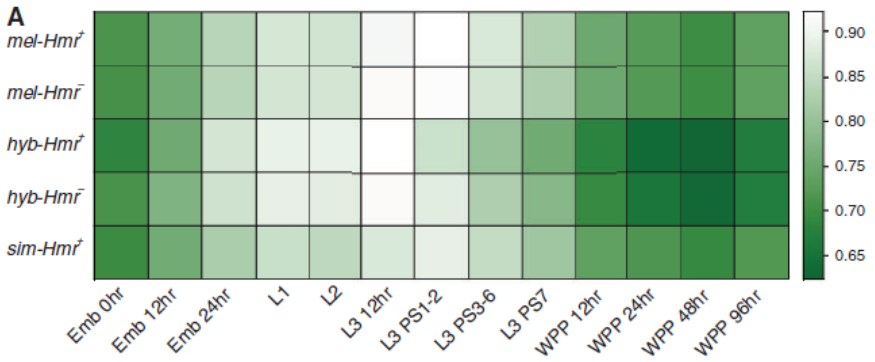


Figure 5.2. Allele-specific regulation in hybrids. (A) A_{mel}-specific expression difference between the hybrids plotted against A_{sim}-specific expression difference. Genes where both alleles are significantly different are indicated in red. Genes where only A_{mel} significantly differs are marked yellow and only A_{sim} significantly differs are marked orange. (B) Number of significant genes from A are plotted. Hatched areas represent the genes deemed different between the hybrids regardless of allele-specific expression (that is, differentially expressed genes in Figure 5.1B).

developmental differences contribute to the observed differential expression, we utilized modENCODE RNA-Seq datasets from multiple developmental stages of *D. melanogaster* (Graveley et al. 2011). Although the datasets are from different genetic backgrounds, we reasoned that the transcriptomes of our samples should be most similar to the modENCODE sample of the closest developmental stage. As expected, *mel-Hmr*$^{+}$, *mel-Hmr*$^{-}$, and *sim-Hmr*$^{+}$ are most highly correlated with the second L3 stage, termed puff stages 1 and 2 (L3.PS1-2; Figure 5.3A). In contrast, both hybrids are most highly correlated with the first L3 stage, collected at 12 hours post-molting (L3.12hr), demonstrating a slower growth rate. Notably, *hyb-Hmr*$^{-}$ is slightly but significantly more similar to L3.PS1-2 than *hyb-Hmr*$^{+}$ is, indicating an additional developmental delay in the lethal genotype (Figure 5.3A, 5.4). We conclude that although both hybrids reach L3 at 72-76 hours (AEL), *hyb-Hmr*$^{-}$ is approaching L3.PS1-2, while *hyb-Hmr*$^{+}$ has only just entered L3.

To identify genes differentially expressed due to the discrepant developmental progress, we looked for overlap between the 622 genes different between *hyb-Hmr*$^{+}$ and *hyb-Hmr*$^{-}$, and those different between L3.12hr and L3.PS1-2 in the ModENCODE data (Figure 5.3B). Of the 412 genes down-regulated in *hyb-Hmr*$^{-}$, 196 (47.6%) are down-regulated in L3.PS1-2. Reciprocally, of the 210 up-regulated in *hyb-Hmr*$^{-}$, 68 (32.4%) are up-regulated in L3.PS1-2. Both numbers are significantly higher than the random expectation ($p < 2.2e-16$ and $p = 0.02053$, respectively). These numbers are likely underestimates, since as noted above, *hyb-Hmr*$^{+}$ and *hyb-Hmr*$^{-}$ do not precisely match L3.12hr and L3.PS1-2, respectively. These results demonstrate that a large proportion of the expression differences between the two hybrid genotypes is due to the growth delay in the dying *hyb-Hmr*$^{+}$. Consistent with this conclusion is the fact that over-represented GO terms include those associated with differences in larval stages (Table 5.1).



B

	Down in *hyb-Hmr⁻*	Up in *hyb-Hmr⁻*
Down in L3 PS1-2	196*	16
Up in L3 PS1-2	24	68*

Figure 5.3. Developmental staging of samples based on ModENCODE data sets. (A) Colour in the heatmap represents correlation (Spearman's ρ) of the samples to modENCODE developmental stages including embryos (Emb); 1st, 2nd, and 3rd larval instars (L1, L2, L3); puff stages of L3 (PS); and white pre-pupae (WPP). (B) 304 genes are different between hybrids (from Figure 5.1B) as well as between L3.12hr and L3.PS1-2. For these genes, the directions of change in the two comparisons are summarized. Numbers that are significantly higher than the genomic average are labeled (* = $p < 0.05$, *** = $p < 0.0001$; FET).

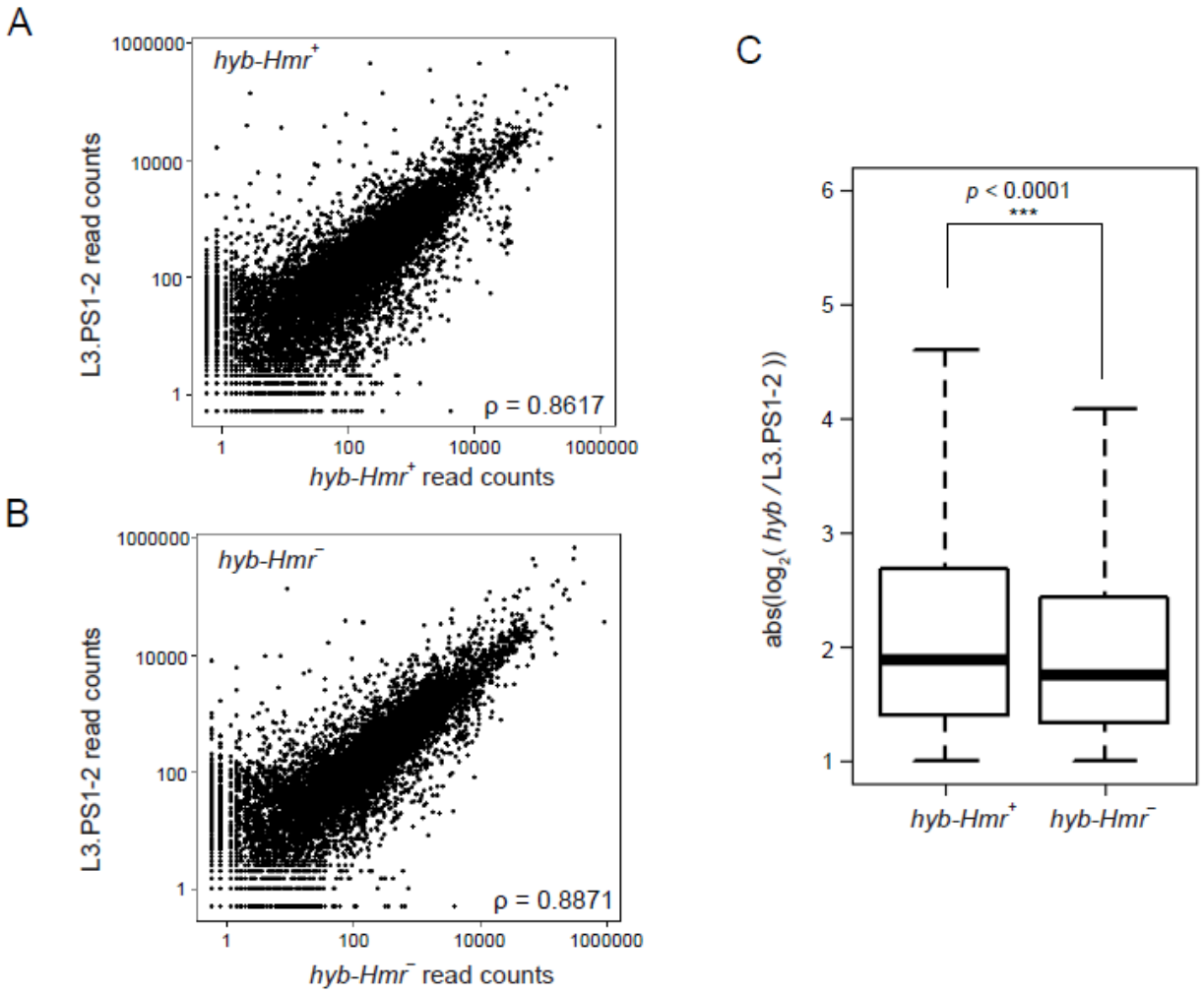


Figure 5.4. Developmental delay in hybrid larvae. A. Correlation between *hyb-Hmr*+ and modENCODE stage L3.PS1-2. B. Correlation between *hyb-Hmr*− and L3.PS1-2. C. Distribution of absolute distance between hybrids and L3.PS1-2. P-value determined by Welch's t-test.

Since *hyb-Hmr*$^{-}$ is closer developmentally to the parents than *hyb-Hmr*$^{+}$ is, we expected gene expression in *hyb-Hmr*$^{-}$ to reflect this similarity. Indeed, the average expression difference between *mel-Hmr*$^{+}$ and *hyb-Hmr*$^{+}$ is significantly greater than between *mel-Hmr*$^{+}$ and *hyb-Hmr*$^{-}$ (Wilcoxon rank sum test, $p < 2.2e\text{-}16$; Figure 5.5A top). This is further exaggerated when we looked at only those genes different between the hybrids (Wilcoxon rank sum test, $p < 2.2e\text{-}16$; Figure 5.5A bottom).

Additivity is a simple null hypothesis where hybrid gene expression is expected to be the average of the two parental species. A potential explanation of differential gene expression between the hybrid genotypes is that one hybrid class deviates from additivity. We therefore analyzed all the differentially expressed genes relative to the additive expectation estimated by averaging expression levels of *mel-Hmr*$^{+}$ and *sim-Hmr*$^{+}$ for autosomal genes and *mel-Hmr*$^{+}$ only for X-linked genes (see Materials and Methods). For the majority of the 412 genes down-regulated in *hyb-Hmr*$^{-}$, the expression in *hyb-Hmr*$^{-}$ is closer to the additive expectation than is the expression in *hyb-Hmr*$^{+}$ (Figure 5.5B and 5.5D). Similarly, the 210 genes up-regulated in *hyb-Hmr*$^{-}$ show expression in *hyb-Hmr*$^{-}$ that more closely reflects additivity (Figure 5.5C and 5.5D). These results clearly indicate that *hyb-Hmr*$^{+}$ deviates from the expected additive level substantially more than *hyb-Hmr*$^{-}$. This departure from additivity in *hyb-Hmr*$^{+}$ is consistent with the developmental differences documented above, and more generally, these results recapitulate that a major cause of differential expression between hybrids is the dissimilarity of *hyb-Hmr*$^{+}$. In addition, we note that for 177 genes, one hybrid genotype is under-dominant while the other is over-dominant ("inter." in Figure 5.5D), suggesting that there are expression differences in both hybrids, but of opposite directions.

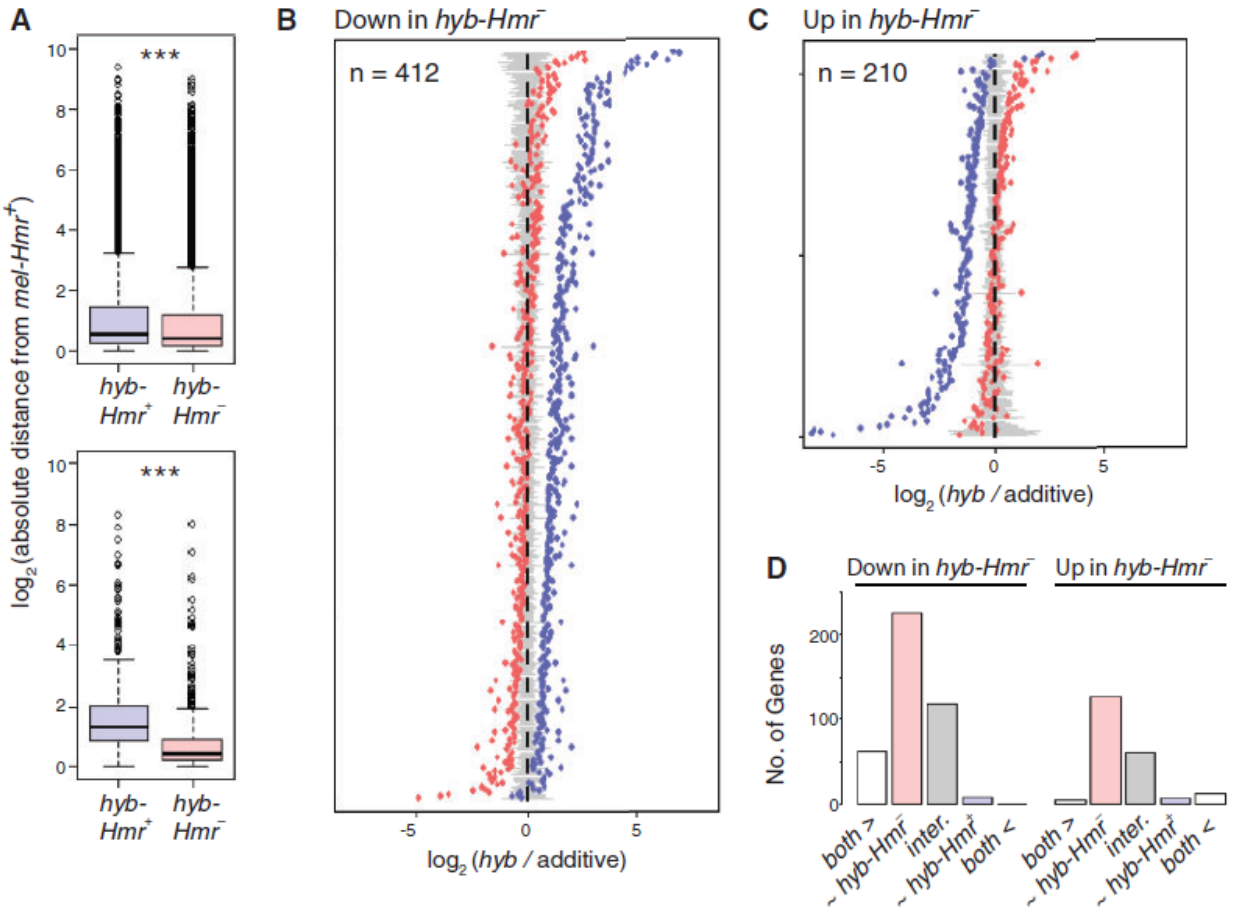
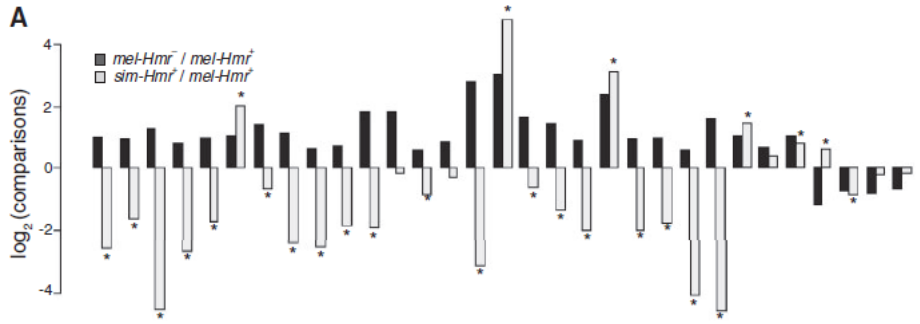


Figure 5.5. Expression of hybrids relative to additive expression. (A) The distribution of absolute expression difference when *mel-Hmr*$^+$ is compared to *hyb-Hmr*$^+$ (indigo) and to *hyb-Hmr*$^-$ (pink) for all genes (top), and for only those genes different between the hybrids (bottom). In both sets, *hyb-Hmr*$^+$ is significantly more different to *mel-Hmr*$^+$ than *hyb-Hmr*$^-$ is. *** = $p <$ 0.0001. (B) For each gene down-regulated in *hyb-Hmr*$^-$, its expression in *hyb-Hmr*$^+$ (indigo) and *hyb-Hmr*$^-$ (red) is plotted relative to the additive expectation centered at 0 in a row. Grey bars around 0 demarcate the range where the expression in hybrids is similar to additivity, set as ± 1/4 of the difference between hybrids for each gene. Genes are ordered based on the difference between the additive level and the average expression of the hybrids. (C) Same as B but for genes up-regulated in *hyb-Hmr*$^-$. (D) Based on B and C, genes are placed into categories according to their expression relative to the additive level: both hybrids are over-dominant ("both >") or both under-dominant ("both <"); one hybrid genotype is within the additive range ("~*hyb-Hmr*$^-$" or "~*hyb-Hmr*$^+$"); or one hybrid is under-dominant while the other is over-dominant ("inter.").

5.2.4 Candidate *Hmr* targets are highly repressed in *D. simulans*

The rapid divergence of *Hmr* between *D. melanogaster* and *D. simulans* raises the question of whether its functions in the respective species have also diverged. Direct examination of *Hmr* function is not possible in *D. simulans* because no mutants are available. However, if *Hmr* has diverged in function, we would expect differences in the regulation of its targets. Therefore, we compared between species the expression of the 40 candidate *Hmr* targets we identified in Figure 5.1A. To guard against mapping biases skewing the results in favor of the better annotated *D. melanogaster* genome, we included only genic sequences with unambiguous orthology between the species (see Materials and Methods), which reduced the number of candidates to 31. Strikingly, almost all (n = 26, 83.9%) show significant interspecific differences (Figure 5.6A), revealing a significant over-representation of genes with divergent expression relative to the entire genome (p = 4.09e-06, FET). Of particular interest, 20 of the 26 genes are expressed lower in *sim-Hmr*$^{+}$ than *mel-Hmr*$^{+}$, 19 of which are repressive targets of *D. melanogaster Hmr* (Figure 5.6B). These results suggest that genes repressed by *Hmr* in *D. melanogaster* experience an even stronger silencing in *D. simulans*. Five of the genes repressed by *Hmr* are expressed higher in *D. simulans*, potentially representing genes that are no longer under *Hmr* regulation. We conclude that expression of *Hmr* target genes has drastically diverged between the species, likely as a consequence of *Hmr*'s rapid divergence, and that the repression of many of these genes has either weakened in the *D. melanogaster* lineage or strengthened in the *D. simulans* lineage.



B

	Down in *mel-Hmr⁻*	Up in *mel-Hmr⁻*
Down in *sim-Hmr⁺*	1	19
Up in *sim-Hmr⁺*	1	5

Figure 5.6. Expression of candidate *Hmr* targets in *D. simulans*. (A) Expression differences of *mel-Hmr*$^{-}$ versus *mel-Hmr*$^{+}$ (black) and *sim-Hmr*$^{+}$ versus *mel-Hmr*$^{+}$ (grey) are plotted for 31 of the candidate *Hmr* targets from Figure 5.1A. Genes with significant interspecific differences are labeled by *. (B) Summary of the direction of change for genes in A with significant interspecific differences.

5.2.5 Hybrid-specific expression differences are limited

To determine expression differences specific to hybrids and not associated with hybrid lethality, we identified 5675 genes that differ in expression by less than 1.3-fold between *hyb-Hmr*$^{-}$ and *hyb-Hmr*$^{+}$ (and are not significant in Figure 1B). We then analyzed these genes relative to the parental species in three different approaches (Table 5.2). First, we compared expression of these genes to each of the parents. When compared to *mel-Hmr*$^{+}$, 974 (17.2% of total) genes have hybrid-specific differences; 663 (68.1% of total) are up-regulated in hybrids while only 311 (31.9%) are down-regulated. More genes (1238) are identified when compared to *sim-Hmr*$^{+}$ but the bias between up and down-regulated is reduced. Second, we compared these genes to the additive expectation. The bias persists but the number of genes differentially regulated is substantially reduced to only 513 (9%). This reduction is not due to underpowered statistical tests as the hybrid expression level is most similar to the additive level (Figure 5.7). Therefore, comparisons solely to the parental species exaggerate hybrid misregulation. Genes whose expression has diverged between the parental species are particularly prone to this over-estimation, as the additive level substantially differs from expression of both parents. Third, we determined genes with transgressive expression, that is, expressed outside of the parental range. This criterion is the most stringent for classifying "misregulation". Only 225 (4%) genes fall under this category, and the bias for up-regulation in hybrids persists.

Between ~27-36% of the total of 5675 genes are different between L3.12hr and L3.PS1-2. A substantial proportion of genes down-regulated in hybrids are also enriched in adult testis expression (Table 5.2), likely due to the fact that hybrids have atrophied gonad discs. These results demonstrate that many apparent expression changes between hybrids and pure-species reflect developmental differences of hybrids rather than hybrid-specific gene misregulation.

Table 5.2. Hybrid-specific expression

	Mis-regulated genes (MR)							
	All mis-regulated			Up in hybrids		Down in hybrid		
	n	%	Development genes (% of MR)[b]	n (% of MR)[a]	X-linked (% of up in hybrid)[c]	n (% of MR)	X-linked (% of down in hybrid)[c]	Testes genes (% of down in hybrid)[d]
hybrids vs. *mel-Hmr*$^{+}$	974	17.2%	275* (28.2%)	663** (68.1%)	30** (4.5%)	311 (31.9%)	26* (7.9%)	160** (51.4%)
hybrids vs. *sim-Hmr*$^{+}$	1238	21.8%	333* (26.9%)	695* (56.1%)	149* (21.4%)	543 (43.9%)	129* (23.8%)	137 (25.2%)
hybrids vs. add. expectation	513	9.0%	168* (32.7%)	336** (65.5%)	40 (11.9%)	177 (34.5%)	23 (13.0%)	94** (53.1%)
transgressive	225	4.0%	81* (36.0%)	163** (72.4%)	21 (12.9%)	62 (27.6%)	11 (17.7%)	46** (74.2%)

Note.— 5675 genes with similar expression in hybrids are compared to the parents and additive levels.

[a] Significance test for bias of direction when compared to downregulated genes

[b] FET for enrichment of genes different between L3 12hr and L3 PS1-2; random expectation = 23.1%.

[c] FET for enrichment/depletion of X-linked genes; random expectation = 15.3%.

[d] FET for enrichment of testis-biased genes; random expectation = 26.6%

* $p < 0.01$, ** $p << 1e\text{-}5$; FET

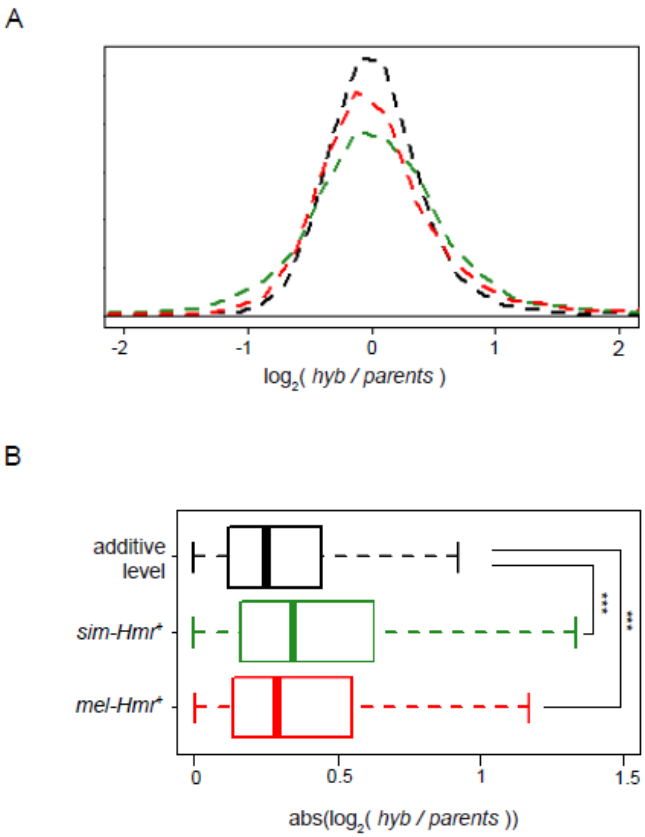


Figure 5.7. Hybrid expression is additive. (A) Distribution of fold-difference between hyb-Hmr+ and parents for 5675 genes that have similar expression between hybrids. Their expression in hybrids is compared to mel-Hmr+ (red), sim-Hmr+ (green), and the additive level (black). The density distribution of the log2 fold-difference is plotted, demonstrating that hybrid expression is most similar to the additive level. (B) Distribution of absolute fold-difference of genes in A when compared to additive and parental levels. *** indicates $p < 0.0001$, Welch's T-test.

Finally, we asked whether there is an excess of X-linked genes with differential expression in hybrids. When compared to *mel-Hmr*$^{+}$, we observed a depletion of X-linked genes. However, the *sim-Hmr*$^{+}$ comparison yielded an enrichment of X-linked genes. This discrepancy is unsurprising, because the hemizygous X in the hybrids is of *D. melanogaster* origin. Indeed, when compared to the additive level, the X is neither enriched nor depleted. Overall our results indicate that after accounting for developmental and tissue differences, gene expression differences in hybrids are surprisingly modest.

5.3 Discussion

5.3.1 Expression profile in hybrids

Interspecific hybrids have been long studied because they often manifest hybrid incompatibilities that cause reproductive isolation between species. More recently, hybrids have been widely used as a genetic background for investigating gene expression divergence. But interpreting and analyzing gene expression in hybrids presents challenges. First, because hybrid gene expression is the combination of the two parental-species alleles, determination of expression level is prone to ascertainment biases because the assembly qualities of the two species are typically not equivalent and/or hybridization probes are designed based on only one of the parental genomes. Second, comparisons of hybrids to parental species are plagued by developmental and physiological defects that are common in hybrids. The net effect is that genes with true regulatory differences are difficult to distinguish from an amalgam of tissue- and developmental stage-specific expression differences. Third, the expected hybrid expression level can be hard to predict for genes that have diverged in expression between species. These analysis

challenges may contribute to previous estimations that the proportion of the genome misregulated in hybrids can be as high as 89% (Ranz et al. 2004).

Here, we attempt to account for these issues when examining gene expression in *D. melanogaster*/*D. simulans* hybrids. To accurately quantitate expression, we analyzed RNA-Seq reads for species-specific SNPs in order to determine the allele-specific expression for all orthologous sites (Degner et al. 2009; McManus et al. 2010; Graze et al. 2012). Using data sets describing developmental stage differences among wild type larvae and tissue-specific gene expression, we evaluated the extent that developmental delay and gonadal degeneration in hybrids contribute to differential expression. For comparison to parental expression, we used several metrics including transgressive expression as in (Llopart 2012), and deviation from additivity.

We found that wild type hybrids (*hyb-Hmr*$^{+}$) have substantial differences from the parental species (Figure 3, supplementary Figure S3). Some of these differences are due to hybrid incompatibility rather than regulatory divergence, because viable *hyb-Hmr*$^{-}$ hybrids are closer in expression to the parental species. Furthermore, *hyb-Hmr*$^{+}$ is more similar to earlier larvae than *hyb-Hmr*$^{-}$, demonstrating that developmental delay is also contributing to gene expression differences. After accounting for these factors, we find that most genes in hybrids conform to additivity and only a limited number are mis-regulated. These results are surprising considering that *D. melanogaster* and *D. simulans* are relatively old species with synonymous divergence of ~10% (Begun et al. 2007). We conclude that regulatory incompatibilities may not be as wide-spread as previously thought (Ranz et al. 2004; Haerty and Singh 2006).

Using a similar framework with RNA-Seq, McManus et al. examined the transcriptome of adult female hybrids between *D. melanogaster* and *D. sechellia* and found extensive non-

additivity (McManus et al. 2010). While the discrepancy could result from the different species pairs used (and/or different sexes), we find this unlikely as *D. sechellia* and *D. simulans* are closely related sister species. Instead, we suggest two other more likely causes. First, the difference may reflect the life-stages investigated. Regulation of gene expression is under stronger purifying selection during development than in adulthood (Castillo-Davis and Hartl 2002; Davis et al. 2005). As a consequence, regulatory incompatibilities may be less likely to accumulate in larvae compared to adults. Second, the hybrid adult females likely have significant physiological differences compared to their parent species because they lack ovaries, which may increase non-additive gene expression when whole adults are sampled.

5.3.2 X-chromosome misregulation

The X chromosome has distinct properties from the autosomes. Its smaller effective population size and hemizygosity in males result in a faster rate of evolution than the autosomes, the so-called fast-X effect (Vicoso and Charlesworth 2006). Additionally, the X accumulates more hybrid sterility loci than the autosomes, known as the large-X effect (Presgraves 2008). One might therefore expect that hybrids have an excess of X-linked misregulation, but the results are mixed. Sterile hybrid male mice show a disproportionate amount of X-linked up-regulation (Good et al. 2010) but sterile Drosophila hybrid males show the opposite, with X-linked misregulation under-represented (Lu et al. 2010; Llopart 2012). These differences between species might reflect different processes of sex chromosome silencing in the male germline. For example, although there is some dispute on the issue, the X chromosome does not appear to be strongly silenced in the Drosophila male germline (Mikhaylova and Nurminsky 2012). Our results here show that hybrid male larvae have neither a higher nor lower proportion of X-linked

genes that differ from additivity. Additionally, the lethality induced by the X-linked *Hmr* is also not associated with more differences among X-linked genes. Together, these findings suggest that increased X-linked misregulation is not a rule in hybrid males.

5.3.3 Discrepancies and similarities with microarray data

A previous microarray study revealed few genes differentially expressed between lethal and viable hybrids (Barbash and Lorigan 2007). In our current study, RNA-Seq offered several advantages, including unbiased determination of allele-specific expression. This allowed us to identify many more genes, and observe allele-specific effects on gene-regulation in hybrids. However, overlap between the two studies is small both in terms of GO terms enriched and specific genes. One likely explanation is that our samples were from older larvae than in the previous study, such that different sets of developmental genes may be affected. Additionally, developmental differences between lethal and viable hybrids will likely become exacerbated over time, resulting in a larger set of differentially regulated genes in our current study. Nonetheless, nearly all genes shared between the two experiments show the same direction of change, potentially revealing genes implicated in causing hybrid lethality. Overall, the small number of genes identified suggests that hybrid lethality is neither caused by nor causes significant changes in gene regulation.

5.3.4 *Hmr* function in hybrids

We find a small set of genes up-regulated in *mel-Hmr*$^{-}$, indicating that *Hmr* functions as a negative regulator in *D. melanogaster*. In contrast, we find that more genes are down-regulated in *hyb-Hmr*$^{-}$, reflecting an activating role for *Hmr* in hybrids. This result is unlikely to be due to

developmental differences between the hybrids, because one would expect to see a similar directional bias in the modENCODE data when comparing different developmental stages. However no such bias exists (n = 1870 and 1726 for genes up and down in L3.PS1-2 when compared to L3.12hr, respectively; p = 0.09398, FET). Additionally, candidate targets of *Hmr* identified in *D. melanogaster* are not differentially expressed between the hybrids. Therefore our results indicate that the repressive effect of *Hmr* is not maintained in hybrids.

This difference between *Hmr* function in *D. melanogaster* versus hybrids was also apparent in an analysis of TEs (Satyaki et al. 2014). *Hmr* is required for TE repression in wild type *D. melanogaster*, yet much higher TE expression occurs in *hyb-Hmr*$^{+}$ compared to *hyb-Hmr*$^{-}$. Together with our findings here, these results strongly argue that *Hmr* has neomorphic function in the hybrid, and that the associated hybrid lethality is a gain-of-function phenotype, as suggested by earlier genetic analyses (Barbash et al. 2000; Orr and Irving 2000). One scenario for this gain of function is that Hmr protein acquires new binding partners in the hybrid background, allowing it to localize to new targets and reverse its repressive activity. This is supported by the observation of mislocalization of Hmr protein in hybrids (Thomae et al. 2013). Given that *Hmr* is required to repress a wide range of heterochromatic repeats (Satyaki et al. 2014), hybrid lethality and the observed misregulation may result from alterations in heterochromatin that affect chromosome function.

Hmr's activating effects in hybrids is particularly intriguing, in light of our observation that hybrids have significantly more up-regulation when compared to the additive expectation. We speculate that the over-expression may be detrimental to hybrids, either broadly affecting the stoichiometry of many complexes and pathways, or through misexpression of a small set of

genes with large effects. The partial mitigation of this effect through down-regulation of some genes in *Hmr*$^{-}$ hybrids may therefore be requisite for hybrid viability.

5.3.5 The role of *Hmr* in allele-specific regulation in hybrids

We showed that in the absence of *Hmr* in hybrids (*hyb-Hmr*$^{-}$), the *D. melanogaster* alleles of many genes are down-regulated while the *D. simulans* alleles are unchanged. Because this excess is exclusive to *D. melanogaster* alleles, it seems unlikely to reflect general misregulation associated with hybrid death. One possible cause of this pattern is intraspecific regulatory differences in developmental genes. Because *hyb-Hmr*$^{+}$ has just reached stage L3 while *hyb-Hmr*$^{-}$ is approaching puff stage 1 of L3, the observed *mel*-specific regulation may be revealing a set of genes that are differentially expressed between the two developmental time points only in *D. melanogaster* but not in *D. simulans*. This differential pattern likely results from cis-regulatory differences between the species, because both alleles are exposed to the same set of trans-factors in hybrids (Wittkopp et al. 2004).

An alternative possibility is that *Hmr* causes allele-specific activation in hybrids. If true, it again points to a neomorphic hybrid function because the genes affected are not regulated by *Hmr* in pure species. Additionally, this allele-specific regulation may indicate that the interacting partner is of *D. melanogaster* origin. One possibility is that it is X-linked as genetic studies have suggested that, in addition to *Hmr*, incompatibility genes on the *D. melanogaster* X are required for fully penetrant hybrid lethality (Barbash et al. 2000). We suggest that such HI genes may be contributing to the allele-specific patterns that we observed.

5.3.6 Divergence of *Hmr* regulation and function

We observed 2.5-fold higher expression of *Hmr* in *D. simulans* male larvae compared to *D. melanogaster* (Figure 5.8A). This result is consistent with Northern blot analysis of mixed-sex larvae (Barbash et al. 2003) and also is apparent, albeit to a lesser extent (1.78-fold higher, Figure 5.8B), in RNA-Seq analysis of white prepupae (Ni et al. 2012). Interestingly, the opposite result is seen at the protein level, with *D. melanogaster* being higher than *D. simulans* (Thomae et al. 2013). The discrepancy between protein and RNA levels strongly argues that *Hmr* levels are controlled by a complex combination of transcriptional and post-transcriptional effects, and that this regulation has changed drastically between the species. These differences are likely to be due at least in part to cis-regulatory divergence because the flanking non-coding regions of *Hmr* show evidence of adaptive evolution (Barbash et al. 2004).

At face value, lower protein level in *D. simulans* predicts weaker suppression of *Hmr* targets. However, to the contrary, we find that most of the genes repressed by *Hmr* in *D. melanogaster* experience a significantly stronger silencing in *D. simulans* (Figure 5.6). One possibility is that the stronger repression of the targets is the result of protein coding differences between the *Hmr* orthologs, either through stronger binding affinity to the targets or stronger recruitment of associated factors. Genetic assays have revealed that *Hmr* has diverged with respect to its hybrid lethal activity, as *D. simulans Hmr* does not cause lethality to hybrid males (Barbash et al. 2004). Our results provide further evidence of the functional consequences of the rapid divergence of *Hmr* between the two species, which is likely the product of both protein-coding and expression-level differences.

Given *Hmr*'s role in regulation of repetitive sequences, its stronger repression of targets in *D. simulans* has intriguing implications for the evolution of heterochromatin in the two

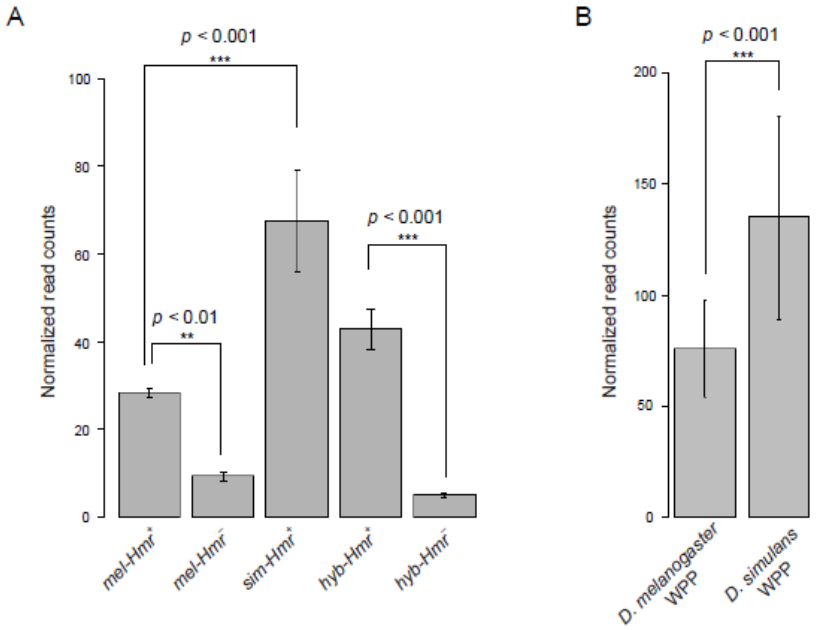


Figure 5.8. *Hmr* expression level. (A) *Hmr* expression in all sequenced samples. Significance determined by DESeq package. B. *Hmr* expression from *D. melanogaster* and *D. simulans* white pre-pupae (WPP) RNA-Seq data (SRP003653); significance determined by FET.

species. Since stronger repression reduces TE and satellite DNA activity, the difference in repressive capabilities of the *Hmr* orthologs may contribute to a stronger defense system against selfish elements in *D. simulans*. This proposal is consistent with the significantly higher TE and satellite DNA content of *D. melanogaster* compared to *D. simulans* (Dowsett and Young 1982; Lohe and Brutlag 1987; Lerat et al. 2011). Furthermore, it underscores that defense against selfish DNA plays a pivotal role in the evolution of genome size and architecture.

5.4 Materials and Methods

5.4.1 Nomenclature

Hybrid, *D. melanogaster*, and *D. simulans* genotypes are designated with the prefixes *hyb*, *mel*, and *sim*, respectively. For example, *hyb-Hmr*$^{-}$ refers to *Df(1)Hmr*$^{-}$*/Y* F1 hybrid males.

5.4.2 Fly stocks and sample collection

The crossing scheme to make *Hmr*$^{+}$ and *Hmr*$^{-}$ hybrids followed that used in (Barbash and Lorigan 2007). Samples were collected as described in (Satyaki et al. 2014). Briefly, virgin females from the *Df(1)Hmr*$^{-}$*, y w v/FM7i, P{w+ mC =ActGFP }JMR* stock (abbreviated as *Hmr*$^{-}$*/FM7i, GFP*) and a background-matched stock with *Hmr*$^{+}$ genotype*, y w v/FM7i, P{w+ mC =ActGFP}JMR* (abbreviated as *Hmr*$^{+}$*/FM7i, GFP*) were crossed to *v/Y D. simulans* males. Hybrid larval sons not carrying the balancer were selected by their *y*$^{-}$ mouth hook and absence of *GFP*. To generate *mel* samples, *Hmr*$^{-}$*/FM7i, GFP* and *Hmr*$^{+}$*/FM7i, GFP* virgin females were crossed to *FM7i, GFP/Y* males. Larval sons not carrying the balancer were selected by *y*$^{-}$ and *GFP*$^{-}$. To generate *sim* samples, *y w D. simulans* virgin females were crossed to *v/Y D. simulans*

males. Larval sons were selected by y^-. Two replicates were collected for every larval genotype, each containing more than 30 larvae pooled from two or more crosses.

5.4.3 Library preparation and sequencing

RNA was extracted using Trizol. cDNA libraries were generated with the TruSeq Kit (Illumina). All samples were barcoded and sequenced on one lane of Illumina HiSeq 2000, single end 100bp. The number of reads generated for each sample is listed in Table 5.6. Illumina sequence data are available from the NCBI website under BioProject number PRJNA236022.

5.4.4 Sequence alignment and determining species-of-origin

For the *D. melanogaster* samples and modENCODE datasets (raw sequences downloaded from the modENCODE website), reads were aligned uniquely to *D. melanogaster* reference r.546 (Flybase) using Tophat, allowing up to 2 mismatches (Trapnell et al. 2009). The read counts of *D. melanogaster* samples generated this way were only used for comparing *mel-Hmr*$^+$ and *mel-Hmr*$^-$ in fig 1. All other analyses involving the *D. melanogaster*, *D. simulans,* and hybrid samples utilized read counts generated after species-of-origin calls described below, to ensure that read counts from all samples are comparable and to minimize mapping biases.

To determine species-of-origin, reads were uniquely mapped to *D. melanogaster* (Release 5.46) and *D. simulans* (Hu et al. 2012) references separately, allowing for up to 4 bp of mismatches per 100bp-read using Tophat. For each read that aligned to both references, we used custom Perl scripts to compare the number of mismatches it has to the two genomes, and designated the reference with the fewer mismatches to be the species-of-origin. Reads mapping to both species with more than 2 mismatches were discarded. Reads mapping equally well to both references

were designated as ambiguous and used for analyzing total expression of a gene but excluded from allele-specific expression analyses. After subjecting the pure-species sequences to this pipeline, we identified the positions of strain-specific polymorphisms that could lead to incorrect species-of-origin calls. Reads containing these polymorphisms in the hybrid samples were then designated as ambiguous. Because annotation of the *D. simulans* genome is lower resolution than *D. melanogaster*, we annotated reads of *D. simulans* origin using the *D. melanogaster* mapping coordinates and annotation from FlyBase (McQuilton et al. 2012). A small number of reads failed to align to one genome, which likely results from either high levels of divergence between species, or gaps in one of the two references. To avoid artifacts associated with the latter scenario, we realigned these reads uniquely to modified references of the two species that have been curated to contain only orthologous exons present in both assemblies (personal communication with E. Kelleher). Reads failing to map to this set of orthologous regions were discarded. Reads with a species-of-origin designation from the two approaches were then summed for the final allele-specific counts. The tally of allele-specific read counts is listed in Table 5.6.

5.4.5 Data analyses

All analyses were carried out in RStudio version 0.98.495 (http://www.rstudio.com/) with R version 3.0.2. The species-of-origin counts were used only for allele-specific analyses. Otherwise, the species-of-origin and ambiguous counts were totaled for each gene. Normalization of samples, fold change, and significance of differential expression were determined using the DESeq package (Wang et al. 2010). Genes with fold-change of greater than 1.5 and false discovery rate adjusted p-value of < 0.1 were considered significantly different. To

Table 5.3. Sequencing and read mapping results.

	No. of raw reads	[a] No. of reads mapped to *D.melanogaster* ref. and annotated	Parent-of-origin mapping				Spearman's Correlation between replicates
			[b] No. of *D. melanogaster* reads annotated	[b] No. of *D.simulans* reads annotated	No. of ambiguous reads annotated	[c] total	
mel-Hmr^{+}	20812819	15380881	9494522	77823	3385882	12958227	0.9936
	20331731	15306838	9347266	80339	3428868	12856473	
mel-Hmr^{-}	9624067	7249471	4436115	41841	1601930	6079886	0.9810
	16669948	12790527	7952729	67449	2776513	10796691	
sim-Hmr^{+}	17669464		5512994	4062090	3762879	13337963	0.9919
	23246647		4175490	2946435	2941303	10063228	
hyb-Hmr^{+}	20970481		4867504	3587435	3584711	12039650	0.9870
	18489589		5997518	4765515	3584421	14347454	
hyb-Hmr^{-}	19199436		222341	8593371	3056594	11872306	0.9750
	20333988		306756	11023472	4001100	15331328	

[a] used only for *mel* comparisons (Figure 5.1A,D,E)

[b] used only for allele-specific comparisons (Figure 5.2)

[c] used for all other analyses

determine genes differentially expressed between modENCODE stages, Fisher's Exact Test (FET) was used instead of DESeq because the modENCODE data are unreplicated (Marioni et al. 2008; Auer and Doerge 2010). Because FET is more prone to false-positives, a stricter cutoff was used where genes were deemed significantly different if the FDR-adjusted p-value is less than 0.001 and the fold change is greater than 2.

Heterochromatin and euchromatin boundaries are designated as in (Satyaki et al. 2014). The additive expectations for each gene were calculated as the average of normalized (by DESeq) *mel-Hmr*$^{+}$ and *sim-Hmr*$^{+}$ expression for autosomal genes, and as only *mel-Hmr*$^{+}$ expression for X-linked genes. The set of genes highly expressed in the testes was downloaded from Flybase using the RNA-Seq search tool "expression by tissue" using modENCODE data. To determine genes with hybrid specific misregulation (Table 5.2), we first identified genes that are not significantly different and with fold-difference of <1.3 between *hyb-Hmr*$^{-}$ and *hyb-Hmr*$^{+}$. These genes are deemed hybrid-specific misregulations if the expression in one or both of the two hybrids is significantly higher or lower by > 1.5-fold compared to either one of the parental or additive levels. Genes were deemed transgressive if the expression is significantly greater or lower than both parents by 1.5 fold.

REFERENCE LIST

References

Abad, J.P., Carmena, M., Baars, S., Saunders, R.D., Glover, D.M., Ludena, P., Sentis, C., Tyler-Smith, C., and Villasante, A. (1992). Dodeca satellite: a conserved G+C-rich satellite from the centromeric heterochromatin of Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *89*, 4663–4667.

Abad, J.P., Pablos, B. de, Osoegawa, K., Jong, P.J. de, Martín-Gallardo, A., and Villasante, A. (2004). Genomic Analysis of Drosophila melanogaster Telomeres: Full-length Copies of HeT-A and TART Elements at Telomeres. Mol. Biol. Evol. *21*, 1613–1619.

Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. *12*, R18.

Anderson, J.A., Song, Y.S., and Langley, C.H. (2008). Molecular population genetics of Drosophila subtelomeric DNA. Genetics *178*, 477–487.

Barbash, D.A. (2010). Ninety years of Drosophila melanogaster hybrids. Genetics *186*, 1–8.

Barbash, D.A., and Lorigan, J.G. (2007). Lethality in Drosophila melanogaster/Drosophila simulans species hybrids is not associated with substantial transcriptional misregulation. J. Exp. Zoolog. B Mol. Dev. Evol. *308*, 74–84.

Barbash, D.A., Roote, J., and Ashburner, M. (2000). The Drosophila melanogaster hybrid male rescue gene causes inviability in male and female species hybrids. Genetics *154*, 1747–1771.

Barbash, D.A., Siino, D.F., Tarone, A.M., and Roote, J. (2003). A rapidly evolving MYB-related protein causes species isolation in Drosophila. Proc. Natl. Acad. Sci. U. S. A. *100*, 5302–5307.

Barbash, D.A., Awadalla, P., and Tarone, A.M. (2004). Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus. PLoS Biol. *2*, e142.

Barrón, M.G., Fiston-Lavier, A.-S., Petrov, D.A., and González, J. (2014). Population Genomics of Transposable Elements in Drosophila. Annu. Rev. Genet. *48*, 561–581.

Bayes, J.J., and Malik, H.S. (2009). Altered Heterochromatin Binding by a Hybrid Sterility Protein in Drosophila Sibling Species. Science *326*, 1538–1541.

Begun, D.J., and Aquadro, C.F. (1993). African and North American populations of Drosophila melanogaster are very different at the DNA level. Nature *365*, 548–550.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.-P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., et al. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans. PLoS Biol. *5*, e310.

Biessmann, H., Carter, S.B., and Mason, J.M. (1990). Chromosome ends in Drosophila without telomeric DNA sequences. Proc. Natl. Acad. Sci. U. S. A. *87*, 1758–1761.

Biessmann, H., Valgeirsdottir, K., Lofsky, A., Chin, C., Ginther, B., Levis, R.W., and Pardue, M.L. (1992a). HeT-A, a transposable element specifically involved in "healing" broken chromosome ends in Drosophila melanogaster. Mol. Cell. Biol. *12*, 3910–3918.

Biessmann, H., Champion, L.E., O'Hair, M., Ikenaga, K., Kasravi, B., and Mason, J.M. (1992b). Frequent transpositions of Drosophila melanogaster HeT-A transposable elements to receding chromosome ends. EMBO J. *11*, 4459–4469.

Biessmann, H., Kasravi, B., Jakes, K., Bui, T., Ikenaga, K., and Mason, J.M. (1993). The genomic organization of HeT-A retroposons in Drosophila melanogaster. Chromosoma *102*, 297–305.

Biessmann, H., Kasravi, B., Bui, T., Fujiwara, G., Champion, L.E., and Mason, J.M. (1994). Comparison of two active HeT-A retroposons of Drosophila melanogaster. Chromosoma *103*, 90–98.

Biggin, M.D., and Tjian, R. (1988). Transcription factors that activate the Ultrabithorax promoter in developmentally staged extracts. Cell *53*, 699–711.

Blackburn, E.H., Greider, C.W., and Szostak, J.W. (2006). Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nat. Med. *12*, 1133–1138.

Blower, M.D., and Karpen, G.H. (2001). The role of Drosophila CID in kinetochore formation, cell-cycle progression and heterochromatin interactions. Nat. Cell Biol. *3*, 730–739.

Blumenstiel, J.P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. Trends Genet. TIG *27*, 23–31.

Bolkan, B.J., Booker, R., Goldberg, M.L., and Barbash, D.A. (2007). Developmental and cell cycle progression defects in Drosophila hybrid males. Genetics *177*, 2233–2241.

Bosco, G., Campbell, P., Leiva-Neto, J.T., and Markow, T.A. (2007). Analysis of Drosophila Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species. Genetics *177*, 1277–1290.

Boussy, I.A., Healy, M.J., Oakeshott, J.G., and Kidwell, M.G. (1988). Molecular analysis of the P-M gonadal dysgenesis cline in eastern Australian Drosophila melanogaster. Genetics *119*, 889–902.

Britten, R.J., and Kohne, D.E. (1968). Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science *161*, 529–540.

Bzymek, M., and Lovett, S.T. (2001). Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. Proc. Natl. Acad. Sci. U. S. A. *98*, 8319–8325.

Carvalho, A.B., and Clark, A.G. (2005). Y Chromosome of D. pseudoobscura Is Not Homologous to the Ancestral Drosophila Y. Science *307*, 108–110.

Castillo, D.M., Mell, J.C., Box, K.S., and Blumenstiel, J.P. (2011). Molecular evolution under increasing transposable element burden in Drosophila: a speed limit on the evolutionary arms race. BMC Evol. Biol. *11*, 258.

Castillo-Davis, C.I., and Hartl, D.L. (2002). Genome evolution and developmental constraint in Caenorhabditis elegans. Mol. Biol. Evol. *19*, 728–735.

Cavalier-Smith, T. (1978). Nuclear volume control by nucleoskeletal DNA, selection for cell volume and cell growth rate, and the solution of the DNA C-value paradox. J. Cell Sci. *34*, 247–278.

Charlesworth, B. (1991). The evolution of sex chromosomes. Science *251*, 1030–1033.

Charlesworth, B., and Charlesworth, D. (2000). The degeneration of Y chromosomes. Philos. Trans. R. Soc. B Biol. Sci. *355*, 1563–1572.

Charlesworth, B., and Langley, C.H. (1989). The Population Genetics of Drosophila Transposable Elements. Annu. Rev. Genet. *23*, 251–287.

Charlesworth, B., Sniegowski, P., and Stephan, W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature *371*, 215–220.

Chmátal, L., Gabriel, S.I., Mitsainas, G.P., Martínez-Vargas, J., Ventura, J., Searle, J.B., Schultz, R.M., and Lampson, M.A. (2014). Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Curr. Biol. CB *24*, 2295–2300.

Choi, J.Y., Bubnell, J.E., and Aquadro, C.F. (2015). Population genomics of infectious and integrated Wolbachia pipientis genomes in Drosophila ananassae. Genome Biol. Evol. evv158.

Chueh, A.C., Wong, L.H., Wong, N., and Choo, K.H.A. (2005). Variable and hierarchical size distribution of L1-retroelement-enriched CENP-A clusters within a functional human neocentromere. Hum. Mol. Genet. *14*, 85–93.

Clark, A.G., Eisen, M.B., Smith, D.R., Bergman, C.M., Oliver, B., Markow, T.A., Kaufman, T.C., Kellis, M., Gelbart, W., Iyer, V.N., et al. (2007). Evolution of genes and genomes on the Drosophila phylogeny. Nature *450*, 203–218.

Comings, D.E. (1972). The structure and function of chromatin. Adv. Hum. Genet. *3*, 237–431.

Corbett-Detig, R., Jacobs-Palmer, E., Hartl, D., and Hoekstra, H. (2015). Direct Gamete Sequencing Reveals No Evidence for Segregation Distortion in House Mouse Hybrids. PloS One *10*, e0131933.

Csink, A.K., and Henikoff, S. (1998). Something from nothing: the evolution and utility of satellite repeats. Trends Genet. TIG *14*, 200–204.

Daniels, S.B., and Strausbaugh, L.D. (1986). The distribution of P-element sequences in Drosophila: the willistoni and saltans species groups. J. Mol. Evol. *23*, 138–148.

Daniels, S.B., Peterson, K.R., Strausbaugh, L.D., Kidwell, M.G., and Chovnick, A. (1990). Evidence for Horizontal Transmission of the P Transposable Element between Drosophila Species. Genetics *124*, 339–355.

David, J.R., and Capy, P. (1988). Genetic variation of Drosophila melanogaster natural populations. Trends Genet. *4*, 106–111.

Davis, J.C., Brandman, O., and Petrov, D.A. (2005). Protein evolution in the context of Drosophila development. J. Mol. Evol. *60*, 774–785.

Degner, J.F., Marioni, J.C., Pai, A.A., Pickrell, J.K., Nkadori, E., Gilad, Y., and Pritchard, J.K. (2009). Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinforma. Oxf. Engl. *25*, 3207–3212.

Dernburg, A.F. (2011). In situ hybridization to somatic chromosomes in Drosophila. Cold Spring Harb. Protoc. *2011*.

Doolittle, W.F., and Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature *284*, 601–603.

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. *38*, 1440–1445.

Dover, G. (1982). Molecular drive: a cohesive mode of species evolution. Nature *299*, 111–117.

Dowsett, A.P., and Young, M.W. (1982). Differing levels of dispersed repetitive DNA among closely related species of Drosophila. Proc. Natl. Acad. Sci. U. S. A. *79*, 4570–4574.

Dunning Hotopp, J.C., Clark, M.E., Oliveira, D.C.S.G., Foster, J.M., Fischer, P., Muñoz Torres, M.C., Giebel, J.D., Kumar, N., Ishmael, N., Wang, S., et al. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science *317*, 1753–1756.

Ellis, L.L., Huang, W., Quinn, A.M., Ahuja, A., Alfrejd, B., Gomez, F.E., Hjelmen, C.E., Moore, K.L., Mackay, T.F.C., Johnston, J.S., et al. (2014). Intrapopulation genome size variation in D. melanogaster reflects life history variation and plasticity. PLoS Genet. *10*, e1004522.

Ferree, P.M., and Barbash, D.A. (2009). Species-Specific Heterochromatin Prevents Mitotic Chromosome Segregation to Cause Hybrid Lethality in Drosophila. PLoS Biol *7*, e1000234.

Fishman, L., and Saunders, A. (2008). Centromere-associated female meiotic drive entails male fitness costs in monkeyflowers. Science *322*, 1559–1562.

Fishman, L., and Willis, J.H. (2005). A Novel Meiotic Drive Locus Almost Completely Distorts Segregation in Mimulus (Monkeyflower) Hybrids. Genetics *169*, 347–353.

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

Fondon, J.W., III, Martin, A., Richards, S., Gibbs, R.A., and Mittelman, D. (2012). Analysis of Microsatellite Variation in Drosophila melanogaster with Population-Scale Genome Sequencing. PLoS ONE *7*, e33036.

Fujiwara, H., Osanai, M., Matsumoto, T., and Kojima, K.K. (2005). Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *13*, 455–467.

Fulcher, N., Teubenbacher, A., Kerdaffrec, E., Farlow, A., Nordborg, M., and Riha, K. (2015). Genetic architecture of natural variation of telomere length in Arabidopsis thaliana. Genetics *199*, 625–635.

Gall, J.G., Cohen, E.H., and Polan, M.L. (1971). Repetitive DNA sequences in Drosophila. Chromosoma *33*, 319–344.

Gao, G., Walser, J.-C., Beaucher, M.L., Morciano, P., Wesolowska, N., Chen, J., and Rong, Y.S. (2010). HipHop interacts with HOAP and HP1 to protect Drosophila telomeres in a sequence-independent manner. EMBO J. *29*, 819–829.

George, J.A., DeBaryshe, P.G., Traverse, K.L., Celniker, S.E., and Pardue, M.-L. (2006). Genomic organization of the Drosophila telomere retrotransposable elements. Genome Res. *16*, 1231–1240.

Golubovsky, M.D., Konev, A.Y., Walter, M.F., Biessmann, H., and Mason, J.M. (2001). Terminal retrotransposons activate a subtelomeric white transgene at the 2L telomere in Drosophila. Genetics *158*, 1111–1123.

Good, J.M., Giger, T., Dean, M.D., and Nachman, M.W. (2010). Widespread over-expression of the X chromosome in sterile F_1hybrid mice. PLoS Genet. *6*, e1001148.

Graveley, B.R., Brooks, A.N., Carlson, J.W., Duff, M.O., Landolin, J.M., Yang, L., Artieri, C.G., van Baren, M.J., Boley, N., Booth, B.W., et al. (2011). The developmental transcriptome of Drosophila melanogaster. Nature *471*, 473–479.

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, 1SI – 21SI.

Graze, R.M., Novelo, L.L., Amin, V., Fear, J.M., Casella, G., Nuzhdin, S.V., and McIntyre, L.M. (2012). Allelic imbalance in Drosophila hybrid heads: exons, isoforms, and evolution. Mol. Biol. Evol. *29*, 1521–1532.

Greenberg, A.J., Hackett, S.R., Harshman, L.G., and Clark, A.G. (2011). Environmental and genetic perturbations reveal different networks of metabolic regulation. Mol. Syst. Biol. *7*, 563.

Gregory, T.R. (2001). Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. Biol. Rev. *76*, 65–101.

Gregory, T.R. (2015). Animal Genome Size Database.

Grewal, S.I., and Rice, J.C. (2004). Regulation of heterochromatin by histone methylation and small RNAs. Curr. Opin. Cell Biol. *16*, 230–238.

Grewal, S.I.S., and Moazed, D. (2003). Heterochromatin and epigenetic control of gene expression. Science *301*, 798–802.

Gurushidze, M., Fuchs, J., and Blattner, F.R. (2012). The Evolution of Genome Size Variation in Drumstick Onions (Allium subgenus Melanocrommyum). Syst. Bot. *37*, 96–104.

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.

Hancock, J.M. (2002). Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects. Genetica *115*, 93–103.

Hardy, O.J., Charbonnel, N., Fréville, H., and Heuertz, M. (2003). Microsatellite allele sizes: a simple test to assess their significance on genetic differentiation. Genetics *163*, 1467–1482.

Hartl, D.L. (2000). Molecular melodies in high and low C. Nat. Rev. Genet. *1*, 145–149.

Henikoff, S., Ahmad, K., and Malik, H.S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science *293*, 1098–1102.

Holm-Hansen, O. (1969). Algae: amounts of DNA and organic carbon in single cells. Science *163*, 87–88.

Hoskins, R.A., Carlson, J.W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., Wan, K.H., Park, S., Mendez-Lago, M., Rossi, F., et al. (2007). Sequence finishing and mapping of Drosophila melanogaster heterochromatin. Science *316*, 1625–1628.

Howe, M., Dimitri, P., Berloco, M., and Wakimoto, B.T. (1995). Cis-Effects of Heterochromatin on Heterochromatic and Euchromatic Gene Activity in Drosophila Melanogaster. Genetics *140*, 1033–1045.

Hu, T.T., Eisen, M.B., Thornton, K.R., and Andolfatto, P. (2013). A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Res. *23*, 89–98.

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208.

Kamm, A., Galasso, I., Schmidt, T., and Heslop-Harrison, J.S. (1995). Analysis of a repetitive DNA family from Arabidopsis arenosa and relationships between Arabidopsis species. Plant Mol. Biol. *27*, 853–862.

Karpen, G.H., and Spradling, A.C. (1992). Analysis of subtelomeric heterochromatin in the Drosophila minichromosome Dp1187 by single P element insertional mutagenesis. Genetics *132*, 737–753.

Karpen, G.H., Le, M.H., and Le, H. (1996). Centric heterochromatin and the efficiency of achiasmate disjunction in Drosophila female meiosis. Science *273*, 118–122.

Kern, A.D., and Begun, D.J. (2008). Recurrent deletion and gene presence/absence polymorphism: telomere dynamics dominate evolution at the tip of 3L in Drosophila melanogaster and D. simulans. Genetics *179*, 1021–1027.

Kerrigan, L.A., Croston, G.E., Lira, L.M., and Kadonaga, J.T. (1991). Sequence-specific transcriptional antirepression of the Drosophila Krüppel gene by the GAGA factor. J. Biol. Chem. *266*, 574–582.

Khurana, J.S., Xu, J., Weng, Z., and Theurkauf, W.E. (2010). Distinct functions for the Drosophila piRNA pathway in genome maintenance and telomere protection. PLoS Genet. *6*, e1001246.

Kidwell, M.G. (1983). Evolution of hybrid dysgenesis determinants in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *80*, 1655–1659.

Klenov, M.S., Lavrov, S.A., Stolyarenko, A.D., Ryazansky, S.S., Aravin, A.A., Tuschl, T., and Gvozdev, V.A. (2007). Repeat-associated siRNAs cause chromatin silencing of retrotransposons in the Drosophila melanogaster germline. Nucleic Acids Res. *35*, 5430–5438.

Klutstein, M., Fennell, A., Fernández-Álvarez, A., and Cooper, J.P. (2015). The telomere bouquet regulates meiotic centromere assembly. Nat. Cell Biol. *17*, 458–469.

Koerich, L.B., Wang, X., Clark, A.G., and Carvalho, A.B. (2008). Low conservation of gene content in the Drosophila Y chromosome. Nature *456*, 949–951.

Lachner, M., O'Carroll, D., Rea, S., Mechtler, K., and Jenuwein, T. (2001). Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature *410*, 116–120.

Lai, Z., Gross, B.L., Zou, Y., Andrews, J., and Rieseberg, L.H. (2006). Microarray analysis reveals differential gene expression in hybrid sunflower species. Mol. Ecol. *15*, 1213–1227.

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.

Langley, C.H., Crepeau, M., Cardeno, C., Corbett-Detig, R., and Stevens, K. (2011). Circumventing heterozygosity: sequencing the amplified genome of a single haploid Drosophila melanogaster embryo. Genetics *188*, 239–246.

Langley, S.A., Karpen, G.H., and Langley, C.H. (2014). Nucleosomes Shape DNA Polymorphism and Divergence. PLoS Genet *10*, e1004457.

Larracuente, A.M., and Ferree, P.M. (2015). Simple method for fluorescence DNA in situ hybridization to squashed chromosomes. J. Vis. Exp. JoVE 52288.

Larracuente, A.M., and Presgraves, D.C. (2012). The selfish Segregation Distorter gene complex of Drosophila melanogaster. Genetics *192*, 33–53.

Lee, Y.C.G., and Langley, C.H. (2010). Transposable elements in natural populations of Drosophila melanogaster. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1219–1228.

Legrand, D., Tenaillon, M.I., Matyot, P., Gerlach, J., Lachaise, D., and Cariou, M.-L. (2009). Species-Wide Genetic Variation and Demographic History of Drosophila sechellia, a Species Lacking Population Structure. Genetics *182*, 1197–1206.

Lemos, B., Araripe, L.O., and Hartl, D.L. (2008). Polymorphic Y Chromosomes Harbor Cryptic Variation with Manifold Functional Consequences. Science *319*, 91–93.

Lemos, B., Branco, A.T., and Hartl, D.L. (2010). Epigenetic Effects of Polymorphic Y Chromosomes Modulate Chromatin Components, Immune Response, and Sexual Conflict. Proc. Natl. Acad. Sci. *107*, 15826–15831.

Lerat, E., Burlet, N., Biémont, C., and Vieira, C. (2011). Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. Gene *473*, 100–109.

Le Rouzic, A., and Capy, P. (2005). The First Steps of Transposable Elements Invasion. Genetics *169*, 1033–1043.

Levinson, G., and Gutman, G.A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. *4*, 203–221.

Levis, R.W. (1989). Viable deletions of a telomere from a Drosophila chromosome. Cell *58*, 791–801.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinforma. Oxf. Engl. *25*, 2078–2079.

Liti, G., Haricharan, S., Cubillos, F.A., Tierney, A.L., Sharp, S., Bertuch, A.A., Parts, L., Bailes, E., and Louis, E.J. (2009). Segregating YKU80 and TLC1 alleles underlying natural variation in telomere properties in wild yeast. PLoS Genet. *5*, e1000659.

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.

Lohe, A.R., and Brutlag, D.L. (1986). Multiplicity of satellite DNA sequences in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *83*, 696–700.

Lohe, A.R., and Brutlag, D.L. (1987). Identical satellite DNA sequences in sibling species of Drosophila. J. Mol. Biol. *194*, 161–170.

Lohe, A.R., and Roberts, P.A. (2000). Evolution of DNA in heterochromatin: the Drosophila melanogaster sibling species subgroup as a resource. Genetica *109*, 125–130.

Lohe, A.R., Hilliker, A.J., and Roberts, P.A. (1993). Mapping simple repeated DNA sequences in heterochromatin of Drosophila melanogaster. Genetics *134*, 1149–1174.

Lu, Q., Wallrath, L.L., Granok, H., and Elgin, S.C. (1993). (CT)n (GA)n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the Drosophila hsp26 gene. Mol. Cell. Biol. *13*, 2802–2814.

Lu, X., Shapiro, J.A., Ting, C.-T., Li, Y., Li, C., Xu, J., Huang, H., Cheng, Y.-J., Greenberg, A.J., Li, S.-H., et al. (2010). Genome-wide misexpression of X-linked versus autosomal genes associated with hybrid male sterility. Genome Res. *20*, 1097–1102.

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature *482*, 173–178.

Maheshwari, S., and Barbash, D.A. (2011). The genetics of hybrid incompatibilities. Annu. Rev. Genet. *45*, 331–355.

Malik, H.S. (2009). The centromere-drive hypothesis: a simple basis for centromere complexity. Prog. Mol. Subcell. Biol. *48*, 33–52.

Malik, H.S., and Henikoff, S. (2001). Adaptive evolution of Cid, a centromere-specific histone in Drosophila. Genetics *157*, 1293–1298.

Mason, J.M., Ransom, J., and Konev, A.Y. (2004). A deficiency screen for dominant suppressors of telomeric silencing in Drosophila. Genetics *168*, 1353–1370.

Mason, J.M., Frydrychova, R.C., and Biessmann, H. (2008). Drosophila telomeres: an exception providing new insights. BioEssays News Rev. Mol. Cell. Dev. Biol. *30*, 25–37.

Mason, J.M., Randall, T.A., and Capkova Frydrychova, R. (2015). Telomerase lost? Chromosoma.

Mavrich, T.N., Jiang, C., Ioshikhes, I.P., Li, X., Venters, B.J., Zanton, S.J., Tomsho, L.P., Qi, J., Glaser, R.L., Schuster, S.C., et al. (2008). Nucleosome organization in the Drosophila genome. Nature *453*, 358–362.

McClintock, B. (1950). The Origin and Behavior of Mutable Loci in Maize. Proc. Natl. Acad. Sci. U. S. A. *36*, 344–355.

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.

Mefford, H.C., and Trask, B.J. (2002). The complex structure and dynamic evolution of human subtelomeres. Nat. Rev. Genet. *3*, 91–102.

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.

Mikhaylova, L.M., and Nurminsky, D.I. (2011). Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the Drosophila X chromosome. BMC Biol. *9*, 29.

Mirsky, A.E., and Ris, H. (1951). THE DESOXYRIBONUCLEIC ACID CONTENT OF ANIMAL CELLS AND ITS EVOLUTIONARY SIGNIFICANCE. J. Gen. Physiol. *34*, 451–462.

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.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C.N., Richardson, A.O., Okumoto, Y., Tanisaka, T., and Wessler, S.R. (2009). Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. Nature *461*, 1130–1134.

Ni, X., Zhang, Y.E., Nègre, N., Chen, S., Long, M., and White, K.P. (2012). Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol. *10*, e1001420.

Nolte, V., and Schlötterer, C. (2008). African Drosophila melanogaster and D. simulans Populations Have Similar Levels of Sequence Variability, Suggesting Comparable Effective Population Sizes. Genetics *178*, 405–412.

Novitski, E. (1951). Non-random disjunction in Drosophila. Genetics *36*, 267–280.

Ohno, S. (1972). So much “junk” DNA in our genome. Brookhaven Symp. Biol. *23*, 366–370.

Okumura, K., Kiyama, R., and Oishi, M. (1987). Sequence analyses of extrachromosomal Sau3A and related family DNA: analysis of recombination in the excision event. Nucleic Acids Res. *15*, 7477–7489.

Orgel, L.E., and Crick, F.H. (1980). Selfish DNA: the ultimate parasite. Nature *284*, 604–607.

Orr, H.A., and Irving, S. (2000). Genetic analysis of the hybrid male rescue locus of Drosophila. Genetics *155*, 225–231.

Osanai-Futahashi, M., and Fujiwara, H. (2011). Coevolution of telomeric repeats and telomeric repeat-specific non-LTR retrotransposons in insects. Mol. Biol. Evol. *28*, 2983–2986.

Pak, D.T.., Pflumm, M., Chesnokov, I., Huang, D.W., Kellum, R., Marr, J., Romanowski, P., and Botchan, M.R. (1997). Association of the Origin Recognition Complex with Heterochromatin and HP1 in Higher Eukaryotes. Cell *91*, 311–323.

Palazzo, A.F., and Gregory, T.R. (2014). The Case for Junk DNA. PLoS Genet *10*, e1004351.

Pardue, M.-L., and DeBaryshe, P.G. (2011). Retrotransposons that maintain chromosome ends. Proc. Natl. Acad. Sci. U. S. A. *108*, 20317–20324.

Pardue, M.-L., Rashkova, S., Casacuberta, E., DeBaryshe, P.G., George, J.A., and Traverse, K.L. (2005). Two retrotransposons maintain telomeres in Drosophila. Chromosome Res. Int. J. Mol. Supramol. Evol. Asp. Chromosome Biol. *13*, 443–453.

Paredes, S., and Maggert, K.A. (2009). Ribosomal DNA Contributes to Global Chromatin Regulation. Proc. Natl. Acad. Sci.

Peacock, W.J., Brutlag, D., Goldring, E., Appels, R., Hinton, C.W., and Lindsey, D.L. (1974). The Organization of Highly Repeated DNA Sequences in Drosophila Melanogaster Chromosomes. Cold Spring Harb. Symp. Quant. Biol. *38*, 405–416.

Peng, J.C., and Karpen, G.H. (2007). H3K9 methylation and RNA interference regulate nucleolar organization and repeated DNA stability. Nat. Cell Biol. *9*, 25–35.

Petrov, D.A. (2001). Evolution of genome size: new approaches to an old problem. Trends Genet. TIG *17*, 23–28.

Piñeyro, D., López-Panadès, E., Lucena-Pérez, M., and Casacuberta, E. (2011). Transcriptional analysis of the HeT-A retrotransposon in mutant and wild type stocks reveals high sequence variability at Drosophila telomeres and other unusual features. BMC Genomics *12*, 573.

Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W., Hambleton, S., Burns, S.O., Thrasher, A.J., et al. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinforma. Oxf. Engl. *28*, 2747–2754.

Platero, J.S., Csink, A.K., Quintanilla, A., and Henikoff, S. (1998). Changes in chromosomal localization of heterochromatin-binding proteins during the cell cycle in Drosophila. J. Cell Biol. *140*, 1297–1306.

Pollard, D.A., Iyer, V.N., Moses, A.M., and Eisen, M.B. (2006). Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. *2*, e173.

Pool, J.E., Corbett-Detig, R.B., Sugino, R.P., Stevens, K.A., Cardeno, C.M., Crepeau, M.W., Duchen, P., Emerson, J.J., Saelao, P., Begun, D.J., et al. (2012). Population Genomics of sub-

saharan Drosophila melanogaster: African diversity and non-African admixture. PLoS Genet. *8*, e1003080.

Presgraves, D.C. (2008). Sex chromosomes and speciation in Drosophila. Trends Genet. TIG *24*, 336–343.

Raff, J.W., Kellum, R., and Alberts, B.M. (1994). The Drosophila GAGA transcription factor is associated with specific regions of heterochromatin throughout the cell cycle. EMBO J. *13*, 5977–5983.

Raffa, G.D., Ciapponi, L., Cenci, G., and Gatti, M. (2011). Terminin: a protein complex that mediates epigenetic maintenance of Drosophila telomeres. Nucl. Austin Tex *2*, 383–391.

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.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D., et al. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. Nature *406*, 593–599.

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. *26*, 925–936.

Rieseberg, L.H., and Willis, J.H. (2007). Plant speciation. Science *317*, 910–914.

Rockmill, B., and Roeder, G.S. (1998). Telomere-mediated chromosome pairing during meiosis in budding yeast. Genes Dev. *12*, 2574–2586.

Rodriguez Alfageme, C., Rudkin, G.T., and Cohen, L.H. (1980). Isolation, properties and cellular distribution of D1, a chromosomal protein of Drosophila. Chromosoma *78*, 1–31.

Rong, Y.S. (2008). Telomere capping in Drosophila: dealing with chromosome ends that most resemble DNA breaks. Chromosoma *117*, 235–242.

Rošić, S., Köhler, F., and Erhardt, S. (2014). Repetitive centromeric satellite RNA is essential for kinetochore formation and cell division. J. Cell Biol. *207*, 335–349.

Sandler, L., Hiraizumi, Y., and Sandler, I. (1959). Meiotic Drive in Natural Populations of Drosophila Melanogaster. I. the Cytogenetic Basis of Segregation-Distortion. Genetics *44*, 233–250.

Satyaki, P.R.V., Cuykendall, T.N., Wei, K.H.-C., Brideau, N.J., Kwak, H., Aruna, S., Ferree, P.M., Ji, S., and Barbash, D.A. (2014). The Hmr and Lhr Hybrid Incompatibility Genes Suppress a Broad Range of Heterochromatic Repeats. PLoS Genet. *10*.

Schlötterer, C., Neumeier, H., Sousa, C., and Nolte, V. (2006). Highly Structured Asian Drosophila melanogaster Populations: A New Tool for Hitchhiking Mapping? Genetics *172*, 287–292.

Schmid, K.J., and Tautz, D. (1997). A screen for fast evolving genes from Drosophila. Proc. Natl. Acad. Sci. U. S. A. *94*, 9746–9750.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. Nat. Methods *9*, 671–675.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I.K., Wang, J.-P.Z., and Widom, J. (2006). A genomic code for nucleosome positioning. Nature *442*, 772–778.

Shelby, R.D., Vafa, O., and Sullivan, K.F. (1997). Assembly of CENP-A into centromeric chromatin requires a cooperative array of nucleosomal DNA contact sites. J. Cell Biol. *136*, 501–513.

Shpiz, S., and Kalmykova, A. (2011). Role of piRNAs in the Drosophila telomere homeostasis. Mob. Genet. Elem. *1*, 274–278.

Simkin, A., Wong, A., Poh, Y.-P., Theurkauf, W.E., and Jensen, J.D. (2013). Recurrent and recent selective sweeps in the piRNA pathway. Evol. Int. J. Org. Evol. *67*, 1081–1090.

Siriaco, G.M., Cenci, G., Haoudi, A., Champion, L.E., Zhou, C., Gatti, M., and Mason, J.M. (2002). Telomere elongation (Tel), a new mutation in Drosophila melanogaster that produces long telomeres. Genetics *160*, 235–245.

Slatkin, M. (1995). A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. Genetics *139*, 457–462.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossover. Science *191*, 528–535.

Soltis, P.S., and Soltis, D.E. (2009). The role of hybridization in plant speciation. Annu. Rev. Plant Biol. *60*, 561–588.

Sparrow, A.H., Price, H.J., and Underbrink, A.G. (1972). A survey of DNA content per cell and per chromosome of prokaryotic and eukaryotic organisms: some evolutionary considerations. Brookhaven Symp. Biol. *23*, 451–494.

Stephan, W. (1986). Recombination and the evolution of satellite DNA. Genet. Res. *47*, 167–174.

Stephan, W., and Cho, S. (1994). Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics *136*, 333–341.

Stephan, W., and Li, H. (2006). The recent demographic and adaptive history of Drosophila melanogaster. Heredity *98*, 65–68.

Stone, E.A., and Ayroles, J.F. (2009). Modulated Modularity Clustering as an Exploratory Tool for Functional Genomic Inference. PLoS Genet *5*, e1000479.

Swift, H. (1950). The constancy of desoxyribose nucleic acid in plant nuclei. Proc. Natl. Acad. Sci. U. S. A. *36*, 643–654.

Tamura, K., Subramanian, S., and Kumar, S. (2004). Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. Mol. Biol. Evol. *21*, 36–44.

Teo, C.H., Ma, L., Kapusi, E., Hensel, G., Kumlehn, J., Schubert, I., Houben, A., and Mette, M.F. (2011). Induction of telomere-mediated chromosomal truncation and stability of truncated chromosomes in Arabidopsis thaliana. Plant J. Cell Mol. Biol. *68*, 28–39.

Thomae, A.W., Schade, G.O.M., Padeken, J., Borath, M., Vetter, I., Kremmer, E., Heun, P., and Imhof, A. (2013). A pair of centromeric proteins mediates reproductive isolation in Drosophila species. Dev. Cell *27*, 412–424.

Thomas, C. (1971). The Genetic Organization of Chromosomes. Annu. Rev. Genet. *5*, 237–256.

Thompson, O., Edgley, M., Strasbourger, P., Flibotte, S., Ewing, B., Adair, R., Au, V., Chaudhry, I., Fernando, L., Hutter, H., et al. (2013). The million mutation project: a new approach to genetics in Caenorhabditis elegans. Genome Res. *23*, 1749–1762.

Thomson, G.J., and Feldman, M.W. (1976). Population genetics of modifiers of meiotic drive. III. Equilibrium analysis of a general model for the genetic control of segregation distortion. Theor. Popul. Biol. *10*, 10–25.

Török, T., Harvie, P.D., Buratovich, M., and Bryant, P.J. (1997). The product of proliferation disrupter is concentrated at centromeres and required for mitotic chromosome condensation and cell proliferation in Drosophila. Genes Dev. *11*, 213–225.

Tsukiyama, T., Becker, P.B., and Wu, C. (1994). ATP-dependent nucleosome disruption at a heat-shock promoter mediated by binding of GAGA transcription factor. Nature *367*, 525–532.

Usakin, L., Abad, J., Vagin, V.V., de Pablos, B., Villasante, A., and Gvozdev, V.A. (2007). Transcription of the 1.688 Satellite DNA Family Is Under the Control of RNA Interference Machinery in Drosophila melanogaster Ovaries. Genetics *176*, 1343–1349.

Vasa-Nicotera, M., Brouilette, S., Mangino, M., Thompson, J.R., Braund, P., Clemitson, J.-R., Mason, A., Bodycote, C.L., Raleigh, S.M., Louis, E., et al. (2005). Mapping of a major locus that determines telomere length in humans. Am. J. Hum. Genet. *76*, 147–151.

Vicoso, B., and Bachtrog, D. (2015). Numerous Transitions of Sex Chromosomes in Diptera. PLoS Biol *13*, e1002078.

Vicoso, B., and Charlesworth, B. (2006). Evolution on the X chromosome: unusual patterns and processes. Nat. Rev. Genet. *7*, 645–653.

Villasante, A., Abad, J.P., Planelló, R., Méndez-Lago, M., Celniker, S.E., and de Pablos, B. (2007). Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. *17*, 1909–1918.

Vinogradov, A.E. (1999). Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol. *49*, 376–384.

Walter, M.F., Jang, C., Kasravi, B., Donath, J., Mechler, B.M., Mason, J.M., and Biessmann, H. (1995). DNA organization and polymorphism of a wild-type Drosophila telomere region. Chromosoma *104*, 229–241.

Walter, M.F., Biessmann, M.R., Benitez, C., Török, T., Mason, J.M., and Biessmann, H. (2007). Effects of telomere length in Drosophila melanogaster on life span, fecundity, and fertility. Chromosoma *116*, 41–51.

Wei, K.H.-C., Grenier, J.K., Barbash, D.A., and Clark, A.G. (2014). Correlated variation and population differentiation in satellite DNA abundance among lines of Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *111*, 18793–18798.

Weiler, K.S., and Wakimoto, B.T. (1995). Heterochromatin and gene expression in Drosophila. Annu. Rev. Genet. *29*, 577–605.

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

Wu, C.-I., Lyttle, T.W., Wu, M.-L., and Lin, G.-F. (1988). Association between a satellite DNA sequence and the responder of segregation distorter in D. melanogaster. Cell *54*, 179–189.

Zhu, L., Hathcock, K.S., Hande, P., Lansdorp, P.M., Seldin, M.F., and Hodes, R.J. (1998). Telomere length regulation in mice is linked to a novel chromosome locus. Proc. Natl. Acad. Sci. U. S. A. *95*, 8648–8653.

Zwick, M.E., Salstrom, J.L., and Langley, C.H. (1999). Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin nod with increased rates of nondisjunction in Drosophila melanogaster. Genetics *152*, 1605–1614.