PHYLOGENETIC ANALYSES OF GENOME SIZE EVOLUTION IN DROSOPHILIDAE

A Dissertation

by

CARL ERNEST HJELMEN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	J. Spencer Johnston
Committee Members,	Aaron M. Tarone
	Raul F. Medina
	Mariana Mateos
Head of Department,	David W. Ragsdale

August 2017

Major Subject: Entomology

ABSTRACT

Genome size varies widely across organisms with no tie to organismal complexity. While much of the variation in genome size can be attributed to accumulation of nongenic DNA, the patterns and mechanisms involved are not fully understood. While many hypotheses have been proposed to explain the patterns, they often ignore evolutionary relationships. I address this issue through the analysis of genome size utilizing comparative phylogenetic methods. I compare the results to expectations for different hypothesized modes of genome size change in Drosophilidae. In addition to analyses of whole genome size, I ask how separate components of the genome, sex chromosomes and heterochromatin level, may contribute to these patterns. I conclude the accordion model is the best fit and discuss how chance and adaptations might have acted to produce the genome size variation we observe today.

New and updated genome size for 93 *Drosophila* species were produced using flow cytometry, with a focus on the *Drosophila* subgenus. I used this data, plus additional published data, to produce a phylogenetic analysis of genome size evolution in 152 species of *Drosophila*. Genome size was found to have complete phylogenetic signal and gradual change, with evidence for some change temporally early. These patterns for whole genome size change were consistent throughout the entire genus, with only subtle differences between the subgenera. Components of genome size, however, provided different patterns. While there was no difference between the patterns in females and males, the heteromorphic sex chromosome shows reduced phylogenetic signal with evidence for rapid change along individual branches. This is supportive of a

loss of DNA on the Y chromosome and consistent with hypotheses for Y chromosome turnover through neo-Y systems. Knowing that *Drosophila melanogaster* exhibits thoracic underreplication (a stalling of the first round of replication before heterochromatic DNA is fully replicated), I scored the percentage of underreplication for approximately 100 *Drosophila* species. Phylogenetic patterns for replicated euchromatin and underreplicated heterochromatin were analyzed. Interestingly, I found that the underreplicated DNA exhibits change late in the phylogeny, suggesting an adaptive role. In the future, these patterns must be compared to other life history characteristics.

DEDICATION

This work is dedicated to my parents, Richard and Renee. To my father for constantly entertaining my questions and pushing for me to understand how things work and to my mother for her never-ending love and support.

ACKNOWLEDGEMENTS

I would first and foremost like to acknowledge my advisor and committee chair, Dr. J. Spencer Johnston. Dr. Johnston has gone above and beyond the expectations of a PhD advisor. His passion for science and answering the unknown is infectious. He is the epitome of a hard-working scientist and teacher: a perfect role-model. Without his support and encouragement I would not have reached this point. I would also like to thank his wife, Diane. They have welcomed me to their family with open arms since the beginning, allowing me to feel like I belong.

I would like to acknowledge the other members of my committee, Dr. Mariana Mateos, Dr. Aaron Tarone, and Dr. Raul Medina. Mariana was instrumental in the inception of this work. Through her teaching and advisement, my Quantitative Phylogenetics assignment quickly developed from a semester long project, to a PhD. I would like to thank Aaron for his support and constant feedback. He is always around and willing to discuss results and their implications for the field, always quick to remind me of literature I have forgotten. I would also like to thank Raul for his support in my ability to disseminate the knowledge I have gained. Without him having me repeatedly describe my project in common terminology, I would still, to this day, lack the ability to present my research in a meaningful way to a lay person.

I would like to thank those who have assisted me in these phylogenetic methods, whether it be to understand what the results mean or troubleshoot technical errors. Dr. Jim Woolley was instrumental in the development of this project. Without his

excitement and passion for teaching systematics and phylogenetics, this project would not have developed as it has. I would like to thank Dr. Ana Dal Molin for her insight and feedback on the development of my phylogenies and interpretation of my phylogenetic results.

I would like to thank the Department of Entomology for their constant support, both academically and financially. Through participation in the Linnaean Games team, I was able to attend many meetings, both regionally and nationally, in order to present my research and participate in committees. This would not have been possible without the support of Dr. David Ragsdale and Dr. Juliana Rangel. I would also like to thank Rebecca Hapes for her help and ability to quickly answer all of my questions, especially since I have never officially scheduled a meeting with her.

I would also like to thank the Department of Entomology and Dr. Pete Teel for the teaching opportunities they have given me, totaling six different courses. Through this experience I have had the opportunity to work with a large range of professors and other graduate students, teaching hundreds of undergraduates. I would like to thank the professors and instructors I worked with for their advice and support: Dr. Spence Behmer, Dr. Jim Woolley, Dr. Ordom Huot, Dr. Kevin Heinz, Dr. J. Spencer Johnston, and Dr. Aaron Tarone. I would also like to give special thanks to Ashleigh Faris. With the help of Ashleigh, I have been able to branch out from teaching entomology to also include forensics.

I could not have come this far without the support and comradery from the friends I have made here. I have made countless friends, far too many to mention here.

There are a few, however, like to acknowledge a few that have made a lasting impact on my work and life. Firstly, I would like to acknowledge Xanthe Shirley, my first friend in graduate school. Since we were lab partners in Quantitative Phylogenetics, she has been there for me. Her support ranges from learning new methods and listening to me complain about R to helping me through personal struggles in life. I would also like to thank Tyler Raszick for his friendship and advice throughout graduate school. Tyler never fails to give honest answers to questions and strong support for my decisions, especially as an officer of EGSO and as a friend. I will always respect and consider his suggestions and opinions. I hope to continue this interaction with him as a friend and colleague throughout both of our careers. I would like to thank Chloë Hawkings for her friendship and insightful discussion. Since our friendship began, she has never turned down a coffee or sneaking out early for beer, which has sparked very fruitful scientific discussions and projects. I would like to give a special thanks to my friend and roommate Joseph Melancon, not only for being a great roommate and an excellent chef, but for always being a listening ear regardless of the situation or time of the day. I would like to thank Emma Glenn for her companionship and support through my stressful last semester and the writing of this dissertation, it has been appreciated and will not be forgotten.

I would like to give a special thanks to my best friend, Trevor Carlson. While he has lived in South Dakota for the entirety of my PhD, he has had no less of an impact. Through conversing every day and his eight separate trips to visit me in Texas, he has been a pillar of support since I decided to accept this position.

Finally, I would like to thank my family: My parents, Richard and Renee, and my brothers, J.J., Bryan, and Eli. While it was difficult for some of them to see me move across the country, and all of them but my dad have no idea what I do, they have continued to give their love and support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors J. Spencer Johnston [advisor], Aaron M. Tarone, and Raul F. Medina of the Department of Entomology, as well as Professor Mariana Mateos of the Department of Wildlife and Fisheries Sciences.

Genome size and underreplication data was collected with the assistance of J. Spencer Johnston and numerous undergraduate researchers: Margaret Garrett, Melissa Mynes, Renee Holmes, and Elizabeth Piron.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the Texas A&M Merit Fellowship

TABLE OF CONTENTS

Page

Page

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION: BACKGROUND ON COMPONENTS OF GENOME SIZE EVOLUTION AND THE APPLICATION OF COMPARATIVE PHYLOGENETIC ANALYSES

Overview of genome size

The amount of DNA in an organism, termed genome size (or C-value), varies extensively across the tree of life. The C-value is not the actual sequence of its genome; it is rather the physical measure of DNA, placing this trait at the intersection of genotype and phenotype (Oliver et al. 2007). While many would assume that more DNA would result in a more complex organism, this is not the case. Humans have three billion base pairs making up their genome, yet this is no different than the genome size of the American cockroach, *Periplenata americana* (Venter et al. 2001; Hanrahan and Johnston 2011) and is vastly smaller than the massive 18 billion base pairs making up the genome of a grasshopper (Unpublished). This phenomenon is referred to as the C-value paradox (or enigma).

Eukaryotic genomes all have approximately 10^4 genes (Walbot and Petrov 2001), yet the whole genome size varies up to 200,000 fold (Gregory 2001), with up to 7,000-fold variation in animals (Palazzo and Gregory 2014). Since this is not variation in the amount of coding sequence, the variation in genome size is largely attributed to regions of nongenic and repeat sequences (Kelley et al. 2014). These regions are typically composed of transposable elements, satellite DNA, tandem repeats, and copy number variants (Gregory and Hebert 1999; Kidwell 2002; Sessegolo et al. 2016). All of these

contribute to differences in genome size. Variation in genome size between closely related species of plants, *Drosophila*, and mosquitos have been shown to be due to differential accumulation of transposable elements and repeat sequences (Black and Rai 1988; Bennetzen and Kellogg 1997; Ågren and Wright 2011; Kelley et al. 2014; Śliwińska et al. 2016; Lower et al. 2017). One study specifically compared the two sister species *Drosophila melanogaster* and *D. simulans*. They found that *D. melanogaster*, which has the larger genome of the two species, had significantly more transposable element accumulation than *D. simulans* (Vieira and Biemont 2004). When 26 species of *Drosophila* were compared across the phylogeny, there was a continued pattern of increased genome size with increased accumulation of transposable elements (Sessegolo et al. 2016).

For quite some time, these noncoding regions of DNA have been considered to be unwieldy and unnecessary, commonly termed "junk DNA" (Ohno 1972). While junk DNA was originally used to describe non-functional pseudogenes, it quickly became a catchall term for DNA which lacks coding or regulatory function (Pagel and Johnstone 1992; Palazzo and Gregory 2014). Junk DNA theories hypothesize that the genome tends to increase in the absence of constraining selection (Reviewed in Gregory 2001). Some argue that non-coding DNA, specifically those such as transposable elements, may expand for their own benefit (Dawkins 1976; Pagel and Johnstone 1992). In many cases, however, selfish DNA theories are grouped together with junk DNA. Both of these theories posit that genome size will increase until it becomes too unwieldy for the cell to contain. Therefore the amount of junk DNA is likely determined by a balance

between insertions of noncoding DNA and selective forces acting against the increase in size (Pagel and Johnstone 1992). Since there is a correlation between increased genome size and longer development (Van't Hof and Sparrow 1963; Bennett 1971; Bennett 1972), it is hypothesized that organisms with a slower development time have a higher tolerance for this "junk" DNA (Charlesworth et al. 1994). Recent work in the ENCODE project suggests that much of this "junk" DNA has function, whether regulatory or structural (Reviewed in Palazzo and Gregory 2014). While there is function behind some of this "junk" DNA, it is important to note that a portion of this DNA is not used by the organism, and has no known role outside of increasing or decreasing genome size.

This physical amount of DNA in a cell has been found to have many phenotypic correlates which natural selection may be able to act upon (Oliver et al. 2007). For example, a larger genome has been shown to lead to a larger cell volume/size (Cavalier-Smith 1982) and also an increased cell-division time (Grosset and Odartchenko 1975). These differences in cell volume have also been shown to influence physiological (Smith 1925) and genetic parameters such as cell metabolism, ion and protein exchange, levels of gene transcription, and the overall genomic landscape (Vinogradov 1995; Pritchard and Schubiger 1996; Lang and Waldegger 1997; Vinogradov 1997; Waldegger and Lang 1998; Lynch and Conery 2003). Other studies have shown that an increased genome size is correlated to cell-cycle and overall generation time (Van't Hof and Sparrow 1963; Bennett 1971; Bennett 1972), all of which are potentially targets of natural selection.

At the level of the individual, organisms can actively manipulate their genome size in order to gain an evolutionary advantage. This may occur through endopolyploidy

and by a process known as underreplication, where the genome in a cell undergoes multiple rounds of complete replication or partially complete underreplication, but does not undergo mitotic division. By utilizing this process of endoreduplication in specific tissues, organisms can benefit from the short generation and cell cycle times of a smaller genome, while achieving the physiological benefits of a larger genome cell size mentioned above (Gregory and Hebert 1999). These endopolyploid cells are fairly common in biosynthetically specialized tissues (d'Amato 1984; Lilly and Duronio 2005), such as secretory cells, and muscle cells of hymenopteran males (Aron et al. 2005; Scholes et al. 2013). Some of the best studied instances of endopolyploid cells are those in the giant salivary glands in *Drosophila*, where ten rounds of replication results in the 1024-fold enlarged salivary chromosomes, the study of which have made invaluable contributions to the genetic studies of this organism. Interestingly, the largest reported values of endopolyploidy, over one million ploid, have been observed in the silk glands of *Bombyx mori*, the silkworm moth (Perdix-Gillot 1979), suggesting that the many generations of artificial selection can have an influence on ploidy. Not only are their varying levels of ploidy throughout tissues, there can be varying levels of ploidy in an individual throughout time. Interestingly, ploidy is like other sources of genome size variation in that it does not always increase. In holometabolous insects, ploidy is typically higher in larvae than in adults, and was actually found to decrease significantly in the leg and flight muscles throughout the lifespan of adult worker honey bees (Rangel et al. 2015).

Proposed hypotheses for models of genome size evolution (Explored in Chapters II-IV)

While the C-value paradox has generally been explained through the accumulation of noncoding and "junk" DNA, the mechanisms by which these DNA are accumulated to produce this variation is still debated (Gregory and Hebert 1999; Gregory 2001; Petrov 2001; Gregory 2003a; Vinogradov 2004; Bennett and Leitch 2005; Kraaijeveld 2010; Palazzo and Gregory 2014; Kapusta et al. 2017; Lower et al. 2017). The hypotheses proposed to explain the mechanisms resulting in genome size variation can generally be phylogenetic, those related to evolutionary time, gradual change, and species relationships, or can be nonphylogenetic, where outside forces may cause bursts of change. Numerous studies have been performed to test these hypotheses; however, there continues to be as much contention as there is support for each of these.

The adaptive hypothesis proposes that genome size change is adaptive in response to life history and ecological conditions, and thus may be episodic (Reviewed in Powell 1997, Gregory 2005b, and Gregory and Johnston 2008). An example of this is the increased body size found at low temperatures, commonly referred to as Bergman's rule. Increases in body size are often the result of increased cell number or cell size, which may be controlled by increased genome size (Reviewed in Hessen et al. 2013). The pattern of large genome sizes and instances of polyploidy in cooler temperatures have been documented in salamanders (Xia 1995; Jockusch 1997), fish (Gregory and Hebert 1999), crustaceans (Beaton and Hebert 1988; Traut et al. 2007; Moens et al. 2008), and plants (Johnson et al. 1965; Stebbins 1985; Brochmann et al. 2004; Leitch

and Leitch 2008), suggesting that this extra amount of DNA allows for plasticity. The relationship between genome size and thermal environments has been supported by work in *Drosophila*. When lines of *D. melanogaster* with large and small genomes were raised in different thermal environments it was found that genome size and genome size by temperature interactions were significantly related to many developmental phenotypes (Ellis et al. 2014). Therefore, clinal variation, at least in regards to temperature, can have selective effects on genome size both inter- and intraspecfically. As such, adaptive changes are not expected to present clear phylogenetic patterns. While there is quite a bit of data to support this hypothesis, there are cases where this temperature size rule does not apply. The smallest insect sequenced to date is *Belgica antarctica*, the Antarctic midge at 99 million base pairs, yet it lives in one of the most cold and harsh environments (Kelley et al. 2014). This is opposite from what the temperature rule expects. One possible explanation for its extremely small genome size is that its heat shock proteins are constitutively expressed, which expression has been shown to suppress transposable element activity (Specchia et al. 2010).

The mutational equilibrium hypothesis, proposed by Dmitri Petrov, suggests that genome size change is gradual and due to imbalances between insertions and deletions which eventually reach an equilibrium (Petrov et al. 2000; Petrov 2001; Petrov 2002b; Petrov 2002a). A unique feature of Petrov's model is that it is mechanistic and allows genome sizes to decrease in size, increase in size, or remain at equilibrium over very long periods of time, something that is lacking in many other proposed models. Therefore, genome size will change upward or downward until the loss of DNA from the

small deletions is balanced with the gain of DNA through insertions (Petrov 2002b). This proposed imbalance and equilibrium of insertions, deletions and overall genome size takes many years to produce significant changes; for example, *D. melanogaster* is losing less than a single base pair per generation. Importantly with regard to genome size in *Drosophila*, some genomes are favored to become smaller as they have a higher deletion rate compared to insertion rates (Petrov 2002a). This downward trend is supported by the recent DGRP (*Drosophil*a Genome Research Panel) genome-wide association study, which found that deletions greatly outnumbered the amount of insertions (Huang et al. 2014). The results from this study suggest that there is stronger purifying selection against deletions than insertions.

While there can be increases or decreases in the genome size, the changes are found to be proportional: larger genomes can handle larger changes in size, whereas smaller genomes exhibit smaller changes. Larger genomes will show a more rapid change, suggesting a positive correlation between genome size and the tempo of evolution (Oliver et al. 2007). The mutational equilibrium hypothesis therefore relies on variation in the DNA insertions and deletions rather than natural selection acing directly on genome size, as in the adaptive hypothesis. While this model seems theoretically sound, there are arguments that the mutational equilibrium model is far too slow to influence the variation we see in genome size and that there is not sufficient support for this model in the literature (Gregory 2003b; Gregory 2004). However, a recently proposed hypothesis, the accordion model, may be able to address the above issues (Kapusta et al. 2017). The newly proposed model would show similar phylogenetic

patterns and gradual changes throughout a phylogeny, yet may produce variation at a fast enough rate to influence the variation we see in genome size. Here, large deletions can counteract the insertions from transposable elements, resulting in an accelerated rate of change.

Lynch and Conery have proposed that changes in genome size are occurring primarily during speciation events due to coincidently small species level effective population sizes (Lynch and Conery 2003; Lynch 2007a). Selection is ineffective at lower effective population sizes, therefore potentially maladaptive changes in genome sizes may be able to accumulate and persist. Since large genomes are often thought to be unwieldly and deleterious, it is hypothesized that these maladaptive changes that occur in small populations increase the size of the genome. In contrast, species with large effective population size are less likely to tolerate large changes in genome size, as selection is far more effective. Therefore, there should be a phylogenetic pattern to genome size evolution. Under the effective population size hypothesis change should occur in speciation events. Change in size, contrary to the mutational equilibrium model, will not be gradual, but will have evidence of bursts throughout the phylogeny.

The effective population size hypothesis was supported in a large comparative study which found that larger genome sizes correlated with species with low effective population sizes (Lynch and Conery 2003). While there was found to be a general pattern between these two variables, it is always important to incorporate phylogenetic relationships when making contrasts between species (Felsenstein 1985). When comparative phylogenetic methods were applied to this same dataset, the relationship

between effective population size and gene number, intron size, intron number, transposon number, and overall genome size disappeared (Whitney and Garland Jr 2010; Whitney et al. 2011). Another article suggests that population size is confounded with multiple aspects of organismal biology that may statistically affect the results of the original analyses from Lynch and Conery (Charlesworth and Barton 2004). The disparate results from these analyses with the same data highlight the controversy of this hypothesis for genome size evolution, as well as highlighting the importance of incorporating phylogenetic relationships to test these hypotheses.

Use of phylogenetic comparative methods (Explored in Chapters II-V)

As mentioned above, general comparisons of traits, such as the comparison of genome size to effective population size, can be made, but they may fall victim to autocorrelation by not incorporating phylogenetic information. This problem arises when species are treated as individual, independent data points. Species, however, are not independent; they should be viewed as part of a greater phylogeny, highlighting the relationships of closely related species that may share trait similarities. While the similarity of these traits could be due to independent evolutionary events, it could also be due to similarities between species due to common ancestry (Felsenstein 1985; Harvey and Clutton-Brock 1985). Nested ANOVAs utilizing taxonomic information (ex: order, family, genera) have been used in an attempt to address these issues of relationship (Clutton-Brock and Harvey 1977; Clutton-Brock and Harvey 1979; Harvey and Clutton-Brock 1985; Bell 1989). An issue with this use of taxonomic information to account for

relationships is that the Linnaean classification system assumes direct monophyly and may not account for issues of paraphyly and differences in divergence time among higher order groups that may be discovered using phylogenetic comparative methods (Felsenstein 1985; Lynch 1991). The use of methods such as nested ANOVA, like other commonly used statistical methods, assumes independence of species relationships and could cause issues with overstatement of significance. Garland et al (1993) found that incorporation of phylogenetic comparisons into studies may reduce the probability of finding statistical differences, giving more power to significant results.

Issues of non-independent evolutionary relationships can be avoided with adequate phylogenetic information and by utilizing modern comparative methods of analysis such as Abouheif's C_{mean} (Abouheif 1999), Blomberg's K (Blomberg and Garland 2002), and Pagel's evolutionary parameters (Pagel 1999). These comparative methods detect phylogenetic autocorrelation and phylogenetic signal. Phylogenetic autocorrelation is the tendency for species near each other on a phylogeny to be more similar, not taking into consideration time, measured by branch lengths (Cheverud et al. 1985; Abouheif 1999). Phylogenetic signal is the tendency for species to resemble each other due to common ancestry, implying that species that have diverged recently will be more similar, with disparities between species becoming greater as evolutionary time passes (Blomberg and Garland 2002; Blomberg et al. 2003). The presence of signal in itself does not indicate a mechanism/pattern for evolution, but this information in conjunction with other parameters of evolution may give support to a proposed pattern of evolution.

Abouheif's C_{mean} is a method of detecting phylogenetic autocorrelation through a modification of Moran's I statistic (Moran 1950; Abouheif 1999; Pavoine et al. 2008). Through the use of a test for serial independence (Von Neumann et al. 1941) on a subset of all possible topology arrangements of the phylogeny, a random, normal distribution is developed in order to test the C value from the original topology (Young 1941). The significance of this C value indicates whether the trait values are expected to be random or significantly influenced by the phylogenetic relationships. Since this method does not require branch length information to make a conclusion on phylogenetic autocorrelation, it is generally used as a precursor test before utilizing more exhaustive comparative methods.

Blomberg et al. (2003) developed a more complete method to detect phylogenetic signal for traits that incorporates evolutionary time through inclusion of branch length information. The new statistic of phylogenetic signal, K, they develop allows comparisons of traits across different phylogenies and between separate traits based on a Brownian motion model (Blomberg et al. 2003; Münkemüller et al. 2012). A K value of one indicates that the trait of interest has phylogenetic signal expected for the Brownian motion model, whereas values of K less than one indicate there is deviation from the expected variation. Deviation implies that there may be some level of adaptation influencing the evolution of the trait of interest.

Pagel developed three specific parameters to describe evolutionary patterns in trait evolution λ, κ, and δ (Pagel 1999). The first parameter, λ, describes the level of phylogenetic signal, or phylogenetic dependence, for the trait. A λ value of one

indicates there is complete phylogenetic signal, suggesting that the amount of variation found can be explained by the branch lengths and is consistent with the expected variation from a Brownian motion model of evolution (Felsenstein 1985; Harvey and Pagel 1991). When λ is estimated to be zero it indicates that the phylogeny cannot explain the variation found in the trait values and that the trait values can be described by a star phylogeny, or lack of clear phylogenetic relationships. Values of λ between zero and one indicate that phylogenetic relationships can explain some of the variation between traits, yet cannot explain all of the variation found. Simulation data has found that λ is a statistically powerful parameter of evolution and is robust to incomplete phylogenetic information, which gives support to its use in modern comparative analyses (Freckleton et al. 2002).

The second two parameters Pagel developed, κ and δ, are unique to this comparative method, and can provide information regarding the mode and tempo of trait evolution (Pagel 1999). The κ parameter tests when change occurs on individual branches of a phylogeny, separating gradual change from punctuated change. When κ values are estimated to be one, there is gradual change on branches. Values of κ above one indicate that change occurs late on the branch, whereas values of κ below one indicate change occurs rapidly on individual branches. The δ parameter, is similar to κ, except that it tests for change along the entire phylogeny, along the entire path, root to tip. Values of one indicate gradual change, values below one indicate early change in the phylogeny, and values above one indicate that change occurs late, near the tips of the phylogeny. Therefore, delta and kappa can supplement information about the

phylogenetic signal (λ) by determining a general pattern of evolution throughout the entire tree (δ) and on individual branches (κ).

A recent study found that under the expectations of a Brownian model of evolution, Abouheif's C_{mean} and Pagel's λ performed significantly better and were more reliable at determining phylogenetic signal than Moran's I and Blomberg's K (Münkemüller et al. 2012). The authors utilized simulations in order to determine if species number, polytomies, and missing branch length information had any effect on the reliability and sensitivity of estimations of phylogenetic signal assuming Brownian motion. They found that while the number of species included affected the values estimated for phylogenetic signal, polytomies and missing branch length information did not greatly affect the reliability of the results. Pagel's λ was least affected by increasing species numbers and had no resulting effects from polytomies. They conclude, based on their results, that the comparative analyses should place most emphasis on Pagel's λ and Abouheif's C_{mean} for phylogenetic signal and autocorrelation.

As mentioned above, the hypotheses for models of genome size evolution are hotly debated (Gregory and Hebert 1999; Gregory 2001; Petrov 2001; Gregory 2003b; Vinogradov 2004; Bennett and Leitch 2005; Kraaijeveld 2010; Hanrahan and Johnston 2011; Palazzo and Gregory 2014; Kapusta et al. 2017). Some studies may have support with databases (Lynch and Conery 2003), but after accounting for evolutionary relationships these significant patterns disappear (Whitney and Garland Jr 2010; Whitney et al. 2011). This highlights the importance of utilizing comparative phylogenetic methods for understanding the patterns of genome size evolution. A

number of recent studies have started to utilize these comparative methods to discern patterns of evolution, but there seems to be contradicting patterns, depending on the taxa being studied, suggesting that genome size change may occur differently depending on the group being studied. For example, genome size was not found to have significant phylogenetic signal in the snapping shrimp (Jeffery et al. 2016) and seed beetles (Arnqvist et al. 2015), yet were found to have significant signal in *Drosophila* (Sessegolo et al. 2016) and fireflies (Lower et al. 2017), with significant relationships to transposable element content in the latter two studies. With more studies utilizing these methods, it is important that the reliability of these methods be tested utilizing genome size data to determine how these patterns of change are maintained or differ between closely related groups.

Drosophila as a model for genome size evolution (Explored in Chapters II-V)

The *Drosophila* genus has been widely studied in all disciplines of biology, including phylogenetics, speciation, and evolution of genome architecture (Gregory and Johnston 2008; van der Linde and Houle 2008; van der Linde et al. 2010; Kellermann et al. 2012a; Kellermann et al. 2012b; Hjelmen and Johnston 2017). This genus is separated into subgenera, *Sophophora* and *Drosophila*, which are estimated to have diverged 40-65 million years ago (Russo et al. 1995; Tamura et al. 2004; Obbard et al. 2012). The *Drosophila* genus, however, is replete with taxonomic issues, as the genus and the *Drosophila* subgenus are paraphyletic. Genera such as *Zaprionus*, *Scaptodrosophila*, and *Hirtodrosophila* are all resolved to be found within the major

subgenus *Drosophila* (Reviewed in van der Linde et al. 2010). These taxonomic issues are an example of why phylogenetic relationships, and not just the Linnaean classification system, need to be accounted for when studying traits, such as genome size evolution (Felsenstein 1985; Lynch 1991).

The two subgenera of *Drosophila*, *Sophophora* and *Drosophila*, are distinct karyotypically, with *Drosophila* having the ancestral 6 chromosomes, whereas *Sophophora* has 4, a reduction likely due to fusion events that formed large metacentric autosomes (Painter and Stone 1935, Reviewed in Schulze et al. 2006). This difference provides an opportunity for comparison in *Drosophila*. The relationship between chromosome count and genome size has been controversial. In fishes, there is a significant positive relationship between chromosome number and genome size (Hinegardner and Rosen 1972; Yi and Streelman 2005). On the other hand, while there is some evidence for this pattern in plants (Pandit et al. 2014; Escudero et al. 2015), there are just as many instances where this pattern is not evident (Fleischmann et al. 2014; Gorelick et al. 2014). A recent phylogenetic study of snapping shrimp found that genome size correlates to chromosome size, but not number (Jeffery et al. 2016).

Studies of genome size evolution in the *Drosophila* genus have focused on the *Sophophora* subgenus (Gregory and Johnston 2008; Hjelmen and Johnston 2017). This does not take advantage of the diversity in the genus. The large amount of evolutionary time and karyotypic differences between the subgenera suggest that there may be differences within *Drosophila*, and that additional studies with a focus on the *Drosophila* subgenus were needed. Not only do *Drosophila* have the benefit of being studied so

intensely for decades, but countless species are readily available from the species stock center, allowing rapid increases in the available data. The species availability and the availability of published sequence data for the genus on GenBank makes it possible to quickly increase the statistical power for phylogenetic analyses. This wealth of information and availability of species for the genus allows researchers to pose very large, ambitious evolutionary studies, in this case, for genome size evolution.

Drosophila also provide opportunities to study the evolution of heterochromatin. The available estimates for female and male genome size and the presence of an X-Y sex determination system give unique and interesting possibilities for understanding differences in genome size evolution between sexes due to heteromorphic chromosomes, and more specifically to examine Y chromosome degradation (discussed below). In addition to chromosomal differences between sexes, *Drosophila* species have been shown to underreplicate their genomes, a process we report (Chapter 5) unique to the genus, by which not all of the largely structural heterochromatin is replicated (Lakhotia 1984; Belyaeva et al. 2006; Johnston et al. 2013). This allows us to ask how different structural DNA may influence genome size evolution (explored below and in Chapter 5).

Drosophila as a model of sex chromosome evolution (Explored in Chapters III and IV)

While there have been attempts to study the variation in genome size extensively across organisms (Gregory 2005a; Hanrahan and Johnston 2011), much of this work is related to genome size for the species (average of the male and female) or only from one

sex, primarily the female. This excludes some interesting comparisons. In relation to genome size differences between sexes, it has been found that genome size is positively associated with reproductive fitness in the seed beetle *Callosobruchus maculatus* (Arnqvist et al. 2015) and negatively correlated to the song attractiveness in some male grasshoppers (Schielzeth et al. 2014), suggesting that sexual selection may be acting on genome size evolution between the sexes. These results suggest that males and females may experience very different life history selection parameters, which may, in turn, influence the size of the sex chromosomes. Since we know that variation in genome size within a species can lead to divergent phenotypes (Ellis et al. 2014; Huang et al. 2014) and that genome size variation has impacts on reproductive fitness (Arnqvist et al. 2015), it is important to ask if it matters if we look at males or females. We can also ask: What can the differences in size of the sex chromosomes tell us about the patterns of whole genome size evolution?

Drosophila species are known to determine sex through the heteromorphic X-Y system, meaning males have one X and one Y, while females have two X chromosomes. Heteromorphic sex chromosomes are hypothesized to originate as a homologous pair with female- and male-sterility alleles (Goodfellow et al. 1983; Charlesworth et al. 2005). Over time, the Y chromosome becomes more heterochromatic through the accumulation of transposable elements and inversions, which may reduce the rate of recombination and increase instances of chromosome breaks (Charlesworth et al. 2005; Bachtrog 2013). Accumulation of transposable elements will likely occur in the coding and regulatory regions, inactivating genes, leading to the loss of genes and gene function

on the Y chromosome (Matsunaga 2009). These changes, in conjunction with differential selection on sex specific mutations, will result in the permanent heterozygosity expected for X and Y chromosomes (Muller 1918; Charlesworth et al. 2005; Bachtrog 2013). Interestingly, this process would first result in an increase in size of the Y chromosome before it inevitably degenerates (Charlesworth and Charlesworth 2000).

Instances where the male has a larger size than a female (a situation with a large Y chromosome), indicates the presence of a neo-Y system. One benefit of working with *Drosophila* is that there are well studied instances of these neo-Y systems. For example, the neo-Y system in *D. miranda* is hypothesized to have occurred through a Y-autosome fusion about 1.2 million years ago (Bachtrog et al. 2008; Matsunaga 2009). While it has had a more than 20 fold greater accumulation of repetitive sequences in comparison to the X chromosome, it still harbors many functional genes (Bachtrog et al. 2008). A study on *D. busckii*, another species with a reported neo-Y chromosome, found that almost 60% of the neo-Y genes have become non-functional in less than one million years, suggesting that Y chromosome degradation occurs very quickly after formation (Zhou and Bachtrog 2015). *D. albomicans* and *D. pseudoobscur*a have also been reported to have neo-Y chromosomes, yet be in different stages of Y chromosome degradation (Reviewed in Bachtrog 2013). *D. albomicans* is reported to have a younger Y chromosome than *D. miranda*, and *D. pseudoobscura* to have an older Y chromosome, so the sequence events of Y chromosome degradation can be studied with relative ease. The Y chromosome in *D. pseudoobscura* has been found to be almost

entirely heterochromatic, a process which has taken a relatively short 17 million years (Bachtrog 2013). These instances of neo-Y chromosomes are highly different than the standard example of *D. melanogaster*, which has become almost entirely heterochromatic and tiny (Adams et al. 2000; Skaletsky et al. 2003). During this process of gain and loss of chromosome content, the genome size, in particular the relative size of the X and Y chromosome and the chromatin content of each may change significantly. Whether these chromosomal responses to sex selection extend to genome size variation among species in the two sexes is largely unknown.

Underreplication: A key to the study of evolution of different types of chromatin (Explored in Chapter V)

The genome is composed of two structural forms of chromatin: euchromatin and heterochromatin. Euchromatin is generally considered to be the largely expressing portion of the genome, containing the majority of genes for regular expression. Heterochromatin, however, is mostly structural, containing very few genes and is mostly transcriptionally silent. When DNA replicates, euchromatin is replicated first (Schübeler et al. 2002; McNairn and Gilbert 2003; MacAlpine et al. 2004; Schwaiger and Schübeler 2006), followed by heterochromatin (Lima-de-Faria and Jaworska 1968). Studies of mutations rates found that the early replicating regions of the genome have a lower mutation rate than late replicating regions, possibly explaining why coding regions are clustered in euchromatic portions of the genome (Wolfe et al. 1989; Schwaiger and Schübeler 2006; Makunin et al. 2014).

Interestingly, there is in *Drosophila* an unusual, if not unique, process termed underreplication wherein the replication process stalls during the S-phase, before the replication of heterochromatin is complete (Lakhotia 1984; Hammond and Laird 1985; Leach et al. 2000; Belyaeva et al. 2006). These underreplicated regions, in general, have low gene density, are tightly packed, and are considered to be transcriptionally inert (Belyaeva et al. 2008; Belyaeva et al. 2012). The process of underreplication is best known from studies of the polytene chromosomes of salivary glands (Rudkin 1969; Hammond and Laird 1985) and nurse cells (Painter and Reindorp 1939) in *Drosophila* and other Diptera. Underreplication was found to be controlled through the Suppressor of Underreplication (SuUR) in *Drosophila* species (Belyaeva et al. 1998). In the absence of this gene product, heterochromatin replication was found to occur earlier in S-phase, whereas overexpression of the SuUR gene resulted in more underreplication (less overall replication) (Zhimulev et al. 2003a). The SuUR gene encodes for a 962 amino acid protein which binds to late replicating portions of the heterochromatin, a process which likely physically slows the process of replication in these regions (Makunin et al. 2002). This could be occurring through modification of repressive chromatin complexes and structures (Volkova et al. 2003) or by directly influencing binding efficiency of replication machinery or altering stability of the replication fork (Zhimulev et al. 2003a).

Underreplication was analyzed phylogenetically in 11 species of *Drosophila* by comparing the number of amino acid substitutions across the SuUR gene and its orthologs (Yurlova et al. 2010). It was found that SuUR can be considered "fast-

evolving." Despite these rapid changes in amino acids throughout time, the size and overall charge of the SuUR protein product is mostly the same across the analyzed species, indicating high conservation of the protein function. For example, when comparing the distantly related *D. melanogaster* and *D. grimshawi,* nearly all secondary structures of the SuUR protein were the same. Low homology orthologs were found in three mosquito species (*Anopheles gambiae*, *Aedes aegypti,* and *Culex quinquefasciatus*), yet no orthologs were found outside of Diptera. This suggests that while underreplication may occur in Diptera, the highest conservation is within *Drosophila*.

Since we know that genome size is highly correlated to noncoding sequences, such as transposable elements, repeats, introns, and intragenic spacers (Bennetzen and Kellogg 1997; Gregory and Hebert 1999; Hartl 2000; Gregory 2001; Kidwell 2002; Vieira and Biemont 2004; Ågren and Wright 2011; Kelley et al. 2014; Sessegolo et al. 2016), which are found primarily in heterochromatic regions, it is hypothesized that the process which underreplicates heterochromatin should be related to genome size. A strong relationship between genome size and underreplication of the 16C polyploid ovarian follicle cells was reported in Bosco et al. (2007). The authors concluded that the addition and loss of heterochromatic satellite repeats could explain the large variation found within *Drosophila* species. We expand these studies to include underreplication of DNA in the thorax. Until recently, underreplication has only been found in polytenized cells, such as salivary glands and ovaries. Recently, however, we found that a majority of thoracic cells in *D. melanogaster* exhibited underreplication (Johnston et

al. 2013). This finding could have many consequences for genome structure and interpreting data for thoracic tissue in *Drosophila*, such as: position effects, transcription levels, and general genome architecture (Belyaeva et al. 2003; Zhimulev and Belyaeva 2003; Belyaeva et al. 2006). For our studies, the underreplication of DNA in the thorax is ideal, since the replication stalls between G1 and G2. This makes possible a very precise measure of the extent of underreplication that is free of the confounding effects of multiple rounds of endoreduplication. Since underreplication is known to be conserved in *Drosophila* species (Yurlova et al. 2010), we hypothesized that the phenomenon of thoracic underreplication should occur across the *Drosophila* genus and, similar to the study by Bosco et al. (2007), should have a significant relationship to genome size. Therefore, we hypothesized that there is variation in the amount of replication among species related to genome size, and we can use these values to estimate the effects of different amounts of structural DNA on genome size evolution.

How this dissertation proposes to adds to the knowledge base related to genome size evolution

As discussed above, the proposed models for genome size evolution remain controversial and lack phylogenetic support. Here, we significantly expand the number of genome size estimates for *Drosophila* species and utilize comparative phylogenetic methods to investigate the patterns of genome size evolution across the *Drosophila* genus (Chapter IV). Since *Drosophila* species are so well studied in a phylogenetic aspect and are readily available from stock centers, they are the optimal group for

developing phylogenetic methods for understanding patterns of genome size evolution. We utilize available sequences and genome size data to ask how genome size is changing in female species as well as how the number of species utilized impact the reliability of comparative analyses. This will allow us to know the threshold number of species for utilizing these methods for genome size evolution (Chapter II).

The X-Y sex determination system in these species also allows us to ask how genome size change may be influenced at the chromosome level. While, the difference in sex chromosomes may not show discernable differences between male and female whole genome size evolution, these comparative methods will allow us to see how patterns of chromosome degradation occur across the phylogeny (Chapters III and IV).

The availability of species from stock centers allows us to provide many new estimates for genome size, concentrating on species in the underrepresented subgenus (*Drosophila*) (Chapter 4). The inclusion of this information allows me to ask how closely related species (*Sophophora* vs *Drosophila*) may differ in genome size evolution. Do I see differences that reflect their karyotypic differences? The inclusion of male and female differences here allows us to ask again how the X-Y system evolves. We can therefore see if there are differences in the instances of potential neo-Y systems and better understand the evolutionary processes behind chromosome degradation (Chapter IV).

Utilizing *Drosophila* species not only allows us to investigate genome size change at the whole genome size and chromosome level, but also at the level of structural DNA. The phenomenon of underreplication in thoracic tissue of *Drosophila*

allows us to easily estimate the portion of the genome that is largely heterochromatic versus euchromatic. Since we know genome size variation is primarily influenced by noncoding regions of DNA, this portion of the study provides a unique and new method for investigating change in the genome (Chapter V).

Once these patterns and methods are better developed in this well-studied model organism, we can continue to ask these questions in a wide range of species. We can continue to collaborate with sequencing groups to quickly develop comparable phylogenetic datasets to ask these same questions in other systems. We strive to eventually determine if these phylogenetic patterns are consistent across species with larger genomes, different karyotypes, and if they may differ between different taxonomic rankings, such as family, order, and type of metamorphosis.

CHAPTER II

THE MODE AND TEMPO OF GENOME SIZE EVOLUTION IN THE SUBGENUS *SOPHOPHORA*[1]

Introduction

When considering trait evolution, sequence of the genotype is traditionally inspected for evidence of selection or drift, through methods such as DN/DS ratios. These tests, however, are not easily applied to genome size. Genome size, like gene expression, is an intermediate phenotype; while the trait is directly influenced by the sequences in the genome, there is not a specific sequence tied to it, and it must therefore be analyzed in a phenotypic fashion. Genome size has been found to vary up to 200,000-fold in eukaryotes (Gregory 2001) and up to 7,000-fold in animals (Palazzo and Gregory 2014), and seems to bear no correlation with organismal complexity among eukaryotic taxa. The wide variation in genome size is not generally attributed to coding DNA sequences, but rather to repetitive and nongenic DNA (Gregory and Hebert 1999; Gregory 2001; Kidwell 2002). For these reasons, we are analyzing the variation in DNA content in a comparative phylogenetic context with the goal of establishing an evolutionary model for genome size from proposed hypotheses.

Among the many questions that remain to be answered in the field of genome size evolution is what drives or constrains genome size (Vinogradov 2004). One of the

[1] *Reprinted with permission from "The mode and tempo of genome size evolution in the subgenus *Sophophora*" by CE Hjelmen and JS Johnston, 2017, PloS one, 12.3, 2017, e0173505, Copyright 2017 Hjelmen, Johnston.

most fundamental questions asks, “Is the accumulation of nongenic DNA adaptive or just tolerated by selection?” and, “If the accumulation is adaptive, what benefits does it have? What mechanisms underlie genome size change?” Of the many hypotheses for genome size evolution (reviewed in Gregory 2001), we focus on the effective population size hypothesis (Lynch and Conery 2003; Lynch 2007b), mutational equilibrium hypothesis (Petrov et al. 2000; Petrov 2001; Petrov 2002b), and the adaptive hypothesis of genome size evolution (Powell 1997; Gregory 2003a; Gregory and Johnston 2008).

Lynch and Conery propose that changes in genome size occur primarily during speciation events due to coincident small species level effective population sizes (Lynch and Conery 2003; Lynch 2007b). They argue that selection is ineffective at lower effective population sizes, and therefore potentially maladaptive changes in genome size may accumulate and persist in the population (e.g. increases in genome size from transposable elements). In contrast, species with large effective population sizes will be less likely to tolerate large changes in genome size, due to more effective selection.

The mutational equilibrium hypothesis proposes that genome size change is gradual and is due to an imbalance in indels (insertions and deletions) that through time, eventually achieve a mutational equilibrium (Petrov et al. 2000; Petrov 2001; Petrov 2002b). Some genomes are favored to move towards smaller sizes because they tend to have higher deletion rates compared to insertion rates (Petrov 2002a). The change, whether downwards or upwards, can be considered gradual, yet proportional; larger genomes can handle larger insertions and deletions than smaller genomes. The above hypotheses suggest different rates and modes of phylogenetic change. The effective size

hypothesis would produce change early in the speciation process, when species effective size is small. In contrast, an insertion deletion balance could increase or decrease genome size, yet the change would accumulate gradually over phylogenetic time. While proposed to explain the variation in genome size, some consider this hypothesis to be largely theoretical and yet to be supported by a large dataset (Gregory 2003b; Gregory 2004). Recently, a related accordion model, which proposes that increases in DNA content by transposable elements is balanced by large segment deletions, was supported by data from 10 species of mammals and 24 species of birds (Kapusta et al. 2017). The model differs from the mutational equilibrium model only in that the change involves larger segments and has the potential for faster rates of evolution.

A third hypothesis, that of adaptive genome size evolution, was summarized by Powell (1997) and is reviewed by Gregory (2005b) and Gregory and Johnston (2008). The adaptive hypothesis proposes that genome size should track the environment; environmental change results in genome size change. Because species evolve to utilize habitats uniquely, we would expect adaptation to uncouple genome size and the phylogeny. While there could possibly be some phylogenetic signal throughout time in this hypothesis, when organisms shift to new ecological environments, punctuated shifts in signal are expected. As such, adaptive changes are unlikely to present clear phylogenetic patterns. It is important to note, however, that if these shifts in habitat that drive genome size change are shared by members of a clade, the relationship between adaptation and signal may be difficult to untangle. As discussed below, the parameters κ and δ that were developed to capture signal produced by changes that are either early or

late in the phylogeny and either early or late in individual branches respectively help us pull out some of these relationships.

Even though much data has been accumulated for genome sizes (5,635 species to date according to genomesize.com (Gregory 2005a), the ever present genome size variation has been largely ignored from a phylogenetic standpoint. The importance of this issue is highlighted in a phylogenetic analysis of the data presented as proof of the relationship of species effective size and genome size change (Lynch and Conery 2003; Whitney and Garland Jr 2010). The proposed significant relationship between effective population size and genome size is lost when accounting for the phylogeny, leaving Lynch's effective population size hypothesis for genome size evolution conjectural. In general, the lack of phylogenetic consideration has resulted in a lack of knowledge about how changes in genome size have occurred throughout evolutionary history, whether random or adaptive and selected.

In an effort to address the lack of consideration of phylogenetic relationships among species when analyzing genome size variation, we produced a phylogeny of Drosophilidae, with a focus on *Sophophora,* using aligned sequence data for 87 species. The resulting tree and associated branch lengths was used to generate Pagel's parameters of evolution, Blomberg's K, Moran's I, and Abouheif's C_{mean} , for genome size evolution in *Sophophora* (Moran 1950; Abouheif 1999; Pagel 1999; Blomberg et al. 2003). If Pagel's λ and Blomberg's K are approaching one, the presence of phylogenetic signal would suggest that genome size is not evolving according to the adaptive hypothesis. If there is signal, and Pagel's κ and δ values are approaching one, this would

suggest gradual change of genome size throughout time, supporting the mutational equilibrium hypotheses. If there is signal, with δ and κ values below one, this would suggest early change in branches and the tree, supporting the low effective population hypothesis.

The complete analysis utilized a relatively large number of species (87) and a 3X range of genome size values to generate parameter estimates. To determine the reliability of these phylogenetic analyses with different numbers of taxa, we generated the same parameter estimates with reduced taxa numbers and reduced ranges of genome size. Several, but not all of the parameters are sensitive to taxon number; genome size range had little effect on the results.

Methods

Genome size database

Genome sizes for species were obtained from published datasets (Gregory and Johnston 2008), with additional data from the laboratory database of J. Spencer Johnston. Genome sizes were estimated using the flow cytometric method (Hare and Johnston 2011) for species obtained from the UC San Diego Species Stock Center (http://stockcenter.ucsd.edu) (Table 1).

Gene sequences and alignment

Sequence data for the 16 genes used to create a molecular phylogeny (4 mitochondrial and 12 protein coding genes) (*COI*, *COII*, *COIII*, *Cytb*, *Amy*, *AmyRel*,

Ddc, *boss*, *SNF*, *Marf*, *Sod*, *per*, *Wee*, *HB*, *ADH*, and *fkh*) was downloaded from NCBI Genbank and aligned using MAFFT v.7 online using iterative refinement methods (http://mafft.cbrc.jp/). Aligned sequences were visually inspected for irregularities in amino acid translation in Mesquite 2.75 and corrected by hand as needed.

Model testing

Each sequence alignment was analyzed in JModelTest 2.1.4 to determine the model of sequence evolution that provided the best likelihood value (Darriba et al. 2012). The likelihood search assumed 11 possible substitution schemes, allowing for both invariant sites and gamma distributions. A fixed BIONJ-JC tree was used for the likelihood calculations. All runs returned the same suggested best model for phylogeny construction, a GTR substitution model with a gamma distribution and invariant sites.

Data file preparation and tree construction

All sequences were interleaved to produce a 10,382 bp alignment. Missing sequence data was imputed for taxa that did not have gene sequences for every gene. Missing data does not influence the results of branch lengths or phylogenetic relationships. This resulted in an average of 7 genes per taxa, with a maximum of 15 and a minimum of 3 genes.

Table 1: Genome size estimates for 87 species of Drosophilidae.
Genome sizes were obtained from published literature and the laboratory database of J. Spencer Johnston. Species were obtained from the UC San Diego Stock Center.

Stock Number	Species	Genome Size (Mbp)	Stock Number	Species	Genome Size (Mbp)
20010-2010.00	*C. amoena*	395.2	14028-0651.00	*D. triauraria*	245.5
20000-2631.01	*C. procnemis*	280.1	-	*D. quadraria*	252.1
13000-0081.00	*D. busckii*	139.9	14028-0661.03	*D. rufa*	255.4
-	*D. mauritiana*	157.9	14028-0591.00	*D. mayri*	257.5
14021-0224.01	*D. erecta*	158.9	14029-0011.00	*D. fuyamai*	264.1
14021-0251.195	*D. simulans*	159.6	14028-0481.00	*D. baimaii*	278.4
14028-0541.00	*D. kanapiae*	165.5	-	*D. orena*	280.7
-	*D. teissieri*	166.3	-	*D. lucipennis*	291
14021-0248.25	*D. sechellia*	179.9	14028-0731.00	*D. pectinifera*	297.4
-	*D. varians*	166.7	-	*D. suzukii*	342.8
14021-0261.01	*D. yakuba*	170.7	-	*D. pseudoobscura*	167.7
-	*D. santomea*	171.5	-	*D. miranda*	175.6
4021-0231.36	*D. melanogaster*	174.5	-	*D. tolteca*	179
14025-0441.05	*D. ficusphila*	190.8	-	*D. ambigua*	186.8
-	*D. birchii*	191.2	14011-0111.49	*D. persimilis*	197.1

Table 1 Continued.

Stock Number	Species	Genome Size (Mbp)	Stock Number	Species	Genome Size (Mbp)
14027-0461.03	*D. elegans*	192.2	-	*D. azteca*	199.5
-	*D. pallidosa*	194.1	-	*D. affinis*	200.5
14024-0361.00	*D. atripex*	195.9	-	*D. barbarae*	200.5
14024-0371.13	*D. ananassae*	196.6	-	*D. greeni*	201.5
14020-0011.01	*D. tani*	199	14028-0671.01	*D. jambulina*	202.7
14028-0641.00	*D. punjabiensis*	200.8	-	*D. bifasciata*	205.4
-	*D. phaeopleura*	202.9	-	*D. narragansett*	205.9
14024-0381.19	*D. bipectinata*	204.6	14028-0561.14	*D. kikkawai*	210.2
-	*D. malerkotliana*	204.9	14012-0161.00	*D. algonquin*	211
14022-0311.13	*D. takahashii*	207.3	-	*D. bicornuta*	213.7
-	*D. parabipectinata*	210.8	-	*D. diplacantha*	232.8
14028-0611.01	*D. orosa*	211.9	-	*D. biauraria*	237.2
-	*D. pseudotakahashii*	212.2	15085-1641.03	*D. hydei*	206.8
-	*D. mimetica*	212.7	14041-0831.00	*D. neocordata*	212.3
-	*D. serrata*	213.3	14042-0841.09	*D. emarginata*	214.1

Table 1 Continued.

Stock Number	Species	Genome Size (Mbp)	Stock Number	Species	Genome Size (Mbp)
-	*D. bunnanda*	215.2	-	*D. virilis*	325.4
14028-0671.02	*D. seguyi*	215.2	-	*D. nebulosa*	187.3
14028-0601.00	*D. nikananu*	216.1	14030-0791.01	*D. sucinea*	209.6
14028-0701.00	*D. tsacasi*	217.7	14030-0721.00	*D. capricorni*	211.8
-	*D. lutescens*	219.1	14030-0771.00	*D. paulistorum*	231.8
14028-0711.00	*D. vulcana*	222.7	-	*D. equinoxialis*	247.9
-	*D. ercepeae*	224	-	*H. pictiventris*	162.8
-	*D. pseudoananassae nigrens*	224	80000-2761.03	*S. leonensis*	261.8
-	*D. prostipennis*	227.4	-	*S. stonei*	206.8
-	*D. pseudoananassae*	228.4	11010-0021.00	*S. lebanonensis lebanonensis*	210.3
-	*D. eugracilis*	228.9	-	*S. pattersoni*	213.2
-	*D. paralutea*	230.8	50001-0001.02	*Z. tuberculatus*	197.6
14028-0571.00	*D. lacteicornis*	242.6	50000-2744.02	*Z. sepsoides*	212.8
14028-0471.00	*D. auraria*	245.1	-	-	-
*Minimum = 139.9, Maximum = 395.2, Mean = 215.5, Median = 210.8					

A phylogeny for *Sophophora* was reconstructed using a supermatrix model of phylogeny construction utilizing MrBayes 3.2.3 on the CIPRES supercomputer (http://www.phylo.org/) with four chains and four runs and a GTR gamma + I evolutionary model for 32,835,000 generations (sampling every 1,000) using a Dirichlet prior of (1, 0.5, 1, 1) (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Parameter output was visualized in Tracer v 1.6 to assure the four runs had reached convergence and to determine burn-in. The consensus tree was then visualized in FigTree v.1.4.2. Genome size was mapped onto the phylogeny using the ContMap function from the phytools package from R3.3.0 (Revell 2012). Multiple trees were constructed with varying Bayesian priors to test if there were any issues with branch lengths (Dirichlet (1,1,1,1), exponential (10)) and a maximum likelihood tree.

Tree manipulation and significance tests with different numbers of taxa

To test for the effect of taxa number on significance levels in a phylogenetic signal analyses, multiple reduced taxa phylogenies were made (5, 10, 15, 20, 30 taxa) with 20 different trees for each group. Trees were constructed by randomly trimming the taxa from the original *Drosophila* tree utilizing the drop.tip function in the package ape from R 3.3.0 (Paradis et al. 2004), while maintaining tree topology and branch lengths. Taxa retained for each tree were chosen by random number generation.

Trait analyses

Comparative phylogenetic analyses (Pagel's parameters of evolution, Blomberg's K, Moran's I, and Abouheif's C_{mean}) were run on both the full phylogeny and each reduced taxa phylogeny with genome size a continuous trait. Pagel's lambda (λ) and Blomberg's K test for phylogenetic signal assuming Brownian motion. Pagel's kappa (κ) tests how traits evolve along branch lengths ($\kappa < 1$, early change; $\kappa = 1$, gradual change). Pagel's delta (δ) tests how traits change from the overall path on the tree, from root to tip ($\delta < 1$, rapid early change, $\delta = 1$, gradual change; $\delta > 1$, increasing rate of change). All comparative phylogenetic analyses were completed using functions and packages available in R. Pagel's parameters of evolution were measured using the function PGLS from package caper (Orme 2013). Blomberg's K was estimated using the phylosignal function from package picante (Kembel et al. 2010). Moran's I and Abouheif's C_{mean} values were calculated using the function abouheif.moran with 999 permutations from package adephylo (Jombart and Dray 2008).

Alternative test for adaptive hypothesis of genome size evolution

In order to test the alternative adaptive hypothesis for genome size, climatic data for these *Sophophora* species (critical thermal maximum, maximum temperature, minimum temperature, annual mean temperature, annual precipitation, precipitation from the wettest month, precipitation from the driest month, and latitude) were mined from two Kellerman et al. papers on phylogenetic constraint of climatic variables in *Drosophila* (Kellermann et al. 2012a; Kellermann et al. 2012b). This totaled 38 species

of *Sophophora*. These variables were analyzed with multiple regression and phylogenetic generalized least squares (PGLS) analyses utilizing the function pgls from package caper (Orme 2013).

Statistical analyses

Statistical tests of fit for each comparative phylogenetic analysis is provided with output of each. For taxa number analysis, the phylogenetic values for each of the 20 runs with taxa number dataset were visualized with boxplots. The effect of taxa number on estimated phylogenetic values was tested using ANOVA in R with number of taxa treated as a random variate. In order to understand how Pagel's λ was affected by increasing taxa number, statistical differences (p-values for H_o: λ = 1.0 or 0.0) were plotted in a boxplot using values from each of the 20 runs at each level of reduction. These p-values were then compared with ANOVA using number of taxa as a random variate. Genome size range was also used as a covariate in an ANCOVA in R (λ p-value = Genome size range + taxa number +genome size range*taxa number) in order to see if there was an interaction between range in genome size and taxa number effect on λ significance values.

Results

Genome size

Genome size for the female of each species is given in Table 1. Genome size for the *Sophophora* subgenus of *Drosophila*, plus a few outgroups from *Drosophila*

subgenera (*Zaprionus, Scaptodrosophila, Scaptomyza, Hirtodrosophila* and *Chymomyza*) ranges from 139.9 Mb to 395.2 Mb with an average of 215.5 and a median of 210.8 (Table 1).

Sophophora phylogeny

The overall *Sophophora* phylogeny is well supported with high posterior probability values at each node, most being 1, with the lowest support value being 0.53 (Figure 1). The relationships in this phylogeny are supported by other large *Drosophila* phylogenies in the literature (Consortium 2007; Da Lage et al. 2007; Gregory and Johnston 2008; van der Linde et al. 2010). These results suggest that the constructed phylogeny is representative of the true relationships found in *Sophophora*, and should have reliable branch lengths. No significant differences were found among trees constructed with varying Bayesian priors.

Figure 1: Bayesian phylogeny of *Sophophora.*
Phylogeny of 87 Drosophilidae reconstructed using MrBayes 3.2.3 with a focus on *Sophophora*. Nodes with posterior probabilities lower than 80 are indicated with '*'. Genome size is visualized in color: smaller sizes in red, larger in blue, and intermediate in green.

Tests of rate and mode of genome size variation

Model fit of the complete dataset (87 genome sizes) in a phylogenetic context using the above phylogeny with branch lengths shows that phylogenetic relatedness is a significant component of genome size variation among *Sophophora*. All tests for phylogenetic signal/dependence (λ, Blomberg's K, Abouheif's C_{mean}, and Moran's I) indicate complete signal with high significance values (Table 2), most notably λ=0.987. Genome size across the phylogeny was also found to have a κ value of 0.971 and a δ value of 0.589.

Table 2: Comparative phylogenetic output for *Sophophora* phylogeny.
Pagel's λ, Blomberg's K, Abouheif's C_{mean}, and Moran's I all found significant phylogenetic dependence for genome size in *Sophophora*. The δ value suggests an early burst of change followed by deceleration in genome size change. The κ value suggests that genome size is changing gradually on individual branches.

Test	Value	Lower Sig.	Sig. from 1
λ	0.987	6.88E-15	0.166
δ	0.589	4.23E-10	0.174
κ	0.971	5.03E-11	1
Blomberg's K	1.373	0.001	-
Abouheif C_{mean}	0.240	0.001	-
Moran's I	0.180	0.001	-

Multiple regression and PGLS for climatic variables

After incorporating phylogenetic relationships, climatic variables failed to be significantly related to genome size variation in 38 species of *Sophophora.* Multiple

regression analysis indicated that genome size was significantly influenced by the climatic variables (p = 0.015, Adj. R-squared = 0.30). When this model was analyzed utilizing PGLS to incorporate phylogenetic relationships, the pattern disappeared (p = 0.602, Adj. R-squared = 0.044, Table 3).

Table 3: Phylogenetic generalized least squares results for genome size and climatic variables.

	Estimate	**Std Error**	**t value**	**Pr(>|t|)**
(Intercept)	41.726	379.297	0.11	0.913
Critical Thermal Maximum	3.865	8.024	0.48	0.634
Minimum Temperature	2.303	5.032	0.46	0.651
Maximum Temperature	8.941	5.524	1.62	0.116
Annual Precipitation	-0.005	0.022	-0.25	0.807
Precipitation of Wettest Month	0.110	0.137	0.81	0.427
Precipitation of Driest Month	0.167	0.314	0.53	0.599
Latitude	-1.874	1.652	-1.13	0.266
Annual Mean Temperature	-12.545	8.622	-1.46	0.156
Residual S.E. = 166.8, Multiple R-squared = 0.1821, Adj. R-squared = -0.04358, F-statistic = 0.8068, DF = 29, p-value = 0.602				

Taxa number analyses

Effects on mean values

When subsets of taxa are analyzed, means for λ, Blomberg's K, Moran's I, and Abouheif's C_{mean} all increased with an increase in taxon number, indicating an increased signal of phylogenetic dependence with increased taxa number (Table 4). However, a significant differences in the estimated parameter value for different taxon numbers was

only found in Moran's I ($p = 1.67e\text{-}05$) and Abouheif's C_{mean} ($p = 0.0469$). No significant effect was found with increasing taxa number for λ values (Figure 2A).

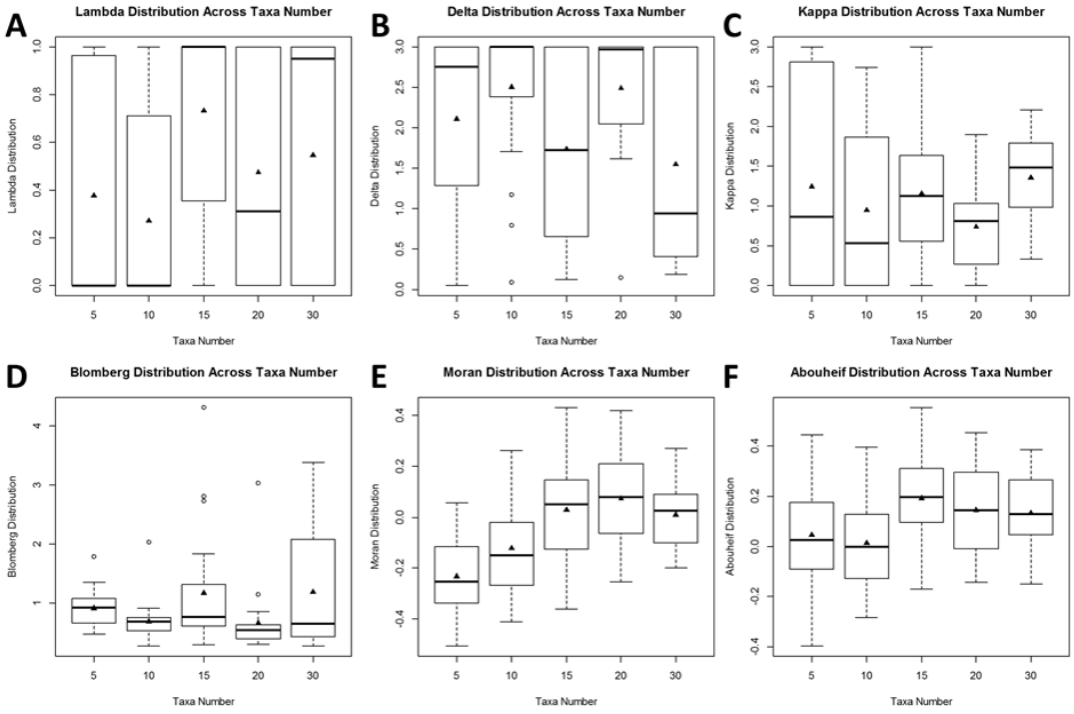

Figure 2: Boxplots for each phylogenetic analysis.
Raw values from comparative phylogenetic tests are plotted for each group of taxa. There is no clear pattern with increasing taxa number for Pagel's parameters of evolution or Blomberg's K; however, there is an increase in values for both Moran's I and Abouheif's C_{mean}. These differences are tested statistically in Table 4.

Table 4: ANOVA results and means for each phylogenetic analysis.
Phylogenetic value means from each phylogenetic analyses for taxa number (5, 10, 15, 20, and 30 taxa) datasets were compared using ANOVA. Given the results, Moran's I and Abouheif's C_{mean} are significantly affected by taxa number.

λ						**δ**					
	Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)		Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)
Taxa	1	0.4	0.4005	1.69	0.197	Taxa	1	3.72	3.72	3.291	0.0727
Residuals	98	23.22	0.2369			Residuals	98	110.75	1.13		
Taxa	5	10	12	20	30	Taxa	5	10	12	20	30
Mean	0.408	0.445	0.482	0.519	0.592	Mean	2.32	2.208	2.095	1.983	1.759
κ						**Blomberg's K**					
	Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)		Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)
Taxa	1	0.11	0.1114	0.136	0.713	Taxa	1	0.85	0.8523	1.605	0.208
Residuals	98	79.98	0.8161			Residuals	98	52.05	0.5311		
Taxa	5	10	12	20	30	Taxa	5	10	12	20	30
Mean	1.045	1.065	1.084	1.103	1.142	Mean	0.806	0.86	0.914	0.967	1.075
Abouheif's C_{mean}						**Moran's I**					
	Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)		Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)
Taxa	1	0.15	0.14967	4.05	0.0469	Taxa	1	0.74	0.7399	20.51	1.67E-05
Residuals	98	3.621	0.03965			Residuals	98	3.535	0.0361		
Taxa	5	10	12	20	30	Taxa	5	10	12	20	30
Mean	0.0564	0.0789	0.1014	0.1239	0.1688	Mean	-0.158	-0.108	-0.058	-0.008	0.092

Effects on significance

In contrast to the effect of taxon number on the means, the number of significant differences of the λ estimate from the boundary conditions, 0.0 or 1.0, increased with taxa number. As shown in the boxplots, λ p-values decreased and had lower variation with increasing taxa numbers, indicating that higher taxa numbers convey higher confidence in the results for each test (Figure 3). The variation among p-values for different taxa numbers was statistically significant (p=0.000771, ANOVA, Table 5). A significant decrease of the p-value for Blomberg's K was also (Figure 3) observed with increasing taxa number (p = 3.65e-07, Table 5).

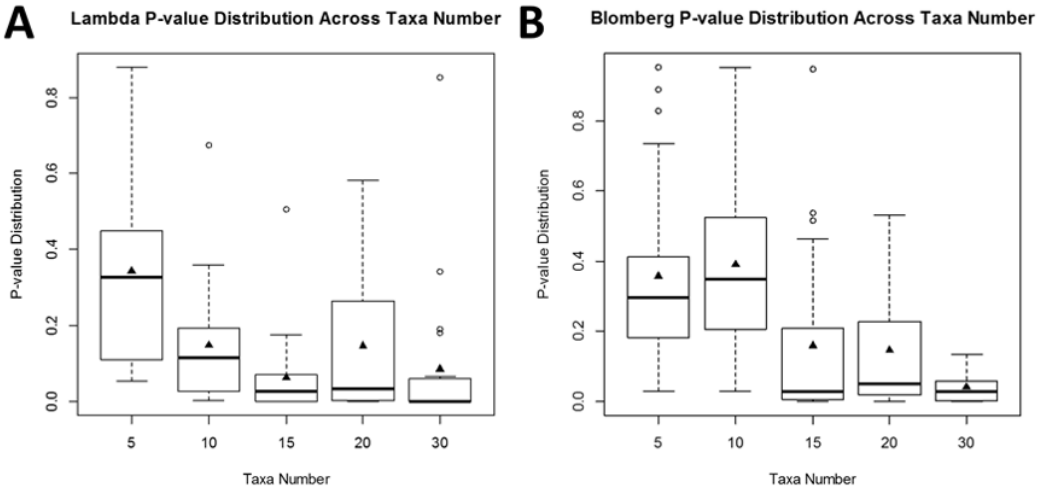

Figure 3: Boxplots of significance values for λ and Blomberg's K analyses. Plotted significance values from phylogenetic signal tests of λ and Blomberg's K decrease as the number of taxa in the analyses increase, most notably above 15 taxa. These are tested for significance in Table 5.

Table 5: ANOVA and means for P-values for λ and Blomberg's K analyses.
P-values from the opposite bounds for Pagel's λ and Blomberg's K for the taxa

λ P-value					
	Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)
Taxa	1	0.468	0.4682	12.05	0.000771
Residuals	98	3.807	0.0388		
Taxa	5	10	12	20	30
Mean	0.2452	0.2054	0.1657	0.1259	0.0464
Blomberg's K P-Value					
	Df	Sum Sq.	Mean Sq.	F Value	Pr(>F)
Taxa	1	1.5	1.4999	29.78	3.65E-07
Residuals	98	4.936	0.0504		
Taxa	5	10	12	20	30
Mean	0.376	0.305	0.234	0.162	0.02

Because experimental error is expected to make up an ever larger proportion of the total variation when the measured range of genome sizes is small, we tested whether the tests of phylogenetic signal are sensitive to genome size variation among taxa. Genome size range was used as a covariate in an ANCOVA in order to determine if the range in genome size contributed to the significance of the λ results among the taxa number datasets. While the ANCOVA model was significant ($p < 0.01$), there was no significant interaction between genome size range and taxa number ($p = 0.263$). Genome size range did not contribute significantly to the model ($p = 0.516$), while the taxa number contribution was highly significantly ($p < 0.001$, Table 6).

Table 6: ANCOVA results for genome size range compared to λ p-values across taxa numbers.
An ANCOVA indicates that there is no interaction between taxa number and genome size range. Only taxa number contributed significantly to the model.

ANCOVA for λ P-value vs. Genome Size Range and Taxa Number					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Taxa	1	0.468	0.468	12.02	0.0008***
Gsrange	1	0.017	0.017	0.42	0.5164
Taxa&Gsrange	1	0.049	0.049	1.27	0.2632
Residuals	96	3.741	0.039		
R^2=0.1249, p=0.005, f-statistics=4.569					

Discussion

Here, we use a comparative phylogenetic approach to investigate genome size in *Sophophora* species. We specifically look at measures of phylogenetic signal (Pagel's λ, Blomberg's K, Abouheif's C_{mean}, and Moran's I) and at measures of mode and tempo of evolution (Pagel's δ and κ) in order to test three hypotheses of genome size evolution: low effective population size hypothesis, mutational equilibrium hypothesis, and an adaptive hypothesis.

Genome size was found to have complete phylogenetic signal for *Sophophora* (λ = 0.987, Blomberg's K = 1.373, Abouheif's C_{mean} = 0.240, Moran's I = 0.180, Table 2). Based on our expectations for the adaptive hypothesis, the presence of complete phylogenetic signal suggests that little, if any, of the genome size variation is evolving in an adaptive fashion. This conclusion is also supported by the results of the PGLS analysis, which found that genome size is not significantly related to climatic variables (Table 3). Interestingly, when these climatic variables were phylogenetically analyzed

by Kellerman et al., they were found to have complete phylogenetic signal (Kellermann et al. 2012a; Kellermann et al. 2012b). Since, genome size and climatic variables both have phylogenetic signal, we can assume that any patterns we see between these characteristics in *Drosophila* in non-phylogenetic aspects are due to constraints of the phylogeny, not a direct relationship. Also, when inspecting the trait mapped on the phylogeny, there is not significant evidence for bursts of change in concordance with the adaptive hypothesis, aside from the decrease in the *D. melanogaster* clade (Figure 1). Rather, genome size evolution is reliant upon phylogenetic patterns. These results are supported by recent work on genome size evolution in *Drosophila* species (Sessegolo et al. 2016). Here, Sessegolo *et al.* investigated the impact of phylogeny on genome size and transposable elements for 26 species of *Drosophila* utilizing available sequences and a de novo transposable element assembly approach. They found a significant correlation between genome size and global transposable element content, with strong phylogenetic signal for each. While simple repeats accounted for up to 1% of the repeatome, LTRs and LINE elements were found to be major components. These data suggest that the genome size variation of *Drosophila* species are largely driven by transposable elements. The current study, while not including information on proportions or dynamics of repeat sequences, has largely expanded the number of taxa used an earlier study from 26 to 87. By increasing the number of taxa, we can hope to determine if the overarching patterns of genome size evolution in *Drosophila* remain consistent and better identify which species may be of interest for full sequence investigation.

Genome size was also found to have a κ value of 0.971 (Table 2), indicating that genome size evolves in a gradual fashion and reflects individual branch lengths. The phylogenetic signal, the gradual manner in which genome size is changing, and the relationship of branch length and amount of change supports the mutational equilibrium hypothesis (Petrov 2002b). However, there have been concerns expressed about this hypothesis, as it seems to be largely theoretical and has yet to have a large enough dataset to support it (Gregory 2003b; Gregory 2004). Small imbalances between insertions and deletions are not likely to move fast enough to change genome size dramatically, especially when it seems as if genome size is being driven by relatively large insertions of transposable elements (Sessegolo et al. 2016). The recent accordion model shows that dynamic changes in transposons may be associated with large deletions and lead to apparent stasis of genome size (Kapusta et al. 2017). However, transposition and large deletions, if imbalanced, would drive genome size evolution at an accelerated rate. However, neither stasis nor an imbalance of transposition and large deletions would necessarily produce the phylogenetic signal observed for these 87 species.

Interestingly, a δ value of 0.589 (Table 2) indicates that the rate of genome size change was not always constant. This δ value suggests change occurred rapidly early in the phylogeny, likely at the formation of the *Drosophila* genus, with a decrease in that rate as time went on throughout *Sophophora*. The early change in genome size could be due to low population sizes, which would appear to support the effective population size hypothesis (Lynch and Conery 2003). However, genome size in this group has moved

towards smaller sizes rather than increased sizes (Petrov 2002a), contradicting the hypothesis that lower effective population sizes lead to larger genome sizes. The original burst could therefore be adaptive. If so, the smaller genome sizes could have been due to selection on one of their phenotypic correlates (Oliver et al. 2007). Specifically, selection could have acted on smaller cell and body sizes or shorter development and cell cycle times (Van't Hof and Sparrow 1963; Bennett 1971; Bennett 1972; Grosset and Odartchenko 1975; Cavalier-Smith 1982). It is also possible that the low effective population was inefficient at selecting out a slightly deleterious, non-adaptive trait, such as increased deletion rate. After early, relatively large genome size changes, the rate of evolution could have slowed to the current, gradual rate. It is important to note that the change in rate would have had to happen quickly to not be reflected in the κ values.

It is important to ask if further sampling will change the conclusions above. If the change in *Sophophora* genome size does actually fit the mutational equilibrium hypothesis, it is possible that heavier sampling of the genus or subgenus could fill in the gaps for the large change early in the tree. While this is a possibility, taxon sampling issues addressed in this study, suggest that the significance values and the magnitude of these values vary little when overall study size reaches n = 30. The number of taxa examined here (n = 87) is well above that. The importance of a large enough sample size for tests of phylogenetic signal cannot be ignored. Increases in significant measures of phylogenetic signal with taxa number were found to increase with taxon number in both Abouheif's C_{mean} and Moran's I. This suggests the results of these two methods are

sensitive to taxa number and they should be used sparingly, more so as preliminary tests for comparative studies. On the other hand, while Abouheif's C_{mean} and Moran's I were sensitive to increasing taxa number, there were no significant effects of taxa number on either Pagel's λ or Blomberg's K, suggesting the results of these methods are less sensitive to taxa number. At the same time, while there are no clear patterns for the magnitude of the parameter values of phylogenetic signal, there is a significant difference in the p-values obtained at different taxa numbers. Since the p-values are measures of significant differences from the bounds (signal vs. no signal), they can be considered proxies for test reliability. Based on these analyses, sample sizes of at least 15 are necessary to achieve reliable results in terms of significance for Pagel's λ and Blomberg's K (Figure 3). The pattern of increased reliability (statistical p-value from the bounds) continues as the taxa number increases; the best results are obtained with larger taxon sampling. These results are supported by a previous study that tested the effectiveness detecting phylogenetic signal using simulated taxa with ranges of Brownian motion (Münkemüller et al. 2012).

The number of taxa is important, yet the range in the trait value across the tree could also affect the reliability of the phylogenetic signal results. Narrow or wide ranges in variation could skew the interpretation of these comparative results. However, we found that sampling from the range of genome size in *Sophophora*, had a non-significant effect ($p = 0.5164$) and no significant interaction was found between genome size range and taxa numbers ($p = 0.263$). Only taxa number was found to be significantly contributing to the fit of the ANCOVA model (Table 6). Reduced taxa results, in

conjunction with previous results using simulated datasets (Münkemüller et al. 2012), show the strength of these tested methods for calculating phylogenetic signal. Most emphasis should be put into Blomberg's K and Pagel's parameters of evolution, as they are least sensitive to taxa number affecting the calculated phylogenetic signal value. However, these two methods must have at least a minimal sample size (15-20) to achieve reliable results. While there are some taxa number effects on phylogenetic signal estimates for Abouheif's C_{mean} and Moran's I, they still are good quick, preliminary measures for phylogenetic signal before the use of more robust comparative methods, such as Pagel's λ and Blomberg's K.

While the signal detected here rejects a non-phylogenetic model of change, it has yet to fully support one of the proposed phylogenetic patterns of change (effective population size vs. mutational equilibrium). The early burst of change ($\delta = 0.589$) would seem to fit the small species effective size hypothesis, yet the trend is to a decrease rather than an increase in genome size, suggesting that this change could be due to adaptation or selection. The gradual change ($\kappa = 0.971$) in genome size after that burst suggests a model similar to the mutational equilibrium hypothesis with large deletions balancing out the large insertions due to transposable elements. We argue therefore, that the rapid early change in *Sophophora* may represent an increase in deletion rate, and possibly an adaptive radiation associated with selection for rapid development rate and small size. Subsequent change is gradual as expected of a deletion insertion balance.

CHAPTER III

A PHYLOGENETIC ANALYSIS OF GENOME SIZE IN MALE *SOPHOPHORA* AND ITS IMPLICATIONS FOR Y CHROMOSOME EVOLUTION

Introduction

Genome size (also termed C-value) varies widely across organisms, with up to 7,000 fold variation in animals alone, yet does not correlate with complexity in eukaryotes (Mirskey and Ris 1951; Palazzo and Gregory 2014). This has been commonly referred to as the C-value paradox (or enigma) (Gregory 2001). The C-value paradox has been generally explained by the variation in amounts of nongenic and repeat regions, resulting from transposable elements, satellite DNA, tandem repeats, and even copy number variation, rather than differences in the amounts of coding sequences (Gregory and Hebert 1999; Kidwell 2002; Kelley et al. 2014). Differential accumulation of transposable elements has been reported to explain much of the variation in genome size between closely related species in both plants and *Drosophila* (Bennetzen and Kellogg 1997; Ågren and Wright 2011; Śliwińska et al. 2016). When comparing the sister species *D. melanogaster* and *D. simulans*, there was found to be considerably less evidence for transposable elements in *D. simulans*, the species with a smaller genome (Vieira and Biemont 2004). A more recent example found a significant relationship between genome size and global transposable element content when looking at 26 species of *Drosophila* in a phylogenetic context. (Sessegolo et al. 2016).

While the C-value paradox has been explained through the accumulation of noncoding DNA, the mechanisms and patterns for the resulting variation are still

debated. Many hypotheses for the long term mechanisms of change have been proposed, ranging from imbalances in insertions and deletions, population genetic forces, to outside adaptive forces. There is support in some way for each of these hypotheses, yet there does not appear to be one answer to the question of genome size variation. The mutational equilibrium hypothesis, which proposes that genome size changes slowly over time due to an imbalance between insertions and deletions (Petrov et al. 2000; Petrov 2001; Petrov 2002b; Petrov 2002a), has been criticized for lacking support in large datasets and because the mechanisms behind the change are too slow (Gregory 2003b; Gregory 2004). The low effective population size hypothesis, which suggests that larger, more deleterious genomes are not selected out of populations in instances of low effective population sizes (Lynch and Conery 2003), lost its statistical support when phylogenetic analyses were introduced (Whitney and Garland Jr 2010). However, there has been a recently proposed hypothesis, known as the accordion model, which seems promising for explaining genome size variation (Kapusta et al. 2017). Here large insertions due to transposable elements may be counteracted by larger deletions. This may provide a similar pattern to the mutational equilibrium model, but evolve at a fast enough rate to fit large datasets within its predictions.

While there have been attempts to study the variation in genome size extensively across organisms (Gregory 2005a; Hanrahan and Johnston 2011), much of this work is only related to genome size for the species (average of male and female) or only from one sex, primarily the female. Recent large studies of *Drosophila melanogaster* have found that there is significant variation within the species, suggesting that there is

potential for differences in genome size to lead to divergent phenotypes (Ellis et al. 2014; Huang et al. 2014). In fact, when *D. melanogaster* lines were raised in different thermal environments, it was found that genome size and genome size by temperature interactions were significantly related to many important developmental phenotypes, such as time to pupation and survival to pupation and adulthood (Ellis et al. 2014). While there is much support for the neutral models of genome size evolution, these results suggest that the variation in genome size in a species may have significant impact on the evolutionary ecology of species, suggesting an adaptive model.

Interestingly, in relation to genome size differences between sexes, it has been found that genome size is positively associated with reproductive fitness in the seed beetle *Callosobruchus maculatus* (Arnqvist et al. 2015). Genome size has also been found to be negatively correlated to the song attractiveness in some male grasshoppers, suggesting that sexual selection may be acting on genome size evolution between sexes (Schielzeth et al. 2014). This suggests that males and females may have very different life history selection parameters, which can, in turn, influence the size of the sex chromosomes. Since we know that variation in genome size within a species can lead to divergent phenotypes (Ellis et al. 2014; Huang et al. 2014) and that genome size variation has impacts on reproductive fitness (Schielzeth et al. 2014; Arnqvist et al. 2015), it is important to ask: Does it matter if we look at males or females? What can the difference in size of the sex chromosome tell us about the patterns of whole genome size evolution?

Heteromorphic sex chromosomes are hypothesized to originate as a pair of homologous chromosomes which contain male-sterility and female-sterility alleles (Goodfellow et al. 1983; Charlesworth et al. 2005). Over time, there is an increase in the amount of heterochromatin on the Y chromosome from the accumulation of transposable elements maintained in part by inversions, which may reduce the amount of recombination and also increase the instances of chromosome breaks (Charlesworth et al. 2005; Bachtrog 2013). Addition of transposable elements is also likely to occur in coding and regulatory regions, which could inactivate genes leading to eventual loss of genes and gene function (Matsunaga 2009). These changes, along with the differential selection on X and Y specific mutations, will result in the permanent chromosomal heterozygosity and reduce selection on the Y chromosome (Muller 1918; Charlesworth et al. 2005; Bachtrog 2013). Interestingly, this process would first result in an increase in size of the Y chromosome before it inevitably degenerates (Charlesworth and Charlesworth 2000). In the case of X-Y systems, males with a larger genome size than females indicate the presence of a neo-Y system. During this process of gain and loss, the genome size, in particular the relative size of the X and Y chromosome and the amounts of heterochromatin and euchromatin in each, may change significantly. Whether these chromosomal responses to sexual selection generally extend to genome size variation among species in the two sexes is unknown.

Here, we report the genome sizes for both males and females of 87 species of Drosophilidae, with a focus on the subgenus *Sophophora*. Values from both sexes are compared utilizing modern phylogenetic comparative methods. Female *Sophophora*

have already been analyzed in a phylogenetic context (Hjelmen and Johnston 2017), yet it is unknown whether the slight differences in size of male and female whole genomes exhibit the same or different evolutionary patterns. Since the difference in size due to heteromorphic sex chromosomes is small in relation to the size of the entire genome, the phylogenetic study of male whole genome size is not expected to provide dramatically different results from the previous study. However, knowledge of both male and female genome size also allows estimation of differences due to Y-chromosome size. This estimate can serve as a proxy for Y-chromosome size, which can then be analyzed in a similar fashion in order to answer questions related to Y-chromosome evolution. Once the autosomal portion of the genome is removed from the analysis, the size difference in heteromorphic sex chromosomes becomes proportionally more important. The X-Y difference should show phylogenetic patterns indicating a more "adaptive" pattern of evolution (explained below) as selection is debatably acting more on the sequences in the sex chromosome rather than amount of sequence. This adaptive pattern would be supported by a reduced amount of phylogenetic signal ($\lambda < 1$), and departure from gradual change throughout ($\delta \neq 1$, $\kappa \neq 1$). These differences in mode and tempo of change in size would potentially pinpoint the formation of neo-Y chromosomes and the rapid degradation of Y-chromosomes. The inclusion of colorized trait phylogenies (Revell 2012), in addition to Pagel's parameters (Pagel 1999), will help visualize where large amounts of change occur as well as where the variation in size is more conserved, showing us where variation arises.

Methods

Genome size database

Genome sizes were obtained for male and female *Sophophora* from published datasets (Gregory and Johnston 2008), with additional species from the laboratory database of J. Spencer Johnston. Genome sizes were estimated using flow cytometry (Hare and Johnston 2011) for species from the UC San Diego Species stock center (http://stockcenter.ucsd.edu). Sex difference for each species was calculated by subtracting double the male genome size (2A + X/Y) from double the female genome size (2A + X/X). Negative values suggest the species has a Y chromosome which is larger than the X chromosome. A t-test was performed in order to determine if there are statistical differences between male and female genome sizes. A Kolmogorov-Smirnov test was also performed in order to see if there is a significant difference in the distribution of female and male genome sizes.

Gene sequences and alignment

The *Sophophora* phylogeny used in this study was obtained from a phylogenetic analysis of genome size in female *Sophophora* (Hjelmen and Johnston 2017). This phylogeny utilized 16 genes (4 mitochondrial and 12 protein coding genes) to create a molecular phylogeny using a supermatrix method (using *COI*, *COII*, *COIII*, *Cytb*, *Amy*, *AmyRel*, *Ddc*, *boss*, *SNF*, *Marf*, *Sod*, *per*, *Wee*, *HB*, *ADH*, and *fkh*). Genes were obtained from NCBI Genbank, aligned using MAFFT v.7 online (http://mafft.cbrc.jp/), and corrected by hand after inspection in Mesquite 2.75.

Model Testing

Each alignment was analyzed in JModelTest 2.1.4 to determine the best model of sequence evolution according to the best likelihood value (Darriba et al. 2012). This likelihood assumed 11 possible substitution schemes, which accounted for invariant sites and gamma distributions. A fixed BIONJ-JC tree was used for all likelihood calculations.

Data file preparation and tree reconstruction

The interleaved sequence alignment of 10,382 bp was prepared with missing data input for genes which were not available for all species. Missing data is coded to not influence phylogenetic relationships or branch lengths. This resulted in an average of seven genes per taxa, with a maximum of 15 genes and a minimum of 3 genes.

The phylogeny was reconstructed utilizing MrBayes 3.2.3 on the CIPRES supercomputer (http://www.phylo.org/) with four chains and four runs and a GTR gamma + I evolutionary model for 32,835,000 generations (sampling every 1,000 generations) using a Dirichlet prior of (1, 0.5, 1, 1) (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Parameter output was visualized in Tracer v1.6 to assure the four runs had reached convergence and to determine burn-in. A consensus tree was constructed using a burn-in period of 10%. The tree was then visualized in FigTree v.1.4.2. Genome size (for males and females) and sex differences due to the Y chromosome were then mapped onto the phylogeny using the ContMap function from the phytools package from R 3.3.0 (Revell 2012). Multiple trees were constructed with

varying priors to test if there were any issues with branch length. No variability was found between these models.

Male genome sizes were analyzed on the phylogeny using Pagel's parameters of evolution for comparison to the results found in Hjelmen and Johnston (2017) for female genome size (Pagel 1999). These analyses were run assuming a Brownian motion model. In addition to whole genome size comparison, the difference in the genome due to the heteromorphic Y chromosome was calculated by taking the difference between the male and female genome size. This difference value was then plotted on the phylogeny for the same comparative analyses as the male genome size. All comparative phylogenetic analyses were completed in R utilizing the function PGLS from package caper (Orme 2013).

Results

Genome size information

Genome size and difference in genome size due to the heteromorphic Y chromosome for each species can be found in Table 7. The distribution of genome sizes for females and males was visualized in histogram form (Figure 4). The average female genome size was found to be 217.5 Mbp and the average male genome size was found to be 211.2 Mbp. No significant statistical differences were found between the sexes ($t = 0.955$; $p = 0.34$). There was also no statistical difference between male and female genome size distributions when tested with Kolmogorov-Smirnov ($D = 0.12644$, $p = 0.49$). Instances where the value for sex difference is below zero indicates genome size

of males were estimated to be larger than females'. The average of the estimated sex difference due to the Y chromosome was found to be 12.7 Mbp, indicating that female genomes are larger than male genomes on average. The distribution of this sex difference is plotted in Figure 5. As expected, instances of Neo-Y chromosomes are rare, and there are very few instances of species where the difference due to sex is less than zero (Table 7), resulting in a very small tail on the left side of the distribution (Figure 5).

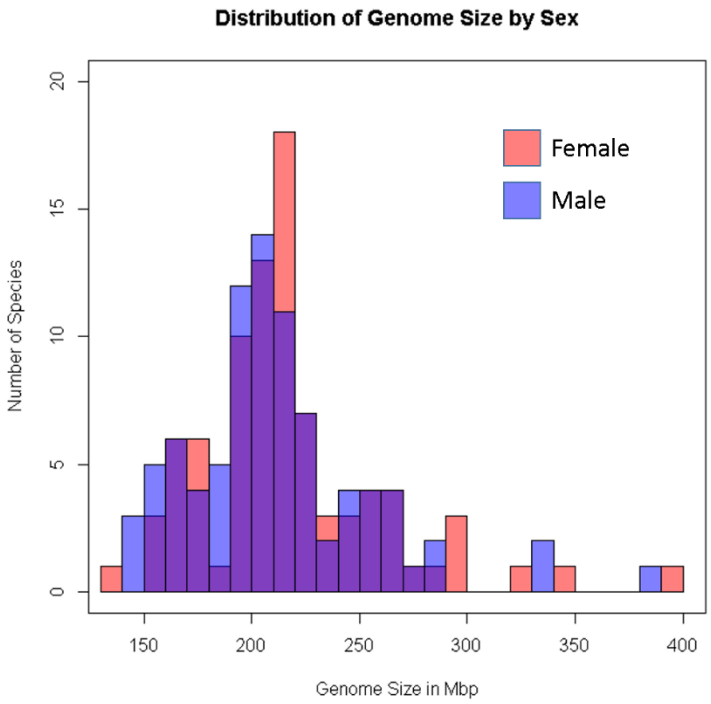

Figure 4: Distribution of Female and Male Genome Size in *Sophophora.*
Genome size is plotted in histogram form to visualize the differences between females and males of *Sophophora* species. Females have a mean of 217.4 Mbp, whereas males have a slightly smaller mean of 211.3 Mbp. These differences are not statistically different (t-test, p = 0.34). These distributions were also not statistically different (Kolmogorov-Smirnov, D = 0.12644, p = 0.49).

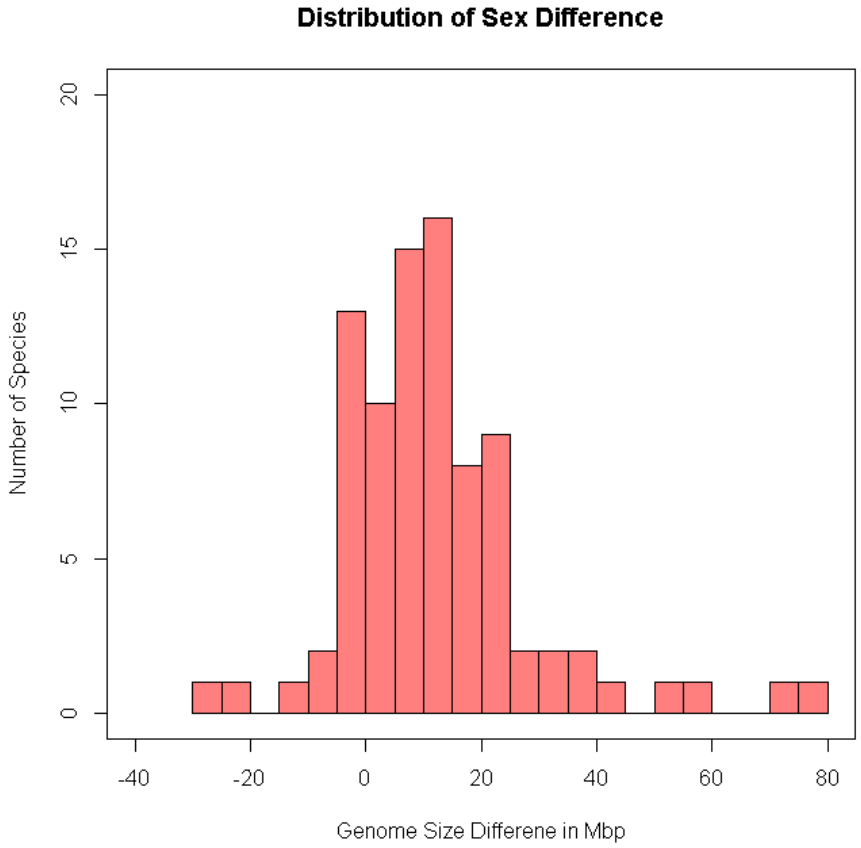

Figure 5: Distribution in Difference in Genome Size due to Y chromosome in *Sophophora.*
The difference between genome size between females and males was calculated by subtracting double the male genome size from double the female genome size. These values were then visualized in histogram form. On average the difference was found to be 12.7 Mbp, with a minimum value of -25.6 Mbp and a maximum value of 77.3 Mbp.

Table 7: Female and male genome sizes for 87 species of Drosophilidae and measured sex difference.
Genome size (Mbp) for each species of Drosophilidae included in the phylogenetic analyses. Sex difference was calculated by subtracting male size from female size (2*XX – 2*XY). Positive values indicate female genome size is larger than males, whereas negative values indicate male genome size is larger than females.

Stock Number	Species	Female Genome Size (Mbp)	Male Genome Size (Mbp)	Sex Difference (Mbp)
20010-2010.00	*C. amoena*	395.2	384.8	20.8
20000-2631.01	*C. procnemis*	298.7	261.5	74.4
-	*D. affinis*	200.5	182.4	36.2
14012-0161.00	*D. algonquin*	205.2	194.2	22.0
-	*D. ambigua*	186.8	175.1	23.5
14024-0371.13	*D. ananassae*	196.6	188.8	15.6
14024-0361.00	*D. atripex*	198.2	194.8	6.8
14028-0471.00	*D. auraria*	254.4	248.4	12.0
-	*D. azteca*	199.5	192.2	14.7
14028-0481.00	*D. baimaii*	279.2	273.6	11.2
-	*D. barbarae*	200.5	197.1	6.8
-	*D. biauraria*	237.2	211.2	51.8
-	*D. bicornuta*	213.7	211.7	3.9
-	*D. bifasciata*	205.4	201.5	7.8
14024-0381.19	*D. bipectinata*	204.6	195.3	18.7
-	*D. birchii*	191.2	190.4	1.6
-	*D. bunnanda*	215.2	208.2	14.0
13000-0081.00	*D. busckii*	139.9	140.1	-0.2
14030-0721.00	*D. capricorni*	211.8	205.4	12.9
-	*D. diplacantha*	232.8	230.3	4.9
14027-0461.03	*D. elegans*	192.2	197.1	-9.8

Table 7 Continued.

Stock Number	Species	Female Genome Size (Mbp)	Male Genome Size (Mbp)	Sex Difference (Mbp)
14042-0841.09	*D. emarginata*	214.1	203.5	21.2
14030-0741.00	*D. equinoxialis*	264.5	266.4	-3.8
-	*D. ercepeae*	224.0	221.0	5.9
14021-0224.01	*D. erecta*	158.9	157.0	3.9
-	*D. eugracilis*	228.9	226.9	3.9
14025-0441.05	*D. ficusphila*	190.8	182.1	17.6
14029-0011.00	*D. fuyamai*	266.3	262.5	7.6
-	*D. greeni*	201.5	194.6	13.7
15085-1641.03	*D. hydei*	206.8	201.0	11.7
14028-0671.01	*D. jambulina*	202.7	201.9	1.6
14028-0541.00	*D. kanapiae*	168.6	164.8	7.6
14028-0561.14	*D. kikkawai*	210.2	210.6	-0.8
14028-0571.00	*D. lacteicornis*	248.6	230.7	35.8
-	*D. lucipennis*	291.0	252.3	77.3
-	*D. lutescens*	219.1	213.7	10.8
-	*D. malerkotliana*	204.9	198.5	12.7
-	*D. mauritiana*	157.9	153.6	8.7
14028-0591.00	*D. mayri*	262.2	256.2	12.0
4021-0231.36	*D. melanogaster*	174.5	172.1	4.9
-	*D. mimetica*	212.7	208.3	8.8
-	*D. miranda*	175.6	177.0	-2.9
-	*D. mojavensis*	165.6	168.7	-6.1
14030-0761.00	*D. nebulosa*	211.5	203.1	16.8
14041-0831.00	*D. neocordata*	219.0	219.8	-1.6
14028-0601.00	*D. nikananu*	219.7	220.6	-1.8
-	*D. orena*	280.7	280.7	0.0

Table 7 Continued.

Stock Number	Species	Female Genome Size (Mbp)	Male Genome Size (Mbp)	Sex Difference (Mbp)
14028-0611.01	*D. orosa*	215.3	216.2	-1.8
-	*D. pallidosa*	194.1	186.8	14.7
-	*D. parabipectinata*	210.8	211.2	-1.0
-	*D. paralutea*	230.8	218.6	24.5
14030-0771.00	*D. paulistorum*	244.7	250.1	-10.8
14028-0731.00	*D. pectinifera*	297.4	280.4	34.1
14011-0111.49	*D. persimilis*	197.1	167.7	58.7
-	*D. phaeopleura*	202.9	187.3	31.3
-	*D. prostipennis*	227.4	224.0	6.8
-	*D. pseudoananassae nigrens*	224.0	224.0	0.0
-	*D. pseudoananassae*	228.4	218.6	19.6
-	*D. pseudoobscura*	167.7	157.9	19.6
-	*D. pseudotakahashii*	212.2	202.9	18.6
14028-0641.00	*D. punjabiensis*	200.0	194.9	10.2
-	*D. quadraria*	252.1	248.4	7.4
14028-0661.03	*D. rufa*	256.8	246.0	21.6
-	*D. santomea*	171.5	168.2	6.6
14021-0248.25	*D. sechellia*	179.9	175.6	8.7
14028-0671.02	*D. seguyi*	249.1	245.3	7.6
-	*D. serrata*	213.3	207.2	12.2
14021-0251.195	*D. simulans*	159.6	147.2	24.8
14030-0791.00	*D. sucinea*	209.6	202.2	14.9
-	*D. suzukii*	342.8	333.3	18.9
14022-0311.13	*D. takahashii*	207.3	193.7	27.1
14020-0011.01	*D. tani*	211.5	200.8	21.5
-	*D. teissieri*	166.3	158.9	14.7

Table 7 Continued.

Stock Number	Species	Female Genome Size (Mbp)	Male Genome Size (Mbp)	Sex Difference (Mbp)
-	*D. tolteca*	179.0	168.7	20.5
14028-0651.00	*D. triauraria*	256.7	252.4	8.6
14028-0701.00	*D. tsacasi*	224.8	211.5	26.6
-	*D. varians*	166.7	156.0	21.5
-	*D. virilis*	325.4	338.2	-25.6
14028-0711.00	*D. vulcana*	226.0	221.1	9.8
14021-0261.01	*D. yakuba*	170.7	168.3	4.9
-	*H. pictiventris*	162.8	142.3	41.1
80000-2761.03	*Sc. leonensis*	261.8	260.4	2.8
11010-0021.00	*S. lebanonensis lebanonensis*	210.3	208.8	2.9
-	*S. pattersoni*	213.2	224.5	-22.5
-	*S. stonei*	206.8	208.3	-2.9
50000-2744.02	*Z. sepsoides*	212.8	214.8	-4.0
50001-0001.02	*Z. tuberculatus*	197.6	199.0	-2.8

Phylogeny reconstruction

The overall reconstructed phylogeny for *Sophophora* was found to be well supported by the posterior probability values at each nodes. Most nodes were found to have a value of 1, whereas the lowest node had a value of 0.56 (Figure 6). These relationships are well supported by previously published phylogenies in the literature (Consortium 2007; Da Lage et al. 2007; Gregory and Johnston 2008; van der Linde et al. 2010). The resulting tree is therefore thought to be representative of the true relationships found in *Sophophora*.

Figure 6: Bayesian reconstruction of *Sophophora* phylogeny.
A phylogeny for *Sophophora* was reconstructed using a supermatrix method of phylogeny reconstruction utilizing MrBayes 3.2.3 on the CIPRES supercomputer with 32,835,000 generation (sampling every 1,000) using a Dirichlet prior of (1, 0.5, 1, 1). A consensus tree was constructed using a burn-in period of 10%. The consensus tree was then visualized in FigTree v.1.4.2. The relationships throughout the phylogeny all have strong support, with most nodes having a 1 for posterior probability. The lowest posterior probability is 0.56. These relationships are supported by phylogenies previously published in the literature.

Comparative phylogenetic results

Male whole genome size was found to have significant phylogenetic signal (λ= 1) with high levels of support (Table 8). The male genome size was also found to change gradually along branch lengths (κ= 1.247) with high support for change early in the phylogeny (δ= 0.657) (Table 8). The values for male genome size are visualized on the phylogeny (Figure 7). Here we can see that closely related species have similar sizes, represented by similar colors, while there is a large change early in the phylogeny. Overall, there is a gradual change in size downwards. This supports the values obtained by Pagel's parameters.

Table 8: Comparative phylogenetic values for male *Sophophora* genome size. Male genome size was found to have complete phylogenetic signal. While κ was found to be above one, it was not significantly different than one, indicating gradual change on individual branches. The δ value below one indicates that change likely happened somewhat early in the phylogeny.

Male *Sophophora* Genome Size		
Test	**Value**	**Significance**
λ	1	<2.22e-16(From 0), 1(From 1)
δ	0.658	2.43E-10 (From 0), 0.58574 (From 0.5), <2.22e-16 (From 3)
κ	1.247	2.55E-15(From 0), 0.0629 (From 1), <2.22e-16 (From 3)

The difference due to heteromorphic sex chromosomes was found to have only partial phylogenetic signal ($\lambda = 0.827$) with significant departure from full phylogenetic signal (H_0: $\lambda = 1$, Table 9). The sex difference was found to have rapid early change in branches ($\kappa = 0.380$) with longer paths in the tree contributing more to change ($\delta = 1.688$). These tests of mode and rate of change have departed significantly from gradual change (H_0 κ and $\delta = 1$, Table 9). These patterns can be visualized on a color phylogeny (Figure 8). Most of the phylogeny has a similar color, which likely gives the phylogenetic signal. When there is change, it occurs late in the tree, in individual species, supporting the δ value.

When the phylogenetic values for male genome sizes are compared to those of females, there is not a substantial difference in the results (Table 10). The values found for the difference in size due to the heteromorphic sex chromosomes were found to be different than those found for whole genome size. These values are visualized side-by-side in Table 10.

Table 9: Comparative phylogenetic values for the sex differences in *Sophophora.* Pagel's parameters of evolution for the difference in size due to the Y chromosome suggest partial phylogenetic signal and early change in branch lengths. Change was found to occur late in the phylogeny.

Sex Difference due to Y Chromosome		
Test	**Value**	**Significance**
λ	0.827	2.79e-06(From 0), 0.0145 (From 0.5), 2.22e-16 (From 1)
δ	1.688	4.16e-13(From 0), 0.0904(From 1), 2.09e-04(From 3)
κ	0.380	0.0009 (From 0), 1.55e-12 (From 1), <2.22e-16 (From 3)

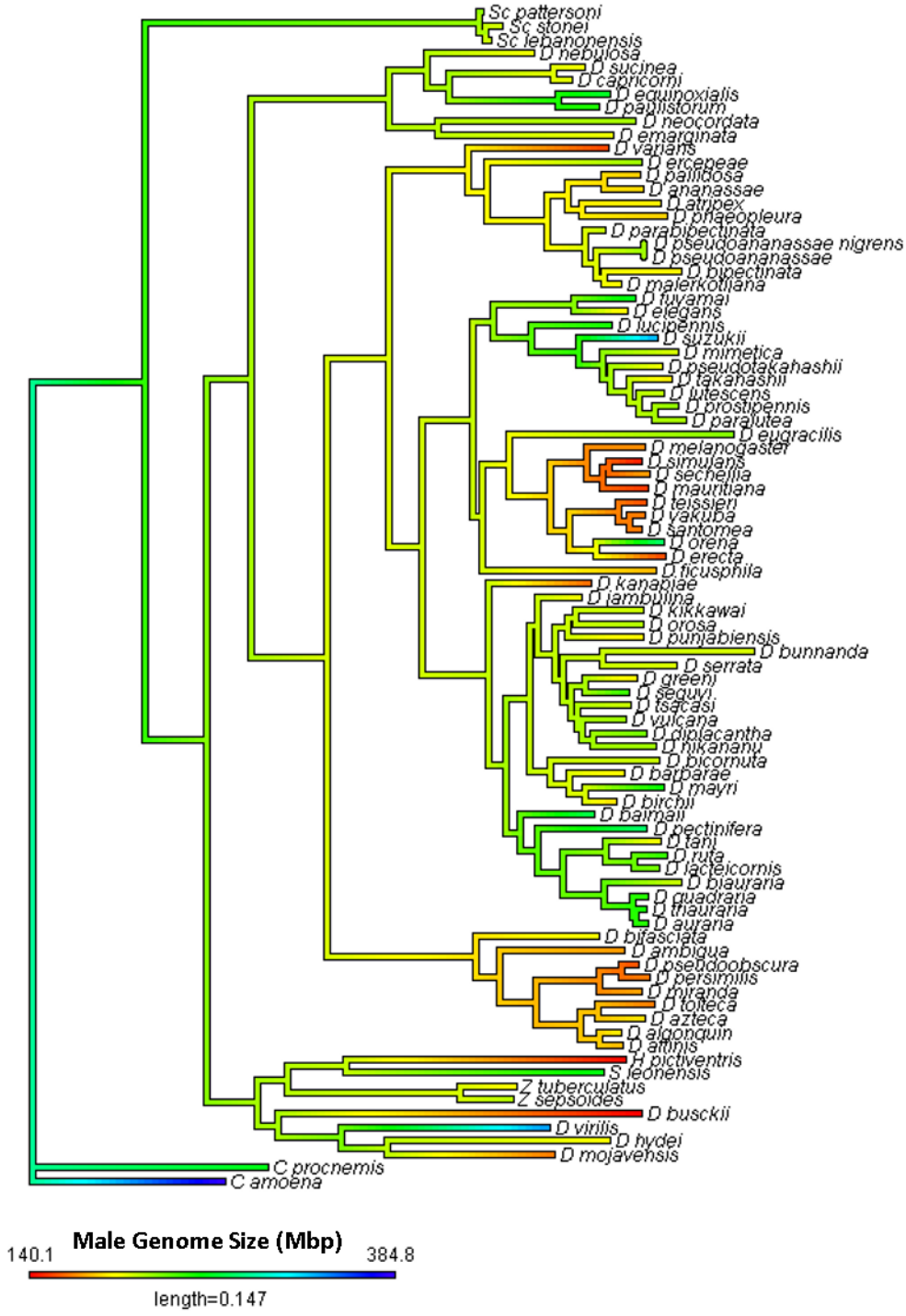

Figure 7: Male *Sophophora* genome size plotted on phylogeny.
The phylogeny of 87 Drosophilidae reconstructed with MrBayes 3.2.3 with a focus on *Sophophora*. Male genome size is visualized in color, with larger genome sizes in blue, smaller in red, and intermediate in green.

Figure 8: Difference due to the sex chromosome plotted on *Sophophora* phylogeny.
The phylogeny of 87 Drosophilidae reconstructed with MrBayes 3.2.3 with a focus on *Sophophora*. The difference in genome size due to the Y-chromosome is visualized in color, with larger size differences in blue, smaller in red, and intermediate in green. Positive values indicate females of the species are larger than males, whereas negative values indicate males have larger genomes than the females. Negative values suggest that there is a neo-Y system present in this species.

Table 10: A comparison of Pagel's parameters between sexes.
When Pagel's parameters are compared for genome sizes between sexes there are no differences in evolutionary patterns. The sex difference due to the Y chromosome had different phylogenetic patterns in comparison to whole genome size.

Test	Female	Male	Sex Difference
λ	0.987	1	0.827
δ	0.589	0.658	1.688
κ	0.971	1.247	0.380

Discussion

Here, we report the genome sizes and differences due to heteromorphic sex chromosomes for 87 species of Drosophilidae (Table 7). While there are instances of males with a larger size than the corresponding females of the species (likely due to neo-Y chromosomes), it is generally expected that females should have larger genome sizes than males on average. However, there was no statistical difference between the average of females and the averages of males in the species (t-test, $p = 0.32$). If species which had larger males were removed from the analysis, there was still not a significant difference between the average of female and male genome size (t-test, $p = 0.2321$).

Male genome size was found to have complete phylogenetic signal ($\lambda = 1$), with change early in the phylogeny ($\delta = 0.658$) and a gradual change along branches ($\kappa = 1.247$). These results suggest a large amount of change early in the phylogeny, and from that point on gradual change, suggesting some sort of imbalance of insertion and

deletion of DNA. The overall trend of this change is downwards, which is supported by work by Petrov (Petrov 2002a). While there are instances of genome size increase, there is a general, gradual trend downwards in size (Figure 7). As expected, these results are not different from those found in the females of *Sophophora* (Table 10) (Hjelmen and Johnston 2017). While gradual change and phylogenetic signal in genome size is supportive of the mutational equilibrium model, the rate by which this change would occur with small insertions and deletions is likely too slow to give us results such as these (Gregory 2003b; Gregory 2004). Therefore, it is more likely that these results support the accordion model of genome size evolution, where large deletions are able to balance out instances of large insertions, such as transposable elements (Kapusta et al. 2017).

The difference in genome size due to the relative sizes of the X and Y chromosome behaved very differently over the evolutionary time span. The difference was found to have incomplete phylogenetic signal ($\lambda = 0.827$), more change happening on long paths in the tree ($\delta = 1.688$), and early rapid change on branches ($\kappa = 0.380$) (Table 9). The differences can be visualized on the colorized phylogeny (Figure 8). Most of the species have approximately the same differences, with a few remarkable exceptions, yet there does not seem to be a clear visual pattern. This lack of pattern supports the λ value of less than 1.0 and an incomplete phylogenetic signal. While these results are dramatically different than those found when looking at the evolution of whole genome size (Table 4), they are not unexpected given the hypothesized modes of Y chromosome evolution and degradation (Charlesworth et al. 2005; Bachtrog 2013) .

Based on the proposed model of Y-chromosome degradation (Reviewed by Charlesworth and Charlesworth 2000; Charlesworth et al. 2005; Bachtrog 2013), there should be a trend towards a smaller Y chromosome. This general trend across species may be what causes the presence of an incomplete phylogenetic signal ($\lambda < 1$). The decrease in size, however, has limits. Over time, the X and Y become almost entirely different, which can then result in the loss of the older X-Y system and result in a new neo-Y/neo-X system.

When a neo-Y system emerges, the neo-Y chromosome may result in males with larger genomes than the female of the species. This increase in size may also occur with the inevitable initial inflation of the Y with transposable elements. An example of this is the *Drosophila miranda* neo-Y chromosome. This neo-Y system is hypothesized to have been formed by a Y-autosome fusion about 1.2 million years ago (Bachtrog et al. 2008; Matsunaga 2009). This chromosome still harbors many functional genes, yet has more than 20 fold greater accumulation of repetitive sequences than the X-chromosome (Bachtrog et al. 2008). The genome sizes reported here for male and female *D. miranda* supports this accumulation, with the male genome larger than that of the female by 2.9 Mbp. This difference is unlike *D. melanogaster* where the Y chromosome, which has become almost fully heterochromatic (Adams et al. 2000; Skaletsky et al. 2003), with the male genome smaller than that of the female by 4.9 Mbp. *D. albomicans* and *D. pseudoobscura* have also been reported to have neo-Y chromosomes, yet be at different stages of Y chromosome evolution (Reviewed in Bachtrog 2013). *D. albomicans* has a younger Y chromosome than that of *D. miranda*, so the sequence of events in

degeneration can be studied in more detail. Early results suggest that transcriptional down-regulation occurs before the degeneration of protein-coding genes. One possible explanation for this would be the formation of heterochromatin (Bachtrog 2013). *D. pseudoobscura* has the oldest Y-chromosome of these three species and has been found to be almost entirely heterochromatic, like the *D. melanogaster* Y chromosome. The process of becoming heterochromatic from an autosome is estimated to have occurred within 17 million years, a relatively short period of time (Bachtrog 2013). One study on *D. busckii*, a species that also has been reported to have neo-sex chromosomes, found that almost 60% of the neo-Y genes have become non-functional in less than the one million years since it was formed. (Zhou and Bachtrog 2015). This suggests that Y degeneration occurs very quickly after the sex chromosomes become established, supporting the low kappa values found here.

Newly formed Y-chromosomes, such as that in *D. miranda*, will rapidly accumulated deleterious mutations and transposable elements, which will then accumulate heterochromatin and result in the rapid decrease in size, as seen in *D. pseudoobscura* (Bachtrog 2013). The rapid decrease in size of the Y chromosome is therefore supported by the low kappa value ($\kappa < 1$), suggesting that change in the Y chromosome happens rapidly in branches followed by stasis or gradualism. The high delta value found suggests that more change occurs on long paths, or that more change occurs later in the tree. Because of the rapid changes in size for the sex difference for those species with neo-Y systems, it is not surprising to find evidence for large amounts of change later in the tree. There truly have been dramatic shifts from the ancestral,

largely heterochromatic Y-chromosome. Since the change is occurring late in the phylogeny, where most of the radiation of species has occurred, it does suggest that the change may be related with speciation. The formation of a neo-Y system could potentially increase instances of reproductive isolation, leading to the formation of new species. This idea is supported in sticklebacks, where it was found that a neo-sex chromosome had a role in a speciation event (Kitano et al. 2009). The neo-X chromosome had loci for male courtship displays which would lead to behavioral, and eventually reproductive, isolation. While this question has not been studied much outside fish, the authors argue that sex-chromosome divergence should be considered as an important mechanism which results in reproductive isolation (Kitano et al. 2009).

It can therefore be concluded that comparing patterns of whole genome size evolution, at least in a phylogenetic sense, using male or female sizes do not show that sex has a large impact on results. However, when the differences between male and female genomes are mapped on the phylogeny, they do give support to the proposed models of neo-Y chromosome formation and subsequent degradation of the neo-Y chromosome. When dissecting components of genome size evolution, such as heterochromatin, repeat, and transposable element content, the differences in sex may result in significant differences. The X-Y sex chromosome system results in significantly different levels of heterochromatin and presumably increased transposable element content. It is also important to note that these patterns seem to hold up throughout the *Sophophora* subgenus, with a few additional outgroups. It will be interesting to see if these patterns hold up when analyzing *Drosophila* as an entire genus

with the inclusion of an equal number of *Drosophila* subgenus species. It will also be of interest to analyze these same questions not only in systems with X-Y sex chromosomes, but also in Z-W and homomorphic sex chromosomes.

CHAPTER IV

COMPARISON OF PHYLOGENETIC PATTERNS OF GENOME SIZE EVOLUTION AMONG 152 *SOPHOPHORA* AND *DROSOPHILA* SPECIES, WITH NEW AND REVISED ESTIMATES OF GENOME SIZE FOR 93 SPECIES OF *DROSOPHILA*

Introduction

Genome size, or the physical amount of DNA contained within the nuclei of an organism, has been found to vary widely across species (Gregory 2001; Palazzo and Gregory 2014). Further, while there is this extensive variation, there is no correlation to complexity in eukaryotes: More DNA does not mean a more complex organism (Mirskey and Ris 1951). This phenomenon is commonly referred to as the C-value paradox (Reviewed in Gregory 2001). The C-value paradox has generally been explained by variation in amount of nongenic and repeat regions rather than the amount of coding sequences. The variation in these regions is typically the result of changes in transposable elements, satellite DNA, tandem repeats, and copy number variation (Gregory and Hebert 1999; Kidwell 2002; Kelley et al. 2014; Sessegolo et al. 2016). Among closely related species of plants and *Drosophila,* much of the variation in genome size has been explained by the differential accumulation of transposable elements (Bennetzen and Kellogg 1997; Ågren and Wright 2011; Śliwińska et al. 2016). For example, *Drosophila melanogaster* has a significantly greater accumulation of transposable elements in comparison to *D. simulans,* and has a larger genome size (Vieira and Biemont 2004). This same pattern of increased transposable element load,

has been shown to be significant, when analyzed in a phylogenetic manner across 26 species of *Drosophila* (Sessegolo et al. 2016).

While the C-value paradox has been explained through the accumulation of noncoding DNA, the evolutionary process by this type of DNA accumulates or shrinks to influence genome size variation is still hotly debated (Gregory and Hebert 1999; Gregory 2001; Petrov 2001; Gregory 2003a; Vinogradov 2004; Bennett and Leitch 2005; Kraaijeveld 2010; Hanrahan and Johnston 2011; Palazzo and Gregory 2014; Kapusta et al. 2017). Many hypotheses have been proposed for the patterns one would expect to see in genome size evolution, ranging from that produced by adaptive forces, imbalance between insertions and deletions and chance population genetic effects. The analysis of many datasets, has provided some support for each of these, yet there does not appear to be one answer, at this point, to the question of genome size variation.

The adaptive hypothesis suggests that an outside force may select a correlate of genome size, such as body size, cell size, etc., in order to modify genome size (Reviewed in Powell 1997). To site one example, one might expect that the further you get from the equator, the larger body sizes (according to Bergman's rule) would result in larger genome sizes (Hessen et al. 2013). This is not however the case with the only insect native to Antarctica, *Belgica antarctica*, which has the smallest insect genome sequenced to date (Kelley et al. 2014). Fortunately, the adaptive hypothesis can be tested. If the adaptive hypothesis were true, we would expect to see a lack of phylogenetic signal and no clear patterns throughout the phylogeny, as there should be bursts of change in genome size in response to local environmental conditions. These

patterns may be less clear if the phylogeny has a strong relationship with geography and other adaptive factors. And yet, while climate and latitude have been found to have phylogenetic signal (Kellermann et al. 2012a; Kellermann et al. 2012b), when these variables are analyzed in comparison to *Drosophila* genome sizes using phylogenetic methods, there is no relationship (Hjelmen and Johnston 2017). It is therefore important to consider alternative hypotheses, and these too can be tested.

The mutational equilibrium hypothesis proposes that genome size changes gradually throughout time due to an imbalance of small insertions and deletions (Petrov and Hartl 1998; Petrov et al. 2000; Petrov 2001; Petrov 2002b). This hypothesis has been criticized for being too theoretical, too slow, and lacking support in large datasets (Gregory 2003b; Gregory 2004). Nonetheless, a recently proposed hypothesis, known as the "accordion" model, may address this concern. Here, large deletions counteract the large insertions due to transposable elements (Kapusta et al. 2017). The larger sizes of DNA added and removed are expected to make this process faster and will likely have more support with the literature. Here we would expect to see phylogenetic signal, as well as evidence for gradual change across a phylogeny and along branch lengths.

A third hypothesis that can be tested is the low effective population size hypothesis. It proposes that small effective population sizes are less effective at selecting out deleterious and unwieldy genomes, resulting in an increase in average genome size (Lynch and Conery 2003). While there was support for this in the initial publication, when phylogenetic information was included into the analysis, the significant pattern disappeared, and the size effect is currently being debated (Whitney and Garland Jr

2010; Whitney et al. 2011). In order to support this hypothesis, we would expect to find phylogenetic signal and evidence for bursts of change, early in the tree and/or branches throughout the phylogeny. The bursts of change are expected due to low species effective population size during speciation. The ineffective selection in these instances will result in a rapid change after a species has split in the phylogeny.

Drosophila as a genus has been widely studied in biology, including phylogenetics and genome size (Gregory and Johnston 2008; van der Linde and Houle 2008; van der Linde et al. 2010; Hjelmen and Johnston 2017). The wealth of information available for this genus allows researchers to develop very ambitious large scale evolutionary studies with ease. Importantly this genus is actually separated into subgenera, *Sophophora* and *Drosophila*, which diverged an estimated 40-65 million years ago (Russo et al. 1995; Tamura et al. 2004; Obbard et al. 2012). The two subgenera can be separated karyotypically by the presence of six chromosomes in *Drosophila* and four chromosomes in *Sophophora*. *Drosophila* have the proposed ancestral karyotype whereas *Sophophora* have a reduced chromosome number due to fusion events that formed large metacentric autosomes (Reviewed in Schulze et al. 2006). A comparison between the subgenera provides both biological replication and a test for the effect of the change in chromosome number. While the *Drosophila* genus is replete with taxonomic issues, as the *Drosophila* subgenus and genus are paraphyletic, with genera such as *Zaprionus*, *Scaptodrosophila*, and *Hirtodrosophila* all resolved to be within the major subgenus *Drosophila* (Reviewed in van der Linde et al. 2010), these issues do not, reduce the value of the subgenera as biological replicates.

While the *Drosophila* genus has been widely studied, much of the emphasis for genome size studies has been placed on species within *Sophophora*, the subgenus which includes the very well-studied *D. melanogaster*. In comparison, the subgenus *Drosophila* has been dramatically underrepresented (Gregory and Johnston 2008; Hjelmen and Johnston 2017). Therefore we update or estimate anew the genome size for females and males of 93 species, with a focus on the *Drosophila* subgenus, including *Zaprionus*. With that, we can make comparisons of genome size variation between the subgenera and compare the phylogenetic patterns that occur for the genome sizes between species in the subgenera and across the genus as a whole. *Sophophora* genome size has been shown to best fit the accordion model hypothesis of evolution, indicating complete phylogenetic signal and mostly gradual change throughout branches and the phylogeny (Hjelmen and Johnston 2017). Even though there has been 40-65 million years of evolution since the divergence, we do not expect, under the accordion model, to find remarkably different patterns between the subgenera.

We examine here the variation of genome size in the subgenus *Drosophila* and across the genus as a whole. If a difference between these subgroups exists, it will indicate the accordion model does not model equally well genome size evolution in the *Drosophila* subgenera. This could possibly be due to the karyotypic difference. Another possibility that can be tested is that sex chromosome evolution is different in the two subgenera. We include in this study genome size values for females and males, which means we can use the differences between males and females as a proxy for Y chromosome evolution. Given the presence of a common XY system in all these species,

we might expect the patterns to be generally the same among subgenera, and be similar to those found in earlier studies (Chapter 3). This expectation however must be tested.

Methods

Genome size estimates

Previous estimates of genome size for 59 *Drosophila* species were found in the published literature (Gregory and Johnston). New genome sizes estimates were produced for 93 additional species of *Drosophila, Chymomyza, Zaprionus, Scaptomyza,* and *Hirtodrosophila,* with a focus in the *Drosophila* subgenus. Individuals for these species were obtained from the UC San Diego Species Stock Center (http://stockcenter.ucsd.edu) (Table 11). Genome sizes were estimated utilizing the flow cytometric method of genome size (Hare and Johnston 2011). Briefly, neural tissue was dissected from samples and placed into 1 mL of Galbraith buffer. All samples were co-prepared with an appropriate standard (YW *D. melanogaster* female = 175 Mbp, Lab strain *D. virilis* female = 328 Mbp). Samples and standards were gently ground with a "loose" A pestle 15 times in order to release nuclei. Samples were then passed through a 41 micron filter before staining with 25µl of 1mg/µl propidium iodide. Samples were allowed to incubate for at least 20 minutes to ensure proper stain saturation had occurred. Samples were then run on a Partec CyFlow SL_3 cytometer with a 532 nm green laser. Fluorescence peaks produced by 2C nuclei of both the sample and the standard were gated, using included Partec software to provide the mean peak positions used to estimate genome size. This process was repeated for at least 5 individuals,

where samples allowed, in order to generate the average genome size estimate and standard error.

Phylogeny reconstruction

Sequence information for 16 genes were downloaded from NCBI GenBank in order to create a molecular phylogeny (4 mitochondrial and 12 protein coding) (*COI, COII, COIII, Cytb, Amy, AmyRel, Ddc, boss, SNF, Marf, Sod, per, Wee, HB, ADH, and fkh*). These sequences were aligned using MAFFT v.7 online with iterative refinement methods (http://mafft.cbrc.jp/). Amino acid translations of these alignments were inspected in Mesquite for irregularities and corrected by hand as needed.

Each sequence alignment was then analyzed in JModelTest 2.1.4 to determine the model of sequence evolution that produced the best likelihood value (Darriba et al. 2012). This likelihood search assumed 11 possible substitution schemes, allowing for invariant sites and gamma distributions. A fixed BIONJ-JC tree was used for all calculations. All genes were found to have the same suggested model for phylogeny reconstruction, a GTR substitution model with gamma distribution and invariant sites.

All sequences were interleaved to produce a 10,382 bp alignment. Missing sequence data was input for taxa that did not have gene sequence data for every gene, as per the supermatrix method (van der Linde et al. 2010). Overall, there was an average of seven genes per taxa, with a minimum of three genes.

Table 11: Genome size for 152 species of *Drosophila.*

Genome sizes estimates were made for females and males 93 species of *Drosophila*. These values were added to 59 previously published estimates. New estimates are those species with values for N and standard deviation. Updated species have symbols next to the species indicating the amount of difference from earlier estimates of female genome size. Those with 0 – 10 Mbp differences are indicated with '=', with 10 – 20 Mbp with 't', and those with greater than 20 Mbp difference with '*'. X-Y differences were calculated by subtracting double the male value (2A + XY) from double the female value (2A + XX). Positive X-Y difference values indicate that the female is larger than the male, whereas negative values indicate that males are larger than females.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
20010-2010.00	*C. amoena*				15081-1371.01	*D. mulleri*			
	Female	395.2	-	-		Female	168.3	1.71	5
	Male	384.8	-	-		Male	158.9	1.31	5
	X-Y Diff	20.8				X-Y Diff	18.8		
20000-2631.01	t*C. procnemis*				15090-1692.11	*D. nannoptera*			
	Female	298.7	1.18	5		Female	245.9	1.66	5
	Male	261.5	2.38	5		Male	250.1	3.15	5
	X-Y Diff	74.4				X-Y Diff	-8.4		
15090-1693.00	*D. acanthoptera*				-	*D. narragansett*			
	Female	155.8	3.00	10		Female	205.9	-	-
	Male	145.6	4.10	7		Male	203.4	-	-
	X-Y Diff	20.4				X-Y Diff	4.9		
14012-0141.00	=*D. affinis*				15112-1781.01	*D. nasuta*			
	Female	208.1	2.14	5		Female	242.4	3.03	5
	Male	184.8	3.55	5		Male	245.7	3.01	5
	X-Y Diff	46.6				X-Y Diff	-3.3		
15112-1751.04	*D. albomicans*				15081-1374.02	*D. navojoa*			
	Female	218.3	-	-		Female	169.6	5.35	7
	Male	220.6	-	-		Male	156.2	1.76	5
	X-Y Diff	-2.3				X-Y Diff	27.6		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
14012-0161.00	=*D. algonquin*				14030-0761.00	**D. nebulosa*			
	Female	205.2	1.17	5		Female	211.5	3.15	5
	Male	194.2	3.28	6		Male	203.1	1.66	5
	X-Y Diff	22.0				X-Y Diff	16.8		
14013-1011.00	=*D. ambigua*				14041-0831.00	=*D. neocordata*			
	Female	187.7	2.38	5		Female	219.0	2.30	5
	Male	176.5	2.38	6		Male	219.8	1.28	5
	X-Y Diff	23.8				X-Y Diff	-1.6		
14024-0371.13	*D. ananassae*				15115-1881.00	*D. neohypocausta*			
	Female	196.6	-	-		Female	165.7	3.17	5
	Male	188.8	-	-		Male	165.6	1.10	5
	X-Y Diff	15.6				X-Y Diff	0.1		
15081-1261.02	*D. anceps*				14028-0601.00	=*D. nikananu*			
	Female	201.8	0.84	5		Female	219.7	2.65	5
	Male	207.5	1.47	5		Male	220.6	1.15	5
	X-Y Diff	-11.4				X-Y Diff	-1.8		
15182-2261.03	*D. arawakana*				15010-1031.04	*D. novamexicana*			
	Female	181.1	2.07	5		Female	251.0	-	-
	Male	186.4	1.24	5		Male	252.7	-	-
	X-Y Diff	-5.3				X-Y Diff	-3.4		
15081-1271.14	*D. arizonae*				-	*D. orena*			
	Female	175.4	-	-		Female	280.7	-	-
	Male	173.9	-	-		Male	280.7	-	-
	X-Y Diff	3.0				X-Y Diff	0.0		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
14024-0361.00	=*D. atripex*				14028-0611.01	=*D. orosa*			
	Female	198.2	2.13	7		Female	215.3	2.21	5
	Male	194.8	3.43	5		Male	216.2	1.62	5
	X-Y Diff	6.8				X-Y Diff	-1.8		
14028-0471.00	=*D. auraria*				15090-1698.01	*D. pachea*			
	Female	254.4	3.22	5		Female	176.5	3.13	5
	Male	248.4	2.60	5		Male	148.3	0.74	5
	X-Y Diff	12.0				X-Y Diff	56.4		
14012-0171.02	=*D. azteca*				15210-2331.01	*D. pallidipennis*			
	Female	202.2	4.57	5		Female	330.9	2.90	5
	Male	192.2	2.43	5		Male	325.0	3.28	5
	X-Y Diff	20.0				X-Y Diff	11.4		
14028-0481.00	=*D. baimaii*				-	*D. pallidosa*			
	Female	279.2	1.75	5		Female	194.1	-	-
	Male	273.6	2.93	6		Male	186.8	-	-
	X-Y Diff	11.2				X-Y Diff	14.7		
-	*D. barbarae*				15130-2001.00	*D. palustris*			
	Female	200.5	-	-		Female	219.1	3.41	5
	Male	197.1	-	-		Male	212.1	3.42	5
	X-Y Diff	6.8				X-Y Diff	14.0		
-	*D. biauraria*				-	*D. parabipectinata*			
	Female	237.2	-	-		Female	210.8	-	-
	Male	211.2	-	-		Male	211.2	-	-
	X-Y Diff	51.8				X-Y Diff	-1.0		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
14028-0511.00	**D. bicornuta*				-	*D. paralutea*			
	Female	245.0	2.25	5		Female	230.8	-	-
	Male	226.6	1.91	5		Male	218.6	-	-
	X-Y Diff	36.8				X-Y Diff	24.5		
14012-0181.02	[=]*D. bifasciata*				15030-1161.03	*D. paramelanica*			
	Female	210.6	2.47	5		Female	192.1	2.94	5
	Male	196.2	4.47	5		Male	191.6	3.91	5
	X-Y Diff	28.8				X-Y Diff	1.0		
14024-0381.19	*D. bipectinata*				14030-0771.00	[t]*D. paulistorum*			
	Female	204.6	-	-		Female	244.7	2.73	5
	Male	195.3	-	-		Male	250.1	3.40	5
	X-Y Diff	18.7				X-Y Diff	-10.8		
-	*D. birchii*				14028-0731.00	*D. pectinifera*			
	Female	191.2	-	-		Female	297.4	-	-
	Male	190.4	-	-		Male	280.4	-	-
	X-Y Diff	1.6				X-Y Diff	34.1		
15010-0961.00	*D. borealis*				14011-0111.49	*D. persimilis*			
	Female	257.8	2.89	5		Female	197.1	-	-
	Male	261.5	1.56	5		Male	167.7	-	-
	X-Y Diff	-7.4				X-Y Diff	58.7		
15085-1682.00	*D. bromeliae*				-	*D. phaeopleura*			
	Female	145.2	3.96	5		Female	202.9	-	-
	Male	137.5	1.16	5		Male	187.3	-	-
	X-Y Diff	7.7				X-Y Diff	31.3		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
-	*D. bunnanda*				15130-2031.00	*D. phalerata*			
	Female	215.2	-	-		Female	240.3	2.20	5
	Male	208.2	-	-		Male	233.3	4.65	6
	X-Y Diff	14.0				X-Y Diff	14.0		
13000-0081.00	*D. busckii*				15100-1711.00	*D. polychaeta*			
	Female	139.9	-	-		Female	202.5	2.35	5
	Male	140.1	-	-		Male	206.0	4.34	6
	X-Y Diff	-0.2				X-Y Diff	-7.0		
15081-1291.01	*D. buzzatii*				15181-2231.00	*D. polymorpha*			
	Female	179.9	2.80	5		Female	179.2	0.83	5
	Male	169.0	1.30	5		Male	171.2	2.77	6
	X-Y Diff	21.8				X-Y Diff	8.0		
14030-0721.00	*D. capricorni*				-	*D. prostipennis*			
	Female	211.8	-	-		Female	227.4	-	-
	Male	205.4	-	-		Male	224.0	-	-
	X-Y Diff	12.9				X-Y Diff	9.0		
15182-2261.03	*D. cardini*				-	*D. pseudoananassae*			
	Female	215.7	2.51	5		Female	228.4	-	-
	Male	212.5	0.89	5		Male	218.6	-	-
	X-Y Diff	3.2				X-Y Diff	19.6		
14028-0586.00	**D. diplacantha*				-	*D. pseudoananassae nigrens*			
	Female	273.2	1.86	5		Female	224.0	-	-
	Male	270.4	1.05	5		Male	224.0	-	-
	X-Y Diff	5.6				X-Y Diff	0.0		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15182-2291.00	*D. dunni*				14011-0121.32	[t]*D. pseudoobscura*			
	Female	286.1	3.78	5		Female	180.8	2.74	5
	Male	287.0	2.32	5		Male	161.3	1.94	5
	X-Y Diff	-0.9				X-Y Diff	39.0		
14027-0461.03	*D. elegans*				-	*D. pseudotakahashii*			
	Female	192.2	-	-		Female	212.2	-	-
	Male	197.1	-	-		Male	202.9	-	-
	X-Y Diff	-9.8				X-Y Diff	18.6		
14042-0841.09	*D. emarginata*				15040-1191.00	*D. pseudotalamancana*			
	Female	214.1	-	-		Female	192.2	3.77	8
	Male	203.5	-	-		Male	194.7	2.12	7
	X-Y Diff	21.2				X-Y Diff	-3.3		
15085-1631.00	*D. eohydei*				14028-0641.00	[=]*D. punjabiensis*			
	Female	231.0	2.07	5		Female	200.0	0.99	5
	Male	218.7	1.25	5		Male	194.9	0.79	5
	X-Y Diff	24.6				X-Y Diff	10.2		
14030-0741.00	[t]*D. equinoxialis*				-	*D. quadraria*			
	Female	264.5	1.49	5		Female	252.1	-	-
	Male	266.4	0.48	5		Male	248.4	-	-
	X-Y Diff	-3.8				X-Y Diff	7.4		
-	*D. ercepeae*				15084-1611.08	*D. repleta*			
	Female	224.0	-	-		Female	189.6	0.73	2
	Male	221.0	-	-		Male	169.0	2.33	5
	X-Y Diff	5.9				X-Y Diff	44.0		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
14021-0224.01	*D. erecta*				15020-1111.01	*D. robusta*			
	Female	158.9	-	-		Female	241.6	1.46	5
	Male	157.0	-	-		Male	249.6	2.62	5
	X-Y Diff	3.9				X-Y Diff	-16.0		
-	*D. eugracilis*				15115-1901.03	*D. rubida*			
	Female	228.9	-	-		Female	218.7	2.95	5
	Male	226.9	-	-		Male	213.4	1.69	5
	X-Y Diff	3.9				X-Y Diff	5.3		
15030-1131.01	*D. euronotus*				14028-0661.03	=*D. rufa*			
	Female	176.8	1.68	5		Female	256.8	1.37	5
	Male	174.3	1.74	5		Male	246.0	0.49	5
	X-Y Diff	2.5				X-Y Diff	21.6		
15010-0971.00	*D. ezoana*				-	*D. santomea*			
	Female	192.9	2.91	5		Female	171.5	-	-
	Male	197.9	1.59	5		Male	168.2	-	-
	X-Y Diff	-10.0				X-Y Diff	6.6		
14025-0441.05	*D. ficusphila*				-	*D. sechellia*			
	Female	190.8	-	-		Female	179.9	-	-
	Male	182.1	-	-		Male	175.6	-	-
	X-Y Diff	17.6				X-Y Diff	8.7		
15010-0981.00	*D. flavomontana*				14028-0671.02	**D. seguyi*			
	Female	259.5	1.30	5		Female	249.1	1.24	5
	Male	259.3	1.32	5		Male	245.3	2.18	5
	X-Y Diff	0.4				X-Y Diff	7.6		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15120-1911.01	*D. funebris*				14028-0681.00	[t]*D. serrata*			
	Female	221.1	3.78	5		Female	226.7	2.72	5
	Male	219.8	3.07	5		Male	215.8	3.02	4
	X-Y Diff	1.3				X-Y Diff	21.8		
14029-0011.00	[=]*D. fuyamai*				14021-0251.195	*D. simulans*			
	Female	266.3	3.34	5		Female	159.6	-	-
	Male	262.5	1.55	5		Male	147.2	-	-
	X-Y Diff	7.6				X-Y Diff	24.8		
14028-0712.00	**D. greeni*				15020-1121.01	*D. sordidula*			
	Female	244.1	3.79	5		Female	227.6	1.59	5
	Male	236.7	0.83	5		Male	229.3	3.61	5
	X-Y Diff	14.8				X-Y Diff	-3.4		
15172-2151.00	*D. guarani*				15130-2071.01	*D. subpalustris*			
	Female	291.3	2.32	5		Female	214.7	1.18	5
	Male	286.1	2.42	5		Male	212.2	1.69	5
	X-Y Diff	5.2				X-Y Diff	5.0		
15130-1971.03	*D. guttifera*				14030-0791.01	[=]*D. sucinea*			
	Female	186.3	1.53	5		Female	215.1	2.06	6
	Male	173.2	3.24	5		Male	211.2	2.19	5
	X-Y Diff	13.1				X-Y Diff	14.9		
15085-1641.03	*D. hydei*				-	*D. suzukii*			
	Female	206.8	-	-		Female	342.8	-	-
	Male	201.0	-	-		Male	333.3	-	-
	X-Y Diff	11.7				X-Y Diff	18.9		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15115-1871.04	*D. hypocausta*				-	*D. takahashii*			
	Female	190.6	3.02	5		Female	207.3	-	-
	Male	185.9	2.27	5		Male	193.7	-	-
	X-Y Diff	4.7				X-Y Diff	27.1		
15111-1731.00	*D. immigrans*				14020-0011.01	[t]*D. tani*			
	Female	265.2	4.22	6		Female	211.5	3.77	6
	Male	262.9	1.29	5		Male	200.7	3.44	5
	X-Y Diff	2.3				X-Y Diff	21.5		
14028-0671.01	[t]*D. jambulina*				14021-0257.01	[=]*D. teissieri*			
	Female	219.0	1.67	5		Female	165.0	3.41	5
	Male	210.4	2.35	5		Male	162.9	3.19	5
	X-Y Diff	17.2				X-Y Diff	4.2		
14028-0541.00	[=]*D. kanapiae*				-	*D. tolteca*			
	Female	168.6	2.23	5		Female	179.0	-	-
	Male	164.8	0.94	5		Male	168.7	-	-
	X-Y Diff	7.6				X-Y Diff	20.5		
15010-1061.00	*D. kanekoi*				14028-0651.00	[t]*D. triauraria*			
	Female	256.4	1.65	5		Female	256.7	2.05	6
	Male	254.4	2.18	6		Male	252.4	0.83	5
	X-Y Diff	4.0				X-Y Diff	8.6		
15112-1761.03	*D. kepulauana*				15220-2401.02	*D. tripunctata*			
	Female	224.6	2.67	5		Female	184.4	1.88	5
	Male	237.3	2.55	5		Male	184.5	2.93	5
	X-Y Diff	-12.7				X-Y Diff	-0.2		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
14028-0561.14	*D. kikkawai*				14028-0701.00	=*D. tsacasi*			
	Female	210.2	-	-		Female	224.8	1.97	5
	Male	210.6	-	-		Male	211.5	2.71	5
	X-Y Diff	-0.8				X-Y Diff	26.6		
15112-1771.01	*D. kohkoa*				-	*D. varians*			
	Female	271.9	3.36	5		Female	166.7	-	-
	Male	255.6	4.27	6		Male	156.0	-	-
	X-Y Diff	16.3				X-Y Diff	21.5		
15020-1101.03	*D. lacertosa*				-	*D. virilis*			
	Female	298.0	2.73	5		Female	325.4	-	-
	Male	303.1	4.94	6		Male	338.2	-	-
	X-Y Diff	-10.2				X-Y Diff	-25.6		
15010-0991.12	*D. lacicola*				14028-0711.00	=*D. vulcana*			
	Female	196.0	3.10	5		Female	226.0	3.09	5
	Male	192.4	0.89	5		Male	221.1	1.57	9
	X-Y Diff	7.2				X-Y Diff	9.8		
14028-0571.00	=*D. lacteicornis*				14021-0261.01	*D. yakuba*			
	Female	248.6	5.45	5		Female	170.7	-	-
	Male	230.7	6.24	5		Male	168.3	-	-
	X-Y Diff	35.8				X-Y Diff	4.9		
15084-1591.02	*D. limensis*				92000-0075.00	*H. duncani*			
	Female	203.1	2.32	5		Female	251.6	0.88	5
	Male	182.9	2.99	5		Male	244.3	1.46	5
	X-Y Diff	20.4				X-Y Diff	14.6		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15010-1001.03	*D. littoralis*				-	*H. pictiventris*			
	Female	248.6	-	-		Female	162.8	-	-
	Male	251.5	-	-		Male	142.3	-	-
	X-Y Diff	-5.8				X-Y Diff	41.1		
-	*D. lucipennis*				33000-2661.01	*S. anomala*			
	Female	291.0	-	-		Female	200.2	3.74	5
	Male	252.3	-	-		Male	193.1	2.95	5
	X-Y Diff	77.3				X-Y Diff	14.2		
15010-1011.07	*D. lummei*				80000-2761.03	*Sa. leonensis*			
	Female	312.7	3.78	5		Female	261.8	-	-
	Male	308.5	4.12	5		Male	260.4	-	-
	X-Y Diff	8.4				X-Y Diff	2.8		
-	*D. lutescens*				11030-0061.01	*Sc. latifasciaeformis*			
	Female	219.1	-	-		Female	200.2	-	-
	Male	213.7	-	-		Male	196.9	-	-
	X-Y Diff	10.8				X-Y Diff	6.6		
-	*D. malerkotliana*				11010-0021.00	*Sc. lebanonensis*			
	Female	204.9	-	-		Female	210.3	-	-
	Male	198.5	-	-		Male	208.8	-	-
	X-Y Diff	12.7				X-Y Diff	2.9		
-	*D. mauritiana*				-	*Sc. pattersoni*			
	Female	157.9	-	-		Female	213.2	-	-
	Male	153.6	-	-		male	224.5	-	-
	X-Y Diff	8.7				X-Y Diff	-22.5		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15081-1397.00	*D. mayaguana*				-	*Sc. stonei*			
	Female	169.9	3.47	5		Female	206.8	-	-
	Male	161.7	1.97	5		Male	208.3	-	-
	X-Y Diff	16.6				X-Y Diff	-2.9		
14028-0591.00	=*D. mayri*				52000-2752.00	*Z. bogoriensis*			
	Female	262.2	3.35	7		Female	224.3	2.50	4
	Male	256.2	2.36	11		Male	226.9	1.68	6
	X-Y Diff	12.0				X-Y Diff	-5.2		
15220-2371.00	*D. mediopictoides*				50000-2743.00	*Z. ghesquierei*			
	Female	188.0	1.50	5		Female	174.5	1.26	5
	Male	184.3	1.35	5		Male	174.9	2.05	5
	X-Y Diff	7.4				X-Y Diff	-0.8		
15030-1141.03	*D. melanica*				50001-1031.09	*Z. indianus*			
	Female	175.9	3.38	5		Female	234.8	2.31	5
	Male	175.3	1.79	5		Male	229.9	1.13	5
	X-Y Diff	0.6				X-Y Diff	9.8		
4021-0231.36	*D. melanogaster*				50000-2746.00	*Z. inermis*			
	Female	174.5	-	-		Female	203.1	2.48	5
	Male	172.1	-	-		Male	202.0	1.56	4
	X-Y Diff	4.9				X-Y Diff	2.2		
15082-1521.00	*D. mercatorum*				50000-2748.00	*Z. kolodkinae*			
	Female	165.5	-	-		Female	199.3	3.47	5
	Male	159.1	-	-		Male	195.5	4.39	7
	X-Y Diff	12.8				X-Y Diff	7.6		

Table 11 Continued.

Stock Number	Species	Average GS	St Dev	N	Stock Number	Species	Average GS	St Dev	N
15030-1151.01	*D. micromelanica*				50002-2750.00	*Z. lachaisei*			
	Female	206.4	2.81	5		Female	194.3	2.70	5
	Male	207.1	3.27	5		Male	187.7	2.39	5
	X-Y Diff	-0.7				X-Y Diff	13.2		
-	*D. mimetica*				50000-2744.02	*Z. sepsoides*			
	Female	212.7	-	-		Female	212.8	-	-
	Male	208.3	-	-		Male	214.8	-	-
	X-Y Diff	8.8				X-Y Diff	-4.0		
14011-0101.13	**D. miranda*				50001-1020.00	*Z. taronus*			
	Female	211.9	1.87	5		Female	203.2	2.49	5
	Male	199.6	3.56	5		Male	199.9	2.81	5
	X-Y Diff	24.6				X-Y Diff	6.6		
-	*D. mojavensis*				50000-2751.00	*Z. tsacasi*			
	Female	165.6	-	-		Female	221.8	3.73	5
	Male	168.7	-	-		Male	221.1	3.63	7
	X-Y Diff	-6.1				X-Y Diff	1.4		
15010-1021.09	*D. montana*				50001-0001.02	*Z. tuberculatus*			
	Female	240.6	3.99	5		Female	197.6	-	-
	Male	242.8	2.24	5		Male	199.0	-	-
	X-Y Diff	-4.4				X-Y Diff	-2.8		

A phylogeny for the 152 species was reconstructed utilizing MrBayes 3.2.3 on the CIPRES supercomputer (http://www.phylo.org/) with four chains and four runs and a GTR gamma + I evolutionary model for 44,119,000 generations (sampling every 1,000 generations) using a Dirichlet prior of (1, 0.5, 1, 1,) (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Outputs for parameters were visualized in Tracer v 1.6 to assure that runs had reached convergence and to determine the appropriate burn-in time. The consensus tree was visualized in FigTree v.1.4.2. In order to analyze the difference between *Sophophora* and *Drosophila*, the phylogeny was trimmed to two smaller trees using the drop.tip function in the package ape from R 3.3.0 (Paradis et al. 2004; Team 2016). This produced phylogenies with each appropriate clade and outgroups. Genome sizes for females and males, as well as the sex differences, were mapped onto the phylogenies using the ContMap function from the phytools package in R 3.3.0 (Revell 2012).

Statistical tests

In order to test for significant differences between the subgenera, species were placed within the subgenus *Sophophora* or *Drosophila* based on the large split (Figure 9A, B) of the phylogeny into 2 major clades. This means, for example, that *Zaprionus* species were included in the *Drosophila* subgenus data. *Sophophora* data included 76 species and *Drosophila* included 71 species. Since species outside of the subgenera were not included here, genera such as *Chymomyza* and *Scaptodrosophila* were excluded. This variation was also visualized in histogram format. T-tests and Kolmogorov-Smirnov tests were used to test for significant differences between the

sexes as well as the differences between the subgenera. These were run for both female and male genome size, as well as sex difference. All t-tests and histograms were run in R 3.3.0.

Comparative phylogenetic analyses

Pagel's parameters of evolution were estimated for each variable of interest, female genome size, male genome size, and the difference in genome size between the male and female of each species (Pagel and Harvey 1989). Each of these analyses was with the entire phylogeny, as well as with the reduced *Sophophora* and *Drosophila* phylogenies. These analyses allow for the comparison of the patterns, mode, and rate of evolution across the entire genus, and between the subgenera. The parameter λ tests for phylogenetic signal of the trait of interest across the phylogeny, assuming Brownian motion ($\lambda = 1$, full phylogenetic signal, $\lambda < 1$, incomplete signal). Complete phylogenetic signal indicates that the variation in the trait can be explained by the evolutionary relationships between the species. The κ parameter tests how traits evolve along individual branches ($\kappa = 1$, gradual change, $\kappa < 1$, rapid early change, $\kappa > 1$, increasing rate of change). Finally, δ tests how the trait evolves along the long paths, or where in the entire phylogeny the change occurs ($\delta = 1$, gradual change, $\delta < 1$, early change in phylogeny, $\delta > 1$, late change in tree). Each of these analyses were completed utilizing the pgls function in the caper package of R 3.3.0 (Orme 2013). These values were then used in conjunction with the colorized trait-map phylogenies for interpretation of evolutionary patterns.

Results

Genome size estimates

Genome size information for females, males, and the sex difference, is given in Table 11. Overall, *Drosophila* (*Drosophila* and *Sophophora* subgenera, *Chymomyza*, *Hirtodrosophila, Samoaia, Scaptomyza, Scaptodrosophila,* and *Zaprionus*) were found to have a female genome size average of 217.7 Mbp and a male genome size average of 212.3 Mbp (n = 152). These overall genome sizes ranged by more than 250 Mbp, from 139.9 Mbp to 395.2 Mbp in females and 137.5 Mbp to 384.8 Mbp in males. The difference between the genome size of the male and female of each species, which is assumed to be due to the difference in the size of sex chromosomes, averaged 10.6 Mbp, indicating that females are larger on average than males. Overall, the differences in sexes measured ranged by 100 Mbp, from 74.4 Mbp in *C. procnemis* to -25.6 in *D. virilis.*

Phylogeny reconstruction

The overall phylogeny for *Drosophila* as shown in Figure 9 is well supported, with the majority of nodes having posterior probabilities of 1.0 with the lowest being 0.56. The relationships found in this phylogeny are supported by those found in other large phylogenetic studies (Consortium 2007; Da Lage et al. 2007; Gregory and Johnston 2008; van der Linde et al. 2010). This congruence suggests that phylogenetic relationships and branch lengths should be reliable in this reconstructed phylogeny.

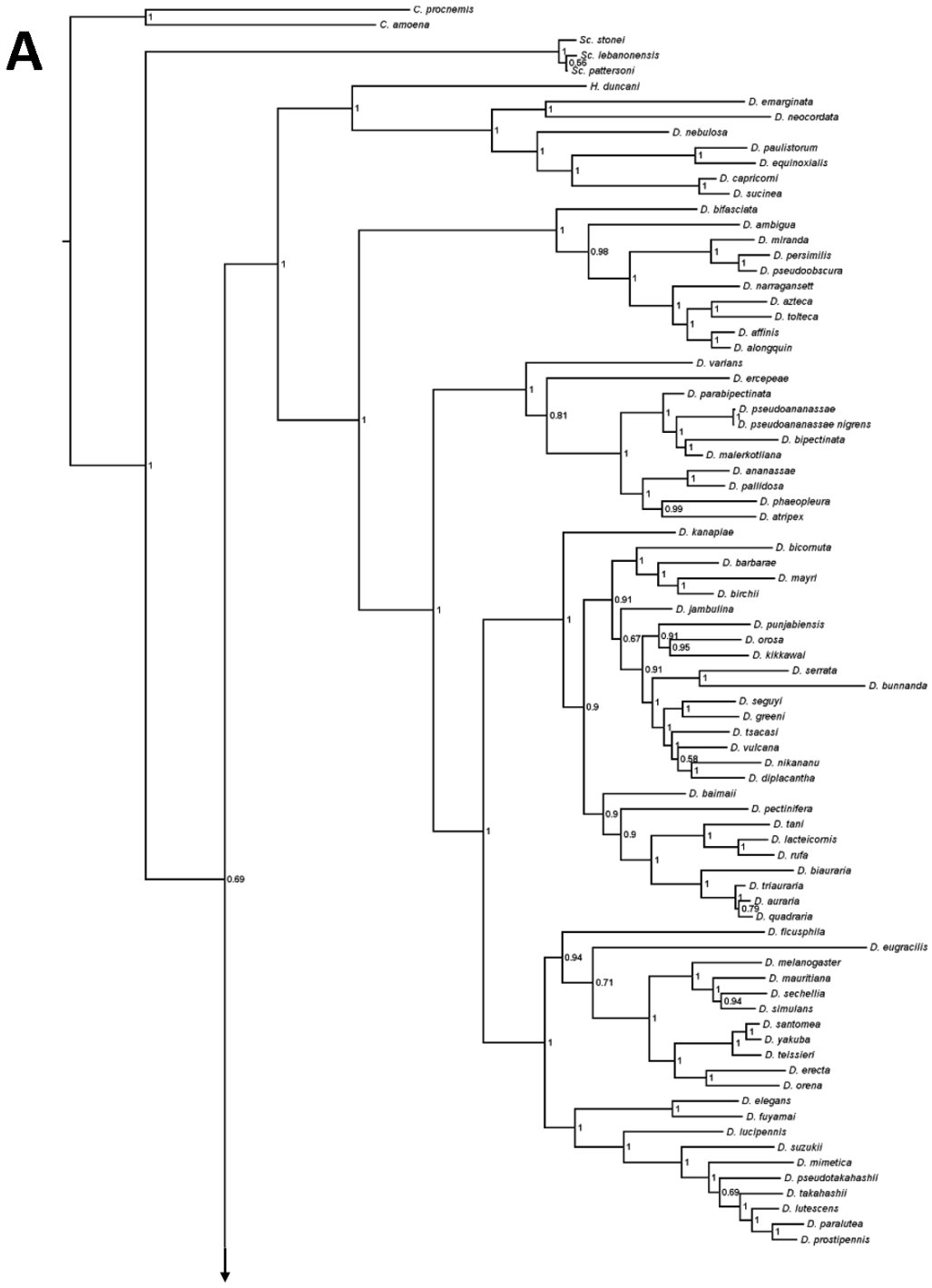

Figure 9: Phylogeny of *Drosophila* reconstructed with MrBayes 3.2.3 using supermatrix methods. The overall phylogeny for *Drosophila* as shown in Figure 1 is well supported, with the majority of nodes having posterior probabilities of 1.0 with the lowest being 0.56. The relationships found in this phylogeny are supported by those found in other large phylogenetic studies. *Sophophora* is in Figure 9A, while *Drosophila* is in Figure 9B.

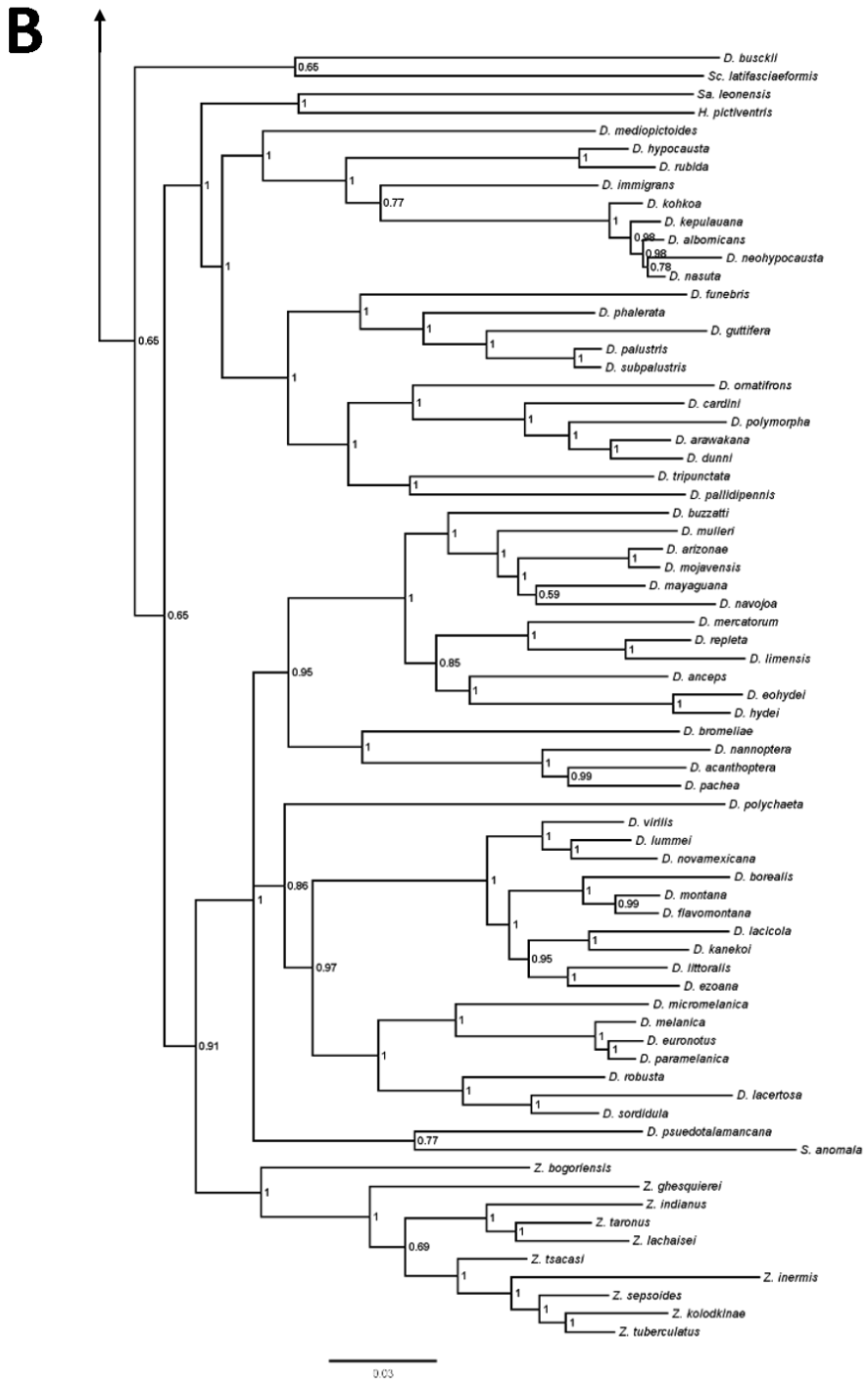

Figure 9 Continued.

Statistical tests

Summary statistics were calculated for the entire data set and for the subspecies *Sophophora* and *Drosophila*. Species were determined to fit within the subgenus *Sophophora* or *Drosophila* based on the large split (Figure 9A, B) of the phylogeny into 2 major clades. This means, for example, *Zaprionus* species are included in the *Drosophila* subgenus. *Sophophora* data included 76 species and *Drosophila* included 71 species. Since species outside of the subgenera were not included here (ex: *Chymomyza*, *Scaptodrosophila*), the total number of species here (n = 147) differs from the comparative phylogenetic analyses. *Sophophora* females and males were found to have average genomes sizes of 218.4 Mbp and 210.7 Mbp, respectively. *Drosophila* females and males had an average genome size of 213.6 Mbp and 210.8 Mbp, respectively. *Sophophora* had an average sex difference of 15.4 Mbp while *Drosophila* had an average sex difference of 5.2 Mbp. The positive values for the difference indicate that female genomes, on average, are larger than male's.

There was no statistical difference between averages genome size of: A) *Sophophora* females and males (Figure 10A), B) *Drosophila* females and males (Figure 10B), C) *Sophophora* and *Drosophila* females (Figure 10C), or D) Sophophora and *Drosophila* males (Figure 10D) (Table 12). On the other hand, while there were no statistical differences between average genome sizes of the subgenera, there is a significant difference between the average sex differences of the subgenera, indicating that *Sophophora* species have proportionally smaller Y and a greater X-Y difference than *Drosophila* (Figure 11, Table 12).

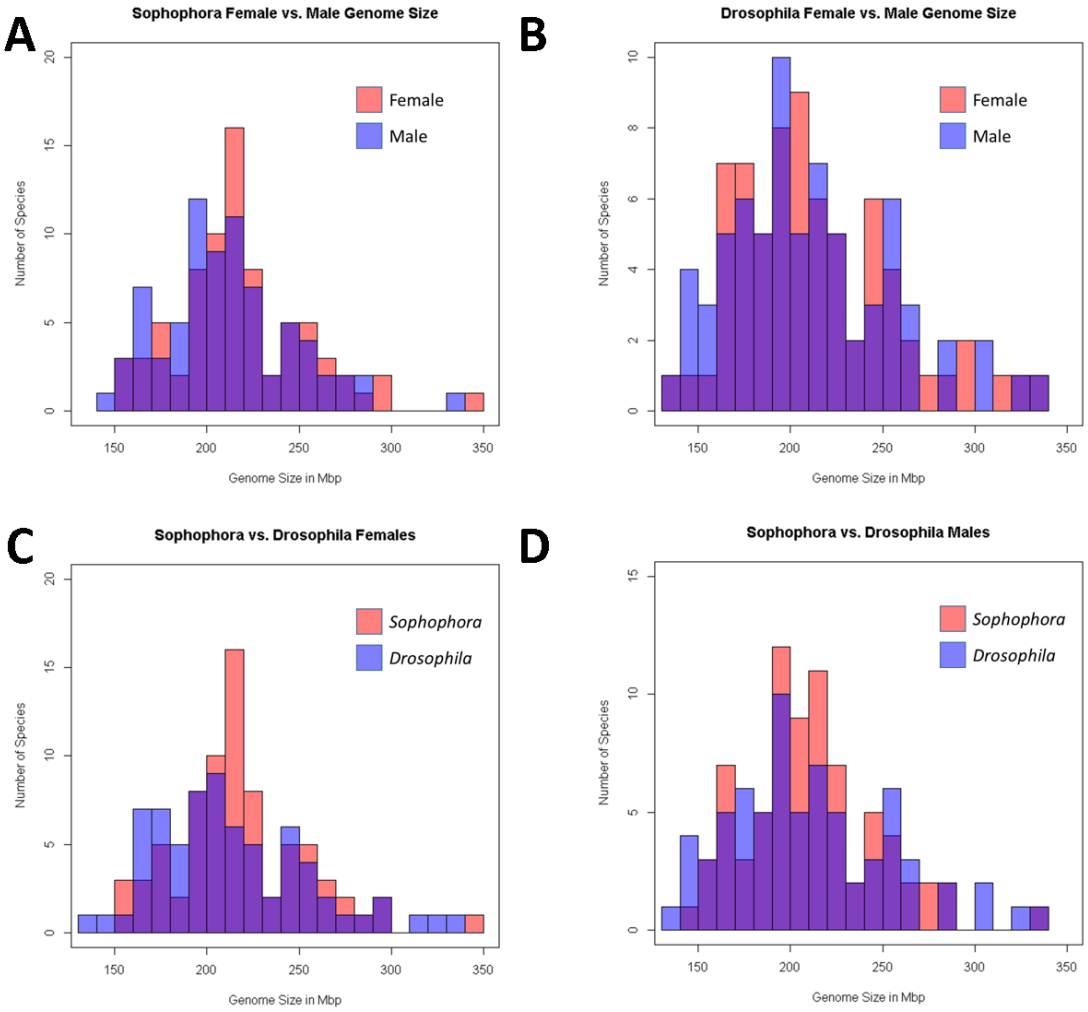

Figure 10: Comparisons of genome size between females and males, *Sophophora* and *Drosophila.*
Distributions of female and male genome size of *Sophophora* (A) and *Drosophila* (B) were made for comparison of sexes. Distributions of females (C) and males (D) between the subgenera were also made. There is significant overlap in all comparisons and none of these comparisons were found to be significantly different with either t-tests or Kolmogorov-Smirnov tests ($p > 0.05$, Table 12).

Table 12: T-test and Kolmogorov-Smirnov results for *Sophophora* and *Drosophila* comparisons.
Comparisons were made between sexes of each subgenus, as well as between subgenera using both t-tests and Kolmogorov-Smirnov tests in R 3.3.0. These same comparisons were utilized to compare the sex differences between the subgenera. The only comparison that was found to be significantly different was the values for sex difference between the subgenera. This indicates that while genome size is not significantly different between the subgenera, the difference due to the X-Y system significantly differs between *Sophophora* and *Drosophila*.

Comparisons	t-test		Kolmogorov Smirnov test	
	t-value	p-value	D-value	p-value
Soph. female and male	1.353	0.178	0.171	0.216
Dros. female and male	0.395	0.694	0.113	0.758
Soph. vs. *Dros*. female	-0.740	0.461	0.205	0.091
Soph. vs. *Dros*. male	0.013	0.990	0.118	0.686
Soph. vs. *Dros*. sex diff	-4.936	2.121E-05	0.390	2.752E-05

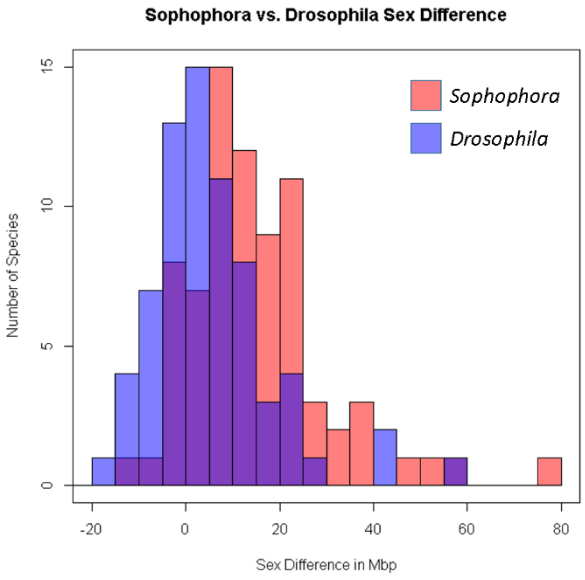

Figure 11: Comparison of sex difference values between *Sophophora* and *Drosophila*.
When the difference between sexes due to the X-Y system were compared between *Sophophora* and *Drosophila*, *Sophophora* (15.4 Mbp) was found to be significantly larger than *Drosophila* (5.2 Mbp) by both a t-test and Kolmogorov-Smirnov test ($p < 0.001$, Table 2). This indicates that in *Sophophora*, the difference between females and males is larger than that in the *Drosophila* subgenus. The positive value for the average of either subgenus indicates that females are, on average, larger than males.

Comparative phylogenetic analyses

Whole genome size for females and males across the entire phylogeny were found to exhibit complete phylogenetic signal (female $\lambda = 1$, male $\lambda = 0.999$), suggesting genome size of a species, regardless of sex, is highly related to evolutionary relationships and time (Table 13). Change in genome size was found to be gradual across individual branches (female $\kappa = 1.216$, male $\kappa = 1.160$), with some evidence for some slightly greater early change in the tree (female $\delta = 0.913$, male $\delta = 0.922$) (Table 13). While the δ values are not significantly different than 1, which would suggest gradual change, they are also not significantly different from 0.5, suggesting that the change is slightly greater early and stabilizes. In order to provide a visual interpretation of these results, female genome size was plotted for all tested species (Figure 12).

Modes of evolution in Sophophora

Estimates of Pagel's parameters based on the subgenus *Sophophora* phylogeny indicates, there is evidence for a complete phylogenetic signal (female $\lambda = 1$, male $\lambda = 1$) and gradual early change in the phylogeny (female $\delta = 0.720$, male $\delta = 0.724$). Similar to the values found when analyzing all species, the δ value is not significantly different than either 0.5 or 1 (Table 13). Males were found to have gradual change on branches ($\kappa = 1.264$), while female had mostly gradual change, with some evidence for an increase in change later in branches ($\kappa = 1.485$). The male κ value was not significantly different than 1, yet the female value was significantly larger than 1 ($p = 0.012$) (Table 13).

Table 13: Pagel's parameters of evolution for female and male genome size and sex difference.

Pagel's parameters of evolution were estimated for the females, males, and the sex difference due to the X-Y system first, for the entire *Drosophila* genus, and next for the *Sophophora* subgenus, and the *Drosophila* subgenus. Genome size for males and females in all tests were found to have complete phylogenetic signal, gradualistic change on individual branches, and some evidence for early change in the tree ($\lambda \sim 1$, $\kappa \sim 1$, $0.5 < \delta < 1$). Sex differences for the entire genus and the *Sophophora* subgenus found incomplete phylogenetic signal, rapid early change in individual branches, and late change in the entire tree ($0.5 < \lambda < 1$, $0 < \kappa$ 0.5, $1 < \delta$). The *Drosophila* subgenus found similar patterns for phylogenetic signal and change on individual branches, yet found evidence for temporally early change in the tree ($0.5 < \delta < 1$).

All Species							
Female		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	1	<2.22E-16	-	1	-	(0.976, 1)
	δ	0.913	1.59E-11	0.142	0.754	<2.22E-16	(0.365, 1.435)
	κ	1.216	2.33E-14	-	0.168	<2.22E-16	(0.907, 1.507)
Male		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.999	<2.22E-16	-	0.866	-	(0.951, 1)
	δ	0.922	1.62E-11	0.138	0.778	<2.22E-16	(0.367, 1.444)
	κ	1.160	7.55E-15	-	0.263	<2.22E-16	(0.877, 1.423)
Sex Difference		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.790	4.91E-12	0.008	<2.22e-16	-	(0.586, 0.914)
	δ	1.472	1.66E-13	-	0.147	9.65E-09	(0.826, 2.048)
	κ	0.336	0.00099	0.078	2.11E-15	<2.22E-16	(0.141, 0.518)
Sophophora							
Female		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	1	<2.22e-16	-	1	-	(0.989, 1)
	δ	0.720	1.85E-10	0.474	0.372	7.42E-14	(0.176, 1.331)
	κ	1.485	6.02E-12	-	0.012	<2.22E-16	(1.116, 1.783)
Male		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	1	2.22E-16	-	1	-	(0.976, 1)
	δ	0.724	1.79E-10	4.63E-01	0.372	2.22E-15	(0.179, 1.323)
	κ	1.264	3.856E-11	-	0.130	<2.22E-16	(0.917, 1.566)
Sex Difference		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.875	2.13E-07	0.00466	5.46E-13	-	(0.652, 0.961)
	δ	1.255	2.79E-11	-	0.559	2.14E-06	(0.417, 2.051)
	κ	0.380	0.00226	0.293	4E-09	<2.22E-16	(0.142, 0.599)
Drosophila							
Female		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.945	6.22E-11	-	0.323	-	(0.760, 1)
	δ	0.851	5.79E-11	0.255	0.635	5.16E-12	(0.270, 1.460)
	κ	1.015	5.93E-07	-	0.940	<2.22E-16	(0.617, 1.414)
Male		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.958	6.45E-10	-	0.193	-	(0.769, 1)
	δ	0.874	5.55E-11	0.235	0.692	2.36E-12	(0.278, 1.487)
	κ	1.002	3.18E-07	-	0.990	<2.22E-16	(0.624, 1.367)
Sex Difference		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.838	1.08E-08	0.005	6.62E-10	-	(0.629, 0.943)
	δ	0.826	1.40E-10	0.346	0.626	1.55E-10	(0.206, 1.523)
	κ	0.390	0.0051	0.400	1.29E-06	<2.22E-16	(0.122, 642)

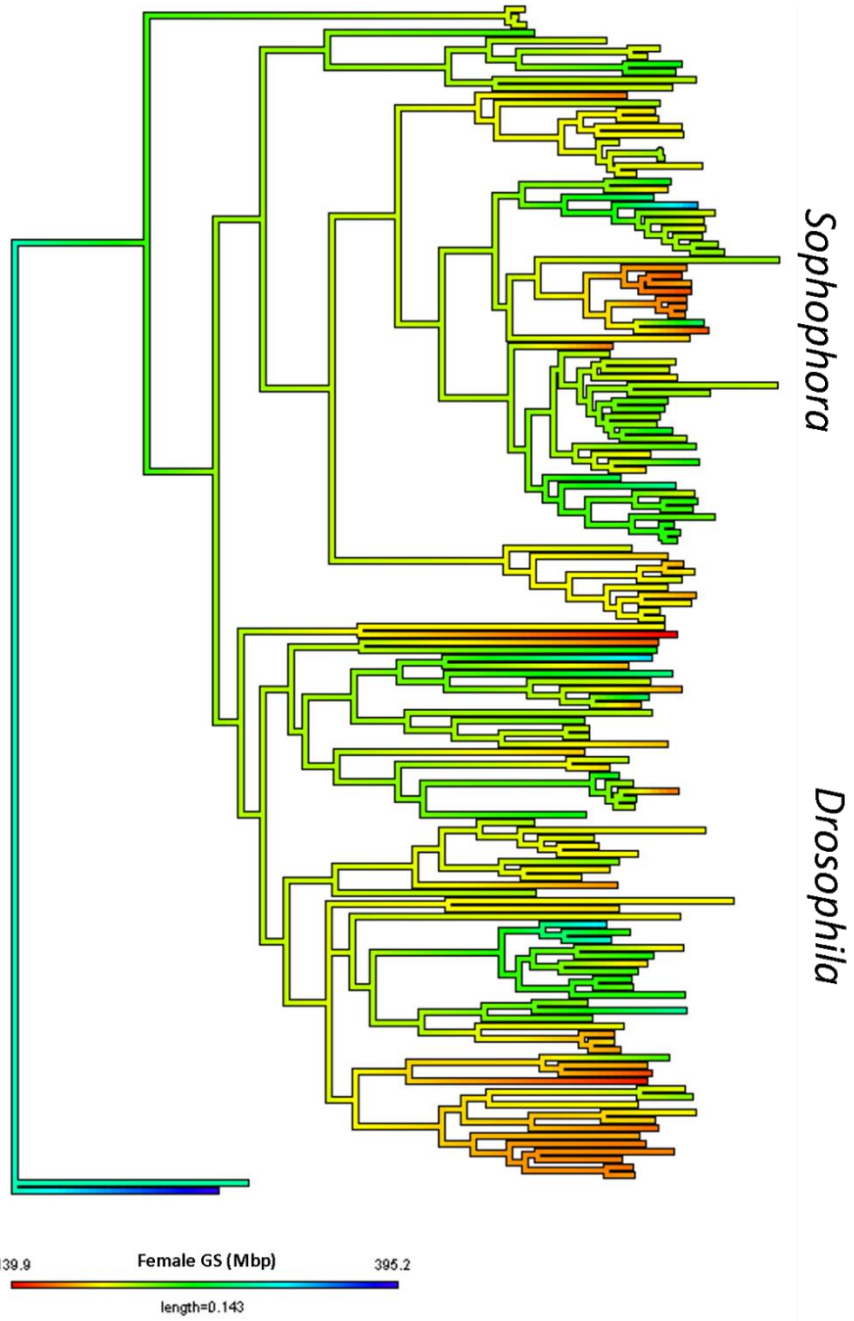

Figure 12: Female genome size color mapped onto full *Drosophila* phylogeny. Female genome size for the *Drosophila* genus was color mapped onto the Bayesian phylogeny utilizing the contmap function in phytools of R 3.3.0. Smaller genome sizes are indicated in red to yellow, where larger genome sizes are indicated in blue. The patterns found in color correspond to the estimates of Pagel's parameters of evolution. There is phylogenetic signal in genome size ($\lambda \sim 1$), gradual change on individual branches ($\kappa \sim 1$), and some evidence for early, yet gradual change on the phylogeny ($0.5 < \delta < 1$)

Modes of evolution in Drosophila

Estimates of Pagel's parameters that were next calculated based on species in the *Drosophila* subgenus have similar patterns for whole genome size. Genome size for females and males was found to have complete phylogenetic signal (female $\lambda = 0.945$, male $\lambda = 0.958$) and completely gradual change along branches (female $\kappa = 1.015$, male $\kappa = 1.002$) (Table 14). Patterns for change along paths throughout the tree were also found to be similar to those found in the entire *Drosophila* phylogeny (female $\delta = 0.851$, male $\delta = 0.874$). While these values are not significantly different than 1, they are also not significantly different than 0.5, indicating that there is a mostly gradual change with some evidence for greater earlier change (Table 14).

Modes of evolution of the sex chromosomes

The difference between sexes due to the sex chromosome size differences was found to have some phylogenetic signal across the species tested ($\lambda = 0.790$) with rapid change in branches ($\kappa = 0.336$). The δ value suggests that the change in sex difference size happens later in the phylogeny ($\delta = 1.472$) (Table 13). These results are similar to those found in Chapter III. However, while *Sophophora* have sex difference phylogenetic values comparable to those of the entire genus indicating change later in the phylogeny ($\lambda = 0.875$, $\delta = 1.255$, $\kappa = 0.380$), *Drosophila* sex difference values indicate that the location of change across the phylogeny is earlier than in *Sophophora* ($\lambda = 0.838$, $\delta = 0.826$, $\kappa = 0.390$) (Table 14). When the sex differences are trait mapped onto the phylogeny for *Sophophora*, there is a visible uniformity across the phylogeny, such uniformity likely underlies the phylogenetic signal. When change does occur in the

phylogeny, it is occurring late in the tree in individual species, which likely drives the increased δ value (Figure 13). In contrast, when the *Drosophila* subgenus sex differences are plotted on the phylogeny, there is a visible, qualitative change in color at the *virilis* group (Figure 14). This change midway through the phylogeny is likely what influences the lower δ value (*Sophophora* δ = 1.255, *Drosophila* δ = 0.826) (Table 14, Figure 14).

Table 14: Comparisons of Pagel's parameters between the subgenera.
Estimates of Pagel's parameters of evolution from Table 4 for the entire genus, the *Sophophora* subgenus, and the *Drosophila* subgenus are consolidated for comparison. There are little differences between the estimates for female and male genome sizes. The estimate of change on the entire phylogeny for the sex difference differs between the subgenera (*Sophophora* δ = 1.255, *Drosophila* δ = 0.826). This indicates that a large change in the X-Y system occurred earlier in *Drosophila* than in the *Sophophora* subgenus.

Female GS			
	All Species	***Sophophora***	***Drosophila***
λ	1	1	0.945
δ	0.913	0.720	0.851
κ	1.216	1.485	1.015
Male GS			
	All Species	***Sophophora***	***Drosophila***
λ	0.999	1	0.958
δ	0.922	0.724	0.874
κ	1.16	1.264	1.002
Sex Difference			
	All Species	***Sophophora***	***Drosophila***
λ	0.79	0.875	0.838
δ	1.472	1.255	0.826
κ	0.336	0.380	0.390

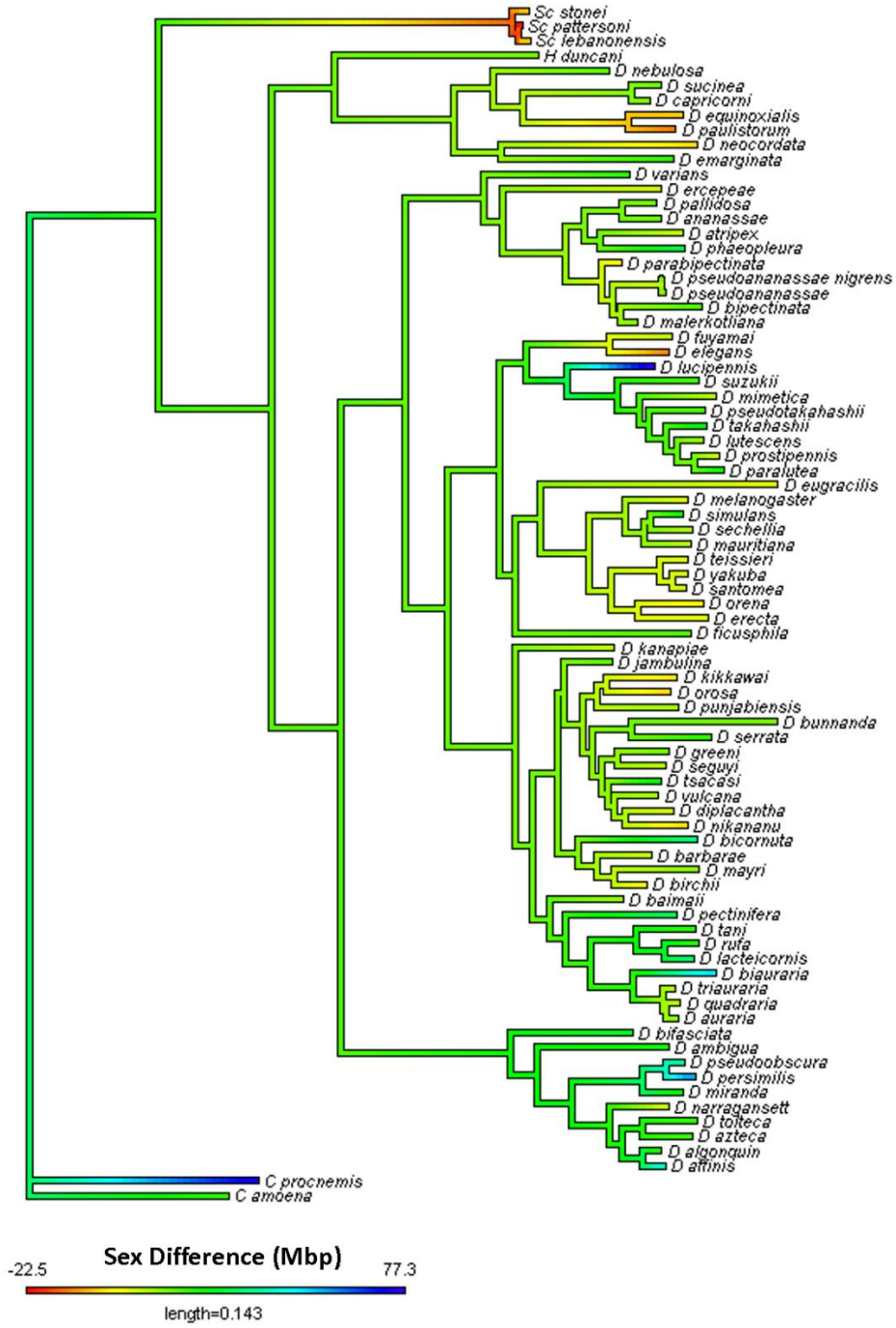

Figure 13: Sex difference values plotted onto *Sophophora* phylogeny.
The differences between the sexes due to the X-Y system were color mapped onto the *Sophophora* phylogeny. Red and yellow values indicate low values (larger males), whereas the blue colors indicate large values (where females are larger than males). This phylogeny visualizes the patterns found by Pagel's parameters of evolution. There is some incomplete phylogenetic signal ($\lambda = 0.875$), rapid early change on individual branches ($\kappa = 0.380$), and late change in the overall phylogeny ($\delta = 1.255$).

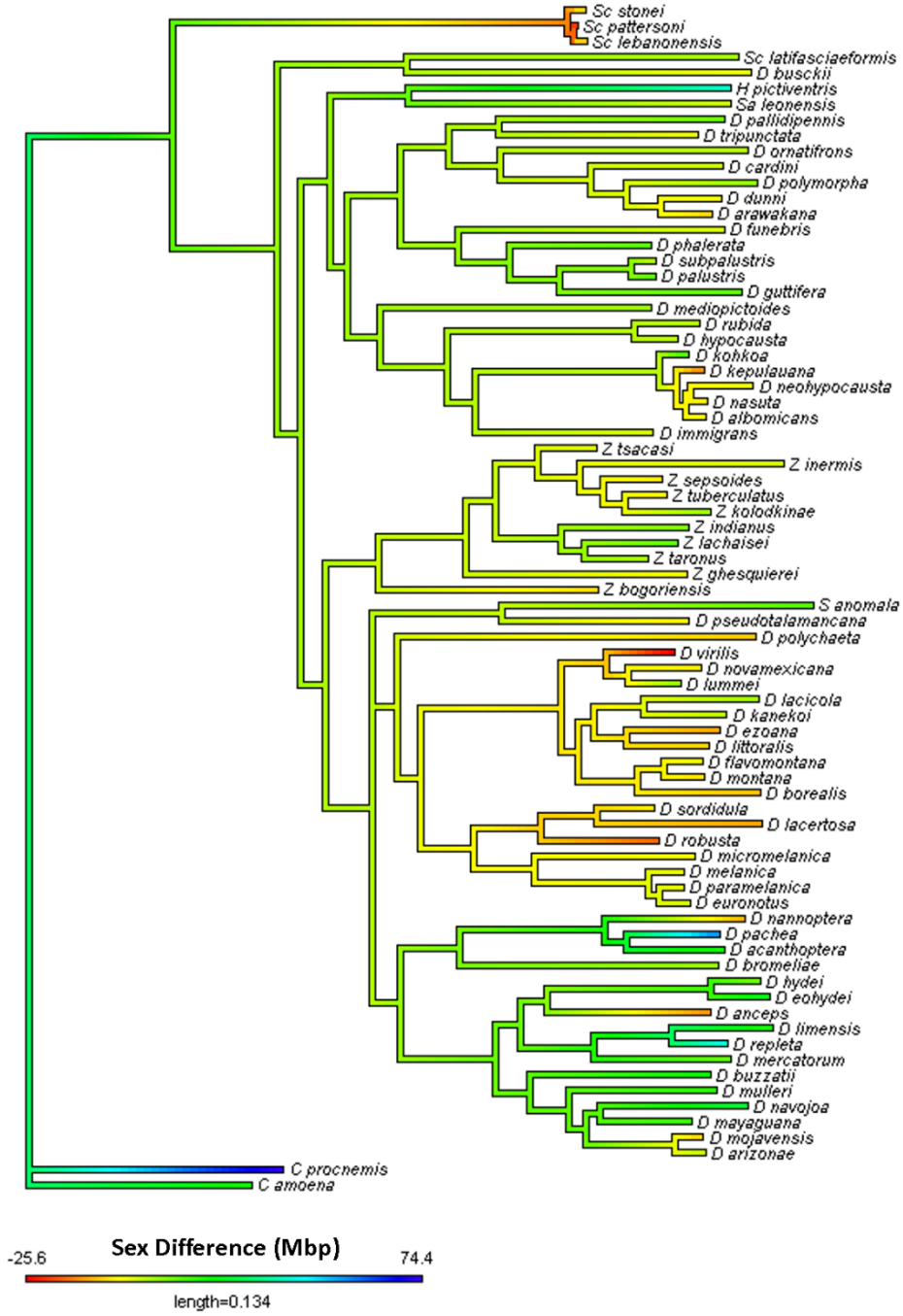

Figure 14: Sex difference values plotted onto *Drosophila* phylogeny.
The differences between the sexes due to the X-Y system were color mapped onto the *Drosophila* phylogeny. Red and yellow values indicate low values (larger males), whereas the blue colors indicate large values (where females are larger than males). This phylogeny visualizes the patterns found by Pagel's parameters of evolution. There is some incomplete phylogenetic signal ($\lambda = 0.838$), rapid early change on individual branches ($\kappa = 0.390$), and somewhat early change in the overall phylogeny ($\delta = 0.826$). This early change seems to correspond to the decrease in values found to the *virilis* and *melanica* groups on the phylogeny.

Discussion

The phylogenetic placement of species across the tested genera (*Drosophila*, *Zaprionus*, *Scaptodrosophila*, *Scaptomyza*, *Hirtodrosophila*) has been the discussion of many systematic studies due to issues with the paraphyly of the *Drosophila* genus (Reviewed in van der Linde et al. 2010 &van der Linde and Houle 2008). In that sense, the phylogeny that was reconstructed in this study is not controversial in regard to the placement of *Zaprionus* and *Scaptomyza* within the *Drosophila* subgenus, nor is the paraphyly of the *Scaptodrosophila* genus (Figure 9) (van der Linde et al. 2010). That being so, in an effort to make the studies of the subgenera comparable while including the related genera, species were treated as *Sophophora* or *Drosophila* depending on their placement in the phylogeny rather than their taxonomic name.

Here we report new and updated female and male genome sizes for 93 species of *Drosophila* and related genera. These are combined with published values for males and females of another 59 species for a total of 152 species. This doubles the previous large scale *Drosophila* species genome size study of 74 species (Gregory and Johnston 2008), and nearly doubles the recently published phylogenetic study of *Sophophora* genome size (Hjelmen and Johnston 2017). These added values dramatically increase the representation of the *Drosophila* subgenus genome size in order to make powerful comparisons to the more heavily studied *Sophophora* species. Genome size varies extensively across the species studied, from 137.5 Mbp in male *D. bromeliae* to 395.2 in female *C. amoena*, nearly a threefold range (Table 11).

It is important to note that some of these new estimations update previously published genome sizes estimates. By updating the previous estimates, we can be sure that all of the species were compared to the same standards and run on the same equipment. The majority of updated species values were within 10 Mbp of earlier estimates. Ranges of differences between the updated and previous values are indicated in Table 11.

Here, the higher number of chromosomes in the *Drosophila* subgenus (6 versus 4 in *Sophophora*) does not significantly influence the genome size of the group. While there is substantial variation in the genome sizes found across these subgenera, there is not a significant difference in the average genome size, female or male ($p > 0.05$, Table 12). Chromosome number has been shown to be positively correlated to genome size when analyzing angiosperms (890 species from 62 genera) (Pandit et al. 2014) and ray fin fishes (1,043 species across 190 families) (Yi and Streelman 2005). This pattern was suggested in early studies in vertebrates where diploid teleost fishes had a significant positive correlation between genome size and chromosome count (Hinegardner and Rosen 1972). While there is much evidence for this trend, this pattern has not been clearly supported across other taxa. The pattern in plants seems to be contradictory. Genome size has been found to correlate to chromosome count in *Carex* (Escudero et al. 2015), yet this correlation was not maintained when comparing *Ginlisea,* a carnivorous plant genus (Fleischmann et al. 2014) or cycads (Gorelick et al. 2014). A recent study has also found that genome size correlates to chromosome size, but not chromosome

number, in snapping shrimp (Jeffery et al. 2016). In that regard, the *Drosophila* results fit with other arthropods and some plants, but do not support the results found in fish.

While genome size between the subgenera is not significantly different, the estimated 40-65 million years since there divergence could have significant impacts on the patterns of genome size evolution that has occurred to achieve these genome sizes (Russo et al. 1995; Tamura et al. 2004; Obbard et al. 2012). However, when patterns of whole genome size evolution are compared between *Sophophora* and *Drosophila*, as well as across the entire genus, there are no remarkable differences (Table 14). Across all tested phylogenies, there is complete phylogenetic signal and a mostly gradual change across branches ($\lambda \sim 1$, $\kappa \sim 1$) (Table 14). Change, in terms of the entire phylogeny, was found to be mostly gradual, with some indication of earlier change, as δ was not significantly different than 0.5 or 1 (Table 13, 14). Since the species of interest are so closely related and genome size has already been shown to have phylogenetic signal in *Sophophora*, it is not surprising that the subgenera provide similar results for whole genome size evolution, (Hjelmen and Johnston 2017), even with dramatically different karyotypes. It is important to note of course that conclusions about these phylogenetic patterns of genome size evolution may differ between organisms. Genome size has been found to have no phylogenetic signal in seed beetles (Arnqvist et al. 2015) and to have different modes and tempos of evolution in fireflies (Lower et al. 2017).

The results of Pagel's parameters can allow us to make conclusions regarding which hypothesis for genome size evolution seems to best fit *Drosophila* genome size. Since there are no dramatic differences between the subgenera and the entire *Drosophila*

phylogeny, we can assume the patterns are upheld throughout the genus (Table 14). The overall results suggest that interspecific genome size across *Drosophila* is not adaptive, but rather follows a phylogenetic pattern (Powell 1997; Gregory and Johnston 2008; Hjelmen and Johnston 2017). It is important to note that there is some evidence for adaptation at the intraspecific level in *D. melanogaster* lines raised at different temperatures (Ellis et al. 2014). The differences we see at interspecific levels could be due to temporal differences. While there is evidence for adaptive changes in the short term, these patterns are not evident in these long term analyses. Further, while there is the possibility that adaptation to new environments may provide phylogenetic signal, especially in species groups, earlier tests of the effects of latitude and climate variables have found no relationship to genome size when analyzed with comparative phylogenetic methods (Hjelmen and Johnston 2017). In contrast, the δ values suggest that there is some early change, which gives some support to a low effective population size hypothesis (Lynch and Conery 2003), whereas the gradualistic κ value suggests that genome size changes is as predicted by the mutational equilibrium hypothesis (Petrov et al. 2000; Petrov 2001; Petrov 2002b).

Previous studies testing the low effective population size hypothesis have also found that the significant relationship reported between genome size and effective population size actually disappears when analyzed in a phylogenetic fashion (Whitney and Garland Jr 2010; Whitney et al. 2011). Here, females of *Sophophora* species were found to have a slightly increasing rate of genome size change on branches (female κ = 1.485, p from 1 = 0.012, Table 13). This increase may be explained by the measured

decrease in genome size found in the literature (Petrov and Hartl 1998; Petrov et al. 2000; Petrov 2002a), yet this increase in rate of genome size change is not upheld in males of the species ($\kappa = 1.264$, Table 13), and males are expected to have the same or a smaller effective population size than females.

Finally, the relatively constant rate of change over time lends some support for the mutational equilibrium hypothesis. While these phylogenetic results support the hypothesis, they are not mirrored in large datasets, with expressed concern regarding the slow rate by which this change would occur (Gregory 2003b; Gregory 2004). Another model of genome size evolution has been proposed, which is not subject to the rate argument and may explain the phylogenetic patterns in *Sophophora* and *Drosophila*: the accordion model (Kapusta et al. 2017). Here, proportionally larger deletions counteract large insertions from transposable elements, giving a gradual change that is faster than that of the mutational equilibrium hypothesis. This newly proposed model is attractive because it would provide the values for the mode of genome size change as seen here, but the change would occur at a speed sufficient to produce the variation we find in the phylogeny.

When the values for sex difference, a proxy for sex chromosome evolution, were analyzed using t-tests and Kolmogorov-Smirnov tests, a significant difference was found between the values for *Sophophora* and *Drosophila* (*Sophophora* 15.4 Mbp, *Drosophila* 5.2 Mbp, $p < 0.001$, Table 12). The positive values for sex differences (female- male genome size for each species) indicate that in the majority of species, the X chromosome is larger than the Y chromosome. The lower average value in the *Drosophila* subgenus

indicates that there is less of a size difference between the X and Y chromosomes in *Drosophila* than in the *Sophophora* subgenus. Since this comparison of raw data was the only significant different between the subgenera, we might expect to see a difference in the phylogenetic parameters and the mode of evolution of the Y chromosome in the subgenera. Incomplete phylogenetic signal ($0 < \lambda < 1$) is expected for sex chromosome evolution, as there is a general trend of Y chromosomes to become smaller over time (Reviewed in Charlesworth and Charlesworth 2000; Bachtrog 2013). The general decrease in size and stability of the Y chromosome results in uniformity of the trait across the phylogeny, as indicated by the similarity in color across the tree (Figures 13 and 14). While there is a trend downwards, there is a limit to how small the Y chromosome can become before a neo-Y event occurs, in which the male determining elements are either moved to another chromosome or the old Y chromosome is fused or translocated to another chromosome (Bachtrog 2013). Once a neo-Y chromosome is formed, it increases in size, through accumulation of transposable element material, which increase may result in the male of a species having a larger genome size (indicated by negative values in Table 11). This increase is then followed by a quick decrease in size due to loss of genetic content, and that was exactly what was seen. There was little difference of phylogenetic signal and change on individual branches between the subgenera (*Sophophora*: $\lambda = 0.875$, $\kappa = 0.380$; *Drosophila* $\lambda = 0.838$, $\kappa = 0.390$). The signal is no longer entirely due to phylogenetic relationships and the expected rapid change in size is likely what drives change to occur early in branches ($\kappa =$

0.380). These patterns are similar to those discussed in Chapter 3, with the exception that the change occurred at different phylogenetic levels in the two subgenera.

In *Sophophora*, the change in sex difference occurs late in the phylogeny (δ = 1.255) where most of the radiation of species has occurred, suggesting that the change may correspond to speciation (Figure 13). The large changes late in the phylogeny are likely due to neo-Y chromosomes, and this neo-Y chromosome formation can be hypothesized to increase instances of reproductive isolation to eventually result in new species. This hypothesis is supported in stickleback fishes, where a recent neo-sex chromosome had a role in speciation (Kitano et al. 2009). In the fish, the neo-X chromosome differences led to behavioral isolation, which eventually led to the reproductive isolation and speciation. Interestingly, the only difference in the evolutionary parameters we found between *Sophophora* and *Drosophila* was the delta value. While there is suggestion for late change in *Sophophora*, there is evidence for change earlier in the phylogeny for *Drosophila* (δ = 0.826, Table 13). This value is not significantly different than 0.5 or 1, suggesting that the change, while mostly gradual, likely occurs in the middle of the phylogeny. Inspection of the colorized trait map phylogeny shows visual evidence for change midway in the phylogeny, including the early reduction (shown as a yellow pattern) for the *virilis* and *melanica* groups (Figure 14). This reduction in size seems to correspond to the radiation of this group, giving further evidence for a potential role of the sex chromosome in speciation. In sum, while the overall patterns (λ and κ) suggest that Y-chromosome evolution occurs in a similar

fashion between the subgenera, δ suggests that the *Drosophila* have experienced sex chromosome changes temporally earlier than those in *Sophophora*.

These results seem to suggest stochastic, gradualistic change in genome size throughout the *Drosophila* genus, supporting the mutational equilibrium (Petrov 2002b) and accordion (Kapusta et al. 2017) models of genome size evolution. However, this conclusion does not seem to be entirely compatible with other published data finding correlation to genome size and life history parameters (Gregory and Johnston 2008; Ellis et al. 2014). Of concern here is the implication that the correlation between fitness effects and genome size, an adaptive characteristic, is not necessarily expected in the case of complete phylogenetic signal and gradualistic change. One complication of this concern is the dramatic difference in the time scale of these observations. The selection seen in laboratory studies occurred in a relatively short timescale, especially in relation to the 40-65 million years since the *Drosophila-Sophophora* subgenera split (Russo et al. 1995; Tamura et al. 2004; Obbard et al. 2012). This temporal difference, in conjunction with the evidence that *Drosophila* genome size is constrained (Wright and Andolfatto 2008; Sella et al. 2009; Koonin and Wolf 2010), suggests that there may be adaptive forces at play, but their patterns may be hard to discern from the phylogenetic patterns and support from the accordion model of evolution. Therefore, future studies should further investigate the role of selection and adaptation using phylogenetic comparisons of genome size and life history characteristics.

In conclusion, we have nearly doubled the genome size estimates for females and males of this group of flies. The additional data has allowed us to make the largest

phylogenetic analysis of genome size evolution to date comparing 152 species of Drosophilidae. And, while there has been 40-65 million years since their divergence, the patterns of genome size evolution is consistent between the subgenera *Drosophila* and *Sophophora* of the genus *Drosophila*. In addition there are no remarkable differences between the patterns found between whole genome sizes for males or females, suggesting that the small difference in Y chromosome evolution does not have a large impact on the evolution of entire genome size. The patterns observed in this group suggest that genome size is changing in a consistent, gradual pattern throughout time, giving most support to the newly proposed accordion model of genome size evolution. Further, while there is not a remarkable difference between sexes in terms of whole genome size, the difference in size of the sex chromosomes provides a unique perspective on Y chromosome evolution. Pagel's parameters of evolution give support for the proposed models of Y chromosome degradation over time. The degradation is reflected by partial, but incomplete phylogenetic signal and rapid change on branches. This same comparison between sexes has allowed us to see that large changes in the Y chromosome occur earlier in the phylogeny in *Drosophila* than it does in *Sophophora*, information that contributes to a better understanding of Y chromosomal changes.

CHAPTER V

UNDERREPLICATION OF MALE AND FEMALE *DROSOPHILA* SPECIES: RELATIONSHIP TO GENOME SIZE, AND IMPLICATIONS FOR GENOME SIZE EVOLUTION

Introduction

The genome has been described in two cytological forms: heterochromatin and euchromatin. Generally, euchromatin is considered to be the largely expressing portion of the genome, containing genes for regular expression. Heterochromatin is mostly structural, containing very few genes and remains relatively transcriptionally silent (Schulze et al. 2006). Oddly enough, heterochromatin has been found to be the last portion of the genome to replicate (Lima-de-Faria and Jaworska 1968), while euchromatin tends to be replicated early (Schübeler et al. 2002; McNairn and Gilbert 2003; MacAlpine et al. 2004; Schwaiger and Schübeler 2006). Early studies of DNA mutation rates found that early replicating regions of the genome have lower mutation rates, which could explain why the coding regions of the genome are clustered in euchromatic regions, with an increased rate of silent mutations and fast evolving genes in heterochromatic portions of the genome (Wolfe et al. 1989; Schwaiger and Schübeler 2006; Makunin et al. 2014).

Underreplication can be described as the stalling of replication during the S-phase before the replication of heterochromatin is complete (Leach et al. 2000). In general, these regions have low gene density, are tightly packed, and are considered to be transcriptionally inert (Belyaeva et al. 2008; Belyaeva et al. 2012). In general,

underreplication is most well-known from studies of polytene chromosomes of salivary glands (Rudkin 1969; Hammond and Laird 1985) and nurse cells (Painter and Reindorp 1939) of *Drosophila* and other Diptera. Underreplication was found to be controlled through the function of Suppressor of Underreplication (SuUR) in *Drosophila* species (Belyaeva et al. 1998). In the absence of SuUR function, intercalary heterochromatin replication was found to occur earlier in S-phase, whereas more SuUR gene product allowed for more underreplication (less total replication) (Zhimulev et al. 2003a). The SuUR gene was found to code for a 962 amino acid protein similar to SNF2/SWI2 proteins (Sucrose Non-Fermentation/Mating-type switching), which binds to late replicating portions of the heterochromatin (Makunin et al. 2002). It is therefore possible that the physical binding of SuUR to heterochromatin could be physically slowing the process of replication in these regions, possibly through modification of repressive chromatin complexes and structure (Volkova et al. 2003),by directly influencing binding efficiency of replication machinery, or by altering stability of the replication fork (Zhimulev et al. 2003a).

Yurlova et al. (2010) analyzed underreplication in 11 *Drosophila* species phylogenetically by comparing the number of amino acid substitutions across SuUR and its orthologs. They found that the SuUR protein product could be classified as "fast-evolving". Despite the numerous amino acids changes that occurred throughout time, the size and overall charge of the SuUR protein is mostly the same across the analyzed species, suggesting high conservation of function. For example, nearly all secondary structures of the protein were the same for the distantly related *D. melanogaster* and *D.*

grimshawi. Low homology orthologs were found in three species of mosquitoes (*Anopheles gambiae*, *Aedes aegypti*, and *Culex quinquefasciatus*), but no orthologs were found outside of Diptera. This suggests that while this phenomenon of underreplication may occur in Diptera, the highest conservation is within the *Drosophila* species.

Until recently, underreplication was a phenomenon only found in highly polytenized cells, such as salivary glands and ovaries. For example, Bosco et al. (2007) found a strong correlation between genome size and the amount of underreplication in the 16C polyploid ovarian follicle cells . They concluded that the addition and loss of heterochromatic satellite repeats could be an explanation for the large variation of genome size in *Drosophila* species. A study, however, reported that a majority of the nuclei of thoracic cells in *Drosophila melanogaster* were, in fact, underreplicated (Johnston et al. 2013). These are not the often reported polytene nuclei. Rather, the underreplication occurs during the S phase, with replication stalling between G1and G2. This finding could have many consequences for genome structure, such as: position effects, transcription levels, and general genome architecture (Belyaeva et al. 2003; Zhimulev and Belyaeva 2003; Zhimulev et al. 2003b; Belyaeva et al. 2006).

Since underreplication affects late-replicating heterochromatin, knowledge of this process and the level at which it occurs in different species provides new and interesting possibilities for understanding genome size evolution. The variation in genome size is known to not be due to coding sequence, but to the amount of non-coding sequences, such as transposable elements, repeats, introns, and intragenic spacers (Bennetzen and Kellogg 1997; Gregory and Hebert 1999; Hartl 2000; Gregory 2001;

Kidwell 2002; Vieira and Biemont 2004; Ågren and Wright 2011; Kelley et al. 2014). Therefore, if there is variation in the amount of replication among species and this variation is related to the genome size of the organism, we can use it to estimate the effects of different amounts of structural DNA on genome size evolution.

In order to address the relationship of genome size to thoracic underreplication as well as its impact on genome size evolution questions, we estimated the level of underreplication in the thorax for 114 species of female Drosophilidae and 99 male Drosophilidae. As SuUR was found to be highly structurally conserved in *Drosophila* species (Yurlova et al. 2010), it is expected that the occurrence of thoracic underreplication should be conserved as well. Since underreplication occurs in heterochromatic portions of the genome and genome size has been found to be related to the amount of noncoding sequences, it is hypothesized that genome size will be negatively correlated to the amount of replication, similar to the results found by Bosco et al. (2007) in follicular tissue. It is also hypothesized that underreplication values should differ between females and males due to the increased levels of heterochromatin on Y-chromosomes (Reviewed in Charlesworth et al. 2005; Bachtrog 2013). Once the amount of replication is estimated, this value can be used to calculate the physical amount of DNA that is replicated using the genome size for the species and sex. It is also expected that amount of DNA that is replicated will differ between sexes due to the variable chromatin level. In order to see if there is an impact on long term genome size evolution, values for the replicated and unreplicated amount of DNA will be analyzed using comparative phylogenetic methods for comparison to the patterns found in whole

genome size. It is expected that the phylogenetic values observed from the replicated and unreplicated amount of DNA will differ from those found from whole genome size. Since we know that the majority of coding information is in the replicated portion of DNA, we expect that this portion of the genome to be more constrained phylogenetically.

Methods

Estimation of underreplication in the thorax of Drosophila

Underreplication was estimated as in Johnston et al. (2013). Thoraces of both male and female species of *Drosophila* were dissected and placed in a 2 mL Kontes dounce tube with 1 mL of Galbraith buffer. Each thorax was then carefully ground with a "loose" A pestle approximately 10-15 times to ensure proper release of nuclei from the thoracic tissue. Each sample was then passed through a 41 micron filter before 25ul of 1 mg/µl propidium iodide was added for staining. Samples were allowed to incubate with the propidium iodide for at least 20 minutes to ensure adequate staining had occurred. Samples were then run on a Partec CyFlow SL_3 cytometer with a 532 nm green laser. Peaks for 2C, 4C and underreplication were gated to generate the mean of each in order to calculate the percent replication that had occurred for each individual (Figure 15). Underreplication was calculated by subtracting the 2C value (mean peak channel number) from the UR value, then divide by the 2C value. This process was repeated for at least 5 individuals in order to achieve an acceptable standard error. Estimates of underreplication were completed on the same individuals for which genome size was

estimated in order to see if there were patterns with genome size and underreplication of individuals within a species as well as between species.

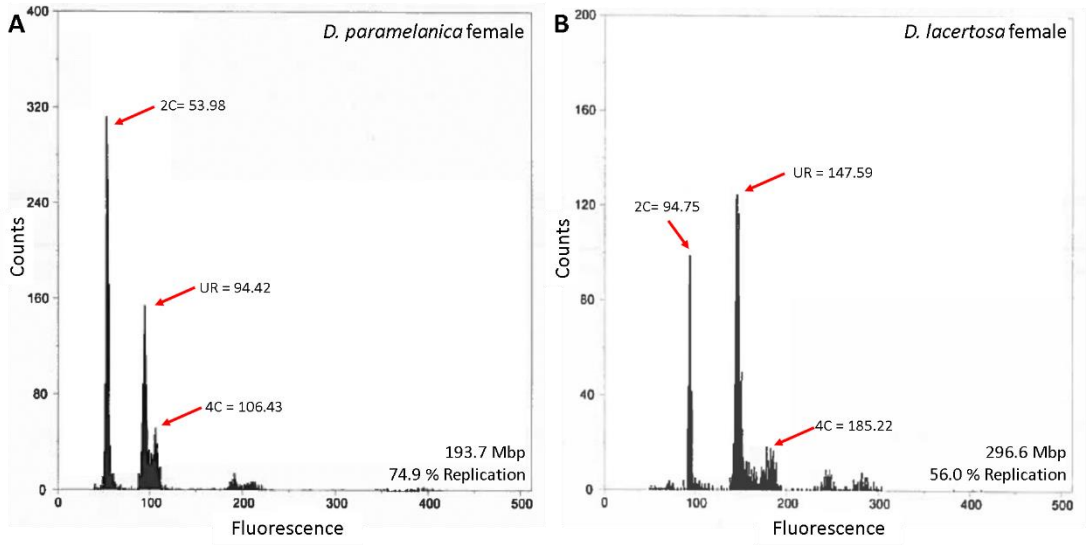

Figure 15: Flow cytometry histogram of underreplication in the thorax of two *Drosophila.*
Underreplication in the thorax is measured through flow cytometry with propidium iodide stained nuclei of thoracic tissue. The amount of underreplication is calculated by taking the difference in mean fluorescence from the UR peak and the 2C peak divided by the 2C mean. The amount of replication between species with varying genome sizes is visually different. Species with smaller genome sizes (*D. paramelanica*, 15A, have a higher level of replication than species with a larger genome size (*D. lacertosa*, 15B

Statistical methods

Regression analyses were completed in R 3.2.3 (Team 2016) using average underreplication values for comparison to the average genome size for each species. This was completed for males and females of each species in order to see if there were

different levels of variation that could be due to the largely heterochromatic Y chromosome.

The amount of DNA replicated was calculated by taking the percentage replication of the estimated whole genome size. Genome size and amount of replication were visualized in histogram format using R 3.3.0 (Team 2016). T-tests were performed on the replication percentage and the amount of DNA replicated between males and females of the species (Team 2016).

Comparative phylogenetic methods

A phylogeny was obtained for the measured species by trimming the large Bayesian phylogeny from Chapter IV. The tree was trimmed to contain *Chymomyza amoena* and *C. procnemis* to ensure proper rooting of the tree (n = 114). Values for replicated and unreplicated amount of genome size were plotted on a colorized phylogeny using the contmap function in the phytools package from R 3.2.3 (Revell 2012).

Pagel's parameters of evolution (λ, κ, and δ) were estimated for female and male genome size, replication percentage, replicated portion of the genome, and the unreplicated portion of the genome utilizing the pgls function from the Caper package in R 3.2.3 (Pagel 1999; Orme 2013).

Results

Replication levels and genome size

The amount of replication was found to vary extensively across the measured species (46.8% - 99.2%) (Table 15). Thoracic underreplication was not found to occur in *Chymomyza*, but was found to occur in all other species of Drosophilidae tested. When analyzed with regression analysis, underreplication was found to be significantly negatively correlated to genome size (Females, p = 2.44e-16, Males, p = 1.163e-11) (Figure 16A, B). Females were found to have a higher R^2 value than males (Female, R^2 = 0.4586, Males, R^2 = 0.385) (Table 16 and 5-3).

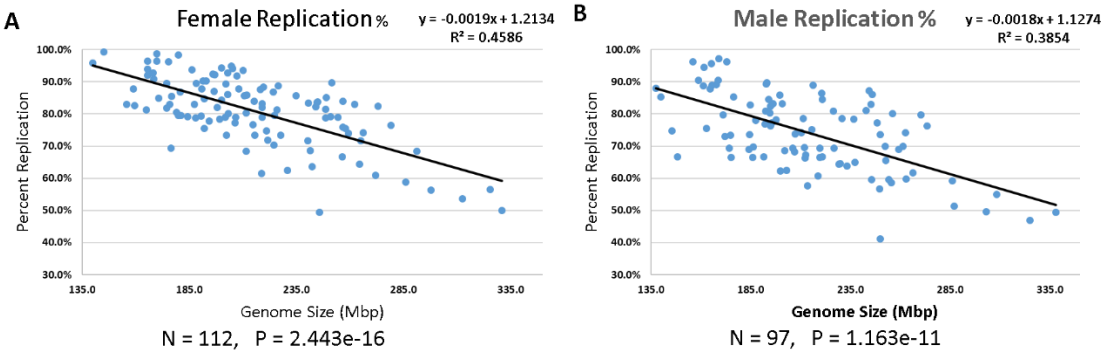

Figure 16: Regression plots for female and male underreplication vs. genome size in *Drosophila* species.
When replication percentage is plotted against genome size, there is a significant relationship (Female p-value = 2.443e-16, Male p-value = 1.163e-11). The overall slope of this relationship is comparable between sex (Female = -0.0019, Male = -0.0018), however the amount of variation in replication percentage explained by genome size is lower in males than in females (Female R^2 = 0.4586, Male R^2 = 0.385). This difference is likely due to the variable levels of heterochromatin on the Y chromosome in males.

Table 15: Genome size and underreplication values for species of *Drosophila.*

Genome size estimates and estimates of percent replication were completed using flow cytometry. The amount of replicated and unreplicated DNA was calculated by taking the percentage of the entire genome size. These were completed for females and males of each species.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
20010-2010.00	*C. amoena*	395.2	384.8	NA	NA	NA	NA	NA	NA
20000-2631.01	*C. procnemis*	298.7	261.5	NA	NA	NA	NA	NA	NA
13000-0081.00	*D. busckii*	139.9	140.1	95.9	85.2	134.2	5.7	119.4	20.7
15040-1191.00	*D. pseudotalamancana*	191.6	194.9	90.3	76.3	173.0	18.6	148.7	46.2
15085-1682.00	*D. bromeliae*	145.2	137.5	99.2	88.0	144.0	1.2	121.0	16.5
15182-2261.03	*D. arawakana*	181.1	186.4	79.6	69.7	144.2	36.9	129.9	56.5
15181-2181.03	*D. cardini*	215.7	212.5	73.4	67.2	158.3	57.4	142.8	69.7
15182-2291.00	*D. dunni*	286.1	287.0	58.7	51.3	167.9	118.2	147.2	139.8
15181-2231.00	*D. polymorpha*	179.2	171.2	80.4	79.8	144.1	35.1	136.6	34.6
15120-1911.01	*D. funebris*	221.1	219.8	74.8	66.4	165.4	55.7	145.9	73.9
15172-2151.00	*D. ornatifrons*	291.3	286.1	68.3	59.2	199.0	92.3	169.4	116.7

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
15130-1971.03	*D. guttifera*	186.3	173.2	93.8	96.3	174.7	11.6	166.8	6.4
15112-1751.04	*D. albomicans*	218.3	220.6	87.6	-	191.1	27.2	-	-
15115-1871.04	*D. hypocausta*	190.6	185.9	79.4	66.5	151.3	39.3	123.6	62.3
15111-1731.00	*D. immigrans*	265.2	262.9	71.6	59.7	189.9	75.3	157.0	105.9
15112-1761.03	*D. kepulauana*	224.6	237.3	70.3	64.9	157.9	66.7	154.0	83.3
15112-1771.01	*D. kohkoa*	271.9	255.6	60.9	58.6	165.6	106.3	149.8	105.8
15112-1781.01	*D. nasuta*	242.4	245.7	63.6	59.5	154.2	88.2	146.2	99.5
15115-1881.00	*D. neohypocausta*	165.7	165.6	94.0	89.0	155.8	9.9	147.4	18.2
15115-1901.03	*D. rubida*	218.7	213.4	61.5	57.6	134.5	84.2	122.9	90.5
15030-1131.01	*D. euronotus*	176.8	174.3	85.4	69.3	151.0	25.8	120.8	53.5
15030-1141.03	*D. melanica*	175.9	175.3	82.9	66.5	145.8	30.1	116.6	58.7

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
15030-1151.01	*D. micromelanica*	206.4	207.1	77.3	68.2	159.5	46.9	141.2	65.9
15030-1161.03	*D. paramelanica*	192.1	191.6	75.4	66.7	144.8	47.3	127.8	63.8
15090-1693.00	*D. acanthoptera*	155.8	145.6	82.9	74.7	129.2	26.6	108.8	36.8
15090-1692.11	*D. nannoptera*	245.9	250.1	49.4	41.2	121.5	124.4	103.0	147.1
15090-1698.01	*D. pachea*	176.5	148.3	69.4	66.6	122.5	54.0	98.8	49.5
15210-2331.01	*D. pallidipennis*	330.9	325.0	50.0	46.8	165.5	165.5	152.1	172.9
15100-1711.00	*D. polychaeta*	202.5	206.0	73.3	69.4	148.4	54.1	143.0	63.0
15130-2001.00	*D. palustris*	219.1	212.1	78.9	69.5	172.9	46.2	147.4	64.7
15130-2031.00	*D. phalerata*	240.3	233.3	71.6	63.7	172.1	68.2	148.6	84.7
15130-2071.01	*D. subpalustris*	214.7	212.2	76.7	66.2	164.7	50.0	140.5	71.7
15081-1261.02	*D. anceps*	201.8	207.5	78.9	73.7	159.2	42.6	152.9	54.6
15081-1271.14	*D. arizonae*	175.4	173.9	96.2	-	168.7	6.7	-	-

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
15081-1291.01	*D. buzzatii*	179.9	169.0	98.4	97.1	177.0	2.9	164.1	4.9
15085-1631.00	*D. eohydei*	231.0	218.7	62.5	60.8	144.4	86.6	133.0	85.7
15085-1641.03	*D. hydei*	206.8	201.0	78.9	-	163.1	43.7	0.0	201.0
15084-1591.02	*D. limensis*	203.1	192.9	92.7	89.3	188.3	14.8	172.3	20.6
15081-1397.00	*D. mayaguana*	170.0	161.7	98.7	94.5	167.8	2.2	152.8	8.9
15082-1521.00	*D. mercatorum*	165.5	159.1	92.1	-	152.3	13.2	-	-
-	*D. mojavensis*	165.6	168.7	96.5	90.5	159.7	5.9	152.7	16.0
15081-1371.01	*D. mulleri*	168.3	158.9	92.7	90.5	156.0	12.3	143.8	15.1
15081-1374.02	*D. navojoa*	170.0	156.2	96.4	96.3	163.9	6.1	150.4	5.8
15084-1611.01	*D. repleta*	187.5	165.5	82.9	95.7	155.4	32.1	158.4	7.1
15020-1101.03	*D. lacertosa*	298.0	303.1	56.3	49.5	167.8	130.2	150.0	153.1

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
15020-1111.01	*D. robusta*	241.6	249.6	68.5	56.6	165.5	76.1	141.3	108.3
15020-1121.01	*D. sordidula*	227.6	229.3	73.3	64.4	166.8	60.8	147.7	81.6
15220-2371.00	*D. mediopictoides*	188.0	184.3	78.8	69.0	148.1	39.9	127.2	57.1
15220-2401.02	*D. tripunctata*	184.4	184.5	79.2	73.5	146.0	38.4	135.6	48.9
15010-0961.00	*D. borealis*	257.8	261.5	75.5	69.9	194.6	63.2	182.8	78.7
15010-0971.00	*D. ezoana*	192.9	197.9	90.2	78.1	174.0	18.9	154.6	43.3
15010-0981.00	*D. flavomontana*	259.5	259.3	74.0	68.9	192.0	67.5	178.7	80.6
15010-1061.00	*D. kanekoi*	256.4	254.4	66.6	59.6	170.8	85.6	151.6	102.8
15010-0991.12	*D. lacicola*	196.0	192.4	92.3	80.8	180.9	15.1	155.5	36.9
15010-1001.03	*D. littoralis*	248.6	251.5	81.4	-	202.4	46.2	-	-
15010-1011.07	*D. lummei*	312.7	308.5	53.7	55.0	167.9	144.8	169.7	138.8
15010-1021.09	*D. montana*	240.6	242.8	83.3	81.0	200.4	40.2	196.7	46.1

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
15010-1031.04	*D. novamexicana*	251.0	252.7	79.2	65.6	198.7	52.3	165.8	86.9
-	*D. virilis*	325.4	338.2	56.5	49.4	183.8	141.5	166.9	171.3
14024-0371.13	*D. anassae*	196.6	188.8	87.0	-	170.9	25.7	-	-
14024-0361.00	*D. atripex*	198.2	194.8	82.0	80.2	162.5	35.7	156.2	38.6
14028-0471.00	*D. auraria*	254.4	248.4	78.9	77.2	200.7	53.7	191.8	56.6
14028-0481.00	*D. baimaii*	279.2	273.6	76.4	76.2	213.3	65.9	208.5	65.1
14024-0381.19	*D. bipectinata*	204.6	195.3	94.9	-	194.2	10.4	-	-
14027-0461.03	*D. elegans*	192.2	197.1	84.7	-	162.7	29.5	-	-
14021-0224.01	*D. erecta*	158.9	157.0	87.7	-	139.4	19.5	-	-
14025-0441.05	*D. ficusphila*	190.8	182.1	87.7	-	167.4	23.5	-	-
14029-0011.00	*D. fuyamai*	266.3	262.5	74.2	74.1	197.6	68.7	194.5	68.0

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
14028-0541.00	*D. kanapiae*	168.6	164.8	90.9	87.7	153.3	15.3	144.5	20.3
14028-0571.00	*D. lacteicornis*	248.6	230.7	78.7	78.6	195.6	53.0	181.3	49.4
14028-0591.00	*D. mayri*	262.2	256.2	83.0	80.0	217.6	44.6	205.0	51.2
4021-0231.36	*D. melanogaster*	174.5	172.1	81.8	73.0	142.7	31.9	125.6	46.5
14028-0601.00	*D. nikananu*	219.7	220.6	88.3	86.4	194.0	25.7	190.6	30.0
14028-0611.01	*D. orosa*	215.3	216.2	84.0	88.9	180.9	34.4	192.2	24.0
14028-0641.00	*D. punjabiensis*	200.0	194.9	84.4	82.3	168.8	31.2	160.4	34.5
14028-0661.03	*D. rufa*	256.8	246.0	84.0	86.1	215.7	41.1	211.8	34.2
14021-0248.25	*D. sechellia*	179.9	175.6	79.6	-	143.11	36.8	-	-
14028-0671.02	*D. seguyi*	249.1	245.3	85.1	83.0	212.0	37.1	203.6	41.7
14028-0681.00	*D. serrata*	226.7	215.8	88.7	75.0	201.1	25.6	161.9	54.0
14021-0251.195	*D. simulans*	159.6	147.2	82.6	-	131.8	27.8	-	-

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
14022-0311.13	*D. takahashii*	207.3	193.7	91.8	-	190.2	17.1	-	-
14020-0011.01	*D. tani*	211.5	200.8	80.3	83.2	169.8	41.7	167.0	33.7
14021-0257.01	*D. teissieri*	165.0	162.9	81.3	75.5	134.1	30.9	123.0	39.9
14028-0651.00	*D. triauraria*	256.7	252.4	75.8	70.0	194.6	62.1	176.7	75.7
14028-0701.00	*D. tsacasi*	224.8	211.5	79.5	78.5	178.7	46.1	166.0	45.5
14028-0711.00	*D. vulcana*	226.0	221.1	81.3	84.6	183.7	42.3	187.1	34.0
14021-0261.01	*D. yakuba*	170.7	168.3	85.0	-	145.0	25.7	-	-
14012-0141.00	*D. affinis*	208.1	184.8	87.8	82.8	182.7	25.4	153.0	31.8
14012-0161.00	*D. algonquin*	205.2	194.2	94.1	84.5	193.1	12.1	-	-
14013-1011.00	*D. ambigua*	188.4	176.5	89.5	85.3	168.6	19.8	150.6	25.9
14012-0171.02	*D. azteca*	202.3	192.2	89.3	76.8	180.7	21.6	147.6	44.6
14028-0511.00	*D. bicornuta*	245.0	226.6	83.8	80.9	205.3	39.7	183.3	43.3

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
14012-0181.02	*D. bifasciata*	210.6	196.2	85.6	83.2	180.3	30.3	163.2	33.0
14028-0586.00	*D. diplacantha*	273.2	270.4	82.4	79.6	225.1	48.1	215.2	55.2
14028-0712.00	*D. greeni*	244.1	236.7	82.3	78.4	200.9	43.2	185.5	51.1
14028-0671.01	*D. jambulina*	219.0	210.4	83.1	74.2	182.0	37.0	156.1	54.3
14028-0561.14	*D. kikkawai*	210.2	210.6	93.6	-	196.6	13.6	-	-
14011-0101.13	*D. miranda*	211.9	199.6	85.9	85.9	182.0	29.9	171.5	28.1
14011-0111.49	*D. persimilis*	197.1	167.7	92.2	89.2	181.6	15.5	149.6	18.1
14011-0121.32	*D. pseudoobscura*	180.8	161.3	86.7	88.6	156.8	24.0	142.9	18.4
14041-0831.00	*D. neocordata*	219.0	219.8	82.0	69.3	179.6	39.4	152.3	67.5
14030-0741.00	*D. equinoxialis*	264.5	266.4	64.3	61.6	170.1	94.4	164.1	102.3
14030-0761.00	*D. nebulosa*	211.5	203.1	68.4	62.5	144.7	66.8	126.9	76.2

Table 15 Continued.

Stock	Species	Female Mbp	Male Mbp	Female Replication %	Male Replication %	Female Replicated Mbp	Female Unreplicated Mbp	Male Replicated Mbp	Male Unreplicated Mbp
14030-0771.00	*D. paulistorum*	244.7	250.1	83.5	73.6	204.3	40.4	184.1	66.0
92000-0075.00	*H. duncani*	251.6	244.3	89.7	87.2	225.7	25.9	213.0	31.3
33000-2661.01	*S. anomala*	200.2	193.1	94.2	89.7	188.6	11.6	173.2	19.9
52000-2752.00	*Z. bogoriensis*	224.3	226.9	86.9	69.4	194.9	29.4	157.5	69.4
50000-2743.00	*Z. ghesquierei*	174.5	174.9	89.5	73.3	156.2	18.3	128.2	46.7
50001-1020.00	*Z. taronus*	203.2	199.9	80.1	62.2	162.8	40.4	124.3	75.6
50000-2746.00	*Z. inermis*	203.1	202.0	86.1	68.6	174.9	28.2	138.6	63.4
50000-2748.00	*Z. kolodkinae*	199.3	195.5	78.4	77.2	156.3	43.0	150.9	44.6
50000-2751.00	*Z. tsacasi*	221.8	221.1	71.9	66.6	159.5	62.3	147.3	73.8
50001-1031.09	*Z. indianus*	234.8	229.9	85.6	64.5	201.0	33.8	148.3	81.6
50002-2750.00	*Z. lachaisei*	194.3	187.7	77.7	77.9	151.0	43.3	146.2	41.5

Table 16: Female genome size vs. underreplication value regression results. Genome size was found be significantly related to underreplication in females of Drosophila ($p < 0.001$). Much of the variation in underreplication was found to be explained by genome size ($R^2 = 0.4586$).

Coefficients				
	Estimate	Std. Error	t value	P(>|t|
(intercept)	121.33809	4.25062	28.546	<2e-16
Female Genome Size	-0.187785	0.01946	-9.653	2.44E-16
Residual S.E. = 7.901, Multiple R-squared = 0.4586, Adj. R-squared = 0.4537, F-statistic = 93.18, p-value = 2.44e-16				

Table 17: Male genome size vs. underreplication value regression results. Genome size was found to be significantly related to underreplication in male of *Drosophila* ($p < 0.001$). Much of this variation in underreplication was found to be explained by genome size ($R^2 = 0.3855$)

Coefficients				
	Estimate	Std. Error	t value	P(>|t|
(intercept)	112.73436	5.09673	22.119	<2e-16
Male Genome Size	-0.18076	0.02341	-7.721	1.16E-11
Residual S.E. = 9.492, Multiple R-squared = 0.3855, Adj. R-squared = 0.3791, F-statistic = 59.61, p-value = 1.16e-11				

The estimated amount of DNA that is replicated in the underreplicated nuclei in the thorax was found to average 171.3 Mbp for females and 153.63 Mbp for males (Table 18, Figure 17A, and B). While the average genome size for females and males is not significantly different (t-test, t = 0.0.88305, df = 225.81, p = 0.3782) (Figure 18),

the average amount of replicated DNA between males and females is significantly different (t-test, t = 4.7915, df = 180.75, p = 3.44e-06).

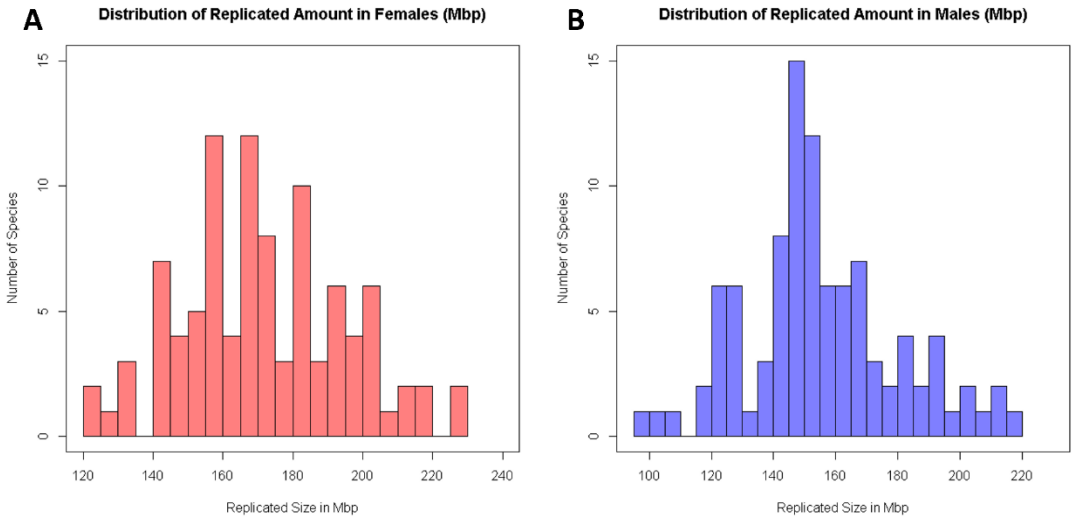

Figure 17: Distribution of replicated portions of the genome for female and male *Drosophila*
There is a significant difference in the average amount of DNA replicated between females and males (t-test, t = 47915, df = 180.75, p = 3.44e-06). A). Females replicate 171.30 Mbp on average, with a minimum of 121.5 Mbp and a maximum of 225.7 Mbp. B). Males replicate 155.33 Mbp on average, with a minimum of 98.77 Mbp and a maximum of 215.3 Mbp.

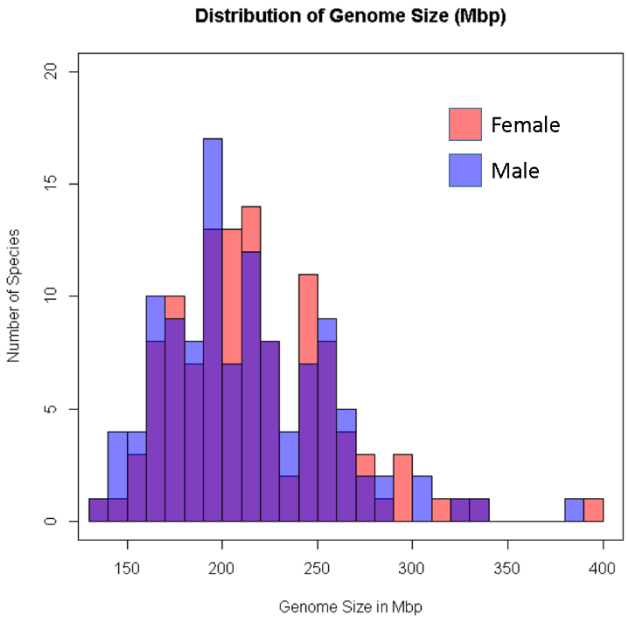

Figure 18: Distribution of genome size for female and male *Drosophila*
The average genome size between females and males of *Drosophila* are not significantly different (t-test, t = 0.88305, df = 225.81, p = 0.3782)

Comparative phylogenetics

Pagel's parameters of evolution for whole genome size for females and males did not differ from previous analyses (Chapters II, III, IV) (Table 18). There was evidence for complete phylogenetic signal (Female $\lambda = 0.947$, Male $\lambda = 0.967$), mostly gradual change on branches (Female $\kappa = 0.84$, Male $\kappa = 0.928$) and some evidence for change early in the phylogeny with gradual change following (Female $\delta = 0.761$, Male $\delta = 0.761$) (Table 18). The differences for the replicated and unreplicated amounts of the genome are remarkably different than those found for the whole genome size. All other tests (female and male underreplication percentage, replicated amount of the genome, and unreplicated amount of the genome) were found to have incomplete, but partial,

phylogenetic signal ($0 < \lambda < 1$), suggesting that the tree and evolutionary relationships have some impact on the variation found in the amount of replication (Table 18). While the partial signal in all of these variables is fairly consistent between females and males, there is slightly more phylogenetic signal in the amount of replicated DNA in female *Drosophila* ($\lambda = 0.837$ vs. $\lambda = 0.723$). All tested variables had δ values of 3, indicating that a majority of the change that is found occurs in the later part of the tree. Values of κ across the variables tested were all found to be above 0.5, but less than 1, suggesting there is a mostly gradual change in the variation of replicated and unreplicated DNA. This pattern for gradual change is supported by the p-values suggesting that all of these κ values are not significantly different than 0.5 or 1 (Table 18). Interestingly, this is not the case for κ in replicated DNA amount in males ($\kappa = 0.512$, p-value = 0.018 from 1). This suggests that the change in variation of replicated DNA in males may occur slightly earlier in branches than the DNA replicated in females.

Table 18: Pagel's phylogenetic parameters of evolution for underreplication in *Drosophila.*

Pagel's phylogenetic parameters between males and females of *Drosophila* were estimated using the pgls function of the Caper package in R 3.3.0. While whole genome size had complete phylogenetic signal, replicated and unreplicated portions of the genomes had lower signal. Change in whole genome size was found to occur relatively early in the phylogeny, whereas the change in replicated and unreplicated DNA was found to occur late in the phylogeny ($\delta = 3$)

Females							
Genome Size		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.947	2.22E-16	4.05E-06	0.175	-	(0.811, 1)
	δ	0.761	1.04E-10	0.375	0.41669	<2.22e-16	(0.218, 1.328)
	κ	0.84	1.03E-04	0.120	0.47535	<2.22e-16	(0.412, 1.284)
Underreplication		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.668	5.28E-05	0.29789	1.52E-08	-	(0.3226, 0.869)
	δ	3	2.50E-09	-	0.00034	1	(2.227, 3)
	κ	0.716	0.00104	0.332	0.209	<2.22e-16	(0.283, 1.161)
Replicated Amount		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.837	1.25E-10	0.0033819	1.05E-05	-	(0.638, 0.944)
	δ	3	2.18E-08	-	0.00223	1	(1.847, 3)
	κ	0.71	0.00105	0.319	0.168	<2.22e-16	(0.293, 1.122)
Unreplicated Amount		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.619	0.00041	0.519	2.60E-07	-	(0.240, 0.863)
	δ	3	6.92E-09	-	0.00065	1	(2.156, 3)
	κ	0.767	0.000321	0.223	0.297	<2.22e-16	(0.3441, 1.208)

Table 18 Continued.

Males							
Genome Size		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.967	1.33E-15	3.63E-06	0.391	-	(0.831, 1)
	δ	0.761	1.03E-10	0.374	0.414	<2.22e-16	(0.218, 1.323)
	κ	0.928	7.56E-05	0.0745	0.769	1.85E-14	(0.459, 1.422)
Underreplication		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.833	3.31E-08	0.025169	0.00259	-	(0.551, 0.970)
	δ	3	3.12E-08	-	0.0023	1	(1.896, 3)
	κ	0.8	0.00022	0.170	0.361	<2.22e-16	(0.372, 1.227)
Replicated Amount		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.723	4.06E-08	0.124	2.27E-06	-	(0.427, 0.893)
	δ	3	4.44E-11	-	2.80E-05	1	(2.424, 3)
	κ	0.512	0.0148	0.955	0.0178	<2.22e-16	(0.101, 0.916)
Unreplicated Amount		**Value**	**From 0**	**From 0.5**	**From 1**	**From 3**	**CI**
	λ	0.696	0.00076	0.35347	0.00017	-	(0.236, 0.919)
	δ	3	1.34E-09	-	0.00021	1	(2.292, 3)
	κ	0.865	6.59E-05	0.0976	0.545	<2.22e-16	(0.434, 1.301)

Discussion

Thoracic underreplication was tested in species of the genera *Drosophila*, *Zaprionus*, *Hirtodrosophila*, and *Scaptomyza* (Table 15). Interestingly, this phenomenon was not found to occur in *Chymomyza*, a genus outside of *Drosophila*¸ yet within the family Drosophilidae. At this point, no other Diptera have been reported to have underreplication take place in the thorax. This is perhaps not surprising, given the remarkable conservation of the SuUR protein throughout *Drosophila*, yet with retention of only low homology orthologs in mosquitos (Yurlova et al. 2010).

The amount of replication was found to vary from 46.8 % to 99.2 %. Since we know that the variation in genome size is not due to differences in the amount of coding sequence, but to variation in the amount of nongenic DNA (Gregory and Hebert 1999; Kidwell 2002; Kelley et al. 2014), we would expect that there would be substantial variation in the heterochromatic regions of the genome where underreplication takes place. As expected, the amount of replication in a species was found to be significantly negatively correlated to the genome size in both males and females (Figure 16 and Tables 16 and 17). As the genome size gets larger, the percentage of the genome that is replicated decreases. Even a cursory glance at histograms from flow cytometry of two species with different genome sizes will show that larger genomes replicate less of their genome (Figure 15). The addition or deletion of this late replicating portion of the genome likely has a large influence on the inflation or decrease in size of the genome. These results are supported by the previous work in underreplication in the follicular tissues of *Drosophila* species (Bosco et al. 2007).

As might be expected, the amount of variation in underreplication explained by genome size is lower in males of the species than in females (Male $R^2 = 0.385$ vs Female $R^2 = 0.459$) (Tables 16 and 17, Figure 16 A, B). The Y chromosome in the males is largely heterochromatic, and non-replicating. The variation in the size of the Y chromosome was earlier shown to have less phylogenetic signal (Chapter III, IV). The same is seen when the underreplicating portion of the Y is added to the other underreplicating DNA. Males also replicate statistically significantly less DNA than females (Males = 153.64 Mbp, Females = 171.30 Mbp, t-test, p-value = 3.44e-06) (Figure 17 A, B). Again, this is what we might expect, as the Y chromosome is largely heterochromatic and late replicating. These differences were obscured when the Y was included with the total DNA from the autosomes (whole genome size). When female and male whole genome sizes are statistically compared there is no difference (t-test, t = 0.88305, df = 225.81, p = 0.3782) (Figure 18).

The differences found between the sexes for the underreplicated portion can likely be attributed to their XY sex determination mechanism. Y chromosomes tend to be small, highly heterochromatic with very few coding regions (Bachtrog 2013). The increase amount of heterochromatic regions would likely increase the amount of late replicating heterochromatin, which would, in turn, decrease the percentage of the genome that is replicated. If males were all to have the same size Y chromosome, the amount of DNA replicated may be smaller than females, but the amount of variation explained by genome size should be similar (Regression R^2). However, the size of the Y is not constant between species, nor is the amount of replicated DNA (Charlesworth and

Charlesworth 2000; Bachtrog 2013) and the reduced R-square for the males confirms that this is so.

Y chromosome size is highly variable, especially with instances of neo-Y chromosomes, where the male determining genes and regions of the chromosomes may move to another chromosome through a fusion event, leading to a Y chromosome that is larger than the X chromosome (Charlesworth and Charlesworth 2000; Bachtrog et al. 2008). As discussed in chapters 2 and 3, the process of Y chromosome degradation can occur rapidly, but would first result in an increase in size of the chromosome through increased transposable elements and repeat sequences. Since Y chromosomes among such a large group of species will likely be in many different places in the process of degradation and formation of neo-Y chromosomes, it is not surprising to see high variation in the amount of heterochromatin and replication. The resulting differences in phylogenetic patterns of replicated versus unreplicated portions of the genome are explored below.

Different patterns were found for replicated and unreplicated amounts of DNA in these species, allowing us to get a more detailed picture of the evolutionary process. All of the variables tested for females and males (replicated and unreplicated DNA) had partial, but incomplete, phylogenetic signal ($0 > \lambda > 1$) (Table 18). This suggests that the evolutionary relationships among species have some impact on the variation found in replication, but does not explain all of it. When the replicated amounts of DNA are plotted on the phylogeny, the colors across the tested species seem to remain fairly consistent (Figures 19 and 20). This indicates that the amount of replicated DNA is

consistent across the phylogeny, with a few instances of change. While there is less overall consistency across the phylogeny in unreplicated amount of DNA, closely related species are consistent with each other. When there is a significant change in the amount of DNA replicated or not replicated (a dramatic shift in coloration), the color change seems to remain consistent between closely related species. Since there is not usually random amounts of change found between closely related species, this amount of change in coloration can visually represent the level of phylogenetic signal that is found. The amount of replicated DNA in males, had a slightly lower level of phylogenetic signal ($\lambda = 0.723$) than was found for females ($\lambda = 0.844$) suggesting that the levels of replication in males are different phylogenetically than in females. While there are no dramatic differences in the coloration of the phylogeny between females and males (Figures 19 and 20), the difference in the range of amount of DNA may lead to this difference (121.5 – 225.7 Mbp in Females, 98.8 – 215.2 Mbp in Males). Again, the average amount of DNA replicated in females vs in males was statistically different (t-test, $t = 4.7561$, $df = 181.4$, $p = 4.011e\text{-}06$). These differences, again, are likely due to the variation in replication due to the highly heterochromatic Y chromosome. The variation in heterochromatin content is more consistent between the X and autosomal chromosomes in females, potentially leading to a more stable and predictable level of replication.

When comparing κ values among the tested variables, all variables had values above 0.5 yet below 1 (Table 18). This differs from the values for whole genome size (Female $\kappa = 0.81$, Male $\kappa = 0.928$). The κ value near 0.5 suggests that while there is somewhat more change earlier in the individual branches of the tree, there is a mostly

gradual change. The suggested pattern of gradual change along branches is supported by the fact that the values for κ for most of the variables is not significantly different from 0.5 or 1 (Table 18). Consistent with the patterns we have found, the only variable that was significantly different from 1 was the replicated amount of DNA in males (κ = 0.512, p-value = 0.018 from1). This suggests that the change in the amount of replicated DNA in males happens early in individual branches, becoming more stable over time. This pattern, again, is likely that expected of Y chromosome evolution. When the Y chromosome changes from an old Y to a neo-Y system, there is a rapid increase in size due to fusion to a new chromosome. A corresponding increase in transposable element activity occurs which results in a rapid loss of gene function and an increase in heterochromatin content. Further, while there is a rapid decrease in size after the formation of heterochromatin, the overall proportion of heterochromatin on the Y chromosome remains high. All of this change would happen early in branches. Early change, followed by relative stability later in Y chromosome evolution would likely pull the κ values upward to reflect the later stability (κ > 0).

The biggest differences found in phylogenetic parameters between whole genome sizes and values from underreplication (replicated and unreplicated amounts) were δ values (Table 18). Whole genome sizes were found to have evidence for some early change in the phylogeny followed by gradual change throughout time (Female δ = 0.761, Male δ = 0.761). All other variables tested (replicated and unreplicated amount for females and males) were found to have δ values equal to 3 (Table 18). Values of 3 suggest that the majority of change in these traits happen late in the phylogeny. When

there is a large amount of change in the amount of replicated or unreplicated DNA, the change tends to occur in clusters of closely related species. These late changes can be visualized on the colorized phylogenies (Figures 19-22). There is a dramatic decrease in the amount of replicated DNA in the *nannoptera* group (*D. nannoptera, D. pachea,* and *D. acanthoptera*), even including the next closest species, *D. bromeliae*. When looking at the amount of replicated DNA, there are fairly rapid increases in the *montium* and *virilis* subgroups (Figure 19 and 20). When inspecting the phylogenies for the amount of unreplicated DNA, there are obvious rapid decreases in the *mulleri, obscura*, and *melanogaster* subgroups (Figures 21 and 22). The results for high δ values therefore could suggest that the change in replication due to varying levels of structural DNA could be related to the radiation of groups of species. The large structural changes in the genome may decrease the amounts of successful hybridization that occurs between populations, leading to reproductive isolation, and eventually speciation due to rapid accumulation of differences (Brown and O'Neill 2010).

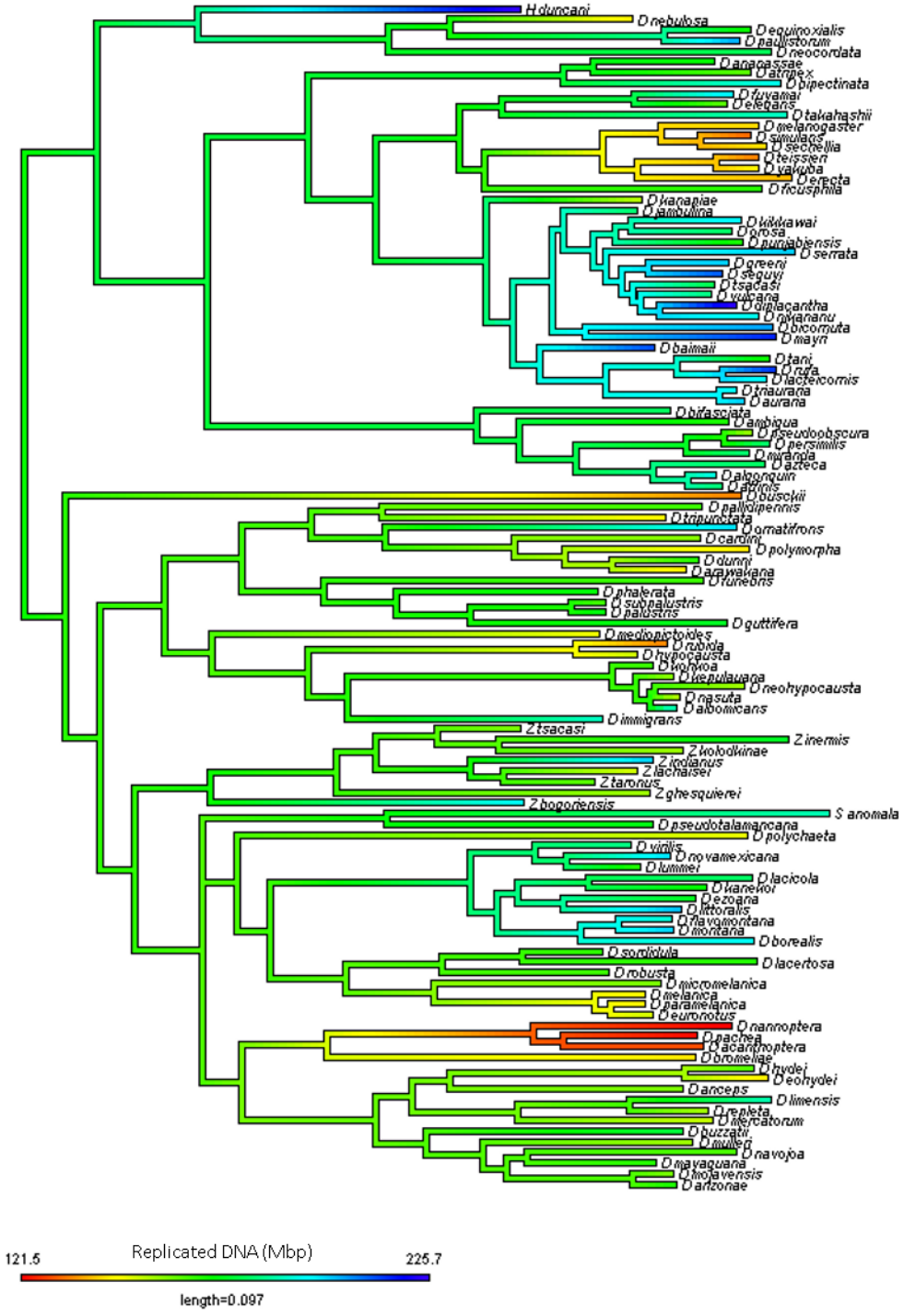

Figure 19: The amount of replicated DNA in female species plotted on the phylogeny
The amount of DNA replicated in females was plotted on the phylogeny using the contmap function in the phytools package of R 3.2.3.

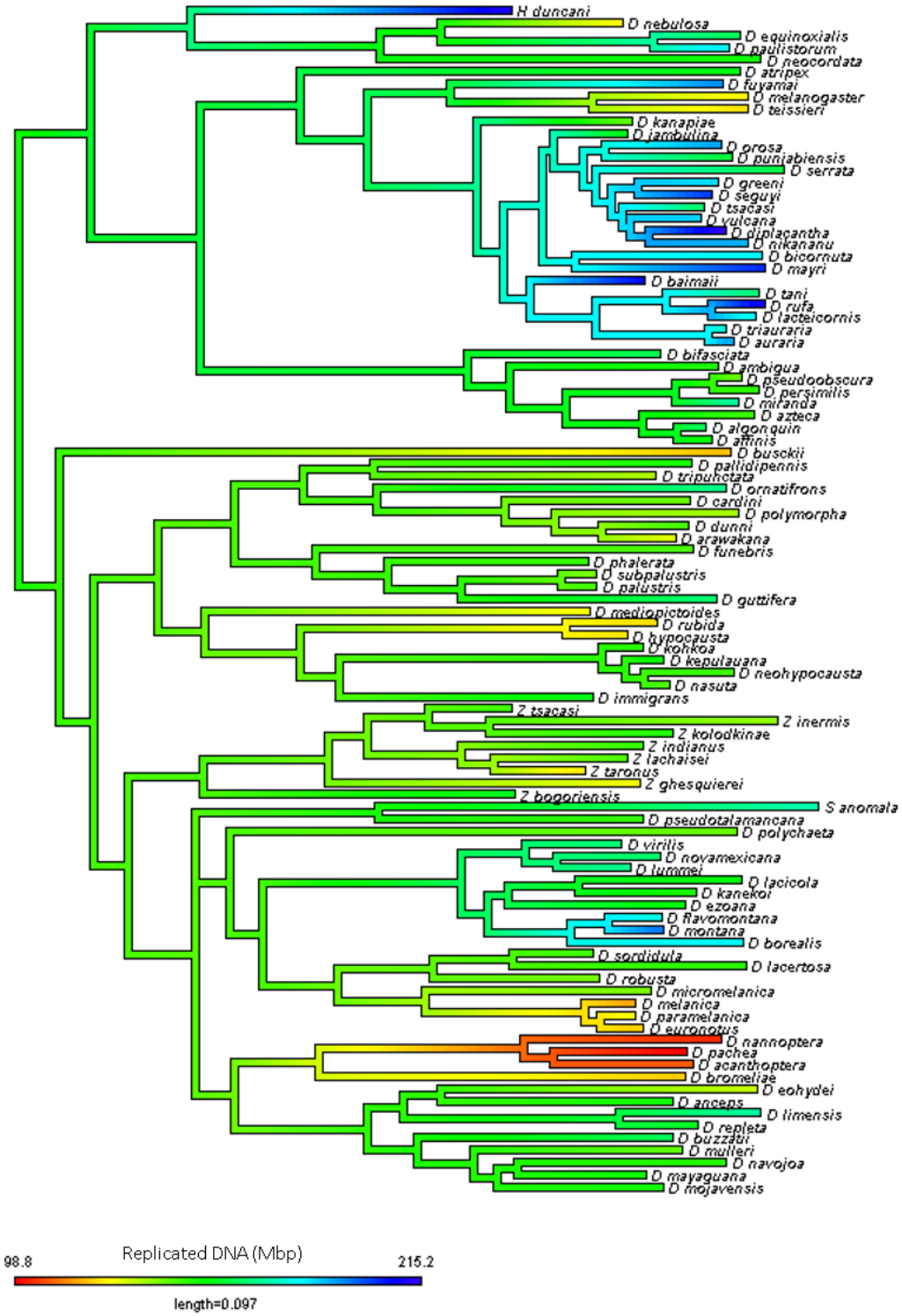

Figure 20: The amount of replicated DNA in male species plotted on the phylogeny
The amount of DNA replicated in males was plotted on the phylogeny using the contmap function in the phytools package of R 3.2.3.

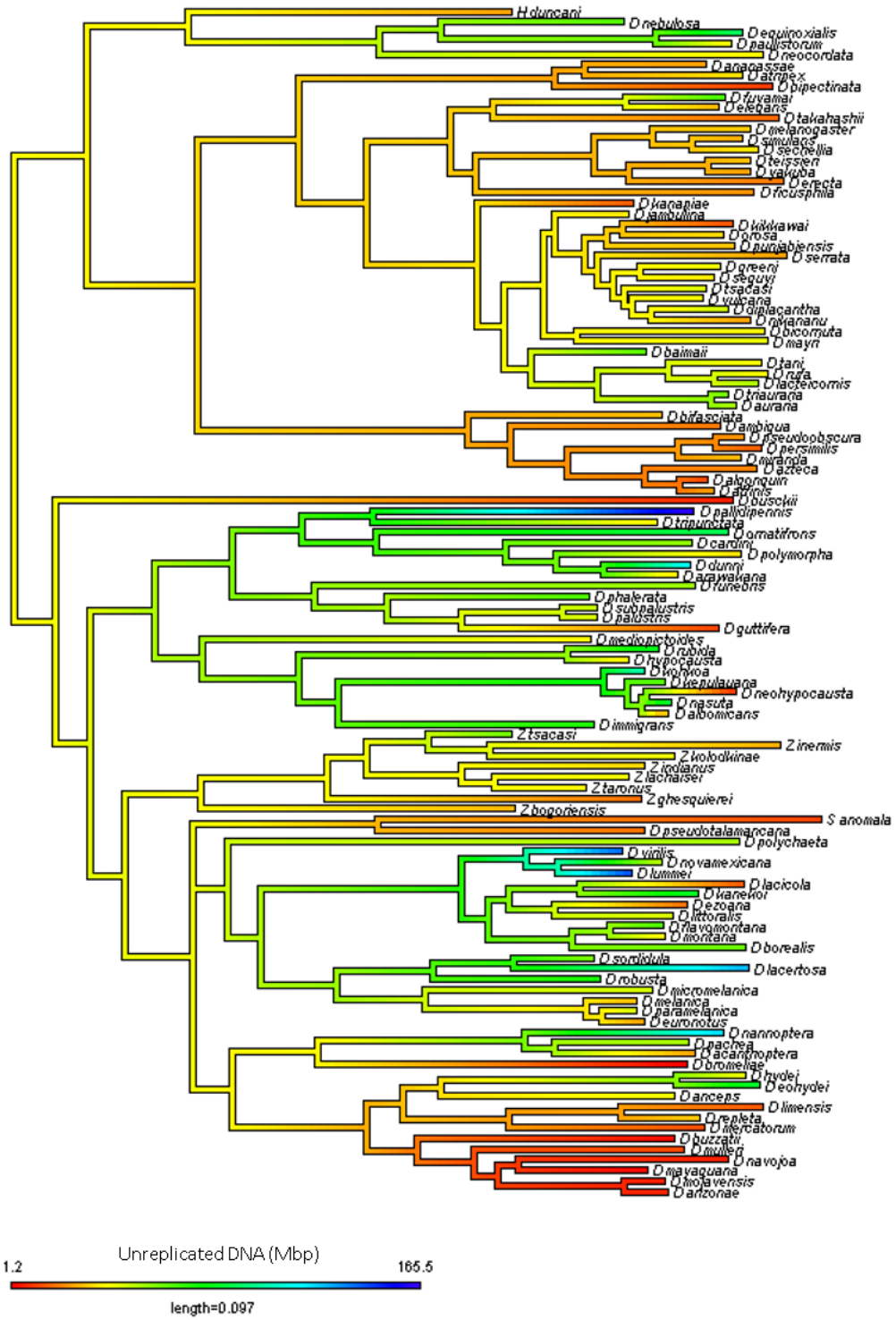

Figure 21: The amount of unreplicated DNA in female species plotted on the phylogeny
The amount of DNA not replicated in females was plotted on the phylogeny using the contmap function in the phytools package of R 3.2.3.

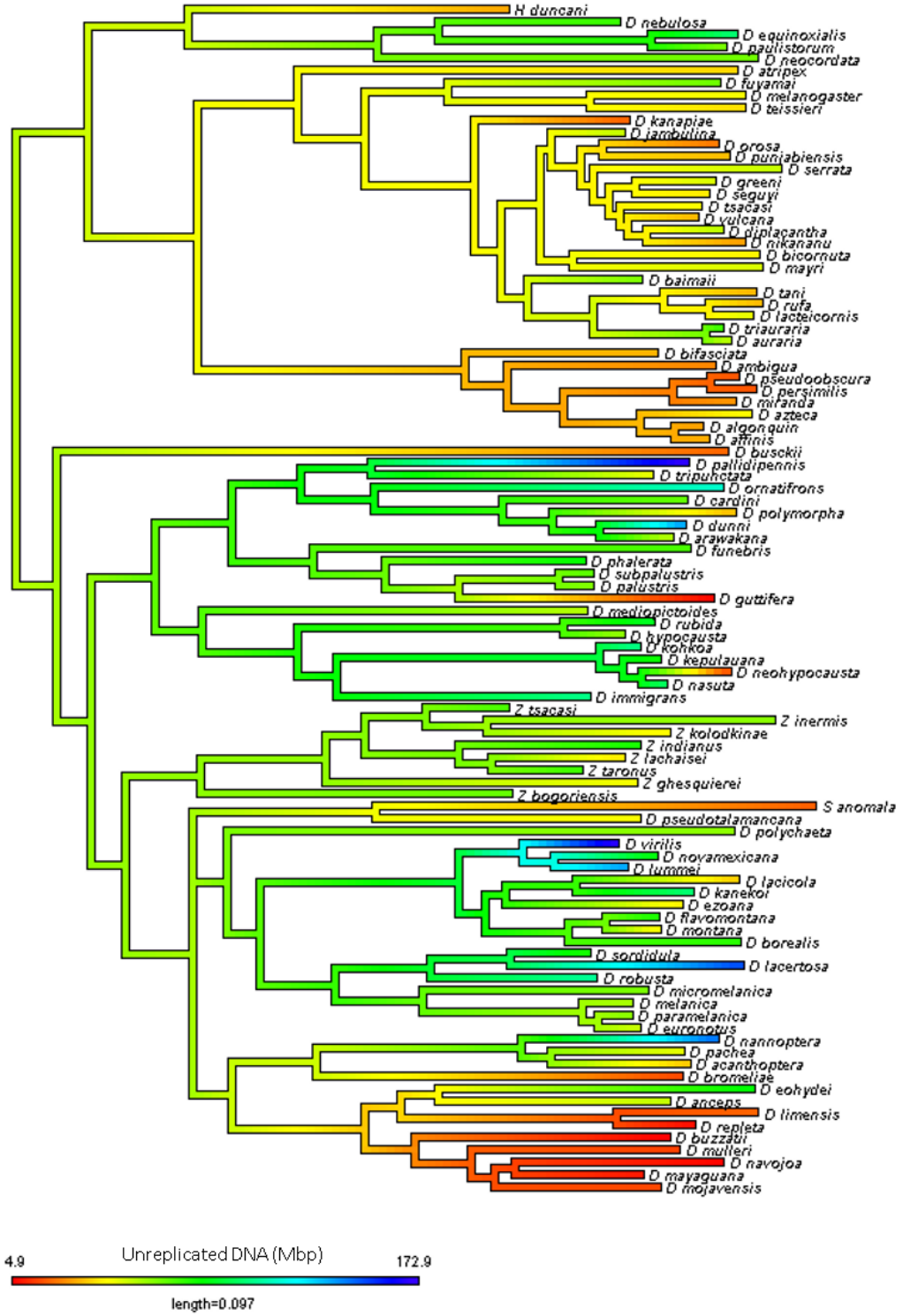

Figure 22: The amount of unreplicated DNA in male species plotted on the phylogeny
The amount of DNA not replicated in males was plotted on the phylogeny using the contmap function in the phytools package of R 3.2.3.

Although there is a statistically strong relationship for genome size and underreplication, there are a few species that do not seem to visibly fit the trend line in the regression (Figure 16). The species that dramatically stand out are the closely related, desert dwelling cactophilic species *D. pachea* (Female = 176.5 Mbp, 69.4% UR; Male = 148.3 Mbp, 66.6% UR), *D. nannoptera* (Female = 245.9 Mbp, 49.4% UR; Male = 250.1 Mbp, 41.2 % UR), and *D. acanthoptera* (Female = 155.8 Mbp, 82.9 % UR; Male = 145.6 Mbp, 74.7% UR) (Table 15). When the amount of replicated DNA is plotted on the phylogeny, these species stand out and are shown in red (low amount of replicated DNA) for both females (Figure 19) and males (Figure 20). Interestingly, when the amount of unreplicated DNA is plotted for females (Figure 22A) and males (Figure 22B), these species do not stand out as dramatically. While the underreplication values may seem surprising compared to other species of *Drosophila*, it is not the first time these species have been noted for being exceptional in their genome architecture. Previous studies analyzing their karyotypes have noted that these three species specifically have a large proportion of heterochromatin in their genome (Ward and Heed 1970) (Figure 23, modified from Ward and Heed (1970)). The karyotypes of these three species differ by four fixed inversions and the addition of heterochromatin. One interesting possibility for these species being so different is they experienced pressures due to their desert dwelling cactophilic lifestyle. The unique characteristics of these species, such as specialization on columnar cacti, asymmetric genitalia, and site of sperm storage, have been studied extensively in relation to ecological speciation, phylogenetic relationships, and specialization (Lang et al. 2014). It is possible that these same

pressures from the environment that shaped the other unique characters could have applied pressure leading to their highly heterochromatic genome. This suggests that looking at the karyotypes prepared for species in the past can give us information on the species that may have exceptional values for underreplication. One species of interest for this would be *D. nasutoides*, which has been reported to have the largest genome size of any *Drosophila* with a very high proportion of heterochromatin (Powell 1997). This species, however, is not currently available for studies.

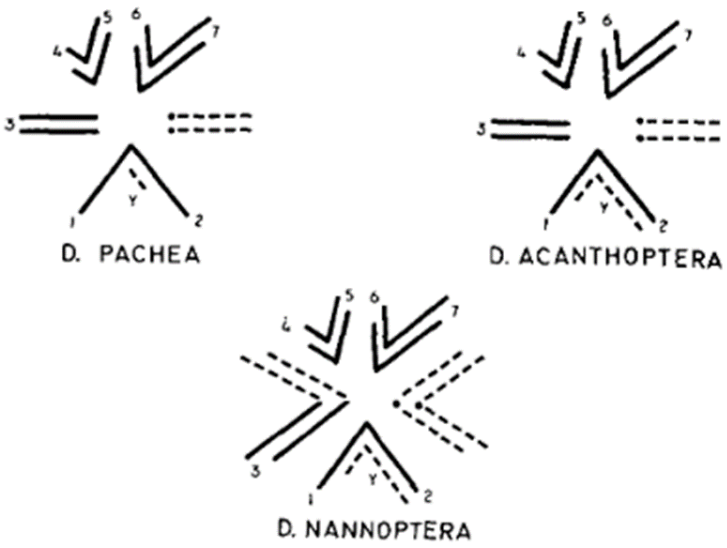

Figure 23: Karyotypes of three species with exceptional underreplication values[2]

Karyotypes of species in the *nannoptera* group have exceptionally large proportions of heterochromatin (indicated by dashed lines). Figure from Ward and Heed 1970.

[2] Reprinted with permission from "Chromosome phylogeny of *Drosophila pachea* and related species." by B.L Ward and W.B. Heed. 1970. *Journal of Heredity* 61.6: 248-258

In conclusion, there is a significant negative relationship between genome size and the amount of replication that occurs in the thorax of *Drosophila* species. Replicated and unreplicated amounts of DNA do not achieve full phylogenetic signal as genome size does, suggesting that there are different forces acting on these separate components of the genome as compared to the summation of them. The presence of partial signal and a gradualistic κ value does, however, suggest that there is a gradual trend in change that would influence the genome size patterns consistently across species. This idea is supported by the consistent relationship between genome size and the amount of replication that occurs in a species (Figure 16). The variation in males and females in κ seems to be explained by the heterochromatic Y chromosome and the process of degradation, suggesting less than gradual change in the Y. The difference in heterochromatin content on the Y also explains why less of the variation in underreplication in males is explained by genome size and why there is significantly less DNA replicated in males than in females. Most interesting, however, are the δ values suggesting that the late replicating heterochromatin in species likely has an impact on radiation of species groups later in the phylogeny. It is important to note, however, that underreplication occurs only in late replicating heterochromatin, not in all heterochromatin. It is possible that the variation found in the early replicating heterochromatin and the variation in all heterochromatin content may be providing different phylogenetic signals.

CHAPTER VI

CONCLUSIONS AND FUTURE DIRECTIONS

The patterns behind genome size, while measurable, are yet to be fully understood. Throughout this dissertation, I focus on hypotheses for patterns such as the effective population size hypothesis (Lynch and Conery 2003), mutational equilibrium hypothesis (Petrov 2002b), the accordion model (Kapusta et al. 2017), and the adaptive hypothesis for genome size change (Powell 1997; Gregory and Johnston 2008). While all of these have support in some form in the literature, there is also data that rejects most of these hypotheses (Gregory 2003b; Gregory 2004; Whitney and Garland Jr 2010; Whitney et al. 2011; Hessen et al. 2013; Kelley et al. 2014). Since these hypotheses all rely on testing differences between species, they should be accounting for phylogenetic relatedness (Felsenstein 1985), yet to this date very few have done so, and those that do have been published since the inception of this dissertation (Arnqvist et al. 2015; Jeffery et al. 2016; Sessegolo et al. 2016; Hjelmen and Johnston 2017; Lower et al. 2017). In order to address the lack of studies incorporating phylogenetic relatedness in analyses of genome size evolution we have implemented and argued for the use of comparative phylogenetic analyses throughout this dissertation. While we test methods such as Moran's I (Moran 1950), Abouheif's C_{mean} (Abouheif 1999; Pavoine et al. 2008), and Blomberg's K (Blomberg and Garland 2002; Blomberg et al. 2003), we put the most focus on Pagel's parameters of evolution (Pagel 1999), because they measure more than the extent of phylogenetic signature; they also measure the mode and tempo of trait evolution.

Since these are relatively novel methods, we first tested the reliability of these methods in regard to significance level and taxa number (Chapter II). While these methods had been tested using simulations (Münkemüller et al. 2012), they had yet to be tested utilizing actual data. Therefore, we used the well-studied, in terms of phylogenetic relatedness and genome size, system of *Sophophora* to test how reliable these analyses are using genome size. Similar to the results from the simulation study, we find that the reliability is related to taxa number (Münkemüller et al. 2012; Hjelmen and Johnston 2017). While there is not a relationship between taxa number and discovery of a phylogenetic signal, we found that there needs to be at least 15-20 species in these phylogenetic analyses to attain statistically reliable results (Chapter II).

When genome size data from databases were used for *Sophophora*, it was found that genome size had complete phylogenetic signal ($\lambda = 0.987$), gradual change along branches and somewhat early change in the phylogeny (Chapter II, Hjelmen and Johnston (2017)). The overall trend found was that *Sophophora* became small temporally early (suggested by the low δ value), and continues to change gradually (suggested by κ value). This downward trend is supported from previous work in the literature (Petrov and Hartl 1998; Petrov 2002a). From our expectations of each hypothesis, this pattern best fit those associated with the mutational equilibrium model of genome size change (Petrov 2002b). However, this hypothesis has been criticized for being too slow to account for the variation we see among species (Gregory 2003b; Gregory 2004). Therefore, the recently proposed accordion model of genome size change (Kapusta et al. 2017) seems to best fit the patterns we see in *Sophophora*.

The use of *Drosophila* species has allowed us to quickly expand these studies in order to see how the patterns hold up across the subgenera, *Sophophora* and *Drosophila* that have had millions of years to evolve. In this dissertation, I have added 93 new and updated genome sizes to expand the study to 152 species within the *Drosophila* genus (Chapter IV). While these subgenera differ in their karyotype (Painter and Stone 1935), they do not have significantly different average genome sizes. Comparing patterns of change across the entire genus, we do not see differences from the study within *Sophophora*. There continues to be complete phylogenetic signal ($\lambda = 1$), gradual change along branches ($\kappa = 1.216$), and some evidence for early change in the phylogeny ($\delta = 0.913$). These results continue to support the accordion model of genome size change. In addition, when these patterns are compared between genome sizes of males and females of the species, the pattern continues to be upheld (Chapters III and IV).

While phylogenetic signal and gradual changes on branches and the tree support the mutational equilibrium (Petrov 2002b) and accordion models (Kapusta et al. 2017) of genome size change, it is important to explore alternative explanations for the phylogenetic signal and change. There are constraints on *Drosophila* genome size (Wright and Andolfatto 2008; Sella et al. 2009; Koonin and Wolf 2010). Without constraint due to selection or a high deletion rate, selfish DNA elements would cause the genome to become bloated (Blumenstiel 2011). And yet, bloating of genome sizes is not common in *Drosophila* species; they are small in comparison to other insect, even within Diptera (Gregory 2005a). It can be argued that the assumption of stochastic, gradual change ignores constraints; published data shows an association of genome size and life

history fitness parameters that suggest an adaptive model. In particular, that assumption appears to ignore the correlations of genome size with fitness characters in *Drosophila* reported in Gregory and Johnston (2008) and Ellis et al. (2014). The former examined genome size evolution across 67 species of *Sophophora,* finding a relationship between genome size and temperature-controlled development. They also find that there may be a slight relationship between body size and sperm length, a reproductive character. The question that must be addressed is, "Would we expect a complete phylogenetic signal if these observations are correct?"

The same question must be addressed for the data of Ellis et al. (2014), which examined genome size evolution in *D. melanogaster*. They scored genome size for 211 inbred strains and compared the life history characteristics of the 25 strains with the largest genome size against the 25 strains with the smallest genome size. Up to 23% of the variation in developmental phenotypes could be accounted for with genome size. Temperature effects on development indicated that the correlation of fitness and genome size may depend on their environmental conditions. I have assumed that adaptation would produce bursts of change in response to a changing environment. Of concern is the possibility that the observed correlation of genome size and fitness parameters could produce complete phylogenetic signal as reported here.

A complication in these considerations is the very large temporal difference involved in the phylogeny as opposed to the time required for adaptation to changes in habitat or environment. The phylogenetic time span for *Drosophila* and *Sophophora* is approximately 50 million years (Russo et al. 1995; Tamura et al. 2004; Obbard et al.

2012). Changes in climate and events that may spur dispersal across distance, determine species distributions and show an adaptive response to local environments occur on a much shorter time span. And yet, while phylogenetic analysis looks at change over a very long period of time, this does not mean we cannot detect adaptation with these methods. Adaptation is evident if we pay attention to where the change happens on individual branches and where it occurs in terms of the entire phylogeny.

These studies indicate genome size evolution in *Drosophila* exhibits complete phylogenetic signal. This means that the variation in the trait among species is expected based on the Brownian motion model of evolution. Species that are closely related have lower levels of variation than those that are more distantly related. Complete phylogenetic signal is expected if genome size evolves stochastically, with mutations and sampling error (genetic drift) the only forces acting to produce genome size change. And while the direction of change cannot be predicted, the rate of change is very predictable. Change will be gradual and the rate of change will be effectively constant over the very long evolutionary time periods represented by a phylogeny.

I have found that genome size change in *Drosophila* is very close to the expectations expected of stochasticity, but not exactly so. The δ and κ parameters, which look at rates of change along branches, detect departures from a gradualistic expectation (Pagel 1999). δ detects the departure from the expected direction of change over the full extent of the phylogeny, from root to tip. The *Drosophila* genome has decreased over evolutionary time (Petrov and Hartl 1998; Petrov et al. 2000; Hjelmen and Johnston 2017) and that directional change is not an expectation of a stochastic

event. That failure to match expectations was reflected in a departure of δ from the expected value of 1.0. It was very interesting to discover that the δ parameter was not identical for the two subgenera (*Sophophora* δ = 0.720, *Drosophila* δ = 0.851). The departure from expectation was greater for *Sophophora* than for *Drosophila*. One possibility for this difference is that *Sophophora* have 4 chromosomes, while *Drosophila* have 6 (Painter and Stone 1935). That difference was not reflected by the mean genome size of the two subgenera (Chapter IV). It was however reflected in δ, and that may suggest that different number of structural elements, and in particular, differences in the number of centromeres and telomeres has had an influence on the direction of change and on the long term rate of that change. The parameters estimated for the heterochromatic portion of the DNA support this, as the location of change occurs late in the phylogeny for late-replicating heterochromatin (Chapter V).

The κ parameter may be the one most likely to detect adaptive change. κ detects change along individual branches of the phylogeny (Pagel 1999). These branches represent shorter periods of time. These shorter time periods are more likely to reflect the changes in habitat that might drive adaptation. Species that move and adapt to a new latitude or altitude may adapt to their new habitat by a change in body size as predicted by the Bergman's rule. While this change in body size may have a relation to genome size change (Hessen et al. 2013), the relationship is not guaranteed (Kelley et al. 2014). Geographical changes are only one possible way life history might change over evolutionary time. Another change in life history might be expected from a change in diet, something that has been commonly shown to promote speciation (Bush 1969; Feder

et al. 1988; Soria-Carrasco et al. 2014). In some cases, species may switch from a relatively constant food source to one that is ephemeral, suggesting an adaptation to a shortened generation time. One can imagine this change occurring in *Drosophila* that have shifted from a pile of rotting fruit to feed on cactus or mushrooms. Prickly pear cactus and mushrooms provide suitable habitat for *Drosophila* for only a short and highly variable period of time. The consequent need for fast development rate might well select for genome size change, a pattern found in plants (Van't Hof and Sparrow 1963; Bennett 1972; Bennett 1976). Cactophilic *repleta* group species have adapted via a very plastic rate of development. When the cactus pad begins to dry up, the stressed cactophilic *Drosophila* will emerge quickly as small individuals (Fanara and Werenkraut 2016). That adaptive change would have to occur over a short period of time, early in the evolution of the species and that would produce a reduced value of κ (<1.0).

A small value for κ values does not necessarily reflect a change in location or habitat, however (Nardon et al. 2005). It may also reflect events associated with the speciation event itself. Among these is small initial species population size. Here, genetic drift is more of a force on change than adaptation. Small populations have less effective selection because genetic drift, which forces stochastic change, will be predominant. The lack of selection and high pressure from genetic drift suggests genome size will be relatively unconstrained. As a new species adapts to its unique role in the tree of life, the species population size is expected to increase dramatically. Coincident with this, drift effects will be reduced, and selection against deleterious effects of genome size change will predominate. The rate of genome size change is no

longer dependent upon stochastic forces alone. Constrained by selection, the rate of genome size change will slow. That reduced rate would be reflected by the departure of the κ parameter from the expected value of 1.0. Two events, a change in population size or a change in location or habitat, could reduce κ. We must therefore ask, "What can we look for to best explain the observed value of κ?"

Speciation events associated with geographical change or a change of habitat could be unique events, but that is not typically so. This was certainly not the case in the adaptation of cactophilic *Drosophila*. The change of habitat rather resulted in a radiation of specialized species. Currently, there are more than 100 closely related cactophilic *repleta* species that develop and feed on rotting prickly pear cactus (O'Grady and Markow 2012). To the extent that many cactophilic *Drosophila* form a clade, and that this clade has adapted in part by a change in genome size, we would expect to see a departure from stochastic expectations. This departure would be reflected in κ, δ and to a smaller extent, λ. Events such as these may help us understand why different phylogenies might produce different parameter values. If we catch a clade in the early stages of species radiation, we can expect a very different set of parameters than if we sample over a much wider range of taxa, with species radiation events represented by only one individual.

Here, we examine not only the whole genome, but also the sex chromosomes and heterochromatin throughout the genome. Examination of Pagel parameters based on XY genome size difference strongly illustrates the strength of those parameters to detect departures from expectation. The evolution of the Y chromosome is driven by sexual

selection and as such does not fit a purely stochastic, gradualistic model. The Y does not evolve at a constant rate, nor does it evolve entirely at random (Chapters III and IV). A neo Y increases in size initially and then decreases rapidly in size as the sequence is silenced and degrades (Charlesworth and Charlesworth 2000; Bachtrog 2013). The parameters reflect exactly this outcome. There is partial phylogenetic signal ($\lambda < 1$) and temporally early change in branches ($0 < \kappa < 0.5$), supporting the rapid increase and decrease in size of neo-Y chromosomes (Chapters III and IV). While there was a significant relationship between genome size and the amount of replication, phylogenetic analyses of these replicated and unreplicated portions of the genome found reductions in phylogenetic signal (Chapter V). The biggest difference between the whole genome and the heterochromatic portion is the location of change: the high delta value suggests that the heterochromatic portion of the genome changes late in the phylogeny. This is potentially related to the radiation of species groups, suggesting a potential role in adaptation. This needs to be further explored with comparisons of underreplication to gene expression studies as well as comparisons of underreplication levels and potentially adaptive traits, such as dispersal distance and capabilities.

In conclusion, the comparative phylogenetic analysis of genome size in *Drosophila* has uncovered phylogenetic signal and gradualistic change across the genus. Different, yet supportive and insightful patterns, are found when separating components of the genome, such as sex chromosomes and heterochromatin. The patterns of change associated with the heterochromatic portions of the genome suggests a relationship to speciation of species groups, potentially associated with adaptation (Chapter V). These

changes, as well as the patterns in whole genome size, need to be compared to other variables for selective change. While temperature and latitude did not have a significant relationship to genome size change when analyzed phylogenetically (Hjelmen and Johnston 2017), there are countless variables that need to be tested to support or reject each hypotheses with more power.

Finally it is important to note that the results found in this dissertation may be exclusive to the *Drosophila* genus, and not all insects. The concept of the *Drosophila* genome being unique in regards to arthropods is explored and reviewed by Wurm (2015). The studies of the pattern of genome size evolution in different species do not always agree. While there has been found to be phylogenetic signal in genome size for *Drosophila* and fireflies, the patterns of change across the phylogeny differ between these two groups (Hjelmen and Johnston 2017; Lower et al. 2017). Even within the order Coleoptera, there are discrepancies between the presence of signal: signal in fireflies (Lower et al. 2017) , yet not in seed beetles (Arnqvist et al. 2015). Therefore, there may not be one clear answer for the patterns of genome size change across all organisms. It is just as likely that there will be combinations of these hypotheses resulting in the patterns and variation we see among closely related and distantly related organisms. It will be important to continue using these comparative methods to ask these questions across a wide range of organisms, large and small (Oliver et al. 2007).

REFERENCES

Abouheif E. 1999. A method for testing the assumption of phylogenetic independence in comparative data. *Evol Ecol Res* **1**: 895-909.

Adams MD Celniker SE Holt RA Evans CA Gocayne JD Amanatides PG Scherer SE Li PW Hoskins RA Galle RF et al. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185-2195.

Ågren JA, Wright SI. 2011. Co-evolution between transposable elements and their hosts: a major factor in genome size evolution? *Chromosome Res* **19**: 777-786.

Arnqvist G, Sayadi A, Immonen E, Hotzy C, Rankin D, Tuda M, Hjelmen CE, Johnston JS. 2015. Genome size correlates with reproductive fitness in seed beetles. In *Proc R Soc B*, Vol 282, p. 20151421. The Royal Society.

Aron S, de Menten L, Van Bockstaele DR, Blank SM, Roisin Y. 2005. When hymenopteran males reinvented diploidy. *Curr Biol* **15**: 824-827.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**: 113-124.

Bachtrog D, Hom E, Wong KM, Maside X, de Jong P. 2008. Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biol* **9**: R30.

Beaton MJ, Hebert PDN. 1988. Geographical parthenogenesis and polyploidy in *Daphnia pulex*. *Am Nat*: 837-845.

Bell G. 1989. A comparative method. *Am Nat*: 553-571.

Belyaeva E, Boldyreva L, Volkova E, Nanayev R, Alekseyenko A, Zhimulev I. 2003. Effect of the Suppressor of Underreplication (SuUR) gene on position-effect variegation silencing in *Drosophila melanogaster*. *Genetics* **165**: 1209-1220.

Belyaeva ES, Andreyeva EN, Belyakin SN, Volkova EI, Zhimulev IF. 2008. Intercalary heterochromatin in polytene chromosomes of *Drosophila melanogaster*. *Chromosoma* **117**: 411-418.

Belyaeva ES, Demakov SA, Pokholkova GV, Alekseyenko AA, Kolesnikova TD, Zhimulev IF. 2006. DNA underreplication in intercalary heterochromatin regions in polytene chromosomes of *Drosophila melanogaster* correlates with the formation of partial chromosomal aberrations and ectopic pairing. *Chromosoma* **115**: 355-366.

Belyaeva ES, Goncharov FP, Demakova OV, Kolesnikova TD, Boldyreva LV, Semeshin VF, Zhimulev IF. 2012. Late replication domains in polytene and non-polytene cells of *Drosophila melanogaster*. *PLoS One* **7**: e30035.

Belyaeva ES, Zhimulev IF, Volkova EI, Alekseyenko AA, Moshkin YM, Koryakov DE. 1998. Su(UR)ES: a gene suppressing DNA underreplication in intercalary and pericentric heterochromatin of *Drosophila melanogaster* polytene chromosomes. *Proc Natl Acad Sci U S A* **95**: 7532-7537.

Bennett MD. 1971. The Duration of Meiosis. *Proc R Soc Lond B* **178**: 277-299.

Bennett MD. 1972. Nuclear DNA content and minimum generation time in herbaceous plants. *Proc R Soc Lond B Biol Sci* **181**: 109-135.

Bennett MD. 1976. DNA amount, latitude, and crop plant distribution. *Environmental and Experimental Botany* **16**: 93-108.

Bennett MD, Leitch I. 2005. Genome size evolution in plants. *The evolution of the genome*: 89-162.

Bennetzen JL, Kellogg EA. 1997. Do plants have a one-way ticket to genomic obesity? *Plant Cell* **9**: 1509.

Black WCt, Rai KS. 1988. Genome evolution in mosquitoes: intraspecific and interspecific variation in repetitive DNA amounts and organization. *Genet Res* **51**: 185-196.

Blomberg SP, Garland T. 2002. Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *J Evolution Biol* **15**: 899-910.

Blomberg SP, Garland T, Ives AR. 2003. Testing for phylogenetic signal in comparative data: behavioral traits are more labile. *Evolution* **57**: 717-745.

Blumenstiel JP. 2011. Evolutionary dynamics of transposable elements in a small RNA world. *Trends in Genetics* **27**: 23-31.

Bosco G, Campbell P, Leiva-Neto JT, Markow TA. 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.

Brochmann C, Brysting A, Alsos I, Borgen L, Grundt H, Scheen AC, Elven R. 2004. Polyploidy in arctic plants. *Biol J Linn Soc* **82**: 521-536.

Brown JD, O'Neill RJ. 2010. Chromosomes, conflict, and epigenetics: chromosomal speciation revisited. *Annual review of genomics and human genetics* **11**: 291-316.

Bush GL. 1969. Sympatric host race formation and speciation in frugivorous flies of the genus Rhagoletis (Diptera, Tephritidae). *Evolution*: 237-251.

Cavalier-Smith T. 1982. Skeletal DNA and the evolution of genome size. *Annu Rev Biophys Bioeng* **11**: 273-302.

Charlesworth B, Barton N. 2004. Genome size: does bigger mean worse? *Curr Biol* **14**: R233-R235.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**: 1563-1572.

Charlesworth B, Sniegowski P, Stephan L. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **371**: 215-220.

Charlesworth D, Charlesworth B, Marais G. 2005. Steps in the evolution of heteromorphic sex chromosomes. *Heredity* **95**: 118-128.

Cheverud JM, Dow MM, Leutenegger W. 1985. The quantitative assessment of phylogenetic constraints in comparative analyses: sexual dimorphism in body weight among primates. *Evolution*: 1335-1351.

Clutton-Brock TH, Harvey PH. 1979. Comparison and adaptation. *Proc R Soc Lond B Biol Sci* **205**: 547-565.

Clutton-Brock TH, Harvey PH. 1977. Primate ecology and social organization. *J Zool* **183**: 1-39.

Consortium *D*G. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203-218.

d'Amato F. 1984. Role of polyploidy in reproductive organs and tissues. In *Embryology of angiosperms*, pp. 519-566. Springer.

Da Lage JL, Kergoat G, Maczkowiak F, Silvain JF, Cariou ML, Lachaise D. 2007. A phylogeny of Drosophilidae using the Amyrel gene: questioning the *Drosophila melanogaster* species group boundaries. *Journal of Zoological Systematics and Evolutionary Research* **45**: 47-63.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**: 772.

Dawkins R. 1976. The Selfish Gene. *NewYorkCity: OxfordUniversityPress*.

Ellis LL, Huang W, Quinn AM, Ahuja A, Alfrejd B, Gomez FE, Hjelmen CE, Moore KL, Mackay TFC, Johnston JS et al. 2014. Intrapopulation genome size in *D. melanogaster* reflects life history variation and plasticity. *PLOS Genetics* **10**.

Escudero M, Maguilla E, Loureiro J, Castro M, Castro S, Luceño M. 2015. Genome size stability despite high chromosome number variation in Carex gr. laevigata. *American journal of botany* **102**: 233-238.

Fanara JJ, Werenkraut V. 2016. Phenotypic plasticity in *Drosophila* cactophilic species: the effect of competition, density, and breeding sites. *Insect science*.

Feder JL, Chilcote CA, Bush GL. 1988. Genetic differentiation between sympatric host races of the apple maggot fly Rhagoletis pomonella. *Nature* **336**.

Felsenstein J. 1985. Phylogenies and the comparative method. *Am Nat*: 1-15.

Fleischmann A, Michael TP, Rivadavia F, Sousa A, Wang W, Temsch EM, Greilhuber J, Müller KF, Heubl G. 2014. Evolution of genome size and chromosome number in the carnivorous plant genus *Genlisea* (Lentibulariaceae), with a new estimate of the minimum genome size in angiosperms. *Annals of botany* **114**: 1651-1663.

Freckleton RP, Harvey PH, Pagel M. 2002. Phylogenetic analysis and comparative data: a test and review of evidence. *Am Nat* **160**: 712-726.

Garland T, Dickerman AW, Janis CM, Jones JA. 1993. Phylogenetic analysis of covariance by computer simulation. *Syst Biol* **42**: 265-292.

Goodfellow P, Banting G, Sheer D, Ropers H, Caine A, Ferguson-Smith M, Povey S, Voss R. 1983. Genetic evidence that a Y-linked gene in man is homologous to a gene on the X chromosome.

Gorelick R, Fraser D, Zonneveld BJ, Little DP. 2014. Cycad (Cycadales) chromosome numbers are not correlated with genome size. *International Journal of Plant Sciences* **175**: 986-997.

Gregory TR. 2001. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev* **76**: 65-101.

Gregory TR. 2003a. The C-value enigma in plants and animals: A review of parallels and and appeal for partnership. *Annals of Botany* **95**: 133-146.

Gregory TR. 2003b. Is small indel bias a determinant of genome size? *Trends in Genetics* **19**: 485-488.

Gregory TR. 2004. Insertion-deletion biases and the evolution of genome size. *Genet Res* **324**: 15-34.

Gregory TR. 2005a. Animal Genome Size Database.

Gregory TR. 2005b. Genome size evolution in animals. In *The Evolution of the Genome*, (ed. TR Gregory), pp. 3-87. Elsevier, San Diego.

Gregory TR, Hebert PD. 1999. The modulation of DNA content: proximate causes and ultimate consequences. *Genome Res* **9**: 317-324.

Gregory TR, Johnston JS. 2008. Genome size diversity in the family Drosophilidae. *Heredity* **101**: 228-238.

Grosset L, Odartchenko N. 1975. Duration of mitosis and separate mitotic phases compared to nuclear DNA content in erythroblasts of four vertebrates. *Cell Tissue Kinet* **8**: 91-96.

Hammond MP, Laird CD. 1985. Control of DNA replication and spatial distribution of defined DNA sequences in salivary gland cells of *Drosophila melanogaster*. *Chromosoma* **91**: 279-286.

Hanrahan SJ, Johnston JS. 2011. New genome size estimates of 134 species of arthropods. *Chromosome Research* **19**: 809-823.

Hare EE, Johnston JS. 2011. Genome Size Determination Using Flow Cytometry of Propidium Iodide-Stained Nuclei. In *Molecular Methods for Evolutionary Genetics*, Vol 772 (ed. V Orgogozo, MV Rockman), pp. 3-12. Humana Press.

Hartl DL. 2000. Molecular melodies in high and low C. *Nature Reviews Genetics* **1**: 145-149.

Harvey PH, Clutton-Brock TH. 1985. Life history variation in primates. *Evolution*: 559-581.

Harvey PH, Pagel MD. 1991. *The comparative method in evolutionary biology*. Oxford university press Oxford.

Hessen DO, Daufresne M, Leinaas HP. 2013. Temperature-size relations from the cellular-genomic perspective. *Biol Rev Camb Philos Soc* **88**: 476-489.

Hinegardner R, Rosen DE. 1972. Cellular DNA content and the evolution of teleostean fishes. *The American Naturalist* **106**: 621-644.

Hjelmen CE, Johnston JS. 2017. The mode and tempo of genome size evolution in the subgenus *Sophophora*. *PloS one* **12**: e0173505.

Huang W, Massouras A, Inoue Y, Peiffer J, Ramia M, Tarone AM, Turlapati L, Zichner T, Zhu D, Lyman RF et al. 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res* **24**: 1193-1208.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**: 754-755.

Jeffery NW, Hultgren K, Chak STC, Gregory TR, Rubenstein DR. 2016. Patterns of genome size variation in snapping shrimp. *Genome* **59**: 393-402.

Jockusch EL. 1997. An evolutionary correlate of genome size change in plethodontid salamanders. *Proc R Soc Lond B Biol Sci* **264**: 597-604.

Johnson AW, Packer JG, Reese G. 1965. *Polyploidy, distribution, and environment*. na.

Johnston JS, Schoener M, McMahon DP. 2013. DNA underreplication in the majority of nuclei in the *Drosophila melanogaster* thorax: evidence from Suur and flow cytometry. *J Mol Biol Res* **3**: p47.

Jombart T, Dray S. 2008. Adephylo: exploratory analyses for the phylogenetic comparative method. URL: ht tp. *cran r-project org/web/packages/adephylo/index html*.

Kapusta A, Suh A, Feschotte C. 2017. Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*: 201616702.

Kellermann V, Loeschcke V, Hoffmann AA, Kristensen TN, Flojgaard C, David JR, Svenning JC, Overgaard J. 2012a. Phylogenetic constraints in key functional traits behind species' climate niches: patterns of desiccation and cold resistance across 95 *Drosophila* species. *Evolution* **66**: 3377-3389.

Kellermann V, Overgaard J, Hoffmann AA, Flojgaard C, Svenning JC, Loeschcke V. 2012b. Upper thermal limits of *Drosophila* are linked to species distributions and strongly constrained phylogenetically. *Proc Natl Acad Sci U S A* **109**: 16228-16233.

Kelley JL, Peyton JT, Fiston-Lavier AS, Teets NM, Yee MC, Johnston JS, Bustamante CD, Lee RE, Denlinger DL. 2014. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat Commun* **5**: 4611.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**: 1463-1464.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**: 49-63.

Kitano J, Ross JA, Mori S, Kume M, Jones FC, Chan YF, Absher DM, Grimwood J, Schmutz J, Myers RM et al. 2009. A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**: 1079-1083.

Koonin EV, Wolf YI. 2010. Constraints and plasticity in genome and molecular-phenome evolution. *Nature Reviews Genetics* **11**: 487-498.

Kraaijeveld K. 2010. Genome size and species diversification. *Evolutionary biology* **37**: 227-233.

Lakhotia S. 1984. Replication in *Drosophila* chromosomes. XII. Reconfirmation of underreplication of heterochromatin in polytene nuclei by cytofluorometry. *Chromosoma* **89**: 63-67.

Lang F, Waldegger S. 1997. Regulating Cell Volume: Maintaining a consistent cellular volume may contribute to a cell's metabolism as well as to its shape. *Am Sci* **85**: 456-463.

Lang M, Polihronakis Richmond M, Acurio AE, Markow TA, Orgogozo V. 2014. Radiation of the *Drosophila nannoptera* species group in Mexico. *Journal of Evolutionary Biology* **27**: 575-584.

Leach TJ, Chotkowski HL, Wotring MG, Dilwith RL, Glaser RL. 2000. Replication of heterochromatin and structure of polytene chromosomes. *Mol Cell Biol* **20**: 6308-6316.

Leitch AR, Leitch IJ. 2008. Genomic plasticity and the diversity of polyploid plants. *Science* **320**: 481-483.

Lilly MA, Duronio RJ. 2005. New insights into cell cycle control from the *Drosophila* endocycle. *Oncogene* **24**: 2765-2775.

Lima-de-Faria A, Jaworska H. 1968. Late DNA synthesis in heterochromatin.

Lower SS, Spencer Johnston J, Stanger-Hall K, Hjelmen CE, Hanrahan SJ, Korunes K, Hall D. 2017. Genome size in North American fireflies: Substantial variation likely driven by neutral processes. *Genome biology and evolution*.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. *Evolution*: 1065-1080.

Lynch M. 2007a. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proceedings of the National Academy of Sciences of the United States of America* **104**: 8597-8604.

Lynch M. 2007b. *The Origins of Genome Architecture*. Sinauer Associates Inc.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* **302**: 1401-1404.

MacAlpine DM, Rodriguez HK, Bell SP. 2004. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev* **18**: 3094-3105.

Makunin IV, Kolesnikova TD, Andreyenkova NG. 2014. Underreplicated regions in *Drosophila melanogaster* are enriched with fast-evolving genes and highly conserved noncoding sequences. *Genome Biol Evol* **6**: 2050-2060.

Makunin IV, Volkova EI, Belyaeva ES, Nabirochkina EN, Pirrotta V, Zhimulev IF. 2002. The *Drosophila* suppressor of underreplication protein binds to late-replicating regions of polytene chromosomes. *Genetics* **160**: 1023-1034.

Matsunaga S. 2009. Junk DNA promotes sex chromosome evolution. *Heredity* **102**: 525-526.

McNairn AJ, Gilbert DM. 2003. Epigenomic replication: linking epigenetics to DNA replication. *Bioessays* **25**: 647-656.

Mirskey AE, Ris H. 1951. The desoxyribonucleic acid content of animal cells and its evolutionary significance. *J Gen Physiol* **34**: 451-462.

Moens P, Rees DJ, Belzile C, Glémet H, Dufresne F. 2008. Large genomes among caridean shrimp. *Genome* **51**: 159-163.

Moran PA. 1950. Notes on continuous stochastic phenomena. *Biometrika* **37**: 17-23.

Muller HJ. 1918. Genetic Variability, Twin Hybrids and Constant Hybrids, in a Case of Balanced Lethal Factors. *Genetics* **3**: 422-499.

Münkemüller T, Lavergne S, Bzeznik B, Dray S, Jombart T, Schiffers K, Thuiller W. 2012. How to measure and test phylogenetic signal. *Methods Ecol Evol* **3**: 743-756.

Nardon C, Deceliere G, Loevenbruck C, Weiss M, Vieira C, Biémont C. 2005. Is genome size influenced by colonization of new environments in dipteran species? *Molecular ecology* **14**: 869-878.

O'Grady PM, Markow TA. 2012. Rapid morphological, ecological and behavioral evolution in *Drosophila*: comparisons between the cactophilic *repleta* species

group and the endemic Hawaiian *Drosophila*. Rapidly evolving genes & genetic systems. New York, NY: Oxford University Press.

Obbard DJ, Maclennan J, Kim K-W, Rambaut A, O'Grady PM, Jiggins FM. 2012. Estimating divergence dates and substitution rates in the *Drosophila* phylogeny. *Molecular Biology and Evolution*: mss150.

Ohno S. 1972. So much "junk" DNA in our genome. In *Brookhaven Symp Biol*, Vol 23, pp. 366-370.

Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM. 2007. The mode and tempo of genome size evolution in eukaryotes. *Genome Res* **17**: 594-601.

Orme D. 2013. The caper package: comparative analysis of phylogenetics and evolution in R. *R package version* **5**.

Pagel M. 1999. Inferring the historical patterns of biological evolution. *Nature* **401**: 877-884.

Pagel M, Harvey P. 1989. Comparative methods for examining adaptation depend on evolutionary models. *Folia Primatologica* **53**: 203-220.

Pagel M, Johnstone RA. 1992. Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proc Biol Sci* **249**: 119-124.

Painter TS, Reindorp EC. 1939. Endomitosis in the nurse cells of the ovary of *Drosophila melanogaster*. *Zeitschrift für Zellforschung und Mikroskopische Anatomie Abt B Chromosoma* **1**: 276-283.

Painter TS, Stone W. 1935. Chromosome fusion and speciation in Drosophilae. *Genetics* **20**: 327.

Palazzo AF, Gregory TR. 2014. The case for junk DNA. *PLoS Genet* **10**: e1004351.

Pandit MK, White SM, Pocock MJ. 2014. The contrasting effects of genome size, chromosome number and ploidy level on plant invasiveness: a global analysis. *New Phytologist* **203**: 697-703.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289-290.

Pavoine S, Ollier S, Pontier D, Chessel D. 2008. Testing for phylogenetic signal in phenotypic traits: new matrices of phylogenetic proximities. *Theor Popul Biol* **73**: 79-91.

Perdix-Gillot S. 1979. DNA synthesis and endomitosis in the giant nuclei of the silkgland of *Bombyx mori*. *Biochemie* **61**: 171-204.

Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet* **17**: 23-28.

Petrov DA. 2002a. DNA loss and evolution of genome size in *Drosophila*. *Genetica* **115**: 81-91.

Petrov DA. 2002b. Mutational equilibrium model of genome size evolution. *Theor Popul Biol* **61**: 531-544.

Petrov DA, Hartl DL. 1998. High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups. *Mol Biol Evol* **15**: 293-302.

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* **287**: 1060-1062.

Powell JR. 1997. Progress and prospects in evolutionary biology: the *Drosophila* model (Oxford series in ecology & evolution).

Pritchard DK, Schubiger G. 1996. Activation of transcription in *Drosophila* embryos is a gradual process mediated by the nucleocytoplasmic ratio. *Genes Dev* **10**: 1131-1142.

Rangel J, Strauss K, Seedorf K, Hjelmen C, Johnston J. 2015. Endopolyploidy changes with age-related polyethism in the honey bee. *Apis mellifera PloS one* **10**.

Revell LJ. 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217-223.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572-1574.

Rudkin GT. 1969. Non replicating DNA in *Drosophila*. *Genetics* **61**: Suppl:227-238.

Russo C, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Molecular biology and evolution* **12**: 391-404.

Schielzeth H, Streitner C, Lampe U, Franzke A, Reinhold K. 2014. Genome size variation affects song attractiveness in grasshoppers: evidence for sexual selection against large genomes. *Evolution* **68**: 3629-3635.

Scholes DR, Suarez AV, Paige KN. 2013. Can endopolyploidy explain body size variation within and between castes in ants? *Ecol Evol* **3**: 2128-2137.

Schübeler D, Scalzo D, Kooperberg C, van Steensel B, Delrow J, Groudine M. 2002. Genome-wide DNA replication profile for *Drosophila melanogaster*: a link between transcription and replication timing. *Nat Genet* **32**: 438-442.

Schulze SR, McAllister BF, Sinclair DAR, Fitzpatrick KA, Marchetti M, Pimpinelli S, Honda BM. 2006. Heterochromatic Genes in *Drosophila*: A Comparative Analysis of Two Genes. *Genetics* **173**: 1433-1445.

Schwaiger M, Schübeler D. 2006. A question of timing: emerging links between transcription and replication. *Curr Opin Genet Dev* **16**: 177-183.

Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila* genome? *PLoS Genet* **5**: e1000495.

Sessegolo C, Burlet N, Haudry A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters* **12**.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T. 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**: 825-837.

Śliwińska EB, Martyka R, Tryjanowski P. 2016. Evolutionary interaction between W/Y chromosome and transposable elements. *Genetica* **144**: 267-278.

Smith HM. 1925. Cell size and metabolic activity in Amphibia. *Biol Bull*: 347-378.

Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA, Feder JL, Bast J, Schwander T et al. 2014. Stick insect genomes reveal natural selection's role in parallel speciation. *Science* **344**: 738-742.

Specchia V, Piacentini L, Tritto P, Fanti L, D'Alessandro R, Palumbo G, Pimpinelli S, Bozzetti MP. 2010. Hsp90 prevents phenotypic variation by suppressing the mutagenic activity of transposons. *Nature* **463**: 662-665.

Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. *Ann Mo Bot Gard*: 824-832.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Molecular biology and evolution* **21**: 36-44.

Team RC. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015, URL h ttp. *www R-project org*.

Traut W, Rees DJ, Dufresne F, Glémet H, Belzile C. 2007. Amphipod genome sizes: first estimates for Arctic species reveal genomic giants. *Genome* **50**: 151-158.

Van't Hof J, Sparrow AH. 1963. A relationship between DNA content, nuclear volume, and minimum mitotic cycle time. *P Natl Acad Sci USA* **49**: 897.

van der Linde K, Houle D. 2008. A supertree analysis and literature review of the genus *Drosophila* and closely related genera (Diptera, Drosophilidae). *Insect Systematics & Evolution* **39**: 241-267.

van der Linde K, Houle D, Spicer GS, Steppan SJ. 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet Res* **92**: 25-38.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The Sequence of the Human Genome. *Science* **291**: 1304-1351.

Vieira C, Biemont C. 2004. Transposable element dynamics in two sibling species: *Drosophila melanogaster* and *Drosophila simulans*. *Genetica* **120**: 115-123.

Vinogradov AE. 1995. Nucleotypic effect in homeotherms: body-mass-corrected basal metabolic rate of mammals is related to genome size. *Evolution*: 1249-1259.

Vinogradov AE. 1997. Nucleotypic effect in homeotherms: body-mass independent resting metabolic rate of passerine birds is related to genome size. *Evolution*: 220-225.

Vinogradov AE. 2004. Evolution of genome size: multilevel selection, mutation bias or dynamical chaos? *Curr Opin Genet Dev* **14**: 620-626.

Volkova EI, Yurlova AA, Kolesnikova TD, Makunin IV, Zhimulev IF. 2003. Ectopic expression of the Suppressor of Underreplication gene inhibits endocycles but not the mitotic cell cycle in *Drosophila melanogaster*. *Mol Genet Genomics* **270**: 387-393.

Von Neumann J, Kent RH, Bellinson HR, Hart BI. 1941. The mean square successive difference. *Ann Math Stat* **12**: 153-162.

Walbot V, Petrov DA. 2001. Gene galaxies in the maize genome. *Proc Natl Acad Sci U S A* **98**: 8163-8164.

Waldegger S, Lang F. 1998. Cell volume and gene expression. *J Membr Biol* **162**: 95-100.

Ward BL, Heed WB. 1970. Chromosome phylogeny of *Drosophila pachea* and related species. *Journal of Heredity* **61**: 248-258.

Whitney KD, Boussau B, Baack EJ, Garland Jr T. 2011. Drift and genome complexity revisited. *PLoS genetics* **7**: e1002092.

Whitney KD, Garland Jr T. 2010. Did genetic drift drive increases in genome complexity? *PLoS Genetics* **6**: e1001080.

Wolfe KH, Sharp PM, Li WH. 1989. Mutation rates differ among regions of the mammalian genome.

Wright SI, Andolfatto P. 2008. The impact of natural selection on the genome: emerging patterns in *Drosophila* and *Arabidopsis*. *Annual Review of Ecology, Evolution, and Systematics* **39**: 193-213.

Wurm Y. 2015. Arthropod genomics beyond fruit flies: bridging the gap between proximate and ultimate causation. *Briefings in Functional Genomics* **14**: 381-383.

Xia X. 1995. Body temperature, rate of biosynthesis, and evolution of genome size. *Mol Biol Evol* **12**: 834-842.

Yi S, Streelman JT. 2005. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends in Genetics* **21**: 643-646.

Young LC. 1941. On randomness in ordered sequences. *Ann Math Stat* **12**: 293-300.

Yurlova AA, Makunin IV, Zhimulev IF. 2010. Phylogenetic analysis of the fast-evolving SuUR gene in insects. *Russ J Genet* **46**: 1127-1129.

Zhimulev IF, Belyaeva ES. 2003. Intercalary heterochromatin and genetic silencing. *BioEssays* **25**: 1040-1051.

Zhimulev IF, Belyaeva ES, Makunin IV, Pirrotta V, Volkova EI, Alekseyenko AA, Andreyeva EN, Makarevich GF, Boldyreva LV, Nanayev RA et al. 2003a. Influence of the SuUR gene on intercalary heterochromatin in *Drosophila melanogaster* polytene chromosomes. *Chromosoma* **111**: 377-398.

Zhimulev IF, Belyaeva ES, Semeshin VF, Shloma VV, Makunin IV, Volkova EI. 2003b. Overexpression of the SuUR gene induces reversible modifications at pericentric, telomeric and intercalary heterochromatin of *Drosophila melanogaster* polytene chromosomes. *Journal of cell science* **116**: 169-176.

Zhou Q, Bachtrog D. 2015. Ancestral Chromatin Configuration Constrains Chromatin Evolution on Differentiating Sex Chromosomes in *Drosophila*. *PLOS Genetics* **11**: e1005331.

APPENDIX

Appendix 1: NCBI Accession numbers for genes used in phylogenetic analyses

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
C. amoena	HQ979021.1	AF478415.1	EU493831.1	EU494086.1	AF059887.1	AY736544.1	-	-	AY124490.1	X61687.1	-	-	-	-	-	-
C. procnemis	GU597449.1	HQ110556.1	GU597512.1	-	AB026521.1	-	-	-	-	AF021821.1	-	-	-	-	-	-
D. busckii	HG798345.1	AF519347.1	AF519379.1	-	AB261141.1	-	AF071037.1	-	AF293733.1	U39445.1	-	-	-	-	-	-
D. pseudotalamancana	KF632610.1	-	-	-	-	KF632685.1	KF632699.1	-	-	-	KF632674.1	-	KF632636.1	KF632624.1	KF632650.1	KF632664.1
D. bromeliae	KF632602.1	AF478418.1	-	-	AY081441.1	AY733049.1	KF632689.1	-	AF324966.1	-	KF632666.1	-	KF632627.1	KF632614.1	KF632640.1	KF632654.1
D. arawakana	HM006881.1	HM006889.1	-	-	AY695384.1	AF491630.1	AF071036.1	-	EU446060.1	-	-	EU447305.1	-	-	-	-
D. cardini	HM006874.1	HM006890.1	AF519351.1	EU494092.1	AY695386.1	AF462599.1	-	-	EU446063.1	-	-	EU447308.1	-	-	-	-
D. dunni	HM006882.1	HM006893.1	EU493838.1	EU494093.1	AY695382.1	-	-	-	HM006868.1	-	-	HM006872.1	-	-	-	-
D. polymorpha	HM006879.1	HM006892.1	-	-	-	AY736495.1	-	-	EU446098.1	-	-	EU447339.1	-	-	-	-
D. funebris	KF632605.1	EF469563.1	EU493840.1	EU494095.1	Y17611.1	AF335557.1	KF632692.1	-	AF293734.1	AF059878.1	KF632669.1	EU447312.1	KF632629.1	KF632617.1	KF632643.1	KF632657.1
D. guarani	EU493582.1	EU493712.1	EU493843.1	EU494098.1	AY081439.1	-	-	-	EU446091.1	-	-	EU450643.1	-	-	-	-
D. guttifera	GU597465.1	AF147119.1	GU597529.1	EU494119.1	AB261139.1	-	-	-	-	AF021826.1	-	-	-	-	-	-
D. albomicans	AY633553.1	-	-	EU494100.1	AY044126.1	AF462595.1	-	AF102153.1	-	-	-	-	-	-	-	-
D. hypocausta	EU493585.1	-	EU493845.1	EU494101.1	AY044131.1	AY733043.1	-	-	-	-	-	-	-	-	-	-
D. immigrans	HQ979108.1	EU493716.1	EU493846.1	EU494102.1	M97638.1	AF491632.1	-	L06337.1	AF293738.1	AF021825.1	-	EU447316.1	-	-	-	-
D. kepulauana	EU493587.1	EU493717.1	EU493847.1	EU494103.1	-	AY733044.1	-	AF102154.1	-	-	-	-	-	-	-	-
D. kohkoa	EU493588.1	EU493718.1	EU493848.1	EU494104.1	AY044124.1	-	-	AF102160.1	-	-	-	-	-	-	-	-
D. nasuta	GU597466.1	GU597498.1	EU493849.1	EU494105.1	AB261137.1	AY733059.1	U31135.1	AF102159.1	-	HQ326583.1	-	-	-	-	-	-
D. neohypocausta	EU493590.1	-	EU493850.1	EU494106.1	AB261134.1	-	-	-	-	-	-	-	-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. rubida	EU493593.1	EU493723.1	EU493853.1	EU494109.1	-	AY736502.1	-	-	-	-	-	-	-	-	-	-
D. euronotus	GU597454.1	EU390743.1	GU597516.1	-	AY750132.1	-	-	-	-	-	-	-	-	-	-	-
D. melanica	KJ463784.1	EU390749.1	EU493871.1	EU494125.1	AY750123.1	AY733056.1	-	-	AF324972.1	AF059880.1	-	-	-	-	-	-
D. micromelanica	AY750096.1	EU390751.1	EU493872.1	EU494126.1	AY750124.1	-	-	-	-	-	-	-	-	-	-	-
D. paramelanica	AY750099.1	EU493742.1	GU597515.1	EU494127.1	AY750127.1	-	-	-	-	-	-	-	-	-	-	-
D. acanthoptera	KF632601.1	EU493728.1	EU493858.1	-		KF632675.1	KF632687.1	-	-	-	JF736442.1	-	JF736382.1	KF632612.1	KF632638.1	KF632652.1
D. nannoptera	DQ471531.1	AF183971.1	AF183976.1	EU494114.1	DQ471666.1	KF632682.1	KF632696.1	-	AF324975.1	AY081433.1	JF736456.1	-	KF632633.1	KF632621.1	KF632647.1	KF632661.1
D. pachea	KF632609.1	KF632600.1	AF183975.1	-	-	KF632683.1	KF632697.1	-	JX067180.1	-	KF632672.1	-	KF632634.1	KF632622.1	KF632648.1	KF632662.1
D. pallidipennis	EF570011.1	AY162982.1	EU493860.1	EU494115.1	-	AY736487.1	-	-	EU446092.1	-	-	EU447332.1	-	-	-	-
D. polychaeta	DQ471555.1	EU493733.1	EU493863.1	EU494118.1	DQ471629.1	AY736494.1	-	-	AF324976.1	-	-	-	-	-	-	-
D. palustris	AF147112.1	AF147122.1	AF147132.1	-	-	AY736538.1	-	-	-	-	-	-	-	-	-	-
D. phalerata	AF147105.1	AF147115.1	AF147125.1	-	JF735884.1	AY736492.1	-	-	AF293745.1	-	-	-	-	-	-	-
D. subpalustris	AF147111.1	AF147121.1	AF147131.1	-	-	AY736539.1	-	-	-	-	-	-	-	-	-	-
D. anceps	DQ471598.1	JF736093.1	-	-	DQ471656.1	-	-	-	-	-	JF736443.1	-	JF736383.1	JF736260.1	JF736208.1	-
D. arizonae	DQ383684.1	DQ436072.1	-	-	AY154863.1	-	-	-	-	-	JF736487.1	-	JF736425.1	JF736304.1	EU341636.1	JF736194.1
D. buzzatii	KF632603.1	AF146169.1	AF146181.1	-	U65746.1	KF632677.1	KF632690.1	-	AF324980.1	-	KF632667.1	-	JF736384.1	KF632615.1	KF632641.1	KF632655.1
D. eohydei	DQ471601.1	JF736124.1	AF145893.1	-	DQ471659.1	-	-	-	AF324970.1	-	JF736490.1	-	JF736429.1	JF736308.1	JF736245.1	JF736197.1
D. hydei	DQ471603.1	AF145888.1	AF145892.1	-	DQ471660.1	AY733042.1	AF071035.1	-	AF293737.1	U37714	JF736448.1	EF559361.1	JF736388.1	JF736265.1	JF736212.1	JF736170.1
D. limensis	JF736082.1	JF736123.1	-	-	-	AY736479.1	AF071046.1	-	-	-	JF736488.1	-	JF736426.1	JF736305.1	JF736242.1	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. mayaguana	DQ202067.1	DQ202027.1	-	-	M62742.1	-	-	-	-	-	JF736451.1	AY155063.1	JF736392.1	JF736269.1	EU341634.1	JF736172.1
D. mercatorum	HG798344.1	EU493737.1	EU493867.1	EU494121.1	DQ471664.1	-	U31133.1	-	AF324973.1	-	JF736484.1	-	JF736422.1	JF736301.1	JF736239.1	JF736191.1
D. mojavensis	DQ383730.1	AY437281.1	EU493868.1	EU494122.1	AY364493.1?	-	AF071041.1	FJ267300.1	-	-	JF736454.1	-	AY437368.1	JF736272.1	AY437326.1	JF736173.1
D. mulleri	DQ437708.1	AY437294.1	DQ437714.1	-	U76483.1	-	-	-	AF324974.1	-	JF736455.1	AY155113.1	JF736396.1	JF736273.1	EU341638.1	JF736174.1
D. navojoa	EU493609.1	DQ436089.1	EU493869.1	EU494123.1	X15585.1	-	-	-	-	-	JF736477.1	-	JF736415.1	JF736294.1	EU341635.1	JF736189.1
D. repleta	DQ471533.1	GQ376044.1	-	-	DQ471667.1	AY736496.1	AF048776.1	-	AF324977.1	-	JF736463.1	-	JF736403.1	JF736281.1	EU341630.1	-
D. lacertosa	AY750095.1	EU493740.1	-	EU494124.1	DQ471652.1	-	-	-	-	-	-	-	-	-	-	-
D. robusta	EU390726.1	GQ244457.1	GU597514.1	EU494128.1	AY750138.1	KF632684.1	KF632698.1	L063401	AF293747.1	AF059883.1	KF632673.1	-	KF632635.1	KF632623.1	KF632649.1	KF632663.1
D. sordidula	EU493580.1	EU390759.1	EU493841.1	EU494096.1	AY750139.1	-	-	-	AF324979.1	-	-	-	-	-	-	-
D. mediopictoides	EU493617.1	EU493746.1	-	EU494131.1	-	AY733055.1	-	-	EU446080.1	-	-	EU447323.1	-	-	-	-
D. tripunctata	EU493619.1	EU493748.1	AF519375.1	EU494133.1	AY081440.1	-	-	-	AF324964.1	-	-	EU447342.1	-	-	-	-
D. borealis	KJ463779.1	GQ244452.1	-	-	U26839.1	-	-	-	-	-	-	-	-	-	-	-
D. ezoana	DQ471600.1	-	-	-	DQ471658.1	-	-	AY854880.1	-	-	-	EF635087.1	-	-	-	-
D. flavomontana	JF735925.1	JF735932.1	-	-	U26838.1	-	-	-	-	-	-	-	-	-	-	-
D. kanekoi	DQ471604.1	GQ244455.1	-	-	JF710263.1	AY736535.1	-	-	-	-	-	-	-	-	-	-
D. lacicola	JF735927.1	JF735933.1	-	-	U26841.1	-	-	-	-	-	-	-	-	-	-	-
D. littoralis	JN019910.1	NC_011596.1	NC_011596.1	NC_011596.1	DQ471662.1	AY733045.1	AF071047.1	EF635096.1	-	-	-	EF635088.1	-	-	-	-
D. lummei	DQ471606.1	AY646746.1	-	AY646770.1	U26843.1	AY733046.1	-	-	-	-	-	-	-	-	-	-
D. montana	DQ426756.1	DQ426799.1	EU493880.1	-	DQ471665.1	-	-	EF635097.1	-	-	-	EF635089.1	-	-	-	-
D. novamexicana	JF735929.1	JF735934.1	-	AY646767.1	AY165542.1	AY736484.1	-	-	AY165545.1	-	-	-	-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. virilis	JQ679111.1	HQ110559.1	-	AY646772.1	DQ471668.1	AF136603.1	U02029.1	X13877.1	AF293749.1	X13831.1	JF736479.1	EF635085.1	JF736417.1	JF736296.1	JF736234.1	XM_002058672.1
D. ananassae	AB032132.1	AF474077.1	BK006336.1	EU601718.1	AB194426.1	AF024691.3	AB003755.1	L06335.1	HQ631615.1	-	-	HQ631746.1	-	-	-	-
D. atripex	FJ795575.1	-	-	EU601556.1	-	U96154.3	-	-	-	-	-		-	-	-	-
D. auraria	AB669696.1	AB243377.1	EU493883.1	EU494137.1	AB669826.1	U96163.2	AB048693.1	-	HQ631616.1	-	-	AF461338.1	-	-	-	-
D. baimaii	AB669751.1	AB669808.1	EU493884.1	EU494138.1	AB669881.1	HQ631506.1	AB048800.1	-	HQ631617.1	-	-	HQ631748.1	-	-	-	-
D. bipectinata	AB032131.1	AY757275.1	-	-	DQ363233.1	AF136936.1	U31145.1	-	DQ363227.1	AJ844862.1	-	HQ631752.1	-	-	-	-
D. birchii	AB669745.1	AB669803.1	-	-	AB669876.1	HQ631507.1	AB048804.1	-	HQ631620.1	-	-	AF461340.1	-	-	-	-
D. bunnanda	AB669747.1	GQ376040.1	-	-	AB669877.1	-	-	-	-	-	-		-	-	-	-
D. elegans	AB032130.1	HQ631610.1	-	-	DQ363230.1	AF136930.2	AY098451.1	-	DQ363224.1	-	-	HQ631756.1	-	-	-	-
D. ercepeae	FJ795576.1	AF461306.1	-	-	AF459784.1	U96155.2	U31137.1	-	-	-	-	AF461354.1	-	-	-	-
D. erecta	JQ679121.1	GQ244453.1	-	-	X54116.1	AF039562.2	M55996.1	AF251239.1	-	AF127156.1	-		-	-	-	-
D. eugracilis	AY098461.1	AF474079.1	-	-	AY279326.1	AF250055.1	AB003764.1	-	DQ363225.1	-	-	HQ631757.1	-	-	-	-
D. ficusphila	AY757285.1	AY757273.1	-	-	DQ363232.1	AF462600.1	AB003765.1	-	DQ363226.1	-	-	HQ631758.1	-	-	-	-
D. fuyamai	AY098460.1	HQ631608.1	-	-	HQ631474.1	HQ631511.1	AB003768.1	-	HQ631624.1	-	-	HQ631759.1	-	-	-	-
D. kanapiae	AB669755.1	AB669812.1	-	-	AB669885.1	-	AB048768.1	-	-	-	-	AF461349.1	-	-	-	-
D. lacteicornis	AB830535.1	AB243388.1	-	-	AB830534.1	-	AB048769.1	-	-	-	-		-	-	-	-
D. lucipennis	HQ631564.1	HQ631607.1	-	-	AF459751.1	AF251138.1	AY098452.1	-	HQ631629.1	-	-	HQ631764.1	-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. lutescens	AB027267.1	AF474090.1	-	-	AY279327.1	AF491637.1	HQ631491.1	-	HQ631630.1	-	-	AF461318.1	-	-	-	-
D. malerkotliana	AY757289.1	EU493756.1	EU493886.1	EU494140.1	AB194422.1	AY733053.1	HQ631504.1	-	HQ631632.1	AJ844839.1	-	AF461326.1	-	-	-	-
D. mauritiana	M57907.1	AF474081.1	NC_005779.1	NC_005779.1	AF284479.1	JF815750.1	D17730.1	AF251240.1	-	AF127158.1	-		-	-	-	-
D. mayri	AB669757.1	AB669814.1	-	-	AB669887.1	HQ631514.1	AB048771.1	-	HQ631633.1	-	-	AF461350.1	-	-	-	-
D. melanogaster	M57910.1	EU493757.1	GQ229522.1	NC_024511.2	U20765.1	NM_057914.4	L22716.1	L22716.1	HQ631634.1	X17332.1	-	HQ631766.1	-	-	-	J03177.1
D. mimetica	AY098454.1	AF474092.1	-	-	AY279328.1	AY733058.1	AY098447.1	-	-	-	-	AF461320.1	-	-	-	-
D. nikananu	AB669744.1	AB669801.1	-	-	AB669874.1	AF251136.1	AB048775.1	-	-	-	-	AF461337.1	-	-	-	-
D. orena	AY757281.1	AY757269.1	-	-	-	U96158.2	D21129.1	AF251242.1	-	AF127155.1	-		-	-	-	-
D. orosa	AB669758.1	AB669815.1	-	-	AB669888.1	-	AB048776.1	-	-	-	-	AF461351.1	-	-	-	-
D. pallidosa	FJ795561.1	AF461280.1	-	EU601680.1	AF459759.1	AF136931.1	-	-	-	-	-	AF461327.1	-	-	-	-
D. parabipectinata	AY757282.1	AY757270.1	-	-	HQ631481.1	AY736489.1	HQ631503.1	-	HQ631637.1	-	-	HQ631769.1	-	-	-	-
D. paralutea	AY757290.1	AY757278.1	-	-	AF459747.1	-	-	-	-	-	-	AF461315.1	-	-	-	-
D. pectinifera	AB669761.1	AB243390.1	-	-	AB669891.1	-	-	-	-	-	-		-	-	-	-
D. phaeopleura	AY757291.1	AY757279.1	-	EU601555.1	AF459756.1	AY736491.1	-	-	-	-	-	AF461324.1	-	-	-	-
D. prostipennis	HQ631570.1	AF474091.1	-	-	HQ631483.1	HQ631519.1	HQ631494.1	-	HQ631641.1	-	-	HQ631773.1	-	-	-	-
D. pseudoananassae nigrens	AJ844809.1	-	-	-	AB194424.1	AY736497.1	-	-		-	-		-	-	-	-
D. pseudoananassae	AY757280.1	GQ376041.1	-	-	AB194425.1	AY736498.1	AB003774.1	-	HQ631639.1	-	-	HQ631771.1	-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. pseudotakahashii	AY757283.1	AY757271.1	-	-	-	AY736499.1	-	-		-	-		-	-	-	-
D. punjabiensis	AB669723.1	AB243401.1	-	-	AB669854.1	U96165.1	AB048772.1	-	HQ631643.1	-	-	HQ631775.1	-	-	-	-
D. quadraria	AB669703.1	AB243392.1	-	-	AB669833.1	AF136934.1	AB048698.1	-	HQ631644.1	-	-	AF461343.1	-	-	-	-
D. rufa	AB669704.1	AB243394.1	-	-	AB669834.1	AF136935.2	AB048779.1	-	HQ631645.1	-	-	AF461353.1	-	-	-	-
D. santomea	JQ679120.1	DQ382822.1	NC_023825.1	NC_023825.1	AY804554.1	AY736503.1	AF280885.1	AF251248.1	-	AY804512.1	-	AY804682.1	-	-	-	-
D. sechellia	M57908.1	GQ244458.1	NC_005780.1	NC_005780.1	AF284481.1	AF039558.1	D17732.1	AF251249.1	-	AF127157.1	-	XM_002031972.1	-	-	-	-
D. seguyi	AB669716.1	AB669775.1	-	-	AB669846.1	HQ631522.1	AB048780.1	-	HQ631646.1	-	-	AF461335.1	-	-	-	-
D. serrata	AB669748.1	AB669806.1	-	-	AB669879.1	AF069756.1	AB048781.1	L06341.1	HQ631647.1	-	-		-	-	-	-
D. simulans	M57911.1	GQ222022.1	NC_005781.1	NC_005781.1	M365811.1	XM_002081767.1	D17734.1	AF251250.1	AY197770.1	X15685.1	-	HQ631776.1	-	-	-	-
D. suzukii	AB032128.1	HQ631606.1	-	-	-	HQ631524.1	AY098448.1	-	HQ631649.1	-	-	HQ631778.1	-	-	-	-
D. takahashii	JF735911.1	AF474089.1	-	-	AF459749.1	U96161.2	U31148.1	-	HQ631650.1	-	-	HQ631779.1	-	-	-	-
D. tani	AB669709.1	AB669770.1	-	-	AB669839.1	HQ631525.1	AB048700.1	-	HQ631651.1	-	-	HQ631780.1	-	-	-	-
D. teissieri	JQ679119.1	DQ382787.1	-	-	X54118.1	AF039557.2	D17736.1	AF251251.1	-	AF127160.1	-	AF461329.1	-	-	-	-
D. triauraria	AB669711.1	AB243399.1	-	-	AB669841.1	AF251141.1	AB048694.1	-	HQ631653.1	-	-	HQ631782.1	-	-	-	-
D. tsacasi	AB669717.1	AB669776.1	-	-	AB669847.1	AF251134.1	AB048783.1	-	-	-	-	AF461346.1	-	-	-	-
D. varians	FJ795579.1	AF461276.1	-	-	AB194427.1	AF136937.2	-	-	-	-	-	AF461323.1	-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. vulcana	AB669718.1	AB669777.1	-	-	AB669848.1	AF251132.1	AB048785.1	-	-	-	-	AF461336.1	-	-	-	-
D. yakuba	X03240.1	X03240.1	X03240.1	X03240.1	X57376.1	AF039561.2	D17738.1	X61127.1	-	AF127159.1	-	AF461330.1	-	-	-	-
D. affinis	HQ979106.1	GQ244451.1	AF519378.1	EF216263.1	AY754404.1	AF037353.1	AF037099.1	-	AY754448.1	U47879.1	-	AF461312.1	-	-	-	-
D. algonquin	JF735923.1	M95144.1	-	U07279.1	AB703316.1	-	-	-	-	-	-		-	-	-	-
D. ambigua	U51610.1	EU493759.1	EU493889.1	EF216264.1	AB703317.1	AF306712.1	AF037092.1	-	-	U47868.1	-	AF461309.1	-	-	-	-
D. azteca	U51605.1	M95146.1	-	U07283.1	AB703319.1	AF150685.1	AF074822.1	-	-	U47866.1	-		-	-	-	-
D. barbarae	AB669728.1	AB669785.1	EU493885.1	EU494139.1	AB669858.1	AF250053.1	AB048802.1	-	HQ631618.1	-	-	HQ631749.1	-	-	-	-
D. biauraria	AB669697.1	AB243383.1	-	-	AB669828.1	AF136932.2	AF393786.1	-	HQ631619.1	-	-	HQ631751.1	-	-	-	-
D. bicornuta	AB669753.1	AB669810.1	-	-	AB669883.1	AF136933.1	-	-	-	-	-	AF461331.1	-	-	-	-
D. bifasciata	AF050749.1	M95147.1	EU493891.1	EU494145.1	AB703320.1	AF251135.1	AF037104.1	-	-	U47869.1	-	AF461313.1	-	-	-	-
D. diplacantha	AB669732.1	AB669789.1	-	-	AB669862.1	AF251142.1	AB048764.1	-	-	-	-	AF461332.1	-	-	-	-
D. greeni	AB669713.1	AB669772.1	-	-	AB669843.1	AF462602.1	AF393792.1	-	-	-	-	AF461325.1	-	-	-	-
D. jambulina	AY757284.1	AB669781.1	-	-	DQ363234.1	AF174489.1	AF393793.1	-	DQ363228.1	-	-	HQ631760.1	-	-	-	-
D. kikkawai	AB669734.1	AB669791.1	-	-	AB669864.1	U96156.3	AB077405.1	-	HQ631625.1	-	-	HQ631761.1	-	-	-	-
D. miranda	AF451104.1	M95148.1	-	EF216276.1	M60998.1	-	U31131.1	AY238807.1	AY754461.1	U47870.1	-	AY754537.1	-	-	-	-
D. narragansett	-	M95149.1	-	-	AB703323.1	-	AF312738.1	-	-	-	-		-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
D. persimilis	AF451100.1	M95143.1	-	EF216278.1	M60997.1	AF129823.1	AF037103.1	L81264.1	-	U47873.1	-	AF461310.1	-	-	-	-
D. pseudoobscura	AB669765.1	AB669821.1	AF519380.1	EF216279.1	M60988.1	AF312739.1	HQ631488.1	X13878.1	AF293746.1	AF059882.1	-	HQ631745.1	-	-	-	-
D. tolteca	-	M95152.1	-	-	AB703325.1	AF306717.1	AF108131.1	-	-	U47867.1	-	AF461314.1	-	-	-	-
D. emarginata	AF045110.1	EU493764.1	EU493894.1	EU494148.1	AF045126.1	-	-	-	-	-	-		-	-	-	-
D. neocordata	AF045104.1	EU493765.1	EU493895.1	EU494149.1	HQ110514.1	AY736480.1	-	-	-	-	-		-	-	-	-
D. capricorni	AY335202.1	EU532079.1	EU493896.1	EU494150.1	AF264073.1	-	-	U51092.1	-	AY335228.1	-	EU532113.1	-	-	-	-
D. equinoxialis	U51593.1	EU493767.1	EU493897.1	EU494151.1	U95268.1	-	-	U51074.1	-	-	-	EU532101.1	-	-	-	-
D. nebulosa	JF735914.1	EU532083.1	-	EU494152.1	HQ110510.1	AY733060.1	-	U51090.1	AF293742.1	AF021830.1	-	EU532116.1	-	-	-	-
D. paulistorum	EU493641.1	HQ110557.1	EU493899.1	EU494153.1	U95272.1	-	-	U51082.1	AF293744.1	AF021829.1	-	EU532111.1	-	-	-	-
D. sucinea	AY335203.1	EU532094.1	-	EU494154.1	AF264080.1	-	-	-	-	AY335229.1	-	EU532114.1	-	-	-	-
H. duncani	JF735916.1	GU597505.1	GU597534.1	-	JF735886.1	-	-	-	-	-	-		-	-	-	-
H. pictiventris	-	AF478434.1	-	-	AB026530.1	AY736493.1	-	U11808.1	AF293736.1	AF021824.1	-		-	-	-	-
Sa. leonensis	EU493683.1	EU493812.1	EU493940.1	EU494185.1	-	EU161100.1	-	-	AF293748.1	-	-		-	-	-	-
Sc. latifasciaeformis	EU493684.1	EU493813.1	-	EU494186.1	-	GQ352255.1	AF071048.1	-	EU446102.1	-	-		-	-	-	-
Sc.lebanonensis	EU493686.1	HQ110572.1	-	EU494188.1	DQ155674.1	-	AB078774.1	-	AF293739.1	AF021822	-		-	-	-	-
Sc.pattersoni	EU493687.1	EU493816.1	DQ155680.1	EU494189.1	DQ155671.1	-	-	-	-	-	-		-	-	-	-
Sc. stonei	JF735922.1	JF735936.1	DQ155678.1	-	DQ155672.1	-	-	-	-	-	-		-	-	-	-
S. anomala	HQ170852.1	HQ170748.1	-	-	AB033647.1	-	-	-	-	-	-		-	-	-	-
Z. bogoriensis	FJ948762.1	GQ352278.1	-	-	-	AY736516.1	-	-	-	-	-		-	-	-	-
Z. ghesquierei	FJ948774.1	EF453719.1	GU597525.1	-	-	AY736518.1	-	-	-	-	-		-	-	-	-

Appendix 1 Continued.

Species	COI	COII	COIII	cytb	ADH	AMYREL	AMY	per	Ddc	Sod	BOSS	hb	snf	wee	marf	fkh
Z. sepsoides	EU493690.1	EU493819.1	GU597526.1	EU494192.1	-	AY736523.1	U31153.1	-	-	-	-		-	-	-	-
Z. taronus	FJ948772.1	EF453705.1	-	-	-	EF458324.1	-	-	-	-	-		-	-	-	-
Z. tuberculatus	EU493691.1	EU595373.1	-	EU494193.1	X63955.1	AY736524.1	U31154.1	U11809.1	AF293751.1	AF021823	-	EU447345.1	-	-	-	-
Z. inermis	FJ948775.1	EF453712.1	-	-	-	AY736519.1	-	-	-	-	-		-	-	-	-
Z. kolodkinae	FJ948778.1	EF453715.1	-	-	-	AY736520.1	AF462593.1	-	-	-	-		-	-	-	-
Z. tsacasi	KC465964.1	EU595374.1	-	-	-	EF458325.1	-	-	-	-	-		-	-	-	-
Z. indianus	KF736195.1	EF632396.1	GU597527.1	-	-	EF458322.1	-	-	EU446103.1	-	-	EU447344.1	-	-	-	-
Z. lachaisei	FJ948766.1	EF453701.1	-	-	-	EF458330.1	-	-	-	-	-	-	-	-	-	-