Inferring evolutionary forces from regional and temporal base composition variation in *Drosophila*

Mishra, Neha

Student ID: 20111858
Department of Genetics, SOKENDAI

Date of Submission: July 24, 2016

Table of Contents

CHAPTER ONE: GENERAL INTRODUCTION

CHAPTER TWO: WITHIN-GENE HETEROGENEITY IN BASE COMPOSITION IN *D. MELANOGASTER*: ASSOCIATION WITH TRANSCRIPTION AND TRANSLATION

List of Tables

List of Figures

Acknowledgements

In the course of my graduate studies at SOKENDAI, Department of Genetics, I am grateful to many people for their support and guidance. I am greatly indebted to my thesis advisor, Professor Hiroshi Akashi for his supervision, advice and patience. I have learnt immensely from his deep knowledge of population genetics and molecular evolution throughout the 5 years that I have worked closely with him. I am also thankful to him for his valuable suggestions about scientific thinking and writing. I would also like to thank past and current Akashi Lab members, Drs. Tomoko Ohta, Naoki Osada, Tomotaka Matsumoto and Boyang Li. I am grateful to Tomoko Ohta for giving valuable advice and suggestions during the lab meetings and journal clubs. I thank Naoki Osada for providing useful advice for the NGS data analysis and Tomotaka Matsumoto for helping with ancestral state reconstruction and divergence analyses. I would also like to thank Kent Kawashima for his support.

I would like to offer my great appreciation to the past and current members of my Progress Committee, Drs. Yasukazu Nakamura, Masanori Arita, Kazuo Ikeo, Masafumi Nozawa, Takashi Gojobori, Asano Ishikawa and Toshiyuki Takano for their invaluable comments and suggestions on my doctoral studies. I have always benefited through discussions with my Progress Committee and they have helped me with their profound knowledge on computational methods and molecular evolution. I would like to extend my gratitude towards Mizue Sei-Macfhearchair, Noriko Yamauchi and Chikako Miura for their help and support.

My stay in Japan was supported by scholarship offered by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan and I am very thankful for it. Finally, I would like to thank my mom, Asha Mishra, my dad, Mukteshwar Nath Mishra, my brother, Sudhanshu Mishra and my husband Deepak Ailani for their unconditional love and support.

Abstract

My research goal is to study the variation in global forces of evolution at the genomic level taking Drosophila as a model system. Although the global forces of genome evolution have largely been established, the variations in their strength within and between genomes are less understood. Patterns of base composition can reveal variation in evolutionary forces, such as selection for codon usage, mutation bias and biased gene conversion. The base composition of synonymous sites is known to evolve primarily under the affect of selection for translational efficiency and accuracy, mutation and drift. Most introns, on the other hand, evolve under much lesser selective constraint. Hence, variations in various evolutionary parameters within and between genomes can be studied using base composition of synonymous sites and introns. Base composition comparison among different nucleotide classes can also help in distinguishing the causes of base composition heterogeneity. I studied the variation in base composition (GC content) within and across genes in the *D. melanogaster* genome and across different lineages in the *D. melanogaster* subgroup.

I found that base composition at synonymous sites and introns varies at a within-gene as well as genome-wide level. Within genes, GC content of small introns decreases from the 5′ end to the 3′ end. The GC gradient near the 5′ end is sensitive to the transcriptional level of the genes with highly transcribed genes having a steeper gradient. The intron divergence exhibits a positive gradient near the 5′ end, which is also steeper for genes with higher transcriptional level. The 5′ GC gradient in small introns is also observed in genes expressed at low levels in the germline, suggesting it to be under selection. On the other hand, GC content at synonymous sites shows a

sharp increase at the 5′ end and then declines towards the 3′ end. The base composition at the synonymous sites near the 5′ end seems to be under strong selection for low GC content, which might be functionally important for translation of mRNA since it is not observed in introns.

At the genome-wide level, base composition is heterogeneous within as well as between chromosome arms. GC content at synonymous sites and introns shorter than 100 bp is significantly higher on the X chromosome compared to the autosomes. GC content at the synonymous sites is the most heterogeneous among all GC classes, suggesting that selection intensity might vary at a shorter scale than mutation in the *D. melanogaster* genome, since base composition of synonymous sites is thought to be evolving predominantly under the selection for codon usage whereas other nucleotide classes are mainly affected by mutation bias and biased gene conversion.

To study the base composition variation across different genomes, I examined the changes in the synonymous base composition patterns in >5000 genes from the 7 *Drosophila melanogaster* subgroup species. I used the existing genome data from five species and also added data of two more species in the *D. melanogaster* subgroup. Next-Generation RNA sequencing and Genome sequencing on the *D. tessieri* and *D. orena* transcriptomes and genomes was performed in the lab. I conducted a rigorous analysis of the RNA-seq and Genome-seq data and developed protocols for *de novo* gene and intron annotation. For this task, I used the available data from the sequenced Drosophila species. I developed several methods and also used some of the softwares that are already available for NGS data analysis. The substitutions occurred in each of the lineages were inferred using maximum likelihood approach. I found that all of the

lineages showed strong departures from the equilibrium states and in some lineages both effective population size and mutation bias seemed to have fluctuated. These findings suggest that magnitude of forces governing base composition at synonymous sites may have varied frequently in a lineage-specific manner.

CHAPTER 1

Introduction

1.1 Weak selection and genome evolution

The central goal of my research is to understand the variation in global forces of evolution, both spatially and temporally, at the genomic level in multicellular eukaryotes taking Drosophila as a model system. Processes, such as mutation, genetic drift, natural selection, and recombination, act constantly within genomes and contribute to their long-term evolution. Although the global forces of genome evolution have largely been established, the variations in their strength within and between genomes are less understood. The strength of these forces can change with time or across genomic breadth due to the fluctuations in various evolutionary parameters, creating heterogeneous substitution patterns (Akashi et al. 2006). Irregularities in substitution patterns can create false signatures of adaptive evolution and affect the accuracy of ancestral inference and phylogenetic inference methods. Even though the variations in evolutionary forces can have a major impact on the interpretations of various molecular evolutionary analyses, they have not been given much attention.

The nearly neutral theory of molecular evolution, proposed by Ohta (Ohta and Kimura 1971; Ohta 1972a; Ohta 1972b; Ohta 1973; Ohta 1974; Ohta 1976), assumes that fitness effects of new mutations form a continuous distribution around the fitness effect of neutral mutations that have a selection coefficient, *s*, of 0. The mutations at the ends of the distribution are either strongly deleterious or highly advantageous. The strongly deleterious mutations are quickly removed from the population whereas highly advantageous mutations can be fixed. A large fraction of mutations have fitness effects near that of neutral mutations and these are weakly selected. The fate of these mutations depends on the product of their selection coefficient and effective population sizes (N_es). In contrast with N_es of advantageous and deleterious mutations, N_es for weakly selected mutations is of the order of 1. Thus, changes in N_e as well as mutation rates and biases can greatly influence the efficacy of selection on such sites.

1.2 Base composition as measure of to test fluctuations in evolutionary forces

Patterns of base composition can reveal variation in evolutionary forces, such as selection for codon usage, mutation bias and biased gene conversion. The base composition of synonymous sites, which reflects codon usage bias, is known to evolve primarily under the affect of translational selection, mutation and drift (Li 1987; Andersson and Kurland 1990; Bulmer 1991; Akashi 1994; Akashi 1997; Akashi 1998; Akashi et al. 1998; Stoletzki and Eyre-Walker 2006; Hershberg and Petrov 2008). The currently accepted model of codon usage bias, referred to as “major codon usage”, states that selection favors major codons whereas mutation bias and genetic drift allow minor codons to persist (Bulmer 1991). This balance of weak forces makes codon bias sensitive to changes in evolutionary forces. Major codons or preferred synonymous codons are usually the ones that are identified by abundant tRNA molecules. This association between major codons and tRNAs has been shown in bacteria, yeast and for a few amino acids in Drosophila (Bennetzen and Hall 1982;

Ikemura 1985; Yamao et al. 1991; Moriyama and Hartl 1993; Kanaya et al. 2001). Higher frequency of major codons results in quicker translation and lesser mis-incorporation of incorrect amino acid (Akashi and Eyre-Walker 1998; Akashi et al. 1998).

However, since most of the major codons in *Drosophila melanogaster* end in G or C, codon usage bias can also be influenced by biased gene conversion (Galtier et al. 2001; Galtier 2003; MARAIS et al. 2003) and selection for other factors such as mRNA secondary structure and translational initiation (reviewed in (Hershberg and Petrov 2008)). Hence, comparison among different nucleotide classes should be used to identify the underlying process that governs base composition heterogeneity (Akashi et al. 2006; Ko et al. 2006).

Most introns evolve under much lesser selective constraint (Halligan 2004; Haddrill et al. 2005). Also, since introns are transcribed but not translated, the base composition variation caused by processes related to transcription can be better-studied using introns. Previous studies in Drosophila have shown that patterns of divergence and polymorphism is sensitive to the intron size (Parsch 2003). Introns shorted than 100 bp have high levels of both interspecific divergence and intraspecific polymorphism (Parsch 2003). Selective constraints might differ between small and long introns (Parsch 2003; Halligan 2004; Haddrill et al. 2005; Parsch et al. 2010)and hence they were analyzed separately.

1.3 Goal of the study

The goal of this dissertation is to study spatial and temporal fluctuations in evolutionary forces in the *Drosophila melanogaster* genome and subgroup. I studied the variation in base composition (GC content) within and across genes in the *D. melanogaster* genome and across different lineages in the *D. melanogaster* subgroup and test its causes.

The spatial variations in evolutionary forces, such as mutation, drift, selection and biased gene conversion, shape the sequence and structure of genes and genomes. Evolutionary forces vary across genomes, based on functional constraints (Kimura 1986; Liu et al. 2008; Rands et al. 2014), expression level (Pál et al. 2001; Krylov et al. 2003; Majewski 2003; Rocha 2003; Comeron 2004; Drummond 2005; Drummond et al. 2005; Cherry 2010)(reviewed in (Akashi 2001; Zhang and Yang 2015)), mutation rates and patterns (Wolfe et al. 1989; Singh 2005a; DURET 2009), and rates of recombination (Hill and Robertson 1966; Birky and Walsh 1988; Aguade et al. 1989; Begun and Aquadro 1992; Kliman and Hey 1993; Charlesworth and Guttman 1996; Hey and Kliman 2002). While studying base composition variation at a genome-wide scale can help in identifying regional differences in evolutionary forces, studying base composition variation within-genes can help in determining the biological processes underlying the heterogeneity in evolutionary forces.

Evolutionary forces also vary temporally due to fluctuations in mutation rate and biases (Akashi 1996; Akashi et al. 2006), recombination rates (Takano-Shimizu

1999), effective population size (Akashi 1996; Akashi et al. 2006) and fitness effects of mutations (Clark et al. 2007; McBride 2007; McBride et al. 2007). Studying lineage-specific evolution of base composition can reveal the changes in evolutionary forces occurred in different lineages.

CHAPTER 2

Within-gene heterogeneity in base composition in *D. melanogaster*: Association with transcription and translation

2.0 Chapter Summary

In this chapter, I study base composition variation within the genes of *D. melanogaster*. Studying base composition variation within gene can provide insights to the biological processes such as transcription and translation underlying the evolution of base composition. Within genes, GC content of small introns decreases from the 5′ end to the 3′ end. The GC gradient near the 5′ end is sensitive to the transcriptional level of the genes with highly transcribed genes having a steeper gradient. The intron divergence exhibits a positive gradient near the 5′ end, which is also steeper for genes with higher transcriptional level. The variation in within-gene base composition is also associated with RNA polymerase II binding levels. On the other hand, GC content at synonymous sites shows a sharp increase at the 5′ end and then declines towards the 3′ end. The base composition at the synonymous sites near the 5′ end seems to be under strong selection for low GC content, which might be functionally important for translation of mRNA since it is not observed in introns.

2.1 Introduction

Sites within genes are subject to different selective constraints or mutational rates based on their function, such as protein coding or intronic (Haddrill et al. 2005), and position in the transcript (Majewski and Ott 2002; Eddy and Maizels 2007; Li et al. 2012; Park et al. 2014), mRNA ((Liljenström and Heijne 1987), reviewed in (Tuller and Zur 2015)) or intron (Halligan 2004). Intragenic base composition at synonymous sites is found to vary in a number of species across various taxa such as bacteria (Bulmer 1988; Hooper and Berg 2000; Qin et al. 2004; Bentele et al. 2013; Hockenberry, Sirer, Amaral, and Jewett 2014a) (Clarke and Clark 2010), yeast (Qin et al. 2004) (Shah et al. 2013) and *Drosophila* (Hey and Kliman 2002; Qin et al. 2004). Base composition heterogeneity at synonymous sites within genes is attributed to Hill-Robertson effect (Qin et al. 2004), selection for reduced ribosomal elongation speed in the beginning of genes (Zhang et al. 1994; Tuller et al. 2010; Hockenberry, Sirer, Amaral, and Jewett 2014b) and reduced mRNA folding for efficient translation initiation (Bentele et al. 2013; Hockenberry, Sirer, Amaral, and Jewett 2014a).

The base composition of synonymous sites is known to evolve under the affect of translational selection, mutation and drift (Li 1987; Andersson and Kurland 1990; Bulmer 1991; Akashi 1994; Akashi 1997; Akashi 1998; Akashi et al. 1998; Stoletzki and Eyre-Walker 2006; Hershberg and Petrov 2008). However, it can also be influenced by biased gene conversion (Galtier et al. 2001; Galtier 2003; MARAIS et al. 2003) and selection for other factors such as mRNA secondary structure and translational initiation (reviewed in (Hershberg and Petrov 2008)). Hence, it is difficult to identify the underlying process that governs base composition heterogeneity in synonymous sites. Most introns, on the other hand, evolve under much lesser selective constraint (Halligan 2004; Parsch et al. 2010). Also, since introns are transcribed but not translated, the base composition variation caused by processes related to transcription can be better-studied using introns. Kliman and Eyre-Walker analyzed the intron base composition within genes using 86 *D. melanogaster* genes and found that it varies along the genes (Kliman and Eyre-Walker 1998). Base composition comparison among different nucleotide classes can also help in distinguishing the causes of base composition heterogeneity (Akashi et al. 2006; Ko et al. 2006).

I studied intragenic variations in base composition of small introns, long introns and synonymous sites. I found that in *D. melanogaster* genes, base composition varies within genes. The GC content within genes decreases with increase in distance from the TSS. I tested the association of base composition with RNA polymerase and compared the base composition variation between introns and synonymous sites to identify the underlying cause of base composition heterogeneity within genes. I found that base composition variation within genes reflects selection associated transcription and translation.

2.2 Materials and Methods

2.2.1 Sequence data

Sequence data and annotations for *Drosophila melanogaster* genome (Release 5.28, June 4, 2010) (Adams et al. 2000), *D. yakuba* genome (Release 1.3) (Clark et al. 2007) and *D. erecta* genome (Clark et al. 2007) were obtained from FlyBase (www.flybase.org). Only genes that were predicted to have single protein isoform were used for the analysis. Genes that had different transcription start sites (TSS) in different isoforms or were not expressed in any of tissues described in the Matsumoto et al, 2016 study (Matsumoto et al. 2016) were also excluded. Additional filters on the gene set were applied for some analyses. These filters are described in the later sections.

2.2.2 Transcript abundance data

The transcript abundance data for various *Drosophila* tissues used in this study were obtained from Matsomoto et al (Matsumoto et al. 2016). These data included transcript abundance data from different developmental stages such as adult flies (male and females), larva and embryo and various adult and larval tissues such as brain, central nervous system, salivary glands, midgut, hindgut and reproductive tissues such as testis, ovary, and accessory glands. The raw transcript abundance data used in Matsumoto et al study were from FlyAtlas (Chintapalli et al. 2007).

2.2.3 Codon bias and GC content measures

Major codon usage (MCU) for 2-fold redundant codon families (except Asp) was used as a measure for codon bias. Most amino acids encoded by 2-fold redundant codon families, except Aspartic acid, show clear preference towards the G or C-ending codon over the A or T-ending codon in highly transcribed genes (Akashi unpublished) and hence 2-fold redundant codons are a good measure of codon bias. MCU is defined as the percentage of major codons in a gene. Major codons classifications were obtained from a previous analysis (Akashi unpublished) for 904 X-linked genes and 5330 autosomal genes.

Introns with length less than 100 bp were classified as small introns and introns with length greater than 100 bp as long introns. Previous studies in Drosophila have shown that introns shorter than 100 bp have high levels of both interspecific divergence and intraspecific polymorphism(Parsch 2003). Selective constraints might differ between small and long introns (Parsch 2003; Halligan 2004; Haddrill et al. 2005; Parsch et al. 2010)and hence they were analyzed separately. The first 10 and the last 30 bases of introns were removed while calculating the GC content in order to exclude potential sites required for splicing and polypyrimidine tract (Halligan 2006).

2.2.4 RNA Pol II enrichment data

RNA PolII enrichement in *D. melanogaster* determined by Gilchrist et al. (2010) was used in this study (Gilchrist et al. 2010). Data from S2 cells were obtained from NCBI Gene Expression Omnibus (GEO) (file: GSE20471_Pol_II_Average_Adelman.txt for RNA Pol II enrichment data).

2.2.5 Within-gene GC content variation

To examine the heterogeneity in evolutionary forces, I analyzed the variation in GC content within individual genes. For this task, I calculated GC content for short segments along each gene with gene length greater than 1500 bp (from the transcription start site to the end of the transcript). The genes that were not transcribed among the tissues described in the Matsumoto et al study (Matsumoto et al. 2016) were excluded. To test for selection on base composition near the TSS, I used the genes that were transcribed at very low levels in the tissues containing germline cells. Hence, genes that were transcribed at the bottom 25% percentile in both ovary and testis and top 75% percentile in at least one other tissue were used. For small and long introns, GC content was calculated for 150 bp segments staring from the transcription start site (TSS). A study by Matsumoto (unpublished data) showed that divergence of intronic sites is sensitive to their position within introns (Matsumoto and Akashi, unpublished). This is especially true for long introns. Since, my objective was to study the variation in GC content with respect to the TSS, I excluded parts of introns that were sensitive to their position within the introns. Hence, I used only 11-100 bp of long introns, which had the least amount of variation within introns in divergence.

Segments with the same relative position from the TSS across genes were grouped into bins. Average GC content was calculated for each bin. I also calculated scaled GC content for each nucleotide class to control for GC variation across genes. Scaled GC content was defined as the GC content of a segment in a gene each minus the GC of the whole gene for a particular nucleotide class. The average scaled GC content across genes was also calculated for each bin. The 95% confidence interval on the average GC content and average scaled GC content of each bin was calculated by resampling the genes present in that bin 1000 times.

To study codon bias variation within genes, I used genes with protein sequence length greater than 300 amino acids. MCU was calculated for segments of 50 codons and 150 bp to study the variation in codon bias with respect to the start codon and TSS, respectively. Segments with the same relative position from the start codon ot TSS were grouped into bins and average as well as scaled MCUs were calculated for each bin. Confidence intervals were calculated by resampling regions of genes present in a bin 1000 times.

Spearman's rank correlations were used to quantify intron GC gradients. To calculate the correlation coefficients between GC content and position of intronic sites along the genes, I divided each gene into segments of 750 bp starting from the TSS since I obtained different patterns in the first 750 bp and post-750 bp from the TSS. Only the segments that had at least 50 intronic sites and at least 200 sites between the first and the last intronic sites were included. Within each segment, each intronic site was assigned a value of either 1 or 0. If the nucleotide at a given site was G or C, it was assigned 1 and vice versa. The dataset used to calculate the Spearman's rank correlation coefficient (r_s) consisted of the position of each intronic site within the segment and its GC content. r_s values were calculated using "distancematrix" function of Bio.Cluster package in python 2.7.8. Segments with the same relative position from the TSS across genes were grouped into bins. Average r_s values were calculated across genes for each bin. Bootstrap procedure was used to calculate 95% confidence interval on average r_s value of each bin by resampling individual r_s values in a bin 1000 times. A similar approach was used for synonymous GC class as well. In case of 2-fold synonymous sites, r_s values were calculated for the first 50 codons and segments of 200 codons along the remaining part of intronless autosomal genes. The minimum number of 2-fold codons in a window was set to 10 for the first 50 codons and 30 for the rest of the gene.

To test the relationship between intron GC content and transcript abundance, I distributed introns into 6 bins based on their transcript abundance in whole adult. The genes that were not transcribed in whole adult were excluded for this analysis. This was done for five intron classes: all introns, small introns, long introns and introns present in the 5′ and 3′ regions of the genes. 5′ region is defined as the region up to 750 bp from the TSS and 3′ region is defined as the region from 751 bp to 1500 bp from the TSS. Introns belonging to the same gene were concatenated. Each bin contained roughly equal number of sites of a given intron class. First, the estimated number of intronic sites in each bin was calculated by dividing the total number of intronic sites of a given intron class by the total number of bins. Each intron belonged to a single bin entirely. If while adding an intron to a bin, the total number of sites per bin exceeded the estimated number, the number of extra nucleotides that need to be accommodated into the bin were calculated. If the number of extra nucleotides was more than half of the intron length, the intron was added to that bin; otherwise it was left for the next bin.

2.2.6 Within-gene RNA PolII enrichment

To calculate the variation in RNAPollII enrichment, I calculated the positions of RNAPollII enrichment domains relative to the transcription start site (TSS) for each gene with length (including introns and UTRs) greater than 1500 bp. The average enrichment across X-linked and autosomal genes was calculated for 150 bp bins along the gene. The inclusion of a domain in a bin was decided using the midpoint of the domain. The 95% confidence interval on the average RNA Pol II enrichment of each bin was calculated by resampling the regions within the bin 1000 times.

To calculate the correlation between small intron GC content and RNA Pol II enrichment, I divided each gene into windows of 150 bp. Average RNA Pol II enrichment and small intron GC content was calculated for each window. All windows with at least 30 small intronic sites were pooled across genes and Spearman's rank correlation coefficient between RNA Pol II enrichment and small intron GC content was calculated for X-linked and autosomal high and low transcript abundance genes separately.

To test the relationship between RNA Pol II enrichment and transcript abundance for 5p750 and post-5p750 regions, I binned genes into 6 bins containing roughly equal number of sites and calculated average RNA Pol II enrichment in the 1-750 bp region and 751-1500 bp region for each bin. The procedure for binning was same as that used for intron GC vs. transcript abundance analysis. The inclusion of a domain in a bin was decided using the midpoint of the domain.

2.3 Results

2.3.1 Variation in small intron GC content with respect to the transcription start site

Genes experience binding of RNA Polymerase during transcription and several kinds of chromatin modifications that differ along the length of the transcript (Bell et al. 2007; Schwaiger et al. 2009; Kharchenko et al. 2012). The processes related to transcription can cause variation in mutational and/or selective pressures within genes resulting in base composition variation (Datta and Jinks-Robertson 1995; Polak and Arndt 2008). To study the variation in base composition within genes, I focused on the GC content of small introns (length up to 100 bp) since these introns have been suggested to evolve under lesser selective pressures than other classes of DNA within genes (Parsch 2003; Halligan 2004; Haddrill et al. 2005; Parsch et al. 2010).

The GC content of small introns within genes decreased from the 5′ to the 3′ end (Figure 2.1 A). The relatively stronger decline in GC content near the 5′ end of the genes was observed for the first 750 bp from the TSS followed by a gradual decline. Significant negative correlation was observed between GC content and intronic positions for the first 750 bp from the TSS (Spearman's r = -0.04, $p < 10^{-3}$).

Hence, I defined the region between the TSS and 750 bp from the TSS as 5p750 region, and the region between 751 bp from TSS to 1500 bp from TSS as 5p751-1500 region. Since, all genes included in this analysis were at least 1500 bp long, the analysis was restricted to until 1500 bp from TSS. The magnitude of the correlation declined with distance from the TSS (Figure 2.2). To control for the GC variation among genes, GC content scaled to the gene averages was also studied. Scaled GC content also declined from 5′ to 3′ end (Figure 2.1 B). The GC content 11-100 bp of long introns also declined with increase in distance from the TSS.

I examined the relationship between transcription patterns and GC content in 5p750 regions. The GC content of small introns in the 5p750 region showed a stronger correlation with transcript abundance (Spearman's $r = 0.3$, $p < 10^{-4}$) than that of the small introns present in the 5p751-1500 region (Spearman's $r = 0.15$, $p = 0.0004$) (Figure 2.3). The difference in the GC content of small introns between genes with high and low transcript abundance was also observed only in the 5p750 region, suggesting that transcription-associated substitution biases might be stronger in the 5p750 region than the post-5p750 region.

2.3.2 Test for selection on base composition near the 5′ end

The variation in the GC content within genes could be a result of variation in mutational patterns, natural selection, or biased gene conversion. Mutational variation can be caused by heterogeneity of DNA repair (Lujan et al. 2014; Li et al. 2015) or accessibility of DNA to damaging agents (Beletskii and Bhagwat 1996; Morey et al. 2000) or both. Heterogeneity in the strength of natural selection to acquire distinct chromatin marks (Dekker 2007; Alekseyenko et al. 2012; Wachter et al. 2014) or to form DNA or RNA secondary structures (Hoede et al. 2006) could also result in base composition variation.

If 5′ GC gradient is a result of variation in transcription-associated mutation or biased gene conversion, the genes that are expressed in the germline are likely to have a stronger pattern since mutations occurring in the germline are likely to pass on to the future generations. On the other hand, if 5′ GC gradient is a result of variation in transcription-associated selection, even the genes that are not expressed or expressed at very low levels in the germline should have a strong pattern. Hence, to test if the 5′ GC gradient is caused by variation in natural selection or mutation patterns and/or biased gene conversion, I analyzed the GC gradient of genes that are transcribed at very low levels in ovary and testis but at moderate to high levels in other tissues. The genes that have low transcript abundance in ovary and testis and moderate to high transcript abundance in other tissues also show a strong 5′ GC gradient (Figure 2.4), suggesting that the high GC near the 5′ reflects transcription-associated selection.

2.3.3 Association of intron GC gradient with RNA Pol II pausing

Chromatin modifications associated with transcription are known to enrich differently along the genes (Kharchenko et al. 2012). Chromatin modifications such as H3K4me2 and H3K4me3 are enriched at the TSS, whereas H3K79me1 and H2B-ub enrich in the middle of the gene bodies in active genes (Kharchenko et al. 2012). RNA Polymerase II is known to pause near the TSS in most genes (Gilmour and Lis 1986; Wirbelauer et al. 2005; Bell et al. 2007; Schwaiger et al. 2009; Kharchenko et al. 2012) and its role in gene regulation has been suggested(Gilchrist et al. 2010).Since the GC gradient within gene reflects transcription-associated selection, I tested if within-gene GC content associates with RNA Pol II binding.

Within genes, RNA Pol II enrichment was high near the TSS and declined with increase in distance from the TSS, as shown in previous studies (Gilmour and Lis 1986; Muse et al. 2007; Zeitlinger et al. 2007; Schwaiger et al. 2009; Kharchenko et al. 2012) (Gilchrist et al. 2010). The gradient of RNA Pol II enrichment was very similar to that of the GC gradient in autosomal as well as X-linked genes since it is steeper in the 5′ region and gradually flattens towards the 3′ end. However, the RNA Pol II gradient is around 200-250 bp closer to the TSS than the small intron GC gradient in both autosomal and X-linked genes (Figure 2.5 A, B). RNA Pol II enrichment was also correlated with the small intron GC content for genes with high as well as low transcript abundance (Table 2.1).

Similar to GC content of small introns, RNA Pol II enrichment in the 5p750 region also showed stronger correlation with transcript abundance (Spearman's $r = 0.2$, $p < 10^{-2}$) than that in the post-5p750 region (Spearman's $r = 0.09$, $p = 0.0045$) (Figure 2.6). This suggests that, similar to intron GC, the relationship between transcriptional level and RNA Pol II binding is more pronounced in the 5′ region.

2.3.3 Codon bias increases near the 5′ end

I have found that intron GC declines near the 5′ end. Since, most major codons in *D. melanogaster* are GC-ending, codon bias should also be expected to decline near the 5′ end if the 5′ GC pattern reflects transcription-associated selection. However, it has been documented in various taxa such as bacteria (Bulmer 1988; Hooper and Berg 2000; Qin et al. 2004; Clarke and Clark 2010; Bentele et al. 2013; Hockenberry, Sirer, Amaral, and Jewett 2014a), yeast (Qin et al. 2004; Shah et al. 2013), and *Drosophila* (Kliman and Eyre-Walker 1998; Qin et al. 2004) that codon bias increases near the start codon.

Consistent with the previous studies (Kliman and Eyre-Walker 1998; Qin et al. 2004), I found that GC content at 2-fold synonymous sites (Major Codon Usage, MCU) and 4-fold synonymous sites showed a sharp increase near the start codon, followed by a gradual decline towards the 3′ end (Figure 2.7). A similar pattern was observed when MCU variation was studied from the transcription start site (TSS) (Figure 2.8). The differences in MCU among genes with different expression level were more pronounced at the 5′ end (Figure 2.9).

To find the regions that show significant positive and negative MCU gradients, I calculated the correlation between MCU and codon positions along the genes. Autosomal genes that do not have any introns were used for this analysis to capture the gradient specific to coding regions. MCU is positively correlated with codon position for the first 50 codons (Spearman's $r = 0.07$, $p < 10^{-4}$) and negatively correlated with codon position for the next 200 codons (Spearman's $r = -0.03$, $p =$

0.002). From 300 codons onwards, the slope is not significantly different from 0.

Previous studies have suggested that reduced codon bias near the beginning of the genes is under selection for reducing translational elongation speed (Tuller et al. 2010; Hockenberry, Sirer, Amaral, and Jewett 2014a) or mRNA folding for translation(Bentele et al. 2013). The first 50 codons of genes also have lower dS than other codons of genes even after controlling for the distance from TSS in both low and highly transcribed genes (Matsumoto and Akashi, unpublished).

To identify the underlying process that governs codon bias near the 5′ end, I compared scaled GC content with respect to the distance from TSS between 2-fold synonymous sites and small introns. Within the first 450 bp, the scaled GC content at 2-fold synonymous sites increases as a function of distance from TSS even though the transcription-associated selection pressure, as indicated by the small intron scaled GC gradient seems to be in the opposite direction (Figure 2.8). These results suggest that the base composition of the synonymous sites near the 5′ end is under strong selection for low codon bias, which could be important for translation . The MCU gradient after the first 450 bp from TSS is similar to that of small introns (Figure 2.8), suggesting that evolutionary forces governing high GC content in small introns in the 5p750 regions might also operate on synonymous sites.

To differentiate between the effect of codon usage bias and transcription-associated selection on synonymous sites, I compared the relationship between transcript abundance and MCU in the 5p750 and 5p750-1500 regions. Since, both codon usage selection and transcription-associated selection predict positive correlation between MCU and expression level, 5p750 region should show a stronger correlation between MCU and transcript abundance than 5p751-1500 region. MCU was strongly positively correlated with transcript abundance in both 5p750 (Spearman's $r = 0.27$, $p < 10^{-4}$) and 5p750-1500 regions (Spearman's $r = 0.26$, $p < 10^{-4}$). The correlation between MCU and transcript abundance was only slightly higher in the 5p750 region and the difference in the MCU of 5p750 and 5p750-1500 regions was apparent only in genes with intermediate transcriptional level (Figure 2.9), unlike small introns. The high correlation between MCU of 5p750-1500 region and transcript abundance, even though the effect of transcription might be weaker on the 5p750-1500 region suggests that translational selection is the major contributor of high GC in the 2-fold synonymous sites.

2.4 Discussion

In this chapter, I show that GC content of introns varies within the genes of *D. melanogaster* genomes and propose underlying causes of GC content heterogeneity. GC content of small introns decreases near the 5′ end of genes and the decline becomes less steep as the distance from the TSS increases. This finding is consistent with a previous study in *Drosophila* (Kliman and Eyre-Walker 1998) where the authors investigated the base composition heterogeneity within 117 genes. Here I have studied a much larger dataset, distinguished between small and long introns since they are known to evolve at different rates (Haddrill et al. 2005; Halligan 2006), and tested causes for base composition heterogeneity. My result is in contrast with the result for base composition at all sites that shows increase in GC from 5′ to 3′ end (Aerts et al. 2004). Negative gradients have also been observed for CG, TG and CA dinucleotides in *Drosophila* introns (Tang et al. 2006). Compositional variations have been observed within genes in the yeast (Stoletzki 2011), *C. elegans* (Khuu et al. 2007), human (Clay et al. 1996; Aerts et al. 2004; Khuu et al. 2007; Polak and Arndt 2008) and plant genomes (Wong et al. 2002; Serres-Giardi et al. 2012; Glémin et al. 2014; Ressayre et al. 2015). Transcription-related mutation or selection bias has been suggested as a potential cause of within-gene compositional gradients based on the correlation between coding and intron GC gradients (Wong et al. 2002), comparison between genes with different breadth of expression (Aerts et al. 2004), association with transcriptional activator binding sites (Khuu et al. 2007) and strand asymmetry in the complementary substitution rate (Polak and Arndt 2008). Some studies argue that compositional gradients reflect biased gene conversion based on correlation of GC content with recombination rate (Stoletzki 2011; Serres-Giardi et al. 2012; Glémin et al. 2014). We, on the other hand, did not find a positive correlation between recombination rate and GC content in the 5p750 region (see Chapter 3), which rules out the possibility of biased gene conversion causing high GC near the 5′ end.

I also show that small intron GC content in the 5p750 region shows a stronger correlation with transcript abundance than that in the 5p751-1500 region. This pattern implies that transcription-related mutation or substitution biases are effecting the base composition of the 5p750 region of the genes, more so than the 5p751-1500 region. Experimental studies in yeast have shown that transcription increases mutation rate (Datta and Jinks-Robertson 1995; Morey et al. 2000; Lippert et al. 2011; Mischo et al. 2011; Takahashi et al. 2011). During the formation for DNA-RNA hybrid during transcription, the non-transcribed strand is transiently single-stranded. The RNA polymerases pause near the 5′ end. Hence, the non-transcribed strand near the 5′ end might remain single-stranded for longer time and be susceptible to lesions. This would result in the increase of mutation rate near the 5′ end. Even if the repair mechanisms are more efficient during transcription, the region near the 5′ end should have high substitution rates if base composition is not under selection. However, small intronic sites near the 5′ end of the genes have low divergence rates (Matsumoto and Akashi, unpublished). The divergence rate increases with the increase in distance from the TSS. This gradient is steeper for genes with higher transcript abundance. Since the divergence of 5p750 region is lower than that of 5p751-1500, either the base composition in the 5p750 region is under higher selective constraint or mutation rate is lower in that region.

Patterns reflecting variation in mutation rate should be observed in genes that are expressed in the germline since mutation occurring in the germline are likely to be passed on to the future generations. I found that genes that are expressed at low levels in tissues containing germline cells but at moderate to high levels in other tissues also have increased GC near the 5′ end, suggesting that transcription-associated selection drives high GC near the 5′ end.

Most of the earlier studies on within-gene variation in base composition have focused on synonymous sites and found that codon bias near the 5′ end is reduced (Bulmer 1988; Eyre-Walker and Bulmer 1993; Kliman and Eyre-Walker 1998; Hooper and Berg 2000; Qin et al. 2004; Bentele et al. 2013; Hockenberry, Sirer, Amaral, and Jewett 2014a). I also found a reduction in codon bias near the 5′ end, but the rest of the pattern shows a similar decline as intron GC. The synonymous sites in the first 50 codons of genes have lower divergence than synonymous sites present in the rest of the codons, even after accounting for the distance from the TSS (Matsumoto and Akashi, unpublished). The first 50 codons of genes have low GC content even though the selection pressure for transcription prefers high GC. These findings suggest that strong selection for reducing GC content is operating in the first 50 codons of genes and provide further evidence to previous studies that suggest that reduced GC content in the first few codons might be important at the mRNA level or during translation.

CHAPTER 3

Genome-wide heterogeneity in base composition in *Drosophila melanogaster*

3.0 Chapter Summary

In the previous chapter, I established that base composition varies within the genes of *D. melanogaster*. At a genome-wide scale, evolutionary forces can vary based on functional constraints, mutation rates and patterns, recombination rates and chromatin states. Hence, to understand variations in evolutionary forces and their underlying causes, I studied base composition variation at a genomic scale.

At the genome-wide level, base composition is heterogeneous within as well as between chromosome arms. GC content at synonymous sites and introns shorter than 100 bp is significantly higher on the X chromosome compared to the autosomes. GC content at the synonymous sites is the most heterogeneous among all nucleotide classes, suggesting that selection intensity might vary at a shorter scale than mutation in the *D. melanogaster* genome, since base composition of synonymous sites is thought to be evolving predominantly under the selection for codon usage whereas other nucleotide classes are mainly affected by mutation bias and biased gene conversion.

3.1 Introduction

As discussed in the previous chapter, variation in base composition within genes could be governed by biological processes such as transcription and translation. However, the sites within a gene are expressed at similar levels, tightly linked and share the regional location within the genome. Hence, to study the affect of variation in expression level, recombination rate and genomic location, base composition variation at the genomic-scale must be studied.

Base composition is known to vary at a genome wide-scale in many species (Sueoka 1962; Bernardi 1989; Gardiner et al. 1990; Carulli et al. 1993; Sharp and Lloyd 1993; Dujon et al. 1994; Feldmann et al. 1994; Deschavanne and Filipski 1995; Tang et al. 2006; Diaz-Castillo and Golic 2007; Jørgensen et al. 2007). The scale of heterogeneity varies among taxa, with the human genome having the highest heterogeneity (Tang et al. 2006). The GC content in the *Drosophila melanogaster* genome varies at a scale of more than 100kb (Carulli et al. 1993). To understand the causes of base composition heterogeneity, variation in base composition of different DNA classes needs to be studied. One of the striking patterns of genome-wide base composition heterogeneity in *D. melanogaster* genome is the difference in base composition between X chromosome and autosomes (Singh 2005b; Singh et al. 2008; Vicoso et al. 2008; Campos et al. 2013). The X chromosome has significantly higher GC content at the synonymous sites than autosomes. This pattern was independent of the gene expression level, gene length, recombination rate, gene density and gene identity and has been implicated as a consequence of more effective selection operating on the X chromosome (Singh 2005b; Singh et al. 2008; Vicoso et al. 2008;

Campos et al. 2013).

In this chapter, I investigate the genome-wide variation in base composition of four nucleotide classes, small introns, long introns, 2-fold synonymous sites and intergenic sites in *Drosophila melanogaster* genome. I show that base composition varies at a genome-wide scale. The base composition of different functional classes of DNA show different patterns enabling me to determine the causes of base compositional heterogeneity. I identified certain regions of the genome that show higher heterogeneity than the rest of the genome. I also report that the higher GC on the X chromosome is only restricted to synonymous sites and small introns.

3.2 Materials and Method

3.2.1 Sequence data

Sequence data and annotations for *Drosophila melanogaster* genome (Release 5.28, June 4, 2010) (Adams et al. 2000) were obtained from FlyBase (www.flybase.org). Only genes that were predicted to have single protein isoforms were used for the analysis.

3.2.2. Recombination rate estimates

The *D. melanogaster* genome recombination rate estimates calculated by Fiston-Lavier et al (Fiston-Lavier et al. 2010) were used in this study. Recombination rate estimates assigned to the locations of single genes on the chromosome were used.

3.2.3 Codon bias and GC content measures

Major codon usage (MCU) for 2-fold redundant codon families (except Asp) was used as a measure for codon bias. MCU is defined as the percentage of major codons in a gene. Major codons classifications were obtained from a previous analysis (Akashi unpublished) for 904 X-linked genes and 5330 autosomal genes.

Introns with length less than 100 bp were classified as small introns and introns with length greater than 100 bp as long introns. The first 10 and the last 30 bases of the small introns were removed while calculating GC content of small introns in order to exclude potential splice sites and polypyrimidine tract (Chapter 2). Nucleotides present in the 11-100 bp region of long introns were used for calculating GC content of long introns (Chapter 2).

Intergenic DNA was defined as the regions that were not annotated as coding regions, introns or UTRs in the *D. melanogaster* genome (Release 5.28, June 4, 2010). Conserved sharp changes in GC content are observed near the 5′ and 3′ ends of Drosophila genes (Zhang et al. 2004). To avoid such potential functional regions, 800 bp around UTRs were also removed while calculating the GC content of intergenic regions.

3.2.4 GC comparison among chromosome arms

GC content was compared among chromosome arms for various nucleotide classes using Mann-Whitney U test (with continuity correction) implemented in R version 3.2.3 (https://www.r-project.org/). Only genes that had more than 30 sites of a given nucleotide class were included in the analysis. All autosome data were pooled for comparison between X and autosomes. For comparison among autosomes, GC measures of each chromosome arm were compared to that of the remaining autosomal chromosome arms. Data from chromosome 4 were not included. Multiple-test correction was performed using Bonferroni-sequential method. Boxplots were generated using “boxplot” function implemented in R version 3.2.3 (https://www.r-project.org/).

3.2.5 Statistical analyses: Heterogeneity estimation

Heterogeneity within a chromosome can be estimated by identifying the proportion of the chromosome that has high or low MCU/GC content relative to the rest of the chromosome. To identify such regions, G-statistic was calculated for each bin using the formula,

$$G = -2\sum O * ln(\frac{E}{O})$$

where, O is the observed frequency and E is the expected frequency. To control for the difference in sample size among bins and nucleotide classes, G-statistic was calculated for bins with equal number of sites. Each chromosome was first divided into non-overlapping blocks of 200Kb nucleotides. Only the blocks that had at least 500 sites of a given nucleotide class were used. To control for difference in the total number of sites among nucleotide classes, 500 sites for each nucleotide class were randomly selected from each block and the GC content of the randomly selected sites was used to represent GC content of the block. G-statistic was calculated for each block. The frequencies of GC and AT sites within each bin were used as the observed frequencies and those outside the bin in the chromosome arm were used as the expected frequencies.

The G-statistic distributions were compared across nucleotide classes and chromosome arms. For the comparison among nucleotide classes, the data for all nucleotide classes across all chromosome arms were first pooled. The 0th, 25th, 50th, 75th and 100th quantiles of the G-statistic of the pooled data were then calculated. The proportion of blocks having G-score between the 0th and 25th quantile, 25th and 50th quantile, 50th and 75th quantile, and 75th and 100th quantile in the individual nucleotide classes were calculated and compared among different nucleotide classes. A similar approach was used to compare G-statistic distributions among chromosome arms. In this case, data for single nucleotide classes across different chromosome arms were used. The distributions were compared using two-sample Kolmogorov-Smirnov test implemented in R version 3.2.3 (https://www.r-project.org/).

3.2.6 Correlation between recombination rate and GC measures

Spearman’s Rank Correlation Coefficient was used to measure correlation between recombination rate and various GC measures using “distancematrix” function of Bio.Cluster package in python 2.7.8. Only the genes that had at least 30 sites of a given nucleotide class were used to calculate the correlation between the GC content of various nucleotide classes. The statistical significance of the correlation was computed by conducting a bootstrap across genes. Genes in the original dataset were resampled to generate 100,000 bootstrap datasets. Correlation coefficients were calculated for each dataset and 95% confidence interval of the distribution of correlation coefficients was computed. Correlation coefficients were also calculated after removing regions of no recombination identified by Kliman and Hey (Kliman and Hey 1993).

3.3 Results

3.3.1 GC content comparison between X chromosome and Autosomes

In the previous chapter, I established that base composition varies at a within-gene scale in *D. melanogaster*. Previous studies have shown that base composition in *Drosophila* also varies at a genome-wide level (Carulli et al. 1993). It is well documented that codon bias is higher for the genes on the X chromosome than for those on the autosomes (Singh 2005b; Singh et al. 2008; Campos et al. 2013). This pattern has been implicated as consequences of more effective selection operating on the X chromosome than on the autosomes. Consistent with these reports, I found that

MCU is significantly higher on the X chromosome compared to autosomes (Table 3.1, Figure 3.1). Among intron GC measures, small intron GC content is significantly higher on X, as reported earlier (Campos et al. 2013) (Table 3.1, Figure 3.1). However, I found that long introns and intergenic DNA have similar GC content on X and autosomes (Table 3.1, Figure 3.1). Intergenic GC content shows variability among autosomes. Right chromosome arms (2R, 3R) tend to have higher intergenic GC content than left chromosome arms (2L, 3L) (Figure 3.1).

Since selection pressures might differ between the 5p750 and post-5p750 regions of the genes, I compared the GC content of 5p750 and post-5p750 regions between X and autosomes separately. MCU and small intron GC content are higher on X compared to autosomes for both 5p750 and post-5p750 regions (Figure 3.2 A, B, Table 3.1). Long introns have similar GC content for both 5p750 and post-5p750 regions on X and autosomes (Figure 3.2 A, B, Table 3.1).

3.3.2 Regional heterogeneity in GC content in the D. melanogaster genome

GC content for synonymous sites and small introns varies across chromosome arms, especially between X and autosomes. To test if the GC content also varies within chromosome arms, the chromosomal GC variation along the chromosome arms was studied.

To estimate the heterogeneity in GC content within chromosome arms, the regions with significantly higher or lower GC content than the rest of the chromosome were identified. MCU varies at a regional-scale on both X chromosome and autosomes (Figures 3.3). The sub-centromeric and sub-telomeric regions present at the ends of the chromosome arms tend to have high MCU. A region with strikingly low MCU is observed between 15-16Mb on the X chromosome. The synonymous GC content of this region is almost as low as that of intergenic regions. The variation in MCU is accompanied by that in small intron GC content, which has peaks and dips in similar regions as MCU (Figures 3.3). A very few regions have significantly different long intron GC content than that of the rest of the chromosome (Figures 3.3). Intergenic GC content, on the other hand, is variable in many regions (Figures 3.3). However, it should be noted that these departures are sensitive to the sample size and intergenic DNA class has larger sample size than any other nucleotide class.

While comparing heterogeneity among nucleotide classes, classes that have larger sample size would have more statistical power than the ones that don't. Hence, to control for the sample size, G-statistic was calculated for equal number of sites, which were randomly chosen from 200Kb non-overlapping blocks of chromosome, for each nucleotide class (see Methods). The distributions of G-statistic were compared among nucleotide classes. MCU was found to be the most heterogeneous GC measure, followed by long intron GC content (Table 3.2, Figure 3.4). Intergenic GC content was the least heterogeneous among all nucleotide classes (Table 3.2, Figure 3.4). Although the average G-statistic for X genes is higher than that for autosomal genes, the distributions of G-statistics on X and autosomes are not statistically different (Table 3.2).

3.2.3 Correlation of GC content with recombination rate

Regional heterogeneity in base composition could be explained by variation in the strength of selection, mutation bias or biased gene conversion. Some previous studies have focused on recombination to explain the variation in codon bias (Comeron et al. 1999; Hey and Kliman 2002). Recombination is expected to increase the efficacy of selection by reducing the linkage between selected sites (Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988). Recombination is also known to associate with mutation bias and biased gene conversion (MARAIS et al. 2003). Overall positive correlation between recombination rate and codon bias has been observed in *Drosophila melanogaster* genes (Comeron et al. 1999; Marais and Mouchiroud 2001; Hey and Kliman 2002; MARAIS and Piganeau 2002; MARAIS et al. 2003; Campos et al. 2013). However, some studies report that recombination rate and codon bias on X were negatively correlated (Singh et al. 2005; Campos et al. 2013). This result is in contrast with the expected pattern and remained unexplained.

Our results from the within-gene analysis indicated that the 5p750 region (up to 750 bp from TSS) and the post-5p750 region (from 751 bp to end of the gene) of the gene experience different mutational and/or selection pressures. Hence, their relationship with recombination rate might also be different. So, I tested the correlation between recombination rate and MCU and GC content for 5p750 and post-5p750 regions separately. MCU of both 5p750 and post-5p750 regions is significantly negatively correlated with recombination rate on X but has no significant correlation with recombination rate on autosomes (Table 3.3). The negative correlation between MCU and recombination rate on X is higher in the 5p750 region than the post-5p750 region. After removing regions of no recombination, the relationship between recombination rate and MCU on X still persists (Table 3.4). Similar to codon bias, small intron GC content is also negatively correlated with recombination rate in the 5p750 region but not in the post-5p750 region on X (Table 3.3). No significant correlation is detected between small intron GC content and recombination rate on autosomes (Table 3.3). In contrast to the small intron GC pattern, the GC content of long introns in post-5p750 region is significantly positively correlated with recombination rate on autosomes but no significant correlation is observed on the X chromosome (Table 3.3). After removing regions of no recombination, no significant correlation between recombination rate and long intron GC content is observed (Table 3.4).

In the previous section, I observed that the region between 15-16Mb on the X chromosome has unusually low codon bias. To test if this region contributes to the negative correlation between codon bias and recombination rate on the X chromosome, I calculated the correlation after excluding the genes present in this region. After removing the genes in this region, no negative correlation is observed between recombination rate and codon bias for either 5p750 or post-5p750 region of genes.

3.4 Discussion

In this chapter, I show that base composition varies at a genome-wide scale in the *D. melanogaster* genome. I observe among as well as within chromosome heterogeneity in base composition, which cannot be explained by the differences in recombination rate. It is already known that codon bias and intron GC content is higher on the X chromosome, compared to the autosomes (Singh et al. 2005; Vicoso and Charlesworth 2006; Singh et al. 2008; Campos et al. 2013). Here, I show that this is only true for MCU and small intron GC content but not long intron GC content or intergenic GC content. I also show that MCU and small intron GC content are higher on X compared to autosomes for both 5p750 and post-5p750 regions. As suggested in previous reports, stronger efficacy of selection on X compared to autosomes could explain higher MCU on X compared to autosomes (Singh et al. 2008). The small intron base composition in the 5p750 region of genes might also be under transcription-associated selection (Chapter 2). Hence, stronger purifying selection could also explain higher 5p750 small intron GC content on X compared to autosomes. The higher post-5p750 small intron GC on X, however, remains to be explained.

Eukaryotic genomes are highly variable in nucleotide composition (Bernardi 2000; Nekrutenko and Li 2000; Tang et al. 2006; Frenkel et al. 2012). Among multicellular eukaryotes, yeast has the least heterogeneity at genomic scale in terms of base composition followed by plants (Nekrutenko and Li 2000). Among vertebrates, primates and ungulates have high composition heterogeneity compared to other vertebrates (Bernardi 2000; Frenkel et al. 2012). Fish have the least heterogeneity in base composition among vertebrates (Frenkel et al. 2012). Human genome has exceptionally high heterogeneity compared to other eukaryotes (Bernardi 1989; Bernardi 1993; Bernardi 2000; Nekrutenko and Li 2000; Frenkel et al. 2012). Previous studies have identified heterogeneity in base composition in the *Drosophila melanogaster* genome as well(Carulli et al. 1993). When compared to other eukaryotes, GC content in *Drosophila* is found to be less heterogeneous than that in mammals but more heterogeneous than that in worms and yeast (Nekrutenko and Li 2000; Tang et al. 2006). I found that among the different GC classes, MCU is the most heterogeneous GC measure and intergenic GC content is the least heterogeneous, suggesting that selective forces are vary at a shorter scale than mutational forces in *D. melanogaster* genome. This was true after accounting for the difference in the sample size of each nucleotide class. Vertebrates show interchromosomal heterogeneity in base composition (Frenkel et al. 2012). However, in *D. melanogaster* the difference in the base composition heterogeneity was not significant among chromosome for most GC classes.

Correlation with recombination rate can help in determining the cause of base composition variation. If base composition variation is caused by variation in selection or biased gene conversion, GC content is expected to positively correlate with recombination rate (Hill and Robertson 1966; Felsenstein 1974; Birky and Walsh 1988; MARAIS et al. 2003). Previous studies have reported a negative correlation of GC content at synonymous sites and introns with recombination rate on X chromosome and a positive correlation on autosomes (Singh et al. 2005; Campos et al. 2013). However, these studies did not distinguish sites by their position in the transcript and the length of the intron they belonged to. I found that recombination rate was negatively correlated with GC content at synonymous sites on the X chromosome in both 5p750 and post-5p750 regions. The magnitude of correlation was higher in the 5p750 regions. Among other GC classes, recombination rate was negatively correlated with small intron GC content only in the 5p750 region. On the autosomes, no significant correlation between any GC class and recombination rate was observed. On the X chromosome as well, the negative correlation seems to be sensitive to certain outlier regions and might not be a meaningful pattern. Hence, recombination rate might not be a strong determinant of the efficacy of natural selection. Expression level, as discussed in the previous chapter, is a better predictor of natural selection at the synonymous sites.

CHAPTER 4

Lineage-specific genome evolution in the *Drosophila melanogaster* subgroup

4.0 Chapter summary

To study the base composition variation across different genomes, I studied lineage-specific codon bias evolution in seven *Drosophila melanogaster* subgroup species. I used existing genome data for five species and added data for two of the *D. melanogaster* subgroup species through Next-Generation RNA sequencing of *D. tessieri* and *D. orena* transcriptomes. I described a protocol for gene annotation of the RNA-seq data using the available data from the sequenced species. Ancestral states were inferred using maximum likelihood approaches that account for both base composition bias and non-stationarity and assigned substitutions to 10 lineages. All lineages showed departures from equilibrium and in some cases multiple factors appeared to have fluctuated. These findings suggest that the magnitudes of forces governing base composition at synonymous sites may have varied frequently in a lineage-specific manner in the *D. melanogaster* subgroup and may need to be taken into account when testing evolutionary mechanisms at other classes of sites.

Comparing classes of DNA evolving under different selective constraints can reveal the underlying evolutionary mechanisms of lineage-specific changes in base composition. Variability of base composition caused by changes in selection intensity would be higher for regions under stronger selection than for regions under weak or

no selection. The effect of mutation, however, would be similar for all DNA classes. We plan to examine the lineage-specific changes in base composition of small introns in seven *Drosophila melanogaster* subgroup species and compare them to that of synonymous sites.

The genome data from five of the sequenced *Drosophila melanogaster* subgroup species will be employed for this study. In addition, we also sequenced the genomes of two more species in the subgroup, *Drosophila tessieri* and *Drosophila orena,* using Next-Generation sequencing techniques. We mapped the RNA-seq data from the two species, which we previously sequenced and analyzed, to their assembled genomes to identify exon-intron junctions. Using the positions of the exon-intron junctions, the intron sequences were extracted from the genomes. We were able to annotate more than 15,000 introns for around 5000 genes from each of the species. We also analyzed sequencing data from 20 inbred lines of *D. simulans* and *D. yakuba* in order to obtain polymorphism data to study recent evolution.

4.1 Introduction

Evolutionary parameters such as mutation rates and biases, recombination rates, effective population size and fitness effects of mutations influence the substitutions occurring in a lineage. If these evolutionary parameters change very slowly, they can cause heterogeneity in the substitution patters due to variation in evolutionary forces (Gillespie 1993; Gillespie 1994). Several studies in Drosophila provide evidence for lineage-specific evolution, suggesting temporal fluctuations in evolutionary forces. For instance, differences in rates of evolution have been observed in closely related species in Drosophila (Takano 1998; Akashi 1999; Akashi et al. 2006; Nielsen et al. 2006). Some studies have also identified several genes evolving in a lineage-specific manner (Clark et al. 2007; McBride 2007; McBride et al. 2007). Lineage-specific base composition evolution have also been observed in coding and non-coding regions for small-scale data (Akashi et al. 2006; Singh et al. 2007) and some species (Kern 2004; Singh et al. 2009) in the *D. melanogaster* subgroup.

I studied lineage-specific codon bias evolution in seven *Drosophila melanogaster* subgroup species to determine in the time-scale and magnitude of fluctuations in evolutionary forces at the genomic level. I examined the changes in the codon bias patterns in more than 5000 genes from seven *D. melanogaster* subgroup species. I employed existing genome data for five species (*D. melanogaster, D. sechellia, D. simulans, D. yakuba* and *D. erecta*) and added data for two more species in the subgroup (*D. teissieri* and *D. orena*) through Next Generation sequencing. Adding sequence data from the two species not only gives me a bigger number of lineages to study but also enables reliable ancestral inference at the internal nodes. All these species show strong patterns of codon usage bias. These species have a well-supported phylogeny and their branch lengths are short enough to make reliable ancestral inference. I inferred ancestral states at the interior nodes and assigned substitutions to ten lineages (seven terminal and three internal). All these lineages showed strong departures from the equilibrium states and in some cases multiple factors seemed to have fluctuated.

4.2 Methods

4.2.1 Coding Sequence Data

In this study, I used available coding sequence data from Flybase (www.flybase.org) for *D. melanogaster* (Adams et al. 2000) (Release 5.28), *D. sechellia* (Release 1.3) (Clark et al. 2007), *D. simulans* (Release 1.3) (Clark et al. 2007), *D. yakuba* (Release 1.3) (Clark et al. 2007) and *D. erecta* (Release 1.3) (Clark et al. 2007). In order to obtain the coding sequence data for *D. teissieri* and *D. orena*, Next-Generation RNA-sequencing of their transcriptomes at four developmental stages: Larva, Pupa, Adult male and Adult female, was performed in the Akashi lab. A paired-end library for each sample was created and sequenced using Illumina Hi-Seq. Two replicates of larval samples (larva rep1 and larva rep2) were sequenced to check for consistency.

I filtered the RNA-seq reads for adapters, which are short nucleotide sequences ligated to both ends of the DNA fragment during library preparation, by using a publically available tool, cutadapt (Martin 2011). I trimmed the sequence at the 3′ end of the reads that overlapped with the adapter sequences by even 1 bp. If a full or partial adapter was found at the 5′ end of a read, the whole read was discarded. The minimum overlap for the partial adapter in this case had to be 9 bp. Table 4.1 shows the summary statistics of the adapter filtering.

I trimmed low quality bases that had a Phred quality score of 25 or less from the reads using the DynamicTrim program of the SolexaQA software package (Cox et al. 2010). This program extracts the longest contiguous sequence in individual reads where all bases have a quality score higher than the threshold (Phred quality score of 25 in my case). After trimming the low quality bases, the reads that had a length of less than 25 bp were discarded long with their paired reads using LengthSort, another program from the SolexaQA package.

I also checked for rRNA contamination in the reads and removed the reads that map to rRNA sequences. I used Bowtie (parameters: default alignment parameters; reported alignments –k 3) (Ben Langmead 2012) to map reads to rRNA sequences from *D. melanogaster*, *D. yakuba* and *D. erecta*. The reads that mapped to at least one rRNA sequence were removed from the analysis. The summary of rRNA filtering of the reads is shown in Table 4.2.

The filtered RNA-seq reads from individual tissue samples of the two species were assembled using a *De novo* transcriptome assembler, Trinity (Haas et al. 2013). Trinity assembles the reads into relatively longer DNA sequences called contigs. The summary of the transcriptome assembly is shown in Table 4.4.

4.2.2 Sequence ortholog assignment for the newly sequenced species

The contigs from the newly sequenced species, pooled together across tissue samples, were matched to amino acid sequences from their closest species, *D. yakuba* and *D. erecta* using blastx (McGinnis and Madden 2004) to identify candidate ortholog sequences. Blastx matches of length greater than 33 amino acids and sequence identity more than 20.0% were used to assign candidate orthologs. To test for type II error (false negatives), I matched *D. erecta* gene sequences to *D. yakuba* amino acid sequences using blastx. I then checked how many real orthologs were discarded by filtering ~~out~~ matches that had less than 20.0% identity or were less than 33 amino acids in length. From this test, I obtained a false negative fraction of 0.06 assuming one-to-one orthologs.

Although *D. teissieri* is phylogenetically closer to *D. yakuba* than *D. erecta* and *D. orena* is phylogenetically closer to *D. erecta* than *D. yakuba*, the candidate orthologs for *D. teissieri* were searched from *D. erecta* sequences and that for *D. orena* were searched from *D. yakuba* sequences. This was done so that the expected pairwise amino acid distance between *D. erecta - D.yakuba* orthologs could be used as a cut off for filtering matches since the estimates for the expected pairwise distance between the *D. orena* and *D. erecta* sequences and that between *D. teissieri* and *D. yakuba* sequences are not available. *D. orena* contigs were aligned to their *D. yakuba* matches and *D. teissieri* contigs to their *D. erecta* matches using MUSCLE (Edgar 2004). The pairwise distance for each amino acid alignment using GRANTHAM distance matrix (Grantham 1974). The GRANTHUM amino acid distance of 12.5, which is the 95% quantile of the distribution of the amino acid distance between *D.*

erecta - D.yakuba orthologs, was defined as a cut off to filter the matches.

If the contigs still had multiple matches, the matches that had the highest conservation were assigned as candidate orthologs. This was done by bootstrapping the alignments between the contigs and their matches 1000 times by resampling the codons and calculating amino acid distance for each replicate. The matches were then ranked by their amino acid distance for each replicate. The match that had rank 1 at least 90% of the times was assigned as the ortholog match (Table 4.3). If the matches could not be resolved, the contig was excluded from the analysis.

4.2.3 Ortholog sequence alignments

The alignments of 8933 orthologous genes from the five already sequenced species (*D. melanogaster, D. sechellia, D. simulans, D. yakuba* and *D. erecta*) were obtained from Hiroshi Akashi. The orthologous contigs of those 8933 genes were identified in *D. teissieri* and *D. orena*. The amino acid alignments of the contigs and their orthologs were converted to nucleotide alignments and added to the five species alignment set after adjusting the gaps and unknown parts of the sequence. For the regions of the genes that mapped to more than one contig, each nucleotide of the overlapping contigs was checked for consistency and the conflicting ones were replaced by ‘N’.

For the downstream analysis, the regions of the alignments (seven species alignment set) that had gaps or unknown nucleotides (Ns) were removed. The alignments that have stop codon within the sequences were trimmed until the stop codon. The genes that had dN or dS values between any two species in the mcstyeo dataset greater than 1 or equal to -1 were also excluded from the analysis. The dN and dS values were calculated using the Nei-Gojobori method (Nei and Gojobori 1986) implemented in the PAML software (Yang 2007).

4.2.4 Binning Data

The genes in the alignments were distributed into bins by their MCU values (Major Codon Usage in *D. melanogaster*). This was done for three categories of codons: all codons, conserved 2-fold codons except Asp codons and conserved 4-fold codons. Before distributing the genes into bins, the genes that had same MCU values were concatenated and treated as a single gene. Each bin should have approximately equal number of codons. So, I first calculated the estimated number of codons in each bin by dividing the total number of codons by the total number of bins. If the number of codons in the bin exceeded the estimated number of codons per bin while adding a codon to the bin, the number of extra codons that need to be accommodated into the bin were calculated. If the number of extra codons was more than half of the codons of the gene, the codon was added to that bin; otherwise it was left for the next bin.

4.2.5 Ancestral Inference

A maximum likelihood approach was used to infer the ancestral states. For maximum likelihood, two types of substitution models implemented in the PAML software (Yang 2007) were used in my analysis: GTR-H and GTR-NH$_b$ (Tavare, 1986) (Matsumoto et al. 2015). GTR substitution model defines transition probabilities between two nucleotides as a function of substitution rate parameters and equilibrium nucleotide frequencies. Under the GTR-H model, the equilibrium nucleotide frequencies are same for all lineages since they do not change over time. The substitution rate parameters are also same for all the lineages. Under the GTR-$\mathrm{NH_b}$ model, both the base compositions and the substitution rate parameters change over time. I used the average of multiple reconstructions weighted by their posterior probabilities. The detailed description of ancestral inference methods can be found in Matsumoto *et. al.*(Matsumoto et al. 2015).

4.2.6 Measure of departure from equilibrium

Under the MCP model, synonymous substitutions could be advantageous (unpreferred to preferred) or slightly deleterious (preferred to unpreferred) (Akashi 1995). In general GC-ending codons are preferred and AT-ending codons are unpreferred in Drosophila. I used a skew of the unpreferred to preferred changes ($d_{up,pu}$), described in Akashi *et al* 2006, to measure the direction and magnitude of departures from equilibrium.

$$d_{up,pu} = \frac{up - pu}{up + pu}$$

where *up* is the number of unpreferred to preferred changes and *pu* is the number of preferred to unpreferred changes.

Computer simulations have shown that the expected $d_{up,pu}$ changes when codon bias changes due to altered N_es or mutation pressure, where N_es is the product of effective population size and selection coefficient for a particular lineage (Akashi et al. 2007).

4.2.7 Genome Sequencing

To obtain the non-coding sequence data for *D. teissieri* and *D. orena*, the genomes of these two species were sequenced using Next-Generation genome sequencing technique. Female flies from *D. teissieri* and *D. orena* were sequenced in the Akashi Lab by constructing Illumina paired-end libraries with peak insert sizes 472 bp and 462 bp, respectively. The DNA fragments were sequenced using Illumina HiSeq 2000. I created the pipeline to analyse the genome sequence data using a combination of available softwares and custom-made codes.

The reads were filtered for adapters using a publically available tool, cutadapt (Martin 2011). The sequence at the 3′ end of the reads that overlapped with the adapter sequences by even 1 bp was trimmed and the reads that had at least 9 bp overlap to the adapter sequence at the 5′ end were discarded. Less than 0.5% of the bases were discarded in this process (Table 4.4). Low quality bases with Phred quality score of less than 25 were trimmed (cut) using DynamicTrim program of the SolexaQA software package(Cox et al. 2010). After the filtering, the reads with lengths less than 63 bp were discarded along with their paired reads. The summary statistics of quality filtering are given in Table 4.5.

The filtered reads were assembled into contigs using SOAPdenovo2(Luo et al. 2012), a *de novo* genome assembler that uses a De Brujin graph approach, with average insert size setting of 346 and 351 for *D. orena* and *D. teissieri* reads, respectively. The minimum aligned length was set to 64 and maximum read length and read length cutoff were set to 100 each. A total of 312,596 and 638,580 contigs (contigs$_G$) were obtained for *D. teisseiri* and *D. orena*, respectively. The total length of *D. teissieri* and *D. orena* assemblies are 157,463,559 bp and 186,280,924 bp and their respective N50s are 15,822 bp and 8972 bp. The contigs$_G$ were assembled into scaffolds using a scaffolding tool, SSPACE (Boetzer et al. 2011) with expected insert size set to 400 and minimum allowed error rate set to 0.25. Several statistics were used to assess the quality of a genome assembly, such as contig or scaffold N50, total length of the assembly, amount of Ns in the assembly, etc (Table 4.6). These statistics were calculated using a genome assembly comparison tool, Quast (version 2.3) (Gurevich et al. 2013).

Another key factor of a good assembly is that most genes should remain unbroken, that is, most genes should be present entirely on one scaffold. This would allow me to annotate introns more easily. Contigs annotated from an RNA-seq experiment (contigs$_R$) were compared against the scaffolds using blastn to find the number of genes that match to only one scaffold. The summary statistics of the genome assembly are given in Table 4.6. The repeat sequences in the scaffolds were masked using RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html) by taking *D. melanogaster* repeat sequences (genome release 5, downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/chromTrf.tar.gz) as a reference. RepeatMasker also masks low complexity sequences along with interspersed repeats.

Another way of assessing the assembly quality is to check large-scale synteny of contigs and scaffolds. In *Drosophila* species, six chromosome arms, known as the Muller's elements, are mostly common across species. To check if the Muller's elements are conserved between *D. teissieri* and *D. orena* and other species of the *D. melanogaster* subgroup, I mapped the *D. teissieri* and *D. orena* contigs$_G$ and scaffolds to the *D. melanogaster* genome using MUMmer (nucmer algorithm with following parameters for scaffolds, --maxgap=500 --mincluster=100, and default parameters for contigs)(Delcher et al. 2002) to check how many of the matched contigs$_G$ and scaffolds match to only one chromosome arm. More than 90% of the matching contigs$_G$ matched to only one chromosome arm in both the species. At the scaffold level, the percent matching to only one chromosome arm was lower than that at the contig level. More than 80% of the scaffolds matched to one chromosome (Table 4.7). In order to get an estimate of the expected inter-chromosomal rearrangement, I also mapped the *D. yakuba* genome to *D. melanogaster* genome. 63% of *D. yakuba* chromosome arms matched to only one *D. melanogaster* chromosome arm (Table 4.7).

4.2.8 Intron identification

I designed a strategy that makes use of the RNA-seq data to identify and annotate introns. To find introns in the scaffolds, I used TopHat2 (Kim et al. 2013), a read mapper that recognizes exon-intron junctions. The minimum length of introns specified was 50 bp, which might be small enough to include most introns. For reference, in *D. melanogaster*, only 282 out of 85191 introns (0.3%) are less than 50 bp in length. TopHat2 outputs the position of the splice junctions along with their flanking sequences. In order to map the introns to the genes, I used the flanking sequences to identify the contigs$_R$ that contain the splice junctions. This was done by matching the junction flanking sequences to annotated contigs$_R$ using blastn. Among the junctions reported by TopHat2, junctions associated with the flanking sequences that did not match to even one annotated contig$_R$ on the same strand with at least 98% identity were removed. The junctions that matched to multiple contigs$_R$ belonging to different genes were also filtered out. The summary statistics of the whole process are shown in Table 4.8.

The junctions that are spanned by very low number of alignments might be unreliable. I address the number of alignments spanning a junction as the junction depth. To decide the cutoff for the junction depth, I used the same intron finding strategy to identify introns in *D. erecta* and *D. yakuba* using their respective larva RNA-seq data. Using the distribution of the junction depth of the introns that were present in the *D. erecta* and *D. yakuba* genomes (release 1.3) (annotated) and of the ones that were not (unannotated) (Figure 4.1), I chose the junction depth of 5 as the cutoff. After removing all introns that had junction depth of 5 or below, around 80% of the remaining introns were annotated introns.

After filtering the intron junctions based on their junction depth, the introns were classified into ones that overlapped with other introns and ones that did not. Among the ones that did not overlap with any other introns, constitutive introns of multiple isoform genes were differentiated from the introns present in single isoform genes (Table 4.9).

4.2.9 Identifying orthologous intron sequences

To identify the orthologs of the putative introns in their closest species, I compared the flanking sequences of the introns to the CDSs from the closest species using discontiguous megablast. Flanking sequences on both sides of each intron (left flanking sequence or LFS and right flanking sequence or RFS) were required to match to continuous segments of the same CDS (Single LFS and RFS connected match). If either of the flanking sequences matched to multiple continuous segments of the same CDS or no continuous segment of the CDS, the intron was discarded. If an intron was present between the segments of the CDS that match to the left and right flanking sequence, it was assigned as the ortholog to the given putative intron. Among the introns overlapping with each other (non-constitutive introns), only the intron that had single LFS and RFS connected matches was assigned ortholog. If multiple introns in a set of overlapping introns had single LFS and RFS connected matches, the whole set was discarded. In *D. melanogaster*, around 98% of the introns have canonical splice sites. I also checked how many putative introns in *D. teissieri* and *D. orena* have canonical splice sites. The summary of the process in described in the Table 4.10.

4.2.10 Intron Alignment

Introns identified in *D. teissieri* and *D. orena* were aligned to orthologous introns from five already sequenced species (*D. melanogaster, D. sechellia, D. simulans, D. yakuba* and *D. erecta*), obtained from a previous study (unpublished data) using MUSCLE (nucleotide alignment option) (Edgar 2004). The introns alignments were classified into small (<=100 bp) and long introns (>100 bp) based on the length of the intron in *D. melanogaster*. The number of introns in each size category is listed in Table 4.11.

The regions of the alignments that had gaps or unknown nucleotides (Ns) were removed. The pairwise percent identity for each intron alignment was calculated. The pairwise percent identity of small introns lies between 45% and 100%. For further analysis, the first two and the last two nucleotides of each intron (splice sites) were trimmed.

4.2.11 Polymorphism data processing

Sequencing reads from Illumina paired-end sequencing for 20 isofemale inbred lines of *D. simlans* and *D. yakuba* were obtained from a study by Rogers et al (Rogers, Cridland, et al. 2014). Adapter sequences that overlapped with the read by even 1 bp at the 3′ end were trimmed using cutadapt (Martin 2011). This resulted in trimming about 27% of R1 and about 35% R2 reads from all samples. The reads from *D. simulans* were mapped to two *D. simulans* reference genomes (Clark et al. 2007; Hu et al. 2013) and those from *D. yakuba* were mapped to one *D. yakuba* reference genome (Clark et al. 2007) using bwa version 0.5.9 (Li and Durbin 2009).

The sam files obtained from bwa were converted into bam files using samtools version 0.1.18(Li et al. 2009). Picard tools (**http://broadinstitute.github.io/picard**) was used to group reads, identify and fix paired reads and mark duplicates. The processed alignments were then analyzed using GATK version 3.6 (McKenna et al. 2010) for indel realignment and variant calling. Base recalibration was skipped because of the absence of database of SNPs for my species. So, I first called variants using the HaplotypeCaller command of GATK and then used these variants to construct a SNP database for the species. After the first round of variant calling, variants with a quality score in the top 10% were extracted from each sample and a high quality variant database was constructed. This database was used for base recalibration and the recalibrated alignments were used to call variants again.

4.3 Results

4.3.1 Transcriptome sequencing and de novo gene annotation

I analyzed the RNA-seq data obtained by sequencing the transcriptome from four developmental stages: larva, pupa, adult female and adult male of *D. teissieri* and *D. orena* using Illumina HiSeq. The RNA-seq reads were filtered for adapters, low quality bases (Phred score < Q25) and rRNA contamination. The filtered reads were assembled using Trinity(Haas et al. 2013). The *de novo* assembly obtained from Trinity comprised of a total of 346,023 and 349,581 contigs across the four developmental stages for *D. teissieri* and *D. orena*, respectively. The summary of the transcriptome assembly for each sample is shown in Table 4.12.

I developed some methods and strategies to assign orthologs to the assembled RNA-seq contigs. The RNA-seq contigs from separate tissue samples were pooled together and annotated using the available data from other Drosophila species in the subgroup. The *D. teissieri* contig sequences were compared against transcript sequences from *D. erecta* and *D. orena* contig sequences were compared against those from *D. yakuba*. I used the available estimate of the distance between *D. yakuba* and *D. erecta* sequences to filter the matches and assign candidate orthologs to the assembled reads (see Methods). Only the genes that had single isoform in *D. melanogaster* were used for comparisons.

I assigned orthologs to 85,906 *D. teissieri* contigs and 85,399 *D. orena* contigs covering 8,050 and 7,732 genes from *D. teissieri* and *D. orena*, respectively. A total of 5560 genes (2,007,995 codons) had orthologs in all seven *D. melanogaster* subgroup species, which were used to construct ortholog alignments and infer ancestral states.

4.3.2 Lineage-specific departures from equilibrium

A total number of 5026 genes were divided into bins containing roughly equal number of codons based on their MCU values,. The binning was done separately for 2-fold codons and 4-fold codons. 2-fold and 4-fold codon families were analyzed separately because these two classes of codons might be under different selection for codon usage. Genes in each bin were aligned among the seven Drosophila species. The substitutions (both silent and replacement) were inferred for a seven terminal (*D. melanogaster, D. sechellia, D. simulans, D. teissieri, D. yakuba, D. erecta, D. orena*) and three internal lineages (*D. sechellia – D. simulans, D. teissieri – D. yakuba, D. erecta – D. orena*) using maximum likelihood under GTR-H and GTR-NH$_b$ models (see Methods). The phylogeny of the *D. melanogaster* subgroup is shown in Figure 4.2. The silent substitutions were classified into preferred-to-unpreferred (*pu*) and unpreferred-to-preferred (*up*) changes. In general, GC-ending codons are preferred in Drosophila. A up skew ($d_{up,pu}$) statistic was calculated for each bin (Akashi et al.

2006).

The $d_{up,pu}$ statistic is expected to change as a function of MCU (major codon usage) when codon bias changes due to altered N_es or mutation pressure (Akashi et al. 2007), where N_e is the effective population size and s is the selection coefficient. $d_{up,pu}$ for conserved 2-fold codon families (except Asp) decreased as a function of MCU in all the lineages, consistent with a decline in N_es (Figure 4.3). The non-zero $d_{up,pu}$ of the lowest MCU class reflects the change in mutation bias. Negative $d_{up,pu}$ of the lowest MCU indicates an overall increase in the mutation bias towards AT and vice versa. *D. melanogaster*, *D. sechellia* and *D. simulans* lineages show evidence for increase in mutation bias towards AT (Figure 4.3 a-c), whereas an increase in mutation bias towards GC was supported in *D. teissieri, D. teissieri – D. yakuba, D. erecta* and *D. orena* lineages (Figure 4.3 e, g, h, i). No evidence for change in mutation bias was observed in *D. sechellia – D. simulans, D. yakuba* and *D. erecta – D. orena* lineages (Figure 4.3 d, f, j). The patterns in most lineages can be explained by a single change in each of the two parameters, N_es and mutation bias (u/v, u and v are mutation rates from AT to GC and GC to AT, respectively), for the whole genome.

The $d_{up,pu}$ statistic for conserved 4-fold codon families also decreased as a function of MCU in all lineages. The pattern obtained for 4-fold codons was generally similar to that for 2-fold codons. However, prominent differences in mutation bias, shown by the difference in the $d_{up,pu}$ values for the lowest MCU class, could be seen between 2-fold codons and 4-fold codons in *D. erecta* and *D. orena* lineages (Figure 4.4 a, b). A possible contributing factor for this pattern could be that for the 2-fold codons the possible silent substitutions are only transitions (G <-> A, C<->T), whereas the silent substitutions for 4-fold codons also include transversions. This, however, does not completely explain the difference. When $d_{up,pu}$ of 4-fold codons, taking into account only the transitions, was compared to that of the 2-fold codons, the difference for the lowest MCU class became smaller but was still present.

4.3.3 Comparison between ancestral inference methods

A simulations study conducted in my lab has shown that ancestral inference differs considerably among GTR-H model, GTR-NH$_b$ model and maximum parsimony (Matsumoto et al. 2015). According to the simulations study, for genome scale data, non-stationary model provides reliable ancestral inference for complex non-stationary scenario. Patterns obtained in my study using both GTR-H and GTR-NH$_b$ models showed departures from equilibrium with GTR-NH$_b$ model showing a stronger departure (Figure 4.3). Although the total number of parameters under the GTR-NH$_b$ model is more than 100, the variance in most lineages is still small.

4.3.4 Genome sequence assembly and intron annotation

In order to obtain non-coding sections of the genome, I sequenced whole genomes of *D. teissieri* and *D. orena* using Illumina HiSeq 2000 and I analyzed the genome sequence data using a combination of available softwares and custom-made codes. The reads were filtered for adapters and low quality bases (Phred score < Q25). I assembled the reads into contigs using a *de novo* genome assembler, SOAPdenovo2 (Luo et al. 2012). I obtained a total of 312,596 and 638,580 contigs for *D. teisseiri* and *D. orena*, respectively. The total length of *D. teissieri* and *D. orena* assemblies are

157,463,559 bp and 186,280,924 bp and their respective N50s are 15,822 bp and 8972 bp. Scaffolds were constructed from those contigs using SSPACE (Boetzer and Pirovano 2014). The summary of the genome assembly is shown in Table 4.6.

I designed a strategy that makes use of the RNA-seq data to identify and annotate introns. I mapped the reads from the RNA-seq experiment to the genome scaffolds using TopHat2 (Kim et al. 2013). The splice junctions reported by TopHat2 were used to extract intron sequences from the genome scaffolds. The flanking sequences of the splice junctions were matched to the annotated RNA-seq contigs to identify the genes that contain the junctions (see Methods). I annotated 21,512 introns for 5,956 genes from *D. teissieri* and 20,609 introns for 5,657 genes from *D. orena* out of the 8,050 and 7,732 genes annotated from each of the species using the RNA-seq data. Since the annotated RNA-seq contigs are orthologous to single isoform genes in *D. melanogaster*, I expect most of them to monocistronic as well. More than 70% of the annotated RNA-seq contigs contained non-overlapping introns, suggesting they have only one isoform (Table 4.9).

Orthologous introns among species were identified by comparing the flanking sequence of the introns of the newly sequences species to the flanking sequences of introns in the closest species (see Methods). Using this method, orthologs from the closest species were assigned to 13,134 *D. teissieri* introns and 12,386 *D. orena* introns, most of which had canonical splice sites. A total of 7,623 introns had orthologs in all seven species (Table 4.11).

4.4 Discussion

In this chapter, I described a protocol for identifying orthologs to known genes in *D. melanogaster* from RNA-seq data using available data from the sequenced species. The *D. melanogaster* orthologs of the genes annotated from *D. teissieri* and *D. orena* covered genes expressed in all developmental stages (Figure 4.5) and tissues (Figure 4.6). The number of genes was also not biased to any particular MCU class (Figure 4.7). Adding data from *D. teissieri* and *D. orena* enabled me to study the ancestral lineages leading to *D. teissieri – D. yakuba* and *D. erecta – D. orena* and to reliably infer ancestral states at these nodes.

This study showed that codon bias in several lineages of *D. melanogaster* subgroup has strong departures from equilibrium. The direction and magnitude of the departures from equilibrium seem to vary among different lineages. Evolutionary forces governing base composition appear to vary frequently, and often strongly, on relatively short time scales.

Similar to the previous study by Akashi *et al* (Akashi et al. 2006), I found that none of the lineages are at equilibrium. However, some patterns obtained after incorporating a larger dataset are different from the previous analysis. Based on the 22 genes analyzed earlier, *D. teissieri* lineage showed evidence of increase in codon bias whereas *D. simulans* and *D. erecta* lineages did not show consistent departures from equilibrium. In this analysis for 5026 genes, I found that all three above-mentioned lineages show consistent departures from equilibrium and declines in codon bias (Figure 4.3 e, c, h).

The lineage leading to *D. melanogaster* showed the strongest departure from the equilibrium (Figure 4.3 a). Under the non-stationary model, the highest MCU class in this lineage has more than 15 unpreferred changes to 1 preferred change. The $d_{up,pu}$ pattern in this lineage is consistent with a scenario in which N_es decreases by 0.3 fold and mutation bias increases by 2.3 fold. The strongest reduction in N_es is observed in *D. orena* lineage where there is a 0.12 fold reduction in N_es (Figure 4.3 i). This might be due to its small distribution within Africa (Lachaise *et al*, 1988), which might correlate to small population size.

I also found that the magnitude of departure from equilibrium under GTR-H and GTR-NH$_b$ models is different. GTR-NH$_b$ model gives stronger departures from equilibrium in all lineages. This suggests that the choice of ancestral inference methods can have a major effect on molecular evolutionary analyses.

Comparison among classes of DNA evolving under different selective constraints have been used to identify evolutionary mechanisms underlying lineage-specific evolution (Akashi et al. 2006). Hence, comparison of base composition between synonymous sites and small introns can help to distinguish whether variation in translational selection or that in transcription-associated selection or biased gene conversion underlie lineage-specific evolution in *D. melanogaster* subgroup. Previous studies have also observed that non-coding regions in Drosophila are not at equilibrium in several lineages (Kern 2004; Akashi et al. 2006; Singh et al. 2009), which can also be tested in a broader species set by adding non-coding regions from the newly sequenced species in the analysis.

To examine the lineage-specific changes in base composition of small introns in seven *D. melanogaster* subgroup species, I expanded my dataset to include introns from the newly sequenced species. Hence, I sequenced the genomes of *D. teissieri* and *D. orena* through Next-Gen sequencing technique and identified orthologous introns in the seven species. Since, my primary objective is to study molecular evolution in genes and introns that have orthologs in all seven species, I only annotated introns that matched to at least one intron in the closest species. Although my method does not find lineage-specific genes, further gene modeling without using other reference species can be done to identify such genes.

The ancestral states of introns will be inferred from the intron alignments using baseml with GTR-NH$_b$ model (Yang 2007; Matsumoto et al. 2015). Base composition changes between small introns in the 5p750 region and silent sites will be compared. Variation in transcription-associated selection or biased gene conversion would predict similar changes for small introns and silent sites. If the changes in codon bias indicate fluctuation in the intensity of translational selection, changes in the base composition of small introns should be smaller than that in codon bias (Akashi et al. 2006).

The changes inferred from the data so far comprise of fixed as well as polymorphic changes. Since many polymorphic mutations could be slightly deleterious (Ohta 1992), they may not necessarily get fixed. Hence, the use of only single alleles from various species can result in overestimating slightly deleterious changes (preferred-to-unpreferred). In order to resolve these issues, data from other

DNA classes and polymorphism data from a few species needs to be incorporated.

Incorporation of polymorphism data will enable me to study patterns of recent evolution and to distinguish between fixed and segregating changes. After successfully identifying variants from *D. simulans* and *D. yakuba* genomes, these variants would be annotated using available genome annotations. The sequences from different populations of *D. simulans* and *D. yakuba* along with available sequences from *D. melanogaster* populations (Pool et al. 2012) will be aligned to references sequences from *D. melanogaster*, *D. simulans* and *D. yakuba* genomes. The ancestral states will be inferred for polymorphic and fixed sites separately and patterns of polymorphic and fixed differences will be compared. The patterns of polymorphic differences will allow me to study recent fluctuations in evolutionary parameters whereas those of fixed differences will reflect long-term changes in evolutionary parameters.

Final conclusions and Discussion

I observed heterogeneity in base composition within and across genes in the *D. melanogaster* genome as well as temporal variations in base composition across different lineages in the *D. melanogaster* subgroup. My study showed that evolutionary forces vary in the genome at within-gene as well as genome-wide scale. Temporally, evolutionary forces have fluctuated within short time-scales in the *D. melanogaster* subgroup.

The GC content of introns decreases in the direction of transcription. The decline in GC content was steeper near the TSS for the first 750 bp. The GC content of the first 750 bp of the genes was found to be sensitive to the level of transcription of the gene. The genes with higher transcript abundance had high GC content in the first 750 bp, suggesting that the base composition near the TSS is associated with the process of transcription. I tested for the underlying evolutionary forces and found that natural selection is a contributor to the elevated GC content near the TSS. The variation in the GC content of small introns associates with the RNA polymerase II enrichment, which suggests that the base composition near the TSS might play a role in RNA polymerase II pausing. In contrast to the GC content of small introns, the GC content of synonymous sites is reduced near the TSS. This is suggested to be under selection for efficient translation. The GC content of synonymous sites located at a distance of 50 codons or more from the start codon shows a similar pattern as of the introns, suggesting that transcription-associated selection also contributes to synonymous GC content.

At the genome-wide scale, I found that base composition varies within as well as between chromosomes. The X chromosome has higher codon bias and small intron GC content than the autosomes. I show that the high GC content on the X chromosome is restricted to synonymous sites and small introns only and long introns and intergenic regions have similar GC content on the X chromosome and autosomes. Within chromosomes, GC content heterogeneity differences among nucleotide classes. Synonymous GC content is the most heterogeneous among all nucleotide classes and intergenic GC content is the least heterogeneous. The heterogeneity in GC content within chromosomes is not explained by the difference in recombination rate.

At the temporal scale, codon bias has decreased in all seven *Drosophila* lineages studied here. All lineages showed departures from equilibrium and in some cases multiple factors appeared to have fluctuated. The strongest departure from equilibrium was observed in the lineage leading to *D. melanogaster*. The pattern observed in the lineages leading to *D. melanogaster, D. sechellia* and *D. erecta* is consistent with an increase in mutation bias towards AT and decrease in N_es. The pattern observed in *D. teissieri, D. teissieri – D. yakuba, D. erecta* and *D. orena* lineages is consistent with decrease in mutation bias towards AT and N_es. The remaining lineages show evidence for decrease in N_es and no change in mutation bias.

Fluctuations in evolutionary forces can cause bias while inferring phylogenetic relationships among species (Yang and Roberts 1995), estimating rates of evolution (Akashi 1996; Takano-Shimizu 1999), identifying evidence for adaptive evolution (Halligan 2004) and inferring changes occurred in lineages (Akashi et al. 2007). Most phylogenetic and evolutionary models assume stationary substitution models. This could lead to generating biased results. For instance, while constructing phylogenetic trees, species with similar substitutional rates (Lake 1994)and biases (Lockhart et al. 1992)or base composition (Galtier and Gouy 1995)can be grouped together, which might deviate from the true phylogeny. Rates of evolution also tend to increase in lineages that are not under equilibrium (Takano-Shimizu 1999). While inferring adaptive changes, synonymous sites or intronic sites are usually used as neutral measures. However, if the base composition of these sites is under selective constraint (Akashi 1995; Halligan 2004)or not in steady-state (Akashi et al. 2006), false signatures of adaptive evolution can be generated (Matsumoto et al. 2016). Possible solutions to the above-mentioned problems would include using substitution models that do not assume stationarity (Galtier and Gouy 1995; Yang and Roberts 1995; Akashi et al. 2007; Matsumoto et al. 2015), accounting for functional constraints on introns and synonymous sites by filtering sites that could be under selection and using complex models that incorporate lineage-specific substitution biases (Matsumoto et al. 2016).

I created a *de novo* genome assembly of *D. teissieri* and *D. orena* genomes. I used a combination of RNA-seq and Genome-seq data to identify intronic regions of the genomes. Using RNA-seq data to annotate introns enables to distinguish between coding and non-coding sequences more accurately. RNA-seq data is being used by other researchers as well to annotate gene models from sequenced genomes (Rogers, Shao, et al. 2014). Coding and non-coding data from *D. teissieri* and *D. orena* add an important resource for studying molecular evolution in *Drosophila* and will be beneficial for the entire *Drosophila* evolutionary genetics community.

References

Adams MD, Celniker SE, Holt RA, Evans CA. 2000. The genome sequence of Drosophila melanogaster.

Aerts S, Thijs G, Dabrowski M, Moreau Y, De Moor B. 2004. Comprehensive analysis of the base composition around the transcription start site in Metazoa. BMC Genomics 5:34.

Aguade M, Miyashita N, Langley CH. 1989. Reduced variation in the yellow-achaete-scute region in natural populations of Drosophila melanogaster. Genetics.

Akashi H, Eyre-Walker A. 1998. Translational selection and molecular evolution. Current Opinion in Genetics & Development 8:688–693.

Akashi H, Goel P, John A. 2007. Ancestral Inference and the Study of Codon Bias Evolution: Implications for Molecular Evolutionary Analyses of the Drosophila melanogaster Subgroup.Fay J, editor. PLoS ONE 2:e1065.

Akashi H, Kliman RM, Eyre-Walker A. 1998. Mutation pressure, natural selection, and the evolution of base composition in Drosophila. Genetica 102-103:49–60.

Akashi H, Ko W-Y, Piao S, John A, Goel P, Lin C-F, Vitins AP. 2006. Molecular evolution in the Drosophila melanogaster species subgroup: frequent parameter fluctuations on the timescale of molecular divergence. Genetics 172:1711–1726.

Akashi H. 1994. Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. Genetics 136:927–935.

Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at “silent” sites in Drosophila DNA. Genetics 139:1067–1076.

Akashi H. 1996. Molecular evolution between drosophila melanogaster and D. simulans reduced codon bias, faster rates of amino acid substitution, and larger proteins in D. melanogaster. Genetics 144:1297–1307.

Akashi H. 1997. Distinguishing the effects of mutational biases and natural selection on DNA sequence variation. Genetics 147:1989–1991.

Akashi H. 1998. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics 151:221–238.

Akashi H. 1999. Within-and between-species DNA sequence variation and the “footprint”of natural selection. Gene 238:39–51.

Akashi H. 2001. Gene expression and molecular evolution. :1–7.

Alekseyenko AA, Ho JWK, Peng S, Gelbart M, Tolstorukov MY, Plachetka A, Kharchenko PV, Jung YL, Gorchakov AA, Larschan E, et al. 2012. Sequence-Specific Targeting of Dosage Compensation in Drosophila Favors an Active

Chromatin Context.Ferguson-Smith AC, editor. PLoS Genet 8:e1002646.

Andersson SG, Kurland CG. 1990. Codon preferences in free-living microorganisms. Microbiol. Rev. 54:198–210.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster.

Beletskii A, Bhagwat AS. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proc. Natl. Acad. Sci. U.S.A. 93:13919–13924.

Bell O, Wirbelauer C, Hild M, Scharf AND, Schwaiger M, MacAlpine DM, Zilbermann F, van Leeuwen F, Bell SP, Imhof A, et al. 2007. Localized H3K36 methylation states define histone H4K16 acetylation during transcriptional elongation in Drosophila. EMBO J. 26:4974–4984.

Ben Langmead CT. 2012. Bowtie.aligner Documentation. :1–6.

Bennetzen JL, Hall BD. 1982. Codon selection in yeast. Journal of Biological Chemistry 257:3026–3031.

Bentele K, Saffert P, Rauscher R, Ignatova Z, thgen NBU. 2013. Efficient translation initiation dictates codon usage at gene start. Molecular Systems Biology 9:1–10.

Bernardi G. 1989. The Isochore Organization of the Human Genome. Annu. Rev. Genet. 23:637–659.

Bernardi G. 1993. The vertebrate genome: isochores and chromosomal bands. In: Chromosomes Today. Dordrecht: Springer Netherlands. pp. 49–60.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17.

Birky CW, Walsh JB. 1988. Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. U.S.A. 85:6414–6418.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. 27:578–579.

Boetzer M, Pirovano W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. 15:1–9.

Bulmer M. 1988. Codon usage and intragenic position. J. Theor. Biol. 133:67–71.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics.

Campos JL, Zeng K, Parker DJ, Charlesworth B, Haddrill PR. 2013. Codon Usage Bias and Effective Population Sizes on the X Chromosome versus the Autosomes in Drosophila melanogaster. 30:811–823.

Carulli JP, Krane DE, Hartl DL, Ochman H. 1993. Compositional heterogeneity and patterns of molecular evolution in the Drosophila genome. Genetics 134:837–845.

Charlesworth B, Guttman DS. 1996. Reductions in genetic variation inDrosophila andE. coli caused by selection at linked sites. Journal of Genetics.

Cherry JL. 2010. Expression Level, Evolutionary Rate, and the Cost of Expression. Genome Biology and Evolution 2:757–769.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nature Genetics 39:715–720.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203–218.

Clarke TF, Clark PL. 2010. Increased incidence of rare codon clusters at 5“and 3”gene termini: implications for function. BMC Genomics.

Clay O, Cacciò S, Zoubak S, Mouchiroud D, Bernardi G. 1996. Human coding and noncoding DNA: compositional correlations. Mol. Phylogenet. Evol. 5:2–12.

Comeron JM, Kreitman M, Aguade M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics 151:239–249.

Comeron JM. 2004. Selective and Mutational Patterns Associated With Gene Expression in Humans: Influences on Synonymous Composition and Intron Presence. Genetics 167:1293–1304.

Cox MP, Peterson DA, Biggs PJ. 2010. SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11:485.

Datta A, Jinks-Robertson S. 1995. Association of increased spontaneous mutation rates with high levels of transcription in yeast. Science 268:1616–1619.

Dekker J. 2007. GC- and AT-rich chromatin domains differ in conformation and histone modification status and are differentially modulated by Rpd3p. Genome Biol. 8:R116.

Delcher AL, Phillippy A, Carlton J, Salzberg SL. 2002. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Research 30:2478–2483.

Deschavanne P, Filipski J. 1995. Correlation of GC content with replication timing and repair mechanisms in weakly expressed E. coli genes. Nucleic Acids Research.

Diaz-Castillo C, Golic KG. 2007. Evolution of Gene Sequence in Response to Chromosomal Location. Genetics 177:359–374.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. Proc. Natl. Acad. Sci. U.S.A. 102:14338–

14343.

Drummond DA. 2005. A Single Determinant Dominates the Rate of Yeast Protein Evolution. Molecular Biology and Evolution 23:327–337.

Dujon B, Alexandraki D, Andre B, Ansorge W, Baladron V, Ballesta JP, Banrevi A, Bolle PA, Bolotin-Fukuhara M, Bossier P. 1994. Complete DNA sequence of yeast chromosome XI. Nature 369:371–378.

DURET L. 2009. Mutation patterns in the human genome: more variable than expected. Plos Biol 7:e1000028.

Eddy J, Maizels N. 2007. Conserved elements with potential to form polymorphic G-quadruplex structures in the first intron of human genes. Nucleic Acids Research 36:1321–1333.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792–1797.

Eyre-Walker A, Bulmer M. 1993. Reduced synonymous substitution rate at the start of enterobacterial genes. Nucleic Acids Research 21:4599–4603.

Feldmann H, Aigle M, Aljinovic G, Andre B, Baclet MC, Barthe C, Baur A, Bécam AM, Biteau N, Boles E, et al. 1994. Complete DNA sequence of yeast chromosome II. EMBO J. 13:5795–5809.

Felsenstein J. 1974. The evolutionary advantage of recombination. Genetics 78:737–756.

Fiston-Lavier A-S, Singh ND, Lipatov M, Petrov DA. 2010. Drosophila melanogaster recombination rate calculator. Gene 463:18–20.

Frenkel S, Kirzhner V, Korol A. 2012. Organizational Heterogeneity of Vertebrate Genomes.Laudet V, editor. PLoS ONE 7:e32076–15.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.

Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends in Genetics 19:65–68.

Gardiner K, Aissani B, Bernardi G. 1990. A compositional map of human chromosome 21. EMBO J. 9:1853–1858.

Gilchrist DA, Santos Dos G, Fargo DC, Bin Xie, Gao Y, Li L, Adelman K. 2010. Pausing of RNA Polymerase II Disrupts DNA-Specified Nucleosome Organization to Enable Precise Gene Regulation. Cell 143:540–551.

Gillespie JH. 1993. Substitution processes in molecular evolution. I. Uniform and clustered substitutions in a haploid model. Genetics 134:971–981.

Gillespie JH. 1994. Substitution processes in molecular evolution. II. Exchangeable models from population genetics. Evolution:1101–1113.

Gilmour DS, Lis JT. 1986. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. Molecular and Cellular Biology 6:3984–3989.

Glémin S, Clément Y, David J, Ressayre A. 2014. GC content evolution in coding regions of angiosperm genomes: a unifying hypothesis. Trends Genet. 30:263–270.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. 185:862–864.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. 29:1072–1075.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols 8:1494–1512.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. Genome Biol. 6:R67.

Halligan DL. 2004. Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila. Genome Research 14:273–279.

Halligan DL. 2006. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. Genome Research 16:875–884.

Hershberg R, Petrov DA. 2008. Selection on Codon Bias. Annu. Rev. Genet. 42:287–299.

Hey J, Kliman RM. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. Genetics 160:595–608.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet. Res. 8:269–294.

Hockenberry AJ, Sirer MI, Amaral LAN, Jewett MC. 2014a. Quantifying Position-Dependent Codon Usage Bias. Molecular Biology and Evolution 31:1880–1893.

Hoede C, Denamur E, Tenaillon O. 2006. Selection Acts on DNA Secondary Structures to Decrease Transcriptional Mutagenesis. PLoS Genet 2:e176.

Hooper SD, Berg OG. 2000. Gradients in nucleotide and codon usage along Escherichia coli genes. Nucleic Acids Research 28:3517–3523.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Research 23:89–98.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. 2:13–34.

Jørgensen FG, Schierup MH, Clark AG. 2007. Heterogeneity in regional GC content and differential usage of codons and amino acids in GC-poor and GC-rich regions of the genome of Apis mellifera. Molecular Biology and Evolution 24:611–619.

Kanaya S, Yamada Y, Kinouchi M, Kudo Y, Ikemura T. 2001. Codon Usage and tRNA Genes in Eukaryotes: Correlation of Codon Usage Diversity with Translation Efficiency and with CG-Dinucleotide Usage as Assessed by Multivariate Analysis. J Mol Evol 53:290–298.

Kern AD. 2004. Patterns of Polymorphism and Divergence from Noncoding Sequences of Drosophila melanogaster and D. simulans: Evidence for Nonequilibrium Processes. 22:51–62.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2012. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. Nature 471:480–485.

Khuu P, Sandor M, DeYoung J, Ho PS. 2007. Phylogenomic analysis of the emergence of GC-rich transcription elements. Proc. Natl. Acad. Sci. U.S.A. 104:16528–16533.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes inthe presence of insertions, deletions and genefusions. Genome Biol. 14:R36.

Kimura M. 1986. DNA and the neutral theory. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 312:343–354.

Kliman RM, Eyre-Walker A. 1998. Patterns of base composition within the genes of Drosophila melanogaster. J Mol Evol 46:534–541.

Kliman RM, Hey J. 1993. Reduced natural selection associated with low recombination in Drosophila melanogaster. 10:1239–1258.

Ko WY, Piao S, Akashi H. 2006. Strong Regional Heterogeneity in Base Composition Evolution on the Drosophila X Chromosome. Genetics 174:349–362.

Krylov DM, Wolf YI, Rogozin IB, Koonin EV. 2003. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. Genome Research 13:2229–2235.

Li H, Chen D, Zhang J. 2012. Analysis of Intron Sequence Features Associated with Transcriptional Regulation in Human Genes.Nurminsky DI, editor. PLoS ONE 7:e46784–e46789.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

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

Li WH. 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J Mol Evol 24:337–345.

Li X, Scanlon MJ, Yu J. 2015. Evolutionary patterns of DNA base composition and correlation to polymorphisms in DNA repair systems. Nucleic Acids Research 43:3614–3625.

Liljenström H, Heijne von G. 1987. Translation rate modification by preferential codon usage: intragenic position effects. J. Theor. Biol. 124:43–55.

Lippert MJ, Kim N, Cho J-E, Larson RP, Schoenly NE, O'Shea SH, Jinks-Robertson S. 2011. Role for topoisomerase 1 in transcription-associated mutagenesis in yeast. Proc. Natl. Acad. Sci. U.S.A. 108:698–703.

Liu J, Zhang Y, Lei X, Zhang Z. 2008. Natural selection of protein structural and functional properties: a single nucleotide polymorphism perspective. Genome Biol. 9:R69.

Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, Mieczkowski PA, Burkholder AB, Fargo DC, Gordenin DA, et al. 2014. Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. Genome Research 24:1751–1764.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. Genome Research 12:1827–1836.

Majewski J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. The American Journal of Human Genetics 73:688–692.

MARAIS G, MOUCHIROUD D, DURET L. 2003. Neutral effect of recombination on base composition in Drosophila. Genet. Res. 81:79–87.

Marais G, Mouchiroud D. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proceedings of the …

MARAIS G, Piganeau G. 2002. Hill-Robertson interference is a minor determinant of variations in codon bias across Drosophila melanogaster and Caenorhabditis elegans genomes. 19:1399–1406.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal 17:pp.10–pp.12.

Matsumoto T, Akashi H, Yang Z. 2015. Evaluation of Ancestral Sequence Reconstruction Methods to Infer Nonstationary Patterns of Nucleotide Substitution. Genetics 200:873–890.

Matsumoto T, John A, Baeza-Centurion P, Li B, Akashi H. 2016. Codon usage selection can bias estimation of the fraction of adaptive amino acid fixations. :msw027.

McBride CS, Arguello JR, O'Meara BC. 2007. Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily. Genetics 177:1395–1416.

McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. Proc. Natl. Acad. Sci. U.S.A. 104:4996–5001.

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research 32:W20–W25.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297–1303.

Mischo HE, Gómez-González B, Grzechnik P, Rondón AG, Wei W, Steinmetz L, Aguilera A, Proudfoot NJ. 2011. Yeast Sen1 helicase protects the genome from transcription-associated instability. Mol. Cell 41:21–32.

Morey NJ, Greene CN, Jinks-Robertson S. 2000. Genetic analysis of transcription-associated mutation in Saccharomyces cerevisiae. Genetics 154:109–120.

Moriyama EN, Hartl DL. 1993. Codon usage bias and base composition of nuclear genes in Drosophila. Genetics 134:847–858.

Muse GW, Gilchrist DA, Nechaev S, Shah R, Parker JS, Grissom SF, Zeitlinger J, Adelman K. 2007. RNA polymerase is poised for activation across the genome. Nature Publishing Group 39:1507–1511.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. 3:418–426.

Nekrutenko A, Li WH. 2000. Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Research.

Nielsen R, Bauer DuMont VL, Hubisz MJ, Aquadro CF. 2006. Maximum Likelihood Estimation of Ancestral Codon Usage Bias Parameters in Drosophila. 24:228–235.

Ohta T, Kimura M. 1971. On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18–25.

Ohta T. 1972a. Evolutionary rate of cistrons and DNA divergence. J Mol Evol 1:150–

157.

Ohta T. 1972b. Population size and rate of evolution. J Mol Evol 1:305–314.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.

Ohta T. 1974. Mutational pressure as the main cause of molecular evolution and polymorphism. Nature 252:351–354.

Ohta T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. Theoretical Population Biology 10:254–275.

Ohta T. 1992. The nearly neutral theory of molecular evolution. Annual Review of Ecology and Systematics:263–286.

Park S, Hannenhalli S, Choi S. 2014. Conservation in first introns is positively associated with the number of exons within genes and the presence of regulatory epigenetic signals. BMC Genomics 15:526–14.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. 2010. On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. 27:1226–1234.

Parsch J. 2003. Selective constraints on intron evolution in Drosophila. Genetics 165:1843–1851.

Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158:927–931.

Polak P, Arndt PF. 2008. Transcription induces strand-specific mutations at the 5′ end of human genes. Genome Research 18:1216–1223.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, Crepeau MW, Duchen P, Emerson JJ, Saelao P, Begun DJ, et al. 2012. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture.Malik HS, editor. PLoS Genet 8:e1003080.

Qin H, Wu WB, Comeron JM, Kreitman M, Li W-H. 2004. Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. Genetics 168:2245–2260.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage.Schierup MH, editor. PLoS Genet 10:e1004525–12.

Ressayre A, Glémin S, Montalent P, Serre-Giardi L, Dillmann C, Joets J. 2015. Introns Structure Patterns of Variation in Nucleotide Composition in Arabidopsis thalianaand Rice Protein-Coding Genes. Genome Biology and Evolution 7:2913–2928.

Rocha EPC. 2003. An Analysis of Determinants of Amino Acids Substitution Rates in Bacterial Proteins. Molecular Biology and Evolution 21:108–116.

Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. 2014. Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans. 31:1750–1766.

Rogers RL, Shao L, Sanjak JS, Andolfatto P. 2014. Revised annotations, sex-biased expression, and lineage-specific genes in the Drosophila melanogaster group. G3: Genes| Genomes| ….

Schwaiger M, Stadler MB, Bell O, Kohler H, Oakeley EJ, Schubeler D. 2009. Chromatin state marks cell-type- and gender-specific replication of the Drosophila genome. Genes & Development 23:589–601.

Serres-Giardi L, Belkhir K, David J, Glémin S. 2012. Patterns and evolution of nucleotide landscapes in seed plants. Plant Cell 24:1379–1397.

Shah P, Ding Y, Niemczyk M, Kudla G, Plotkin JB. 2013. Rate-Limiting Steps in Yeast Protein Translation. Cell 153:1589–1601.

Sharp PM, Lloyd AT. 1993. Regional base composition variation along yeast chromosome III: evoluation of chormosome primary structure. Nucleic Acids Research 21:179–183.

Singh ND, Arndt PF, Clark AG, Aquadro CF. 2009. Strong Evidence for Lineage and Sequence Specificity of Substitution Rates and Patterns in Drosophila. 26:1591–1605.

Singh ND, Bauer DuMont VL, Hubisz MJ, Nielsen R, Aquadro CF. 2007. Patterns of Mutation and Selection at Synonymous Sites in Drosophila. 24:2687–2697.

Singh ND, Davis JC, Petrov DA. 2005. Codon Bias and Noncoding GC Content Correlate Negatively with Recombination Rate on the Drosophila X Chromosome. J Mol Evol 61:315–324.

Singh ND, Larracuente AM, Clark AG. 2008. Contrasting the Efficacy of Selection on the X and Autosomes in Drosophila. 25:454–467.

Singh ND. 2005a. Genomic Heterogeneity of Background Substitutional Patterns in Drosophila melanogaster. Genetics 169:709–722.

Singh ND. 2005b. X-Linked Genes Evolve Higher Codon Bias in Drosophila and Caenorhabditis. Genetics 171:145–155.

Stoletzki N, Eyre-Walker A. 2006. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. 24:374–381.

Stoletzki N. 2011. The surprising negative correlation of gene length and optimal codon use--disentangling translational selection from GC-biased gene conversion in yeast. BMC Evol Biol 11:93.

Sueoka N. 1962. ON THE GENETIC BASIS OF VARIATION AND

HETEROGENEITY OF DNA BASE COMPOSITION. Proc. Natl. Acad. Sci. U.S.A. 48:582–592.

Takahashi T, Burguiere-Slezak G, Van der Kemp PA, Boiteux S. 2011. Topoisomerase 1 provokes the formation of short deletions in repeated sequences upon high transcription in Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. U.S.A. 108:692–697.

Takano TS. 1998. Rate variation of DNA sequence evolution in the Drosophila lineages. Genetics 149:959–970.

Takano-Shimizu T. 1999. Local recombination and mutation effects on molecular evolution in Drosophila. Genetics 153:1285–1296.

Tang CS, Zhao YZ, Smith DK, Epstein RJ. 2006. Intron length and accelerated 3′ gene evolution. Genomics 88:682–689.

Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y. 2010. An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. Cell 141:344–354.

Tuller T, Zur H. 2015. Multiple roles of the coding sequence 5′ end in gene expression regulation. Nucleic Acids Research 43:13–28.

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

Vicoso B, Haddrill PR, Charlesworth B. 2008. A multispecies approach for comparing sequence evolution of X-linked and autosomal sites in Drosophila. Genet. Res. 90:421.

Wachter E, Quante T, Merusi C, Arczewska A, Stewart F. 2014. Synthetic CpG islands reveal DNA sequence determinants of chromatin structure. eLife.

Wirbelauer C, Bell O, Schübeler D. 2005. Variant histone H3.3 is deposited at sites of nucleosomal displacement throughout transcribed genes while active histone modifications show a promoter-proximal bias. Genes & Development 19:1761–1766.

Wolfe KH, Sharp PM, Li W-H. 1989. Mutation rates differ among regions of the mammalian genome. , Published online: 19 January 1989; | doi:10.1038/337283a0 337:283–285.

Wong GK-S, Wang J, Tao L, Tan J, Zhang J, Passey DA, Yu J. 2002. Compositional Gradients in Gramineae Genes. Genome Research 12:851–856.

Yamao F, Andachi Y, Muto A, Ikemura T, Osawa S. 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Research 19:6119–6122.

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. 24:1586–1591.

Zeitlinger J, Stark A, Kellis M, Hong J-W, Nechaev S, Adelman K, Levine M, Young RA. 2007. RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nature Genetics 39:1512–1516.

Zeng K, Charlesworth B. 2009. Studying Patterns of Recent Evolution at Synonymous Sites and Intronic Sites in Drosophila melanogaster. J Mol Evol 70:116–128.

Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. Nat Rev Genet 16:409–420.

Zhang L, Kasif S, Cantor CR, Broude NE. 2004. GC/AT-content spikes as genomic punctuation marks. Proc. Natl. Acad. Sci. U.S.A. 101:16855–16860.

Zhang S, Goldman E, Zubay G. 1994. Clustering of low usage codons and ribosome movement. J. Theor. Biol. 170:339–354.

Tables

Table 2.1. **Correlation between small intron GC content and RNA Pol II enrichment within autosomal genes**

Gene category	X chromosome	Autosomes
Genes with low transcript abundance	0.41*	0.65**
Genes with high transcript abundance	0.44*	0.68**

* P<0.05

** P<0.001

*** P<0.0001

Note: The correlation coefficients were between average small intron GC content and average RNA PolII enrichment across genes for 50 bp non-overlapping windows in the first 1500 bp from the TSS.

Table 3.1. **GC content comparison between X and autosomes using Mann-Whitney U test**

GC measure	Gene region	X mean	Autosome mean	*U*	BS-corrected p-value
MCU	Whole gene	0.67	0.64	877190	**$5.97x10^{-15}$**
	5p750	0.68	0.64	384880	**$4.42x10^{-6}$**
	Post-5p750	0.67	0.64	868720	**$2.28x10^{-14}$**
Small intron GC	Whole gene	0.38	0.34	357510	**$5.72x10^{-15}$**
	5p750	0.41	0.38	29332	**0.0224**
	Post-5p750	0.37	0.32	219570	**$4.69x10^{-13}$**
Long intron GC	Whole gene	0.38	0.36	202500	0.141
	5p750	0.39	0.38	61951	1
	Post-5p750	0.37	0.35	100670	1
Intergenic GC	-	0.41	0.41	748890	0.1

BS-corrected p-values are p-values obtained after multiple test correction for tests comparing parameters between X and autosomes and among autosomes using Bonferroni sequential method.

Note: 5p750 region is defined as the region up to 750 bp from TSS. The first 50 codons are also removed for the calculation of MCU. Post-5p750 region is defined as the region from 751 bp from the TSS to the end of the gene.

Table 3.2. **G-score comparisons between X and autosomes and among nucleotide classes using two-sample Kolmogorov-Smirnov test.**

Parameter 1	Mean G of parameter 1	Parameter 2	Mean G of parameter 2	D	BS-corrected p-value
X MCU	15.71	Autosome MCU	8.73	0.17	0.2479
X siGC	6.09	Autosome siGC	3.79	0.30	0.8508
X liGC	4.92	Autosome liGC	3.90	0.16	1
X intergenic GC	2.86	Autosome intergenic GC	2.08	0.06	0.8495
MCU	9.93	siGC	4.05	0.18	**0.010164**
MCU	9.93	liGC	4.10	0.23	**$3.1x10^{-06}$**
MCU	9.93	Intergenic GC	2.23	0.35	**$2.2x10^{-15}$**
siGC	4.05	liGC	4.10	0.09	0.8426
siGC	4.05	Intergenic GC	2.23	0.19	**0.0037264**
liGC	4.10	Intergenic GC	2.23	0.15	**0.009198**

BS-corrected p-values are p-values obtained after multiple test correction for all tests listed here using Bonferroni sequential method.

Table 3.3**. Spearman’s Rank Correlation coefficients between recombination rate and GC measures**

GC measure	Gene region	R_s on X chromosome	R_s on autosomes
MCU	5p750	-0.217***	0.019
	Post-5p750	-0.153*	0.051
Small intron GC	5p750	-0.315**	-0.056
	Post-5p750	-0.176	-0.02
Long intron GC	5p750	-0.08	0.052
	Post-5p750	-0.011	0.117*

* $P<0.01$

** $P<0.001$

*** $P<0.0001$

Note: 5p750 region is defined as the region up to 750 bp from TSS. The first 50 codons are also removed for the calculation of MCU. Post-5p750 region is defined as the region from 751 bp from the TSS to the end of the gene. P-values are corrected for all tests listed here using Bonferroni-sequential method.

Table 3.4. **Spearman's Rank Correlation coefficients between recombination rate and GC measures after removing regions with no recombination**

GC measure	Gene region	R_s on X chromosome	R_s on autosomes
MCU	5p750	-0.229***	-0.011
	Post-5p750	-0.178**	0.011
Small intron GC	5p750	-0.302*	-0.058
	Post-5p750	-0.207	-0.054
Long intron GC	5p750	-0.046	0.047
	Post-5p750	0.044	0.104

* P<0.01

** P<0.001

*** P<0.0001

Note: 5p750 region is defined as the region up to 750 bp from TSS. The first 50 codons are also removed for the calculation of MCU. Post-5p750 region is defined as the region from 751 bp from the TSS to the end of the gene. P-values are corrected for all tests listed here using Bonferroni-sequential method.

Table 4.1**. Summary of adapter filtering of RNA-seq reads**

Species	Tissue Sample	Read	# Processed reads	% Reads containing adapters	% Bases removed
D. teissieri	Larva rep 1	R1	44,126,512	30.45	0.93
		R2	44,126,512	35.55	0.91
	Larva rep 2	R1	48,648,371	28.86	0.45
		R2	48,648,371	34.09	0.50
	Pupa	R1	46,673,378	29.44	0.45
		R2	46,673,378	33.99	0.50
	Female	R1	47,014,797	28.47	0.42
		R2	47,014,797	33.23	0.47
	Male	R1	53,347,300	30.25	0.44
		R2	53,347,300	33.50	0.46
D. orena	Larva rep 1	R1	38,907,278	30.92	0.95
		R2	38,907,278	36.31	0.97
	Larva rep 2	R1	40,763,583	28.02	0.41
		R2	40,763,583	33.39	0.48
	Pupa	R1	45,457,493	28.58	0.44
		R2	45,457,493	33.98	0.50
	Female	R1	38,153,929	28.69	0.42
		R2	38,153,929	32.81	0.46
	Male	R1	44,019,572	29.73	0.43
		R2	44,019,572	33.14	0.46

Table 4.2**. Summary of rRNA filtering of RNA-seq reads**

Species	Sample	Reads processed	% reads with at least one rRNA match	# reads with no rRNA match
D. teissieri	Larva rep 1	39,979,337	0.98	39,589,469
	Larva rep 2	40,493,773	2.37	39,535,566
	Pupa	38,846,581	2.31	37,947,840
	Female	40,203,297	1.02	39,792,056
	Male	45,692,128	2.48	44,556,905
D. orena	Larva rep 1	35,682,246	0.67	35,442,454
	Larva rep 2	33,851,517	1.51	33,338,685
	Pupa	37,761,137	1.26	37,285,715
	Female	32,463,589	1.12	32,099,584
	Male	37,715,928	1.52	37,141,664

Table 4.3**. Summary of ortholog assignment of RNA-seq contigs**

Species	Type of contig matches	Total matches in rs2	# contigs with rs1 orthologs assigned
D. teissieri	Contigs with single match	177,815	153,933
	Contigs with multiple matches	22,529	8,801
D. orena	Contigs with single match	175,797	152,858
	Contigs with multiple matches	27,746	9,730

❖ The contigs in each species are pooled across tissue samples.

Table 4.4. **Summary of adapter filtering of genomic reads**

Sample	Read	# Processed reads	% Reads containing adapters	% Bases removed for adapters
D. teissieri	R1	172,225,062	27.107	0.397
	R2	172,225,062	35.567	0.474
D. orena	R1	176,570,473	26.607	0.397
	R2	176,570,473	35.548	0.476

Table 4.5. **Length of genomic reads after quality filtering**

Sample	Read	Mean read length	Mean read length after quality filtering	Median read length after quality filtering
D. teissieri	R1	99.62 bp	82.2 bp	99 bp
	R2	99.53 bp	75.9 bp	98 bp
D. orena	R1	99.62 bp	80.7 bp	99 bp
	R2	99.53 bp	74.6 bp	96 bp

Table 4.6. **Summary of genome assembly**

Assembly	*D. teissieri* assembly	*D. orena* assembly
Total number of scaffolds	18,971	28,905
Length of largest scaffold (bp)	537,749	410,257
Total length (bp)	141,688,368	147,271,023
N50	84,073	55,291
#Ns per 100 k bp	2742.55	3534.96
% genes matching to single scaffold	81%	76%

❖ All statistics are based on scaffolds of size >= 200 bp, unless otherwise noted

Table 4.7. **Summary of MUMmer analysis**

	D. teissieri assembly		*D. orena* assembly		*D. yakuba* genome
	Contig$_G$	Scaffold	Contig$_G$	Scaffold	Genome
#Matching to single chromosome	19,570	3,671	34,905	6,121	2,165
#Matching to multiple chromosomes	2,142	729	2,416	816	1,267
%Matching to single chromosome	90.13	83.43	93.53	88.23	63.08

Table 4.8. **Summary of comparison between RNA-seq annotated transcripts and flanking sequences of junctions reported by TopHat2.**

	D. teissieri	*D. orena*
Total number of TopHat junctions in blastn matches	79693	78742
Total number of junctions in blastn matches after filtering for length and identity	78100	77193
Total number of junctions with matches on same strand as contig_R	76133	74933
Total number of junctions in blastn matches with both flanking sequences matching to same contig_R	54631	53928
Total number of junctions in blastn matches to contigs_R in the gene set from other Drosophila species	30911	30862
Total number of junctions in blastn matches after resolving multiple matches	28847	28479

Table 4.9. **Number of introns identified under each intron category**

	D. teissieri		*D. orena*	
	#introns	#genes	#introns	#genes
Single isoform	10991	4259	10449	3996
Constitutive	4721	1394	4627	1392
Non-constitutive	5830	1697	5533	1661

Table 4.10. **Summary of comparison between CDSs of closest species and flanking sequences of junctions reported by TopHat2.**

	D. teissieri introns			*D.orena* introns		
	Single isoform	Constit utive	Non-constitutive	Single isoform	Constit utive	Non-constitutive
#Single LFS and RFS connected matches	8995	3702	998	8400	3623	1014
#Introns with orthologs assigned	8588	3586	960	7962	3485	939
#Small introns (<100 bp) with orthologs	6212	2609	530	5724	2495	522
#Medium introns (100-500 bp) with orthologs	1556	628	228	1470	659	233
#Long introns (>500 bp) with orthologs	820	349	202	768	331	184
#Introns having canonical splice sites (GT-AC) with orthologs assigned	8533	3566	955	7919	3457	930
#Introns with no orthologs assigned	407	116	38	438	138	75

- ❖ LFS: Left Flanking Sequence (5′)
- ❖ RFS: Right Flanking Sequence (3′)

Table 4.11. **Summary of intron alignments**

Intron category	tyeo set	mtyeo set	mcstyeo set
Small introns	7016	6943	5196
Long introns	3038	3072	2427

- tyeo set : Orthologous introns in *D. teissieri, D. yakuba, D. erecta* and *D. orena*
- mtyeo set : Orthologous introns in *D. melanogaster, D. teissieri, D. yakuba, D. erecta* and *D. orena*
- mcstyeo set: Orthologous introns in *D. melanogaster, D. sechellia, D. simulans, D. teissieri, D. yakuba, D. erecta* and *D. orena*

Table 4.12**. Summary of transcriptome assemblies**

Species	Tissue Sample	Total number of contigs	Length of largest contig (bp)	Total length of the assembly (bp)	N50 (bp)
D. teissieri	Larva	119,801	43,821	169,958,830	3065
	Pupa	65,921	22,745	92,177,790	2995
	Male	96,306	29,841	148,126,295	3130
	Female	64,015	39,340	92,149,341	2802
D. orena	Larva	115,283	28,104	146,602,500	2593
	Pupa	60,969	25,544	77,762,550	2502
	Male	102,904	38,905	152,579,757	2933
	Female	70,425	47,656	97,346,415	2630

❖ The larva transcriptome in each of the species is a collection of transcriptomes from two RNA-seq experiments.

Figures

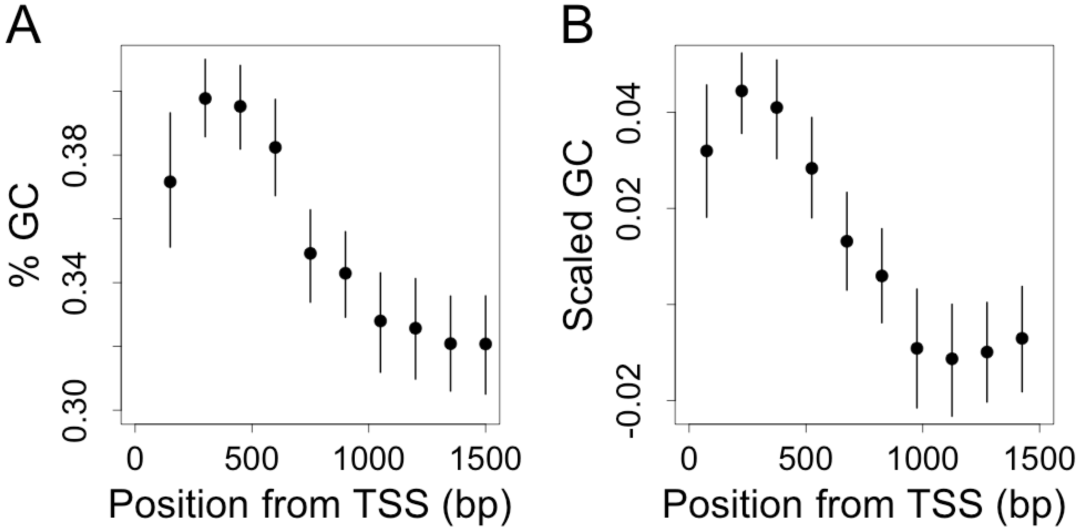


Figure 2.1. **Within gene variation in GC content of small introns.** (A) GC content of small introns with respect to the distance from the TSS, averaged across 2488 autosomal genes. (B) Scaled GC content of small introns with respect to the distance from the TSS, averaged across 2911 genes (autosomal and X-linked). Scaled GC for a given gene is the GC content of small introns in a bin minus the average GC content of small introns of the gene. GC contents and scaled GC contents were calculated for 150 bp non-overlapping bins from the TSS. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered. Error bars represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average distance of each bin from the TSS is plotted on the x-axis and the average GC content in (A) and average scaled GC content in (B), along with the 95% confidence interval, of each bin are plotted on the y-axis.

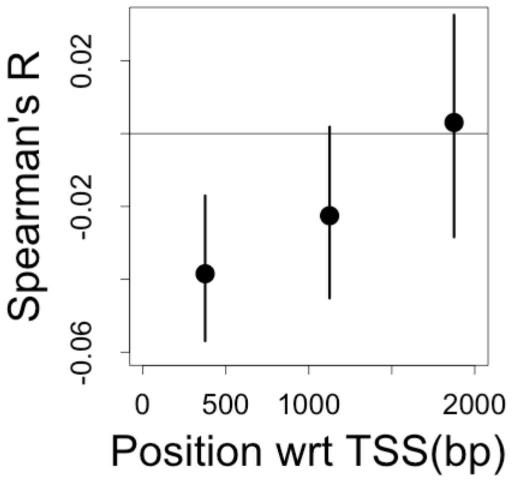


Figure 2.2. **Correlation between small intron GC content and position within the transcript.** Spearman's rank correlation coefficients between GC content and intron site positions in 750 bp bins across small introns of *D. melanogaster* (X and autosomal) genes. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered. Error bars represent bootstrap 95% confidence interval on average correlation coefficients calculated by resampling genes in each bin (1000 replicates). The average position of each bin with respect to the TSS is plotted on the x-axis and the average Spearman's R along with 95% confidence intervals of each bin is plotted the y-axis.

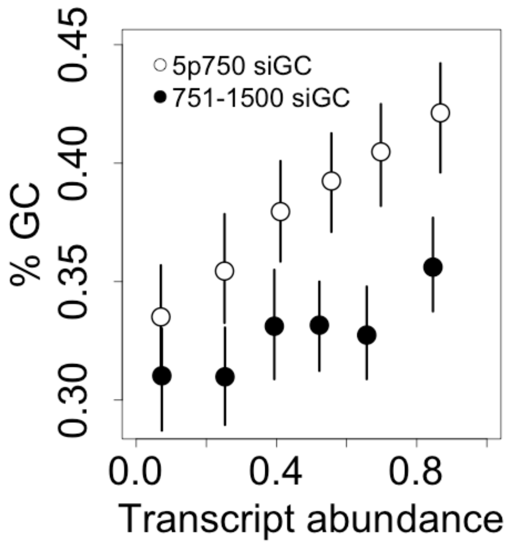


Figure 2.3. **Relationship between intron GC content and transcript abundance.** The relationship between GC content of small introns present in the first 750 bp from TSS (open circles) and small introns present at the distance of 751-1500 bp from TSS (filled circles) with the transcript abundance of genes in whole adult. Only introns belonging to autosomal genes are used. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered. For each intron category, the introns are binned by transcript abundance into 6 bins with roughly equal number of intronic sites of the given category. Error bars in each graph represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average transcript abundance of the genes in a bin is plotted on the x-axis. The average GC content of the introns of a given category present in the genes in a bin is plotted on the y-axis.

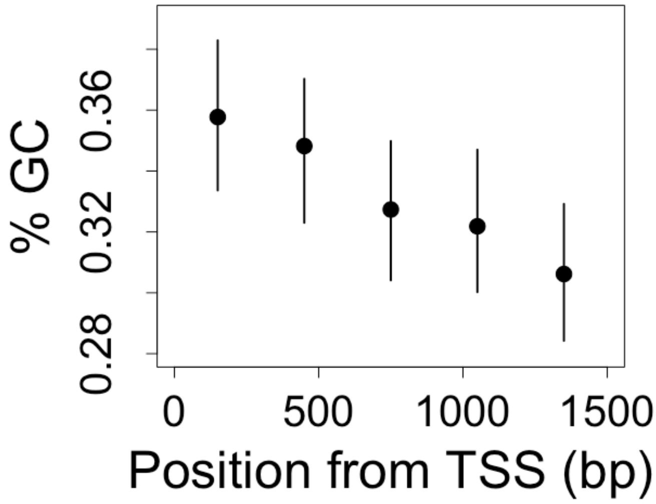


Figure 2.4. **5′ GC gradient of genes lowly expressed in tissues containing germline cells.** GC content of small introns with respect to the distance from the TSS, averaged across genes expressed at the bottom 25% percentile in ovary and testis and top 75% percentile in at least one other tissue. GC contents were calculated for 300 bp non-overlapping bins from the TSS. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered. Error bars represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average distance of each bin from the TSS is plotted on the x-axis and the average GC content, along with the 95% confidence interval, of each bin is plotted on the y-axis.

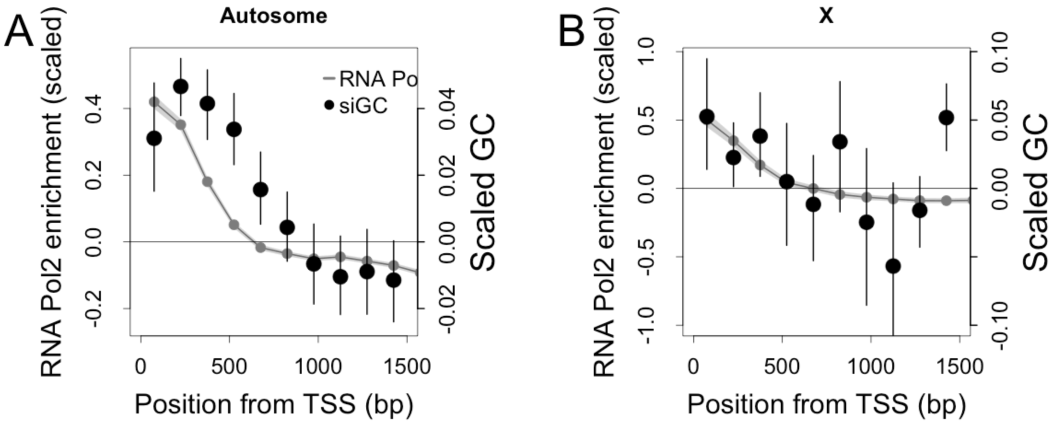


Figure 2.5: **Association of small intron GC variation with RNA PolII enrichment in autosomal genes**. Average scaled GC content (filled circle) of small introns for 150 bp non-overlapping bins plotted alongside average scaled RNA Pol II enrichment (grey line) of 150 bp non-overlapping bins of autosomal (A) and X-linked (B) genes. Error bars represent bootstrap 95% confidence interval on the average scaled GC content calculated by resampling the genes in a bin (1000 replicates). Light grey region around the line representing scaled RNA Pol II enrichment represents bootstrap 95% confidence interval on the average scaled RNA Pol II enrichment calculated by resampling the genes in a bin (1000 replicates). RNA Pol II enrichment data is from S2 cells. X-axis denotes the average nucleotide distance of each bin from TSS in bp and y-axis denotes the scaled average enrichment (left axis) and average scaled GC content (right axis) of each bin. The scale of the y-axis for scaled GC is 1/10th of that for scaled RNA Pol II enrichment. The y=0 point coincides for both the y-axes.

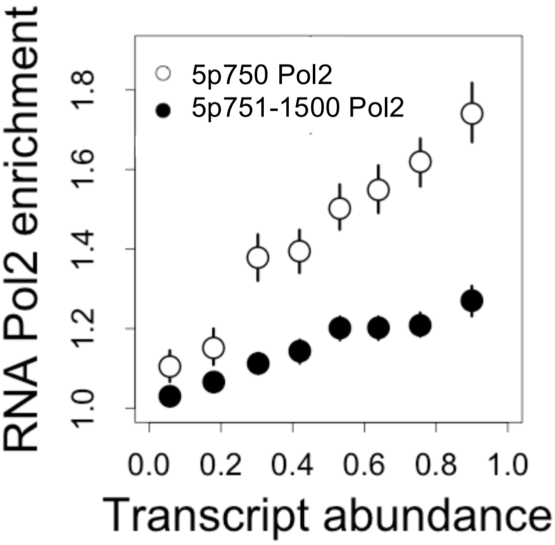


Figure 2.6. **Relationship between RNA Pol II and transcript abundance for 5p750 and 5p751-1500 regions.** RNA Pol II enrichment for 5p750 region of genes (open circles) and 5p751-1500 region of genes (filled circles) with transcript abundance of genes in whole adult. 5p750 region is defined as the region up to 750 bp from the TSS and 5p751-1500 region is defined as the region from 751 bp to 1500 bp from the TSS. Average RNA Pol II enrichment was calculated for 5p750 and 5p751-1500 regions of each autosomal gene. The genes were binned into 8 bins with roughly equal numbers of RNA Pol II bound sites. Error bars in each graph represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average transcript abundance of the genes in a bin is plotted on the x-axis. The average RNA Pol II enrichment of sites in a bin is plotted on the y-axis.

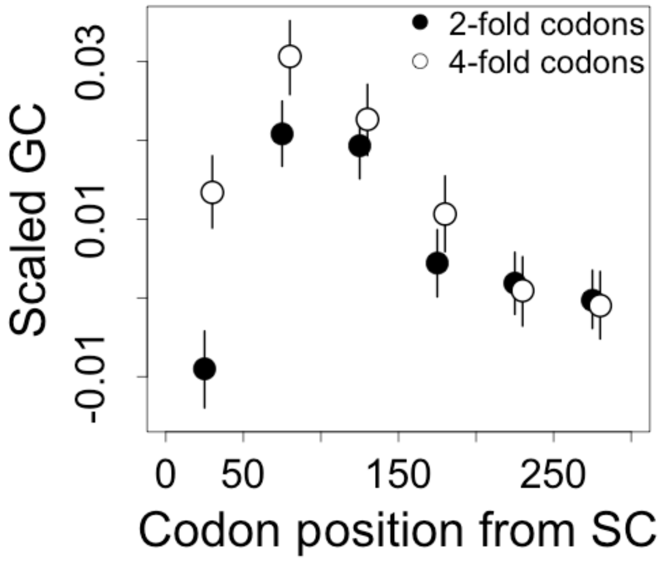


Figure 2.7. **Comparison of GC gradients between 4-fold and 2-fold synonymous sites.** Scaled GC content of 4-fold synonymous sites (open circle) and 2-fold synonymous sites (except Aspartic acid) (filled circle) up to 300 codons from the start codon (SC). Scaled GC contents of 4-fold and 2-fold synonymous sites are calculated for non-overlapping bins of 50 codons. Error bars represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average distance of each bin from the start codon is plotted on the x-axis and the average scaled GC content, along with the 95% confidence interval, of each bin are plotted on the y-axis. The data points for 4-fold synonymous GC content are staggered by 5 codons to avoid overlapping error bars.

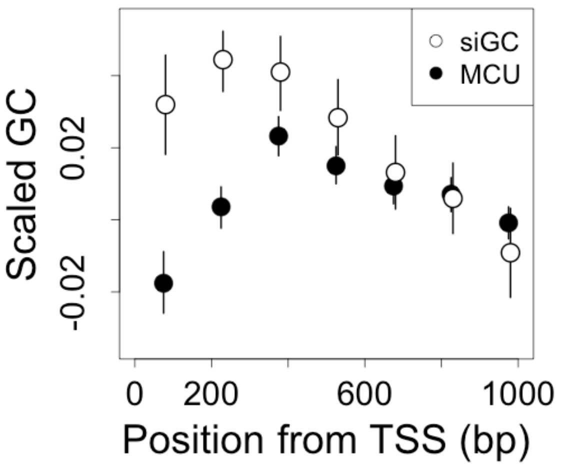


Figure 2.8. **Comparison of GC gradients between introns and synonymous sites.** Scaled GC content of small introns (open circle) and 2-fold synonymous sites (except Aspartic acid) (filled circle) up to 900 bp for the TSS. Scaled GC contents of small introns and synonymous sites are calculated for non-overlapping bins of 150 bp. Error bars represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average distance of each bin from the TSS is plotted on the x-axis and the average scaled GC content, along with the 95% confidence interval, of each bin are plotted on the y-axis. The data points for small intron GC content are staggered by 5 bp to avoid overlapping error bars.

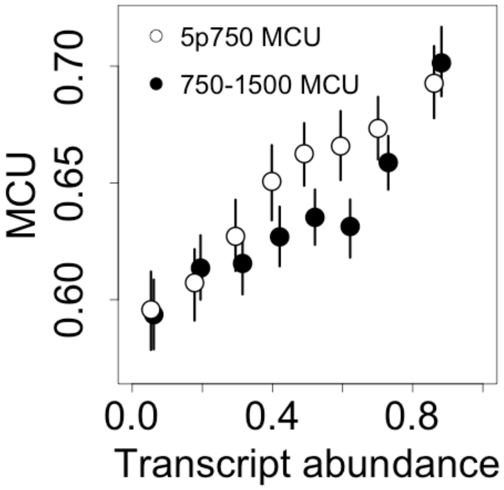


Figure 2.9. **Relationship between synonymous GC content and transcript abundance.** The relationship between MCU at 2-fold synonymous sites present in the first 750 bp from TSS (open circles) and 2-fold synonymous sites present at the distance of 751-1500 bp from TSS (filled circles) with the transcript abundance of genes in whole adult. The first 50 codons of the genes are filtered. For each intron category, the codons are binned by transcript abundance into 8 bins with roughly equal number of codons of the given category. Error bars in each graph represent bootstrap 95% confidence interval on the bin averages calculated by resampling genes in each bin (1000 replicates). The average transcript abundance of the genes in a bin is plotted on the x-axis. The average MCU of the codons of a given category present in the genes in a bin are plotted on the y-axis.

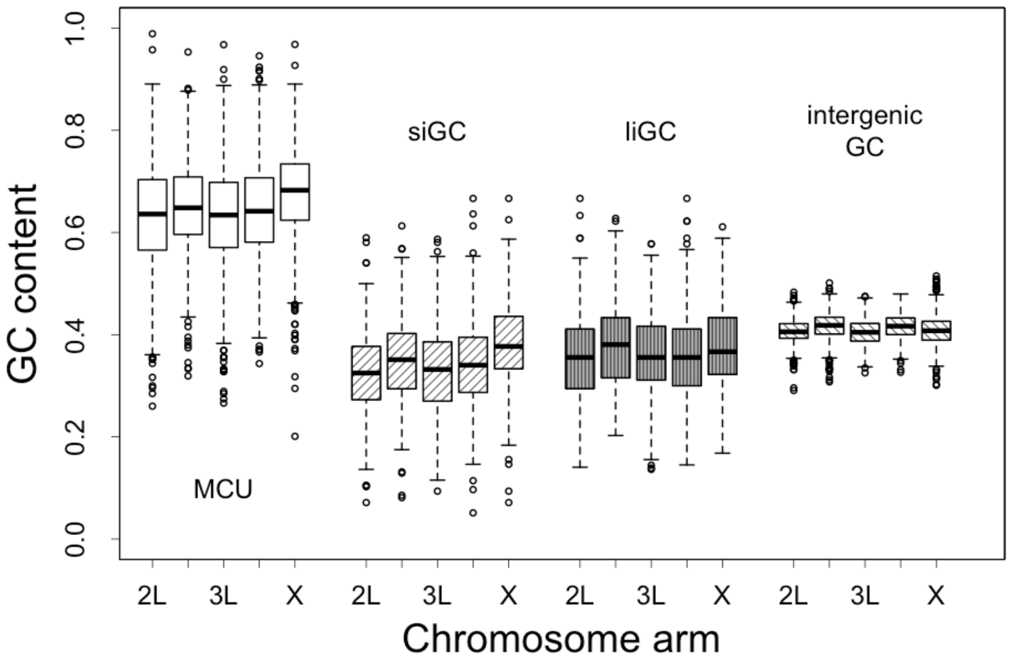


Figure 3.1. **GC content comparison between X and autosomes.** Boxplots comparing MCU (white), small intron GC (right diagonal lines), long intron GC (11-100 bp) (vertical lines) and intergenic GC content (left diagonal lines) among chromosome arms. The end of the error bars denote 1.5 times of the interquartile distance. The end of the boxes denote the first and third quartiles and the black lines within the box give the median of the data. Each boxplot shows that GC distribution for one chromosome arm, specified on the x-axis. The y-axis shows the GC content of synonymous sites (MCU), small introns (siGC), long introns (liGC) and intergenic GC. Only genes that had more than 30 sites of a given nucleotide class were included in the analysis. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered.

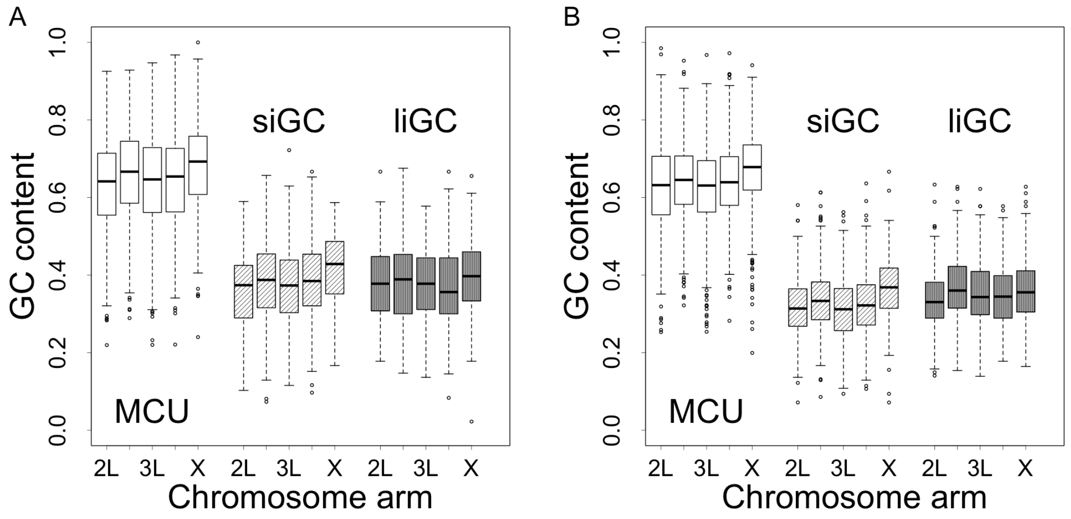


Figure 3.2. **GC content comparison between X and autosomes for 5′ and 3′ regions separately.** A) Boxplots comparing MCU (white), small intron GC (right diagonal lines) and long intron GC (11-100 bp) (vertical lines) of 5′ regions of genes among chromosome arms. B) Boxplots comparing MCU (white), small intron GC (right diagonal lines) and long intron GC (vertical lines) of 3′ regions of genes among chromosome arms. The end of the error bars denote 1.5 times of the interquartile distance. The end of the boxes denote the first and third quartiles and the black lines within the box give the median of the data. Each boxplot shows that GC distribution for one chromosome arm, specified on the x-axis. The y-axis shows the GC content of synonymous sites (MCU), small introns (siGC) and long introns (liGC). Only genes that had more than 30 sites of a given nucleotide class were included in the analysis. 10 bp from the 5′ end and 30 bp from the 3′ end of introns are filtered.

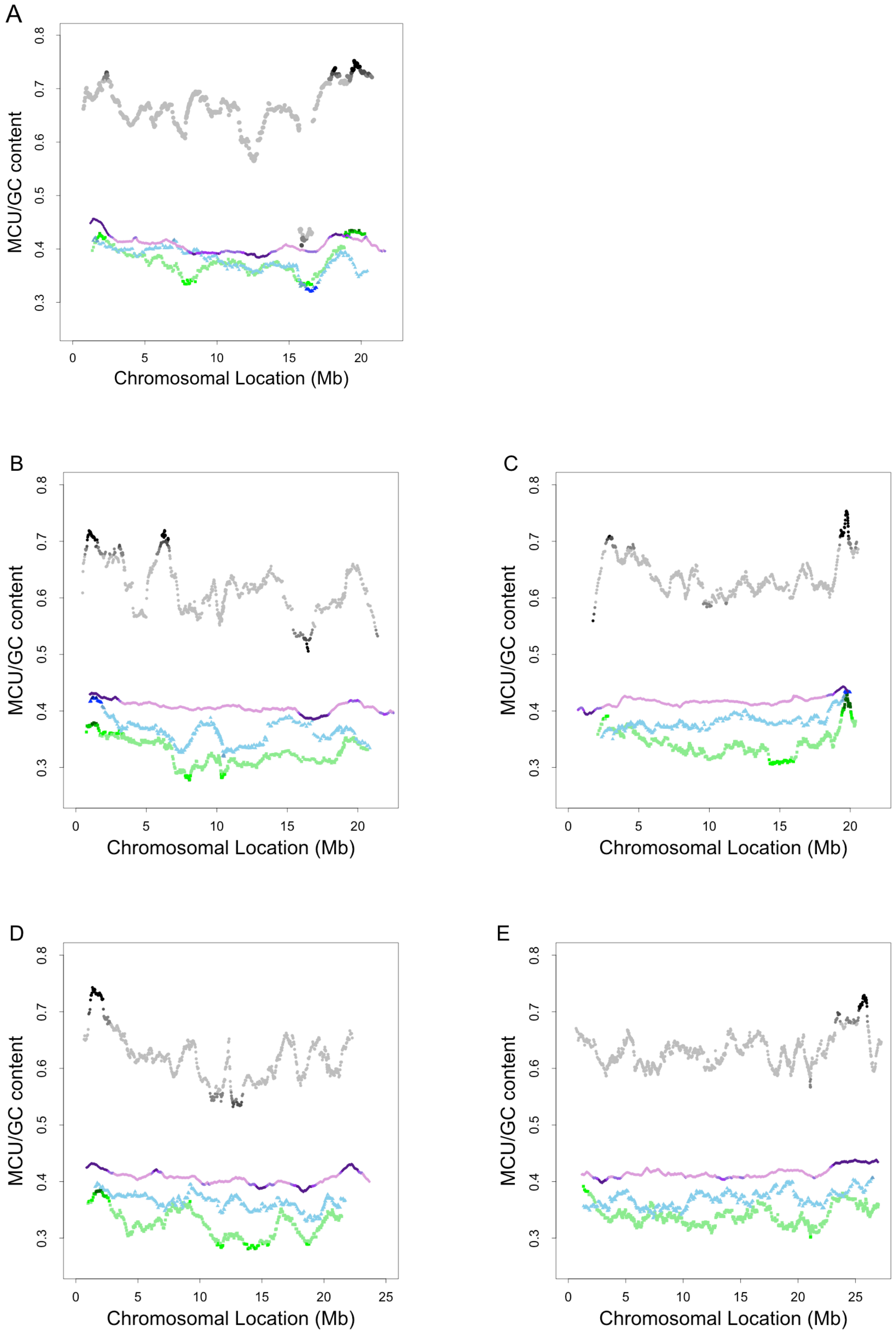


Figure 3.3. **Regional heterogeneity in the base composition in different chromosome arms.** Sliding window plot for MCU (grey), small intron GC content (green), long intron GC content (blue) and intergenic GC content (pink) for genes and regions in X (A), 2L (B), 2R (C), 3L (D), 3R (E) chromosome arms. Each window

contains 30 genes (for MCU, siGC and liGC) or 500Kb blocks (for intergenic GC). Sliding width is 1 gene for MCU, siGC and liGC and 10Kb for intergenic GC. Darker points represent regions with stronger departure in GC content from the null distribution estimated by permutation. The value on the x-axis represent the midpoint of the chromosomal location of each bin and that on the y-axis represents the average MCU or GC content of each bin.

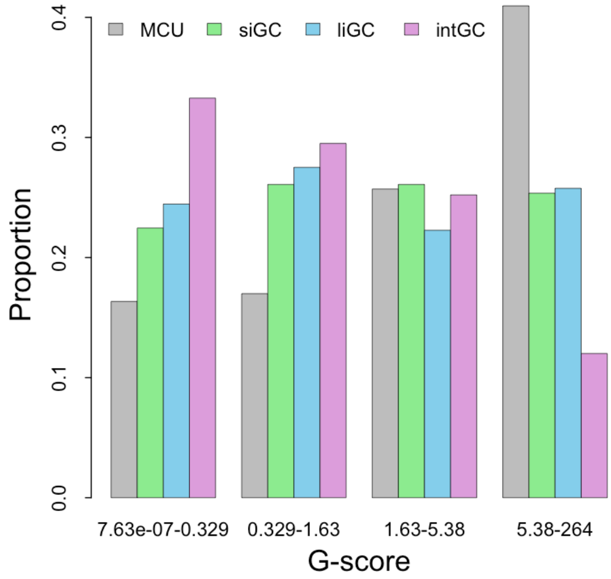


Figure 3.4. **Chromosomal heterogeneity in the base composition of different nucleotide classes.** Barplot showing the proportion of chromosomal blocks that have the G-score value in the given range. The data for chromosomes X, 2 and 3 are used and are partitioned into different nucleotide classes. The G-scores are calculated for non-overlapping chromosomal blocks after controlling for the sample sizes across blocks and nucleotide classes. The G-score ranges are defined by the 0th, 25th, 50th, 75th and 100th quantiles of the combined data of all nucleotide classes and all chromosomes.

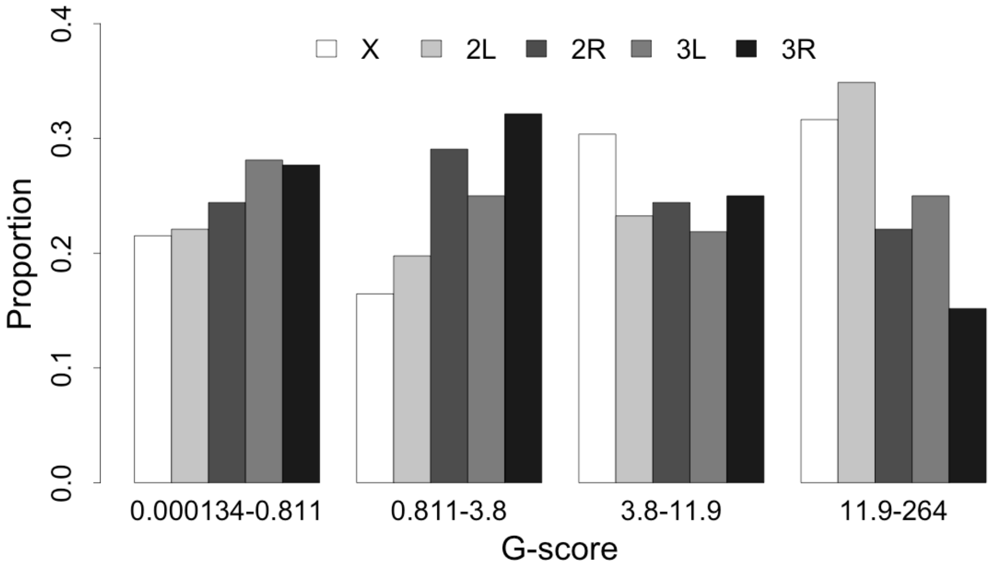


Figure 3.5. **Comparison of heterogeneity in codon bias among chromosome arms.** Barplot showing the proportion of chromosomal blocks that have the G-score value for MCU in the given range. The data are partitioned for chromosome arms. The G-scores are calculated for non-overlapping chromosomal blocks after controlling for the sample sizes across blocks. The G-score ranges are defined by the 0th, 25th, 50th, 75th and 100th quantiles of the combined data of all chromosomes. The first 50 codons of genes were filtered while calculating MCU.

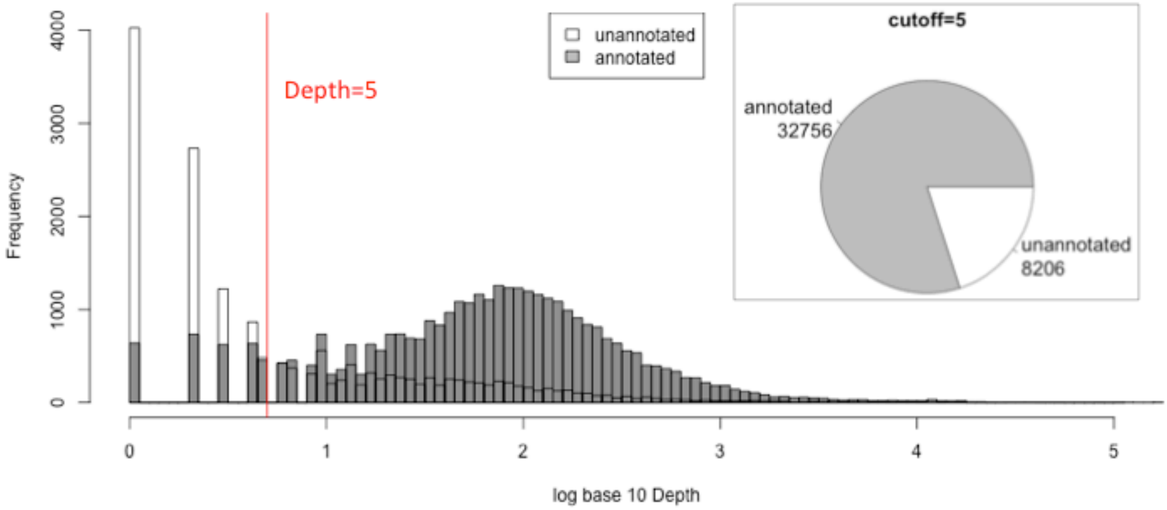


Figure 4.1. **Histogram of annotated and unannotated putative introns.** Histogram of $\log_{10}$ junction depth of annotated and unannotated introns of *D. yakuba*. The red line denotes the junction depth of 5. The pie chart shows the number of annotated and unannotated introns left after removing all introns that had junction depth of 5 or below.

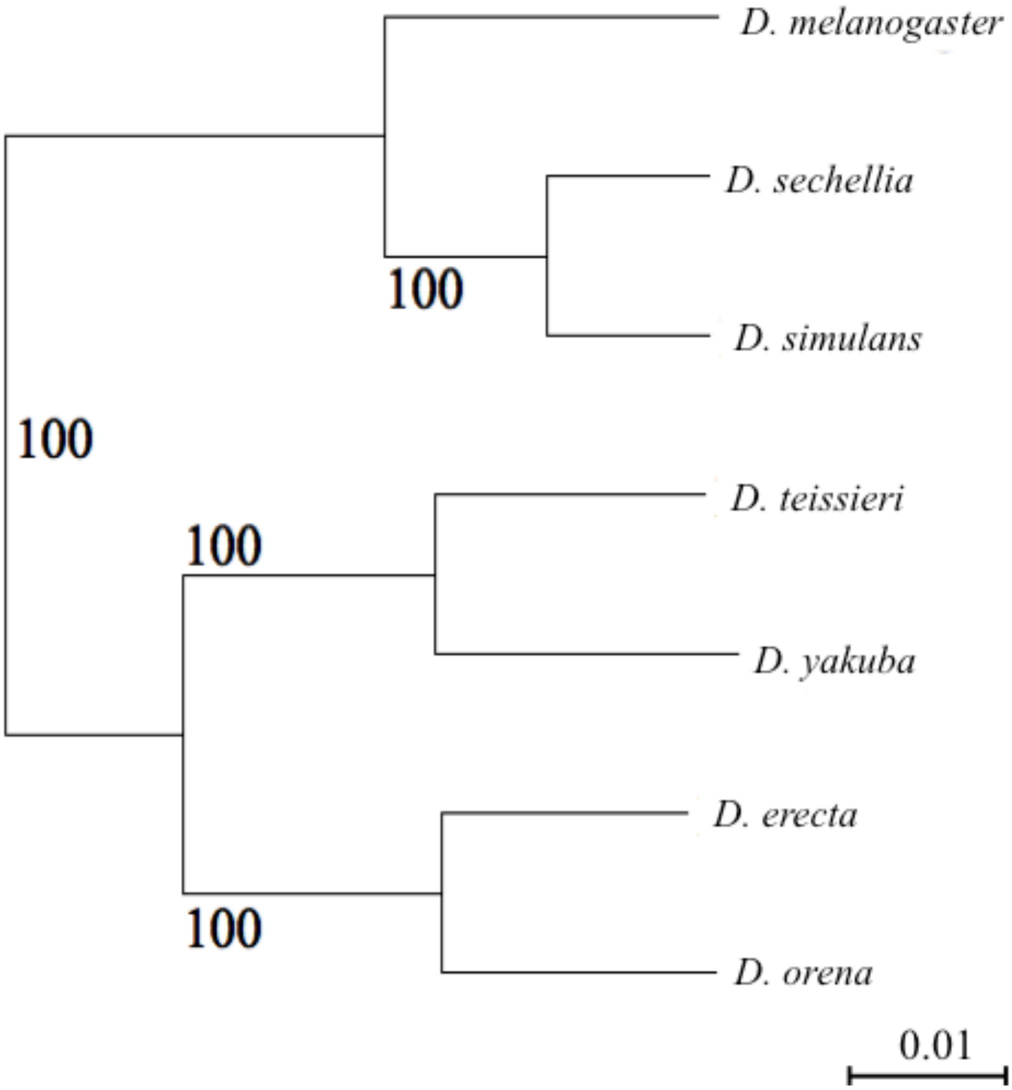
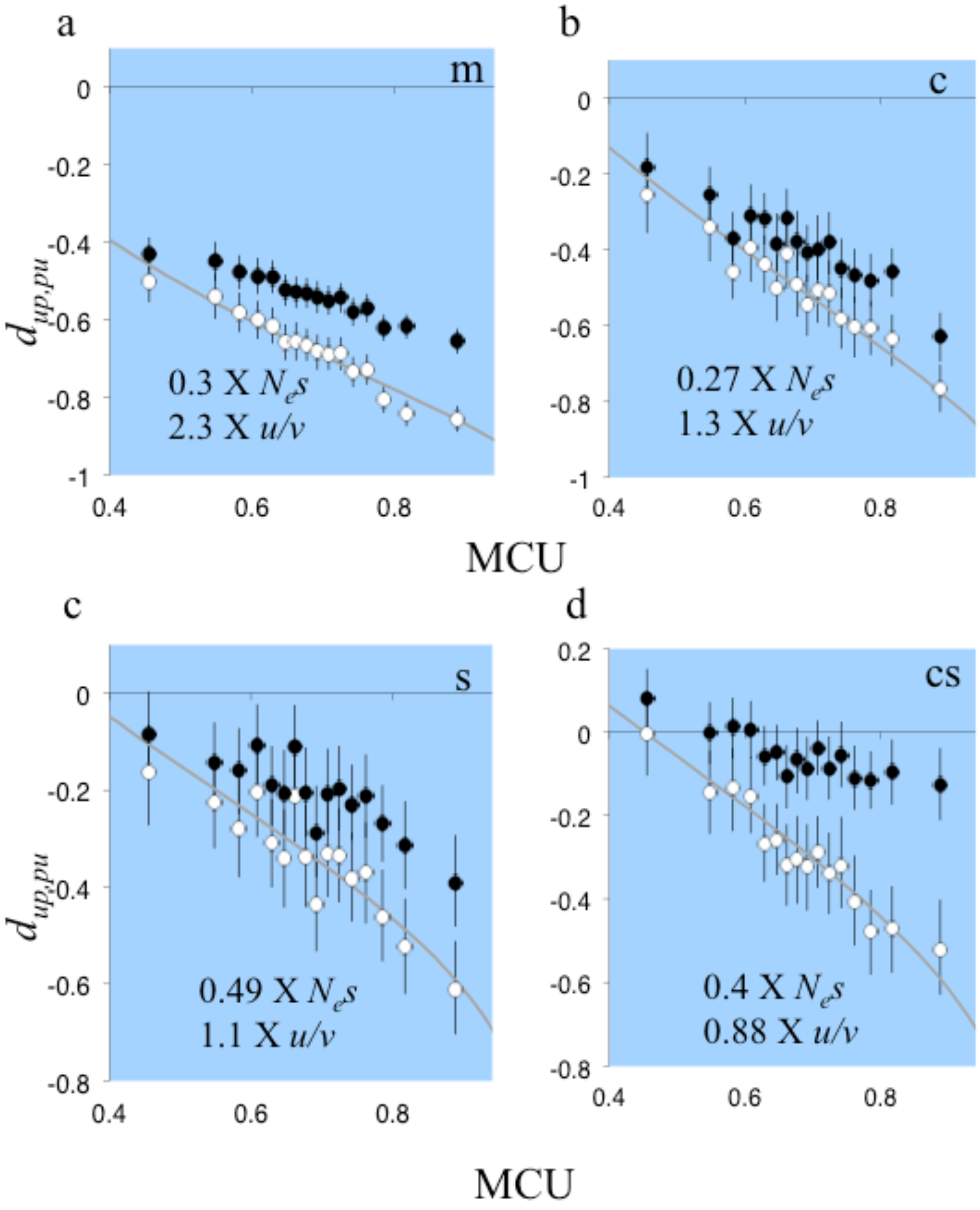
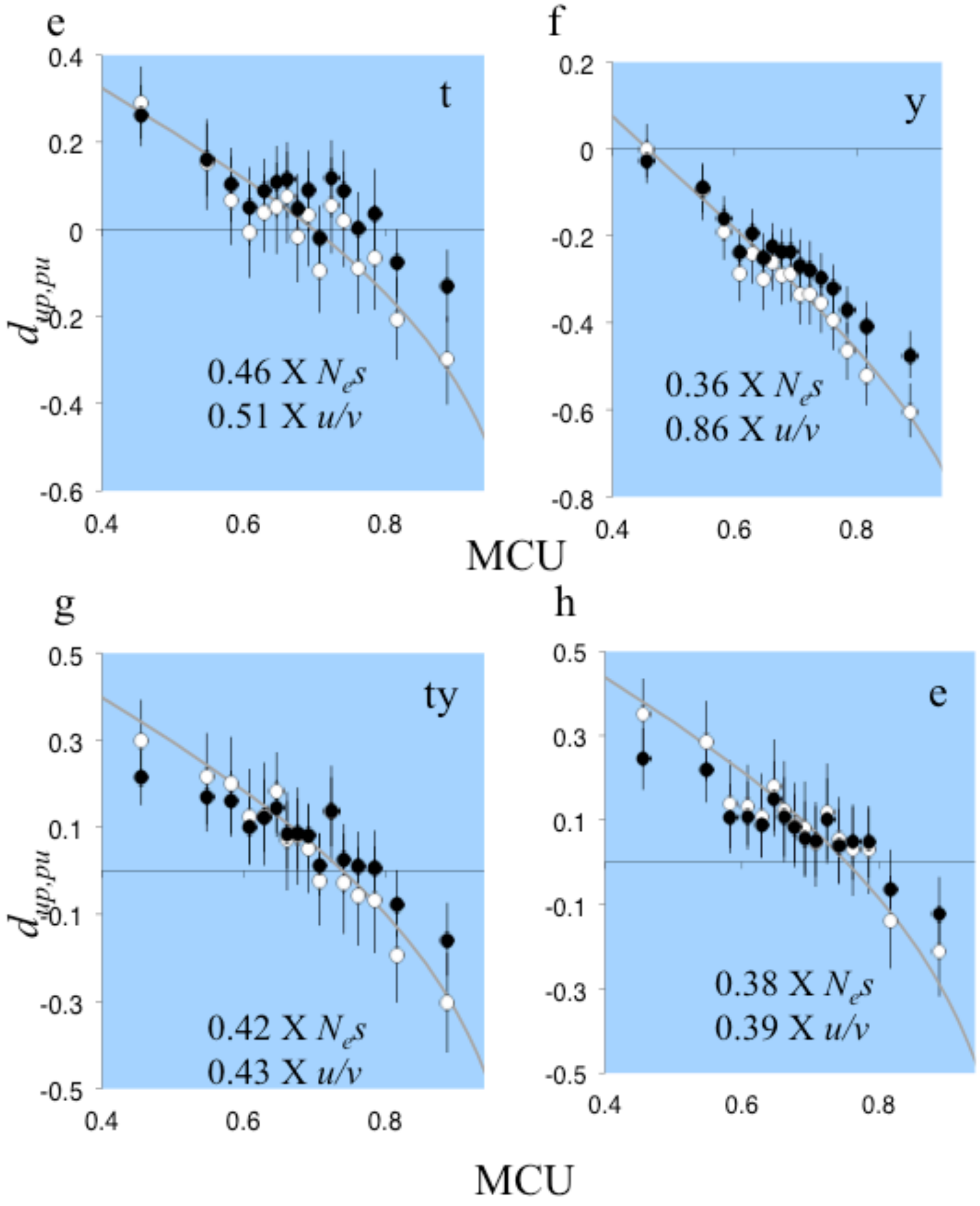


Figure 4.2**. Phylogenetic tree of *D. melanogaster* subgroup.** The tree has been constructed using maximum likelihood method. Branch lengths are the nucleotide distance across 5026 genes.


a
m
0.3 X $N_e s$
2.3 X u/v
b
c
0.27 X $N_e s$
1.3 X u/v
c
s
0.49 X $N_e s$
1.1 X u/v
d
cs
0.4 X $N_e s$
0.88 X u/v
$d_{up,pu}$
MCU
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
0.4
0.6
0.8


e
t
0.4
0.2
0
-0.2
-0.4
-0.6
0.4
0.6
0.8
0.46 X $N_e s$
0.51 X u/v
f
y
0.2
0
-0.2
-0.4
-0.6
-0.8
0.4
0.6
0.8
0.36 X $N_e s$
0.86 X u/v
$d_{up,pu}$
MCU
g
ty
0.5
0.3
0.1
-0.1
-0.3
-0.5
0.4
0.6
0.8
0.42 X $N_e s$
0.43 X u/v
h
e
0.5
0.3
0.1
-0.1
-0.3
-0.5
0.4
0.6
0.8
0.38 X $N_e s$
0.39 X u/v
$d_{up,pu}$
MCU

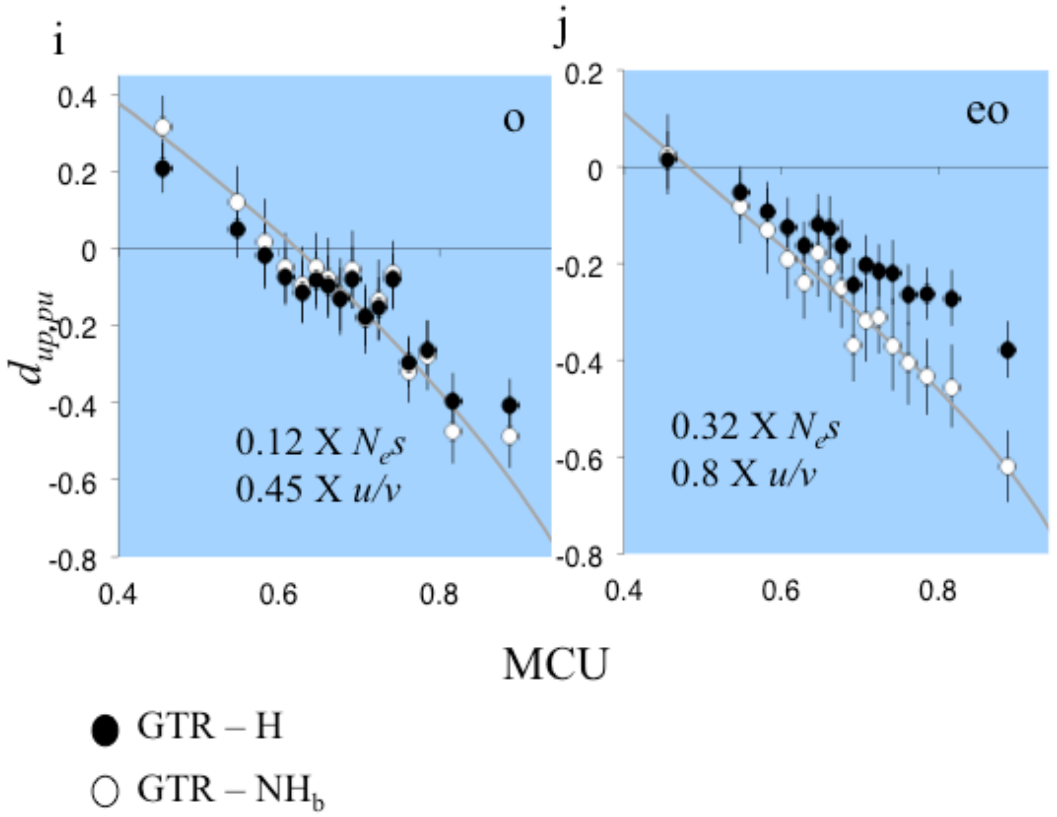


Figure 4.3. **Lineage-specific departures from equilibrium in the *D. melanogaster* subgroup.** Changes in $d_{up,pu}$ as a function of MCU (MCU=#major/(#major+#minor)) in a) *D. melanogaster* (m), b) *D. sechellia* (c), c) *D. simulans* (s), d) *D. sechellia-D. simulans* (cs), e) *D. teissieri* (t), f) *D. yakuba* (y), g) *D. teissieri-D. yakuba* (ty), h) *D. erecta* (e), i) *D. orena* (o), j) *D. erecta-D. orena* (eo) lineages. MCU values are calculated for *D. melanogaster* genes. Each dot represents one bin containing roughly 40,000 2-fold codons. Error bars represent 95% confidence interval of 1000 bootstrap replicates by resampling genes in each bin. Parameters were re-estimated for each replicate. Eyeball estimates of changes in N_es and *u/v* are calculated under the GTR-NH_b model.

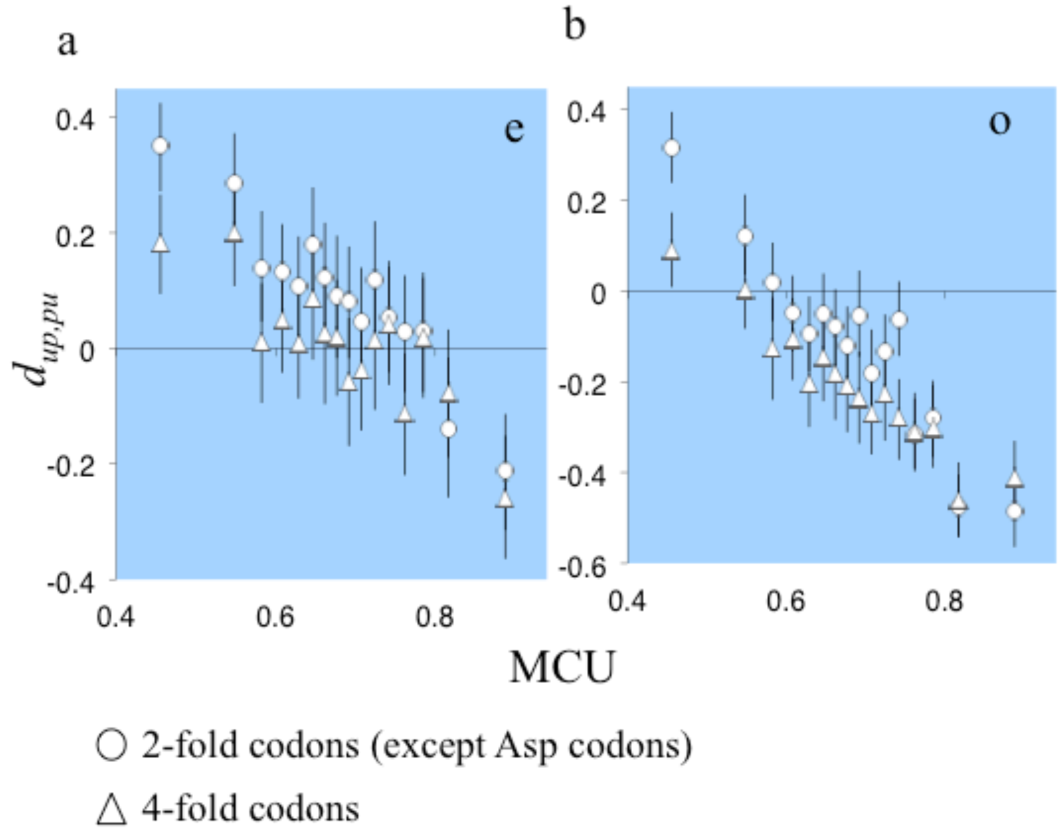


Figure 4.4. **Difference between evolution of 2-fold and 4-fold codons.** Changes in $d_{up,pu}$ under the GTR-NH$_b$ model as a function of MCU (MCU=#major codons/(#major codons+#minor codons)) for 2-fold (except Asp codons) and 4-fold codons in a) *D. erecta* (e), b) *D. orena* (o) lineages. Each point represents one bin containing roughly 40,000 2-fold and 4-fold codons. Error bars represent 95% confidence interval of 1000 bootstrap replicates by resampling genes in each bin. Parameters were re-estimated for each replicate.

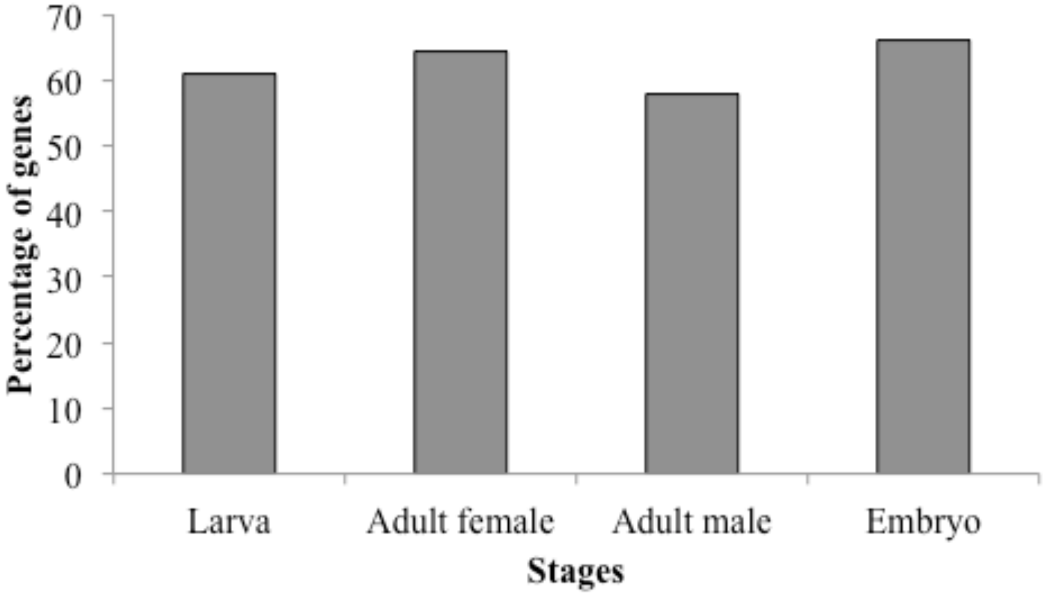


Figure 4.5. **Coverage of genes used in the analysis based on expression across developmental stages.** Percentage of single isoform *D. melanogaster* genes having biased expression in the four developmental stages that had orthologs in all seven species. Stage bias is defined for *D. melanogaster* genes using microarray expression data from FlyAtlas (Chintapalli et al. 2007; Matsumoto et al. 2015).

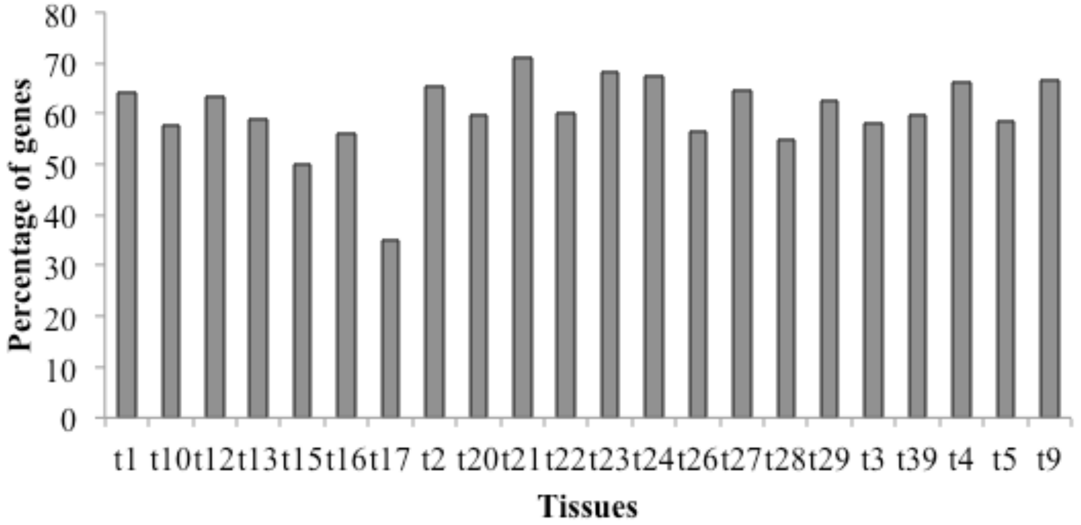


Figure 4.6. **Coverage of genes used in the analysis based on expression across different tissues.** Percentage of single isoform *D. melanogaster* genes having biased expression in the 22 tissues that had orthologs in all seven species. Tissue bias is defined for *D. melanogaster* genes using microarray expression data from FlyAtlas (Chintapalli et al. 2007; Matsumoto et al. 2016). The tissue numbers stand for the following tissues:

t1:Adult hindgut, t2:Adult midgut, t3:Adult male accessory gland, t4:Adult brain, t5:Adult crop, t9:Adult ovary, t10:Adult testis, t12:Adult Salivary gland, t13:Adult carcass, t15:Larval hindgut, t16:Larval midgut, t17:Larval Salivary gland, t20:Larval tubule, t21:Larval fat body, t22:Larval carcass, t23:Larval CNS, t24:Larval trachea, t26:Adult fat body, t27:Adult eye, t28:Adult heart, t29:Adult male ejaculatory duct, t39:Embryo

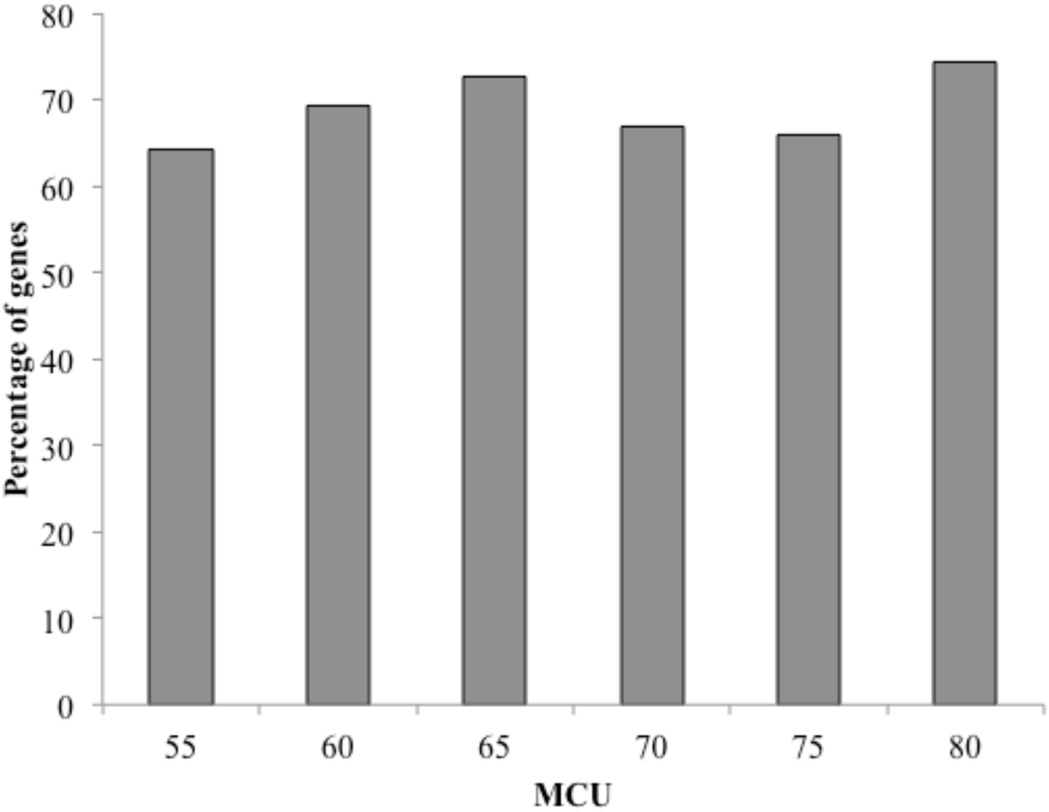


Figure 4.7. **Coverage of genes used in the analysis based on their MCU values.** Percentage of single isoform *D. melanogaster* genes with the given MCU values that had orthologs in all seven species. MCU is calculated for *D. melanogaster* genes