EVOLUTION OF SINGLE AMINO ACID REPEATS IN EUKARYOTIC SPECIES

EVOLUTION OF SINGLE AMINO ACID REPEATS IN EUKARYOTIC SPECIES

By

XIAOYU MU, B. Sc.

A Thesis Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree Master of Science

McMaster University @Copyright by Xiaoyu Mu,

August 12, 2015

McMaster University MASTER OF SCIENCE (2015)
Hamilton, Ontario (Biology)

TITLE: Evolution of Single Amino Acid Repeats in Eukaryotic species

AUTHOR: Xiaoyu Mu, B, Sc (Ocean University of China, P.R.China)
SUPERVISOR: Dr. Brian Golding

NUMBER OF PAGES: vi, 108

ABSTRACT:

A common feature of eukaryotic genomes is the abundance of simple sequences. Single amino acid repeats, which is one kind of simple sequences, are characterized by tandem recurrence of only one amino acid within the proteins and are broadly found among almost all genomes of eukaryotic species. Combined with its abundance, the lack of deterministic function of SAAR makes it intriguing to study on its evolution. In this study, 34 eukaryotic genomes are used and an abundance of SAARs on X/Z chromosomes is observed. Also, amino acid composition and codon usage bias is different between SAARs and non-repetitive regions. We also observe that the conserved number of SAARs is linearly correlated with logarithm of divergence time.

ACKNOWLEDGEMENT:

I would like to thank Dr. Brian Golding for his guidance and support throughout my projects. As an erudite supervisor, his patience helps his students especially me a lot. I would like to thank Dr. Ben Evans for his marvelous and helpful ideas and critiques. As a non-native English speaker, I also would like to thank Dr. Jonathon Stone for his inspiring help for my first TA session. Most of all, my thanks go out to my friends and fellow graduate students. I am really happy and lucky to meet you along this way, especially Yifei Huang, Wilson Sung, Mengxuan Sui, Ben Furman, Nick Yap, Josh Robertson, Yixian Cai, Yue Sun and members of the Golding lab and Evans lab, as well as my friends and family.

Contents

Chapter 1

Introduction

A common feature of eukaryotic genomes is the abundance of simple sequences [Golding, 1999]. This kind of sequence has lower information content because they are composed of only one or few amino acids [Golding, 1999, Huntley and Golding, 2000]. Single amino acid repeats, which is one kind of simple sequence, are characterized by tandem recurrence of only one amino acid within proteins and are broadly found among almost all proteomes of eukaryotic species. In addition, not only do these repeats occur widely, but they also have a huge abundance in some proteomes. For example, in the human genome, 15% to 20% of human proteins have at least one single amino acid repeat with length 5 or longer [Karlina et al., 2002]. In the genomes of mice and rats, this ratio is slightly less than that of human (around 14.9% for mice and 13.7% for rats) but still remains large [Alba and Guigo, 2004]. The human malarial parasite, *Plasmodium falciparum* proteome has an abnormally high frequency of homopolymers [Pizzi and Frontali, 2001]. In addition, among all of the species that have been sequenced and examined, for single amino acid repeats, the genomes of 12 sequenced *Drosophila* species (*D. simulans, D. sechellia, D. melanogaster, D. melanogaster, D. yakuba, D. erecta, D. anaanssae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis, D. grimshawi*) [Stark et al., 2007] have

abundant single amino acid repeats. Because these sequenced *Drosophila* genomes are closely related, they are therefore a great resource to understand the generation, evolution, creation and death of low complexity regions and single amino acid repeats in a comparative way.

Often, based on the 3-dimensional structure, researchers can determine the function of a specific motif in protein. However, low complexity regions and single amino acid repeats have in the past been commonly considered as the protein counterpart of junk DNA. X-ray crystallography can not determine the 3-D structure of low complexity regions and single amino acid repeats, so inference of their function is difficult [Bannen et al., 2008]. If we assume such sequences are non-functional, they are freed from the pressure of selection and should evolve neutrally [Lovell, 2003]. Consistent with the assumption of neutrality, many studies show these sequences can be highly polymorphic and to evolve rapidly among species [Huntley and Golding, 2002, Huntley and Clark, 2007].

However, more and more evidence is emerging that although neutral evolution of noncoding and nonfunctional regions is true, the evolution of single amino acid repeats and low complexity regions is more complex. First, there is evidence that shows that the copy number of single amino acid repeats could be related to some human genetic disorders such as Huntington disease (HD) and spinobulbar muscular atrophy (SBMA) [Usdin, 2008]. As an example, in Huntington disease, the repeat of a CAG codon is normally present in the HTT gene located on chromosome 4 with a length of 6-34 copies. When expansion happens, the CAG codon repeat could reach 36 copies or more in length and patients will have a higher chance to show symptoms. This gene will result in the production of an altered protein form called mHtt and will cause an increase in the decay of neurons in human's brains [Walker, 2007]. In

addition, studies have shown that low complexity regions have a preferred distribution towards TF genes with more than one copy due to whole genome duplication. By comparing the LHX gene family and PHOX gene family, studies have shown that the gain of an alanine repeat in one of the copies significantly increases the capacity of the protein to activate transcription [Trilla et al., 2015].

Second, researchers have found that variation in the length of single amino acid repeats in transcription factors is linked to morphological differences among breeds of dogs [Fondon and Garner, 2004]. A total of 37 different tandem repeats in 17 genes present in the dog genome are related to morphology. In these genes, the shape of dog's skull is partly influenced by expansion and contraction of single amino acid repeats. These genes are those that encode transcription factors that play a role in the formation of the morphologies and lengths of middle faces during development. Also, in these genes, researchers found that the ratio of the lengths of two independent single amino acid repeats within the same gene could also influence the development of the dog's skull. Also, studies have shown that the circadian clock pathway which is controlled by the CLOCK protein has poly-glutamine domains in both *Drosophila melanogaster* and *Mus musculus* that are directly involved in the activation of transcription of downstream genes [King, 1997, Darlington, 1998]. There is evidence that the poly-glutamine repeat located in the CLOCK protein shows changes that reflect local adaptation. O′Malley and Banks [2008] found that the length of poly-glutamine in the CLOCK protein is positively correlated with the latitude of the *Oncorhynchus tshawytscha* populations.

Finally, it is well known that single amino acid repeats have a preference to embed in disordered protein regions [Huntley and Golding, 2002, Simon and Hancock, 2009]. This may show that single amino acid repeats are involved in transient or

low-affinity interactions between different protein substructures which are currently hard to detect, verify and measure using experimental methods [Dunker, 2008]. There is evidence that the distribution of single amino acid repeats could be biased between chromosomes and even different positions of coding regions. For example, in *Drosophila* species, single amino acid repeats are more commonly found near the termini and less often in a central position of proteins [Huntley and Clark, 2007]. This means that if the generation of single amino acid repeats are randomly distributed through proteins, a large proportion of central repeats are eventually eliminated from the population [Huntley and Clark, 2007].

Based on all these observations and studies, it is thought that simple sequences, such as low complexity regions and single amino acid repeats, might be more or less influenced under the action of selection. To search for proof of this hypothesis, Huntley and Golding [2006] used PTDSR which is a kind of phosphatidylserine receptor as an example and prove that the driving force of the contraction and expansion of single amino acid repeats composed of serine amino acids in PTDSR is due to selection rather than slippage [Huntley and Golding, 2006]. To name more examples, the expansion of poly-Q within gene SCA2 in *Homo sapiens* is strongly selected against because this expansion could cause spinocerebellar ataxia type 2 [Pulst et al., 1996]. However, recently, studies also find that the sequence within the single amino acid repeat of SCA2 is under positive selection [Yu et al., 2005].

To detect and measure whether single amino acid repeats are under selection, there are questions that need to be addressed. First, based on the *Drosophila* data, are single amino acid repeats neutrally evolving or under the action of selection? Second, if the single amino acid repeats are under selection, what is the magnitude of selection operating on these repeats? Is it the presence or absence of repeats in genes or is it

the lengths of repeats that are under selection or the codon usage bias within repeats?

The second part of my thesis concerns using single amino acid repeats to explore sex chromosome evolution. In many eukaryotic organisms, the chromosomal complements are different between individuals with different sexes. By definition, the heterogametic individual has a pair of morphologically different chromosomes and the other sex has two identical members of each chromosomal pair of sex chromosome and are called homogametic. Species like human and *Drosophila melanogaster* are male heterogametic and the sex chromosomes are called X and Y (male individuals are XY and female XX). But, there exists another different sex determination system, which is female heterogametic, found in some birds, snakes and insects and the sex chromosomes are referred to as Z and W (male individuals are ZZ and female ZW). Although sex chromosomes are believed to be derived from ancestral autosomes, sex chromosomes have some special characteristics that distinguish them from autosomes. For example, dosage compensation mechanisms have evolved to restore balanced expression of the genome, since sex-linked genes have one copy in heterogametic individuals but two copies in homogametic individuals. In addition, the Y or W chromosomes are male specific or female specific, with quickly evolving regions due to a lack of genetic recombination [Graves, 2006, Goodfellow, 1985].

Since the heterogametic sex individuals only have one copy of X chromosome, compared to two copies of autosomes, this will cause the effective population size of X chromosome linked genes to become lower. If we assume the variation in offspring number for both females and males is purely random, the effective population size of the X chromosome will be three quarters that of autosomes [Charlesworth, 1987, Vicoso and Charlesworth, 2009]. The lower effective population size means X chromosomes can be influenced more by genetic drift. However, the X chromosome

may also experience stronger positive selection than the autosomes since there is no homologous chromosome that could mask recessive mutations so that these mutations will be directly visible to selection and could be directly selected. If a recessive mutation is beneficial, it will take less time for this mutation to be fixed in population. In contrast, if a recessive mutation is deleterious, it will be purged faster from the population. This is known as the "Faster X" effect [Charlesworth, 1987].

From a theoretical perspective, the "Faster X" effect is clearly expected. The empirical results however seem to be complex. Mammals [Torgerson and Singh, 2003, Sequencing and Consortium, 2005, Khaitovich, 2005, Torgerson and Singh, 2006, Baines and Harr, 2007] and birds [Mank et al., 2007] show a "Faster-X" effect. However, some of results from *Drosophila* species cast doubt on the efficacy of "Faster-X" and indicate a negligible or no "Faster-X" effect [Thornton et al., 2006, Vicoso and Charlesworth, 2008]. Other results show small or marginal "Faster-X" effect [Counterman and Noor, 2004, Begun, 2007]. Since the possible genetic cause of "Faster-X" is the same among mammals, birds and *Drosophila*, and since, it is unlikely that the nature of new mutations differs sufficiently among these species to cause a "Faster-X" effect in some clades but not in others, it remains a mystery why different phylogenetic clades could have a different signature of "Faster-X" effect.

So it is intriguing to use data of single amino acid repeats which is sometimes considered neutrally evolving coding region of *Drosophila* species to detect whether there is a signature of "Faster-X" effect in order to try to illustrate the mechanism of "Faster-X" effect clearly.

Besides the 12 sequenced *Drosophila* species from FlyBase, there is also another interesting dataset that could be used to explore the "Faster-X" hypothesis which

is the Neo-X data from a *Drosophila* species called *Drosophila miranda*. *Drosophila miranda* is a close relative of *Drosophila pseudoobscura*, but it has a pair of recently formed neo-sex chromosomes (Neo-X, Neo-Y). The karyotype of the ancestor of all *Drosophila* species is composed of six different Muller-elements (B, C, D, E, F and A; A is the ancestral sex chromosome). About 10-18 MYA, Muller-A and Muller-D of the ancestor of *Drosophila pseudoobscura* and *Drosophila miranda* fused to become one sex chromosome with two arms (XL and XR). Then, about 1MYA in *Drosophila miranda* a fusion event happened between an autosome (Muller-C) and the Y chromosome [Bachtrog and Charlesworth, 2002b]. After this fusion, the *Drosophila miranda* genome now has three different sex chromosomes: one Neo-Y chromosome which is fused to one autosome and the ancestral Y chromosome, one ancestral X chromosome, and the homolog of the fused autosome which is Neo-X chromosome. Over evolutionary time, the Neo-X chromosome has gained the properties of ancestral sex chromosomes such as lacking of recombination with Neo-Y chromosome, dosage compensation and altered effective population size. So based on these properties, we can compare the evolution of single amino acid repeats in genes on the Neo-X chromosome to that of genes on other homologous autosomes in other species to see what the evolution of single amino acid repeats is like in a much younger X chromosome.

My thesis is mainly composed of two parts. All analysis are done concentrating on how single amino acid repeats evolving.

Chapter 2

In chapter 2, using all sequenced genomes found in Ensembl, which mapped protein coding genes to their corresponding chromosomes, we illustrate there is an enrichment of single amino acid repeats on X chromosomes of XY-species. Also,

using this dataset, we show single amino acid repeats have a biased distribution within genes, with biased amino acid composition with relatively longer lengths on sex chromosomes than that of autosomes. For detailed comparative analysis, we used 12 sequenced *Drosophila* species to track how single amino acid repeats are evolving through out the whole *Drosophila* phylogeny. By checking the codon usage bias within both single amino acid repeats and non-repetitive regions, we illustrate that some of the dominant codon for amino acids are changed in SAARs compared to that of non-repetitive regions. Also, using ancestral sequences inferred based on genomes of *Drosophila melanogaster, Drosophila simulans* and *Drosophila yakuba*, we find that SAARs coding regions have a higher ratio of preferred codons change to unpreferred codons versus that of unpreferred codons change to preferred codons compared to that of non-repetitive regions. We study a pairwise comparison between 5 *Drosophila* species pairs and find the number of conserved single amino acid repeats are diminishing in a linear way using the logarithm of divergence time as a predictor. Also, conserved single amino acid repeats are generally longer than non-conserved ones and, the amino acids that compose conserved SAARs are different from those of unconserved ones.

Chapter 3

In chapter 3, with the advent of the next generation sequencing technology, the full picture of sex chromosome evolution is becomming more and more clear after several young sex chromosomes were sequenced. In this chapter, we used the *Drosophila miranda* genome which was recently sequenced, to generate a picture of how single amino acid repeats are evolving on Neo-sex chromosomes and how SAARs could affect the evolution of homologous genes located on Neo-X and Neo-Y chromosomes

after the ceasation of recombination between these two chromosomes. We found that the proportion of genes with SAARs is relatively less on Neo-Y than that on Neo-X. Also, by characterizing the synonymous codon changes in all homologous genes using ancestral sequences inferred based on genomes of *Drosophila pseudoobscura, Drosophila miranda* and *Drosophila persimilis* as outgroup, we compared synonymous codon changes between SAARs coding regions and non-repetitive regions and found there is no significant difference between SAARs and non-repetitive regions on each of the chromosome pairs. In addition, to understand whether SAARs could affect the evolution of Neo-sex chromosome linked genes after the ceasation of recombination, we used PAML to calculate the evolution rates for each gene with homologous genes on all 12 sequenced *Drosophila* species. We found that there are more genes without SAARs on Neo-Y that have high divergence rates than on Neo-X.

Chapter 2

SAAR evolution among eukaryotic species

2.1 Introduction

Single amino acid repeats (SAARs) are a common feature of eukaryotic protein sequences [Golding, 1999]. Previous research has found that SAARs are abundantly distributed in sequenced species especially in proteomes of *Drosophila* species, mammal species and *Plasmodium falciparum* [Huntley and Clark, 2007, Haerty and Golding, 2010]. Previously, single amino acid repeats were mostly considered to be evolving neutrally just like a couterpart of "junk DNA" but located in coding regions. However, several studies have found that single amino acid repeats have potential specific functions or could be potential causes of several abnormal phenotypes. For example, in humans, Huntington's disease is correlated to an expansion of a poly-Q single amino acid repeat in Huntingtin protein [Marcy, 1993]. Similarly, modulated cell adhesion properties of *Saccharomyces cerevisiae* are associated with variation in a single amino acid repeat in the flocculin protein and researchers also proposed associations with repeats counts with the evolution of pathogenicity generated by

genes in *Candida albicans* [Verstrepen et al., 2005, Butler et al., 2009].

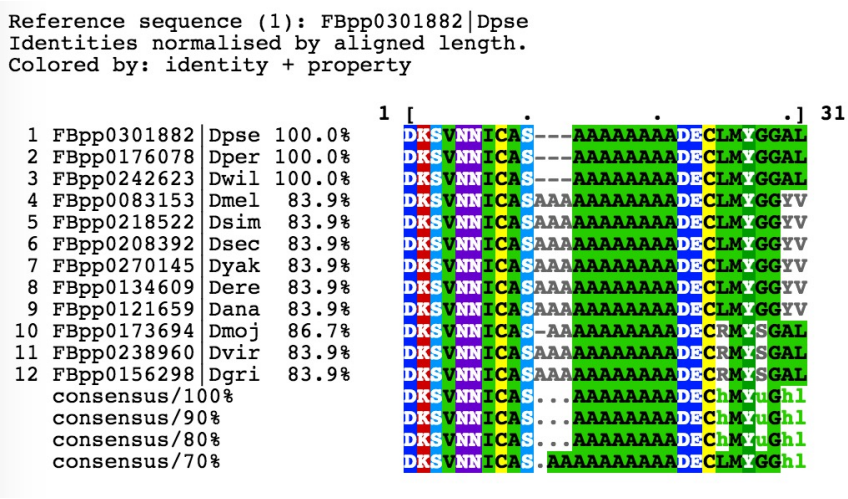


Figure 2.1: Example of a conserved single amino acid repeat (poly-A in the middle) through out the 12 *Drosophila* species, the ID of the protein this segment comes from is FBpp0301882.

It is possible that single amino acid repeats may provide DNA coding sequences or proteins with a source of genetic variability in order to permit rapid adaptation, for example, in an evolutionary arms race between host and parasite [Haerty and Golding, 2011, Marcotte et al., 1999]. In addition, the conservation of a specific single amino acid repeat motif in a protein through evolutionary time could also be a sign of a potential function for the single amino acid repeat. The biased distribution of single amino acid repeats between different part of genes, between different regions of the same chromosome or even between different chromosomes may also be an indicator of functional relevance. Studies have shown that the distribution of single amino acid repeats within proteins in *Drosophila* proteomes is biased to have more single amino acid repeats close to the N-terminus and the C-terminus [Huntley and Clark, 2007]. However, little is known about the distribution of SAARs between different chromosomes within and between species.

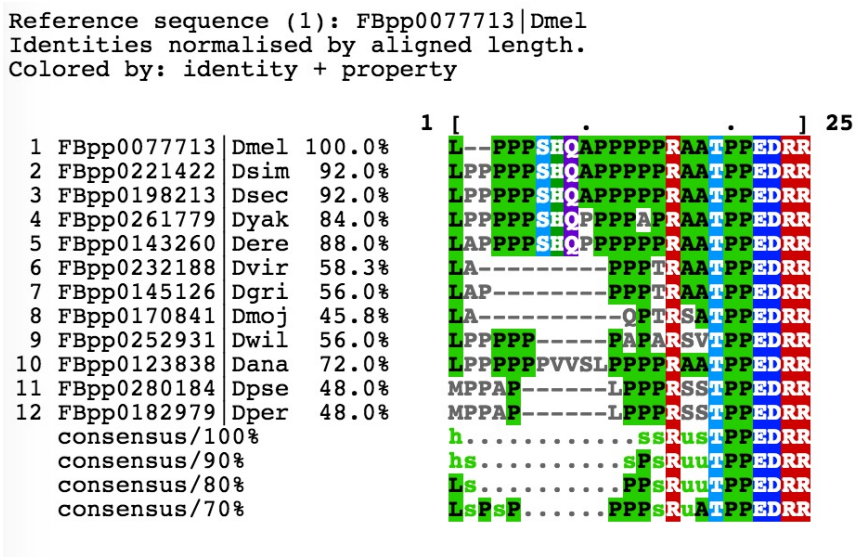


Figure 2.2: Example of an unconserved single amino acid repeat (poly-P in the middle) through out the 12 *Drosophila* species, but conserved in the *melanogaster* group. The ID of the protein this segment comes from is FBpp0077713.

In most eukaryotic organisms, sex chromosomes have a slightly different evolutionary process compared to that of autosomes. Diploid heterogametic individuals have a pair of morphologically different chromosomes (XY male or ZW female). Homogametic individuals possess a pair of identical sex chromosomes (XX female or ZZ male). There are several sex chromosome systems found in eukaryotic species. Two major types are XY-heterogametic system, and ZW-heterogametic system. Examples of species with XY-heterogametic system, are *Homo sapiens* and *Drosophila melanogaster* where male individuals are XY and females are XX. ZW-heterogametic system is commonly found in birds (including finch and chicken), some insects (such as butterflies and moths) and some reptiles (komodo dragon and snake for example). Sex chromosomes are believed to originate from ancestral autosomes. However, sex chromosomes have some properties that distinguish them

from autosomes. For example, when a gene is located on autosomes, it is always present in two copies. If one copy of the gene has a recessive deleterious or recessive beneficial mutation, it's effect can be masked by its counterpart with the dominant allele. However, for sex-chromosome-linked genes there is only one copy in one of the sexes and so, genes with mutations will be favored or selected against directly in that sex and there is no dominant copy to mask the effect of the mutation. Based on this reasoning, if SAARs are acted upon by selection, they might respond differently on sex chromosomes versus autosomes.

In this chapter, we will illustrate that how single amino acid repeats have a biased distribution between sex chromosomes and autosomes among many eukaryotic species. We will present comparisons of length, amino acid composition, location within genes, conservation through phylogenetic levels between autosome-linked and sex-chromosome-linked single amino acid repeats among eukaryotic species. Furthermore, we will use 12 sequenced *Drosophila* species to explore different aspects such as conservation, codon usage bias and sex chromosome bias to illuminate the evolution between sex-chromosome-linked and autosome-linked single amino acid repeats.

2.2 Materials and methods:

2.2.1 Biased distribution of single amino acid repeats between sex chromosomes and autosomes in eukaryotic species

All genomes that were used in the analysis of single amino acid repeats biased distrbution among chromosomes were downloaded from Ensembl (*www.ensembl.org*) with the exception of the 12 *Drosophila* species genomes which were downloaded from FlyBase *(www.flybase.org)*. The versions of these genomes are summarized in Table 2.1. We didn't include some sequenced genomes from Ensembl due to several reasons: 1. Some of these genomes (for example, *Ailuropoda melanoleuca* and *Xenopus tropicalis*) haven't mapped genes to their corresponding chromosomes so that the mapping information can only link genes with contigs or scaffolds instead of chromosomes. 2. Some species (such as *Saccharomyces cerevisiae*, *Ciona intestinalis*) have no sex chromosome so that we excluded from our dataset. After both filtering processes, there remains 31 XY species (including 12 *Drosophila* species) and 3 ZW species. There are two special cases that should be noted. First, since the *Platypus* genome has five sequenced X chromosomes, we considered genes from any one of these five X chromosomes as X-linked genes. In addition, in several *Drosophila* species, X chromosomes could be formed by two different Muller-elements (for example, in *Drosophila pseudoobscura*, *Drosophila persimilis* and *Drosophila willistoni*, Muller-element A and D are fused into one X chromosome with XL and XR arms), genes linked with those kinds of Muller-elements are all considered X-linked. Furthermore, for *Drosophila* species for which scaffolds and contigs are not yet mapped onto chromosomes, Synpipe data from the AAA website (*http://rana.lbl.gov/drosophila/*, Bhutkar et al. [2011]) was used to associate scaffolds or contigs with chromosomes.

Species	Abbreviation	Genome assembly version
Homo sapiens	H.sap	GRCh38.p2
Pan troglodytes	P.tro	CHIMP v2.1.4
Gorilla gorilla	G.gor	gorGor v3.1
Macaca mulatta	M.mul	MMUL v1.0
Callithrix jacchus	C.jac	C jacchus v3.2.1
Papio anubis	P.anu	PapAnu v2
Pongo abelii	P.abe	PPYG v2
Chlorocebus sabaeus	C.sab	ChlSab v1.1
Oryctolagus cuniculus	O.cun	OryCun v2.0
Rattus norvegicus	R.nor	Rnor v5.0
Felis catus	F.cat	Felis catus v6.2
Bos taurus	B.tau	UMD v3.1
Caenorhabditis elegans	C.ele	WBcel235
Equus caballus	E.cab	EquCab v2
Meleagris gallopavo	M.gal	UMD v2
Monodelphis domestica	M.dom	BROADO v5
Ornithorhynchus anatinus	O.ana	OANA v5
Ovis aries	O.ari	Oar v3.1
Gallus gallus	G.gal	Galgal4
Canis familiaris	C.fam	CanFam3.1
Sus scrofa	S.scr	Sscrofa v10.2
Taeniopygia guttata	T.gut	taeGut v3.2.4
Drosophila simulans	D.sim	Dsim r1.4
Drosophila sechellia	D.sec	Dsec r1.3
Drosophila melanogaster	D.mel	Dmel r5.33
Drosophila yakuba	D.yak	Dyak r1.3
Drosophila erecta	D.ere	Dere r1.3
Drosophila ananassae	D.ana	Dana r1.3
Drosophila pseudoobscura	D.pse	Dpse r3.2
Drosophila persimilis	D.per	Dper r1.3
Drosophila willistoni	D.wil	Dwil r1.3
Drosophila mojavensis	D.moj	Dmoj r1.3
Drosophila virilis	D.vir	Dvir r1.2
Drosophila grimshawi	D.gri	Dgri r1.3

Table 2.1: Genomes used to examine in biased chromosome distribution and their genome assembly versions, based on information from Ensembl, FlyBase and WormBase.

For all species, we scanned each genome and extracted all protein coding DNA sequences using BioPython [Cock et al., 2009] based on the gff3 file of each assembly. Then, we used a customized Python script to translate all these protein coding DNA sequences into protein sequences. After that we clustered all proteins that were translated by the same protein coding gene and only kept the longest one so that we could eliminate the biased effect of genes that have multiple isoforms due to alternative splicing. After all these steps, we could use this dataset as our final data to analyze the distribution of single amino acid repeats between X/Z chromosomes and autosomes.

We used a customized Python script to screen all proteins from each genome for single amino acid repeats which were at least five amino acids long. For more stringent analysis, we also increased the length cutoff to seven amino acids long. For each single amino acid repeat, we recorded the composition of amino acid, the length of repeats, the start and end locations corresponding to the N-terminus of the protein sequences which the SAARs are in and which chromosome these SAARs located on.

For all species, we seperated all genes into two different categories, sex-chromosome-linked genes and autosome-linked genes. For each category, we counted the total number of proteins and the proteins with at least one single amino acid repeat and the total number of single amino acid repeats in that category. Then after that, we counted the proportion of genes with at least one single amino acid repeat within these two categories. We then plotted the ratio of average number of SAAR per gene on X/Z chromosome versus that of autosome with cutoff 5 or 7 (Figure 2.4 and Figure 2.5). We then assessed the difference in chromosomal distributions within genomes using a binomial test. The null hypothesis is that the proportion of

genes with at least one single amino acid repeat on autosomes is the same as that on sex chromosomes and the alternative hypothesis is that the proportion of genes with at least one single amino acid repeat is different from autosomes and sex chromosomes.

2.2.2 Biased distribution of single amino acid repeats within protein coding genes

For each gene with at least one single amino acid repeat, we recorded the length of the encoded protein and the begin/end position relative to the N-terminus of the protein. We calculated the middle position of the repeat and divided by the length of protein to calculate the location of middle point of that specific repeat relative to the protein it is located on. Then, we divided the protein length into 25 bins and each bin has four percent of protein length and we counted how many single amino acid repeats are within each bin for both X linked single amino acid repeats and autosome-linked SAARs. Then we plot the counts of single amino acid repeats located in each bin, within proteins for ZW species, for XY species and for sex chromosomes and autosomes. To test whether the general presence of SAARs is terminal-biased, we divided each protein into 5 bins for simplicity. The first and last bin, if we assume SAARs are randomly distributed within each proteins, should in total have 40% of the SAARs. We further calculated the frequency of the presence of all SAARs for each species and compared that observed frequency to 40% which is the expected frequency for SAARs present in the first and last bin.

2.2.3 Lengths of single amino acid repeats among different species

If SAARs are X or Z linked, they might act differently on sex chromosomes versus autosomes, since genes located on X/Z chromosomes are directly exposed to selection. Also, due to the fact that previous studies have shown that the contraction and expansion of single amino acid repeats are correlated with the function of the repeat, it is meaningful to compare the general length difference of single amino acid repeats between X/Z chromosomes and autosomes.

To test whether there are different length distributions of single amino acid repeats between sex chromosomes and autosomes among different species, we used 5 amino acids as a length cutoff to search for all SAARs in the "no-redundency" protein sequences (the longest isoform of each protein) we used in the previous analysis from 34 species to search for single amino acid repeats and record their length and whether it is sex-chromosome-linked or autosome-linked. We then seperated all these single amino acid repeats into two parts based on chromosome linkage. Then, we used the NumPy package [Walt et al., 2011] to calculate the mean and variance of lengths of SAARs both on autosomes and sex chromosomes for each species. To check whether the average of lengths of single amino acid repeats located on sex chromosomes are different from that of SAARs located on autosomes, we used R [R Development Core Team, 2008] to do both two-tail and one-tail two-sample tests between lengths of autosome-linked SAARs and sex chromosome linked SAARs.

2.2.4 Amino acid composition comparison between single amino acid repeats with different lengths and between different chromosomes.

We set up three different length cutoffs (5, 7, 12 amino acids long) to scan for single amino acid repeats both on X chromosomes and autosomes. For each single amino acid repeat found longer than the cutoff, we recorded its amino acid composition. As a control, amino acid content through out the protein was calculated both for autosomes and X chromosomes (Table 2.4). We calculated the proportion of each type of amino acid within the dataset only contains SAARs. We used these proportions on X chromosome or autosomes to answer two questions: first, whether the proportion of one type of amino acid in single amino acid repeats is different from the proportion of the same kind of amino acid from non-repetitive regions? Second, whether the proportion of one specific amino acid in X-linked single amino acid repeats is different from the proportion in autosome-linked single amino acid repeats? We used the two-tail binomial test to answer these questions (Table 2.5, Table 2.6).

We scanned all the single amino acid repeats through out each genome (only keeping the longest isoform for each gene) and recorded the length of the single amino acid repeats with length at least five amino acids long both for X chromosomes and autosomes of all species. Then we used the NumPy package [Walt et al., 2011] to calculate the mean and variance of the lengths of single amino acid repeats (Table 2.6).

To better understand the abundance of amino acid component, we also did a repeat-based analysis. For all of the amino acid repeats, we counted all single amino acid repeats based on the amino acid type with a length cutoff both at least 5 and 7

amino acids. For each amino acid type, we seperately counted the number of single amino acid repeats located on X chromosomes and autosomes. After all counts were finished, these counts were divided by the total number of repeats located on sex chromsomes or autosomes to get the proportion of that type of single amino acid repeats so that the proportion would be comparable between sex chromosomes and autosomes.

2.2.5 Comparison of dominant codons within SAARs in *Drosophila melanogaster*

Biased codon usage in many species results from a balance among mutation, selection and genetic drift [Akashi and Walker, 1998]. The difference of relative importance between these driving powers could shape the evolution of codon usage bias of specific regions. Very few studies have examined the nature of evolution within the sequence of single amino acid repeats. It is meaningful to use codon usage bias as a tool to understand the evolution within single amino acid repeat sequences. Since the codon usage bias can occur on a short evolutionary time scale, we used the genomes of *Drosophila melanogaster* to study codon usage bias within SAARs. All the protein coding genes with no single amino acid repeats (greater than 5 amino acids long) were discarded from this dataset.

To check whether there is potential difference in terms of the composition of codon usage between single amino acid repeats and non-repetitive regions, we concatenated the 5' flanking coding regions from *Drosophila melanogaster* and 3' flanking coding regions, so that there will be no single amino acid repeats contained in these DNA sequences. The non-repetitive sequences and single amino acid repeats

were seperated into two classes: X chromosome linked and autosome linked. Thus, there are four types of protein coding DNA sequences: X-linked single amino acid repeats sequences, X-linked non-repeat sequences, autosome-linked single amino acid repeat sequences and autosome-linked non-repeat sequences. For each we calculated the relative synonymous codon usage (RSCU) value using DAMBE [Xu, 2013]. The RSCU value is defined as the ratio of the observed frequency of codons to the expected frequency given that all the synonymous codons for the same amino acids are used equally [Sharp et al., 1986].

2.2.6 Codon usage bias within SAARs of *Drosophila melanogaster*

Comparison of codon usage bias between genomes could provide us with a different perspective about the forces that contribute to and affect the evolution of codon usage bias. We filtered the longest protein of each gene in *Drosophila melanogaster*, *Drosophila simulans* and *Drosophila yakuba* genomes and used BLAST (version 2.2.28) [Altschul et al., 1990] to search for all homologous longest isoforms. Then we used MAFFT(version 7.221, [Katoh and Standley, 2013]) to align these homologous proteins. After that we scanned through all the alignments and identified all the single amino acid repeats in them and record their positions, lengths and amino acid composition using a customized Python script. If a pair of repeats were found to be homologous to a single amino amino acid repeat in the other species, then we characterized this pair as homologous SAARs. Based on protein alignments and positions and lengths of single amino acid repeats, we can extract the DNA sequence coding for homologous single amino acid repeats. We then aligned corresponding codons based on the protein alignments. Using these codon alignments, we applied

FastML [Ashkenazy et al., 2012] in our study to infer the ancestral state of each codon. To measure the direction and evolution of codon usage bias, we classfied synonymous codons into 2 groups: preferred codons (P) and unpreferred codons (U) based on the RSCU values we calculated in previous section. Codons with RSCU values larger than one are classified as preferred and RSCU values smaller than one as unpreferred codons. Based on these assumptions, if one codon in inferred ancestral sequence changes to another homologous codon in *Drosophila melanogaster*, there are four different scenarios depending on the dominant state of codon before and after: preferred to preferred (no change), unpreferred to unpreferred (no change), preferred to unpreferred (P2U) and unpreferred to preferred (U2P). Here we don't compare codon changes within the same dominant state before and after, so we assumed codon change with same dominant state (both preferred or both unpreferred) to be "no change". Also, even if a synonymous codon with slightly lower/higher RSCU value change to a codon with higher/lower RSCU value, as long as they are all considered "preferred" or "unpreferred", they were characterized as "no change". In the end, we counted all X-linked and autosome-linked fixed codon changes in three characterizations: "No change", "P to U", "U to P" both in single amino acid repeat regions and non-repetitive regions. The results are shown in Table 2.10.

2.2.7 The conservation of single amino acid repeats within *Drosophila* species

Since 12 *Drosophila* species have been sequenced [Clark, 2007], this dataset can provide us a resource for comparative research into the conservation of single amino acid repeats between different evolutionary scales. We used the divergence times estimated by Tamura to group 10 sequenced *Drosophila* species

into 5 groups (*D.melanogaster-D.simulans, D.yakuba-D.erecta, D.pseudoobscura-D.persimilis, D.willistoni-D.ananassae, D.grimshawi-D.virilis*), each group with different divergence times. These 10 groups are chosen based on the phylogeny of all 12 sequenced *Drosophila* species, all pairs of species not only have different divergence times, but also between different pairs, there is no overlapping phylogenetic branch. In this way we can make sure that the evolution of SAARs are independent from other groups. For each group, we aligned the homologous proteins to find homologous single amino acid repeats using MAFFT. We adopt the following filtering rules: First, single amino acid repeats should have at least one amino acid overlap. Second, the flanking regions (with lengths of 10 amino acids) of homologous single amino acid repeats should have 2 mismatches at most. Then we counted the number of single amino acid repeats conserved in each group. Based on the divergence time within each group, we plotted the divergence times as predictor variables versus the counts of conserved single amino acid repeats as response variables.

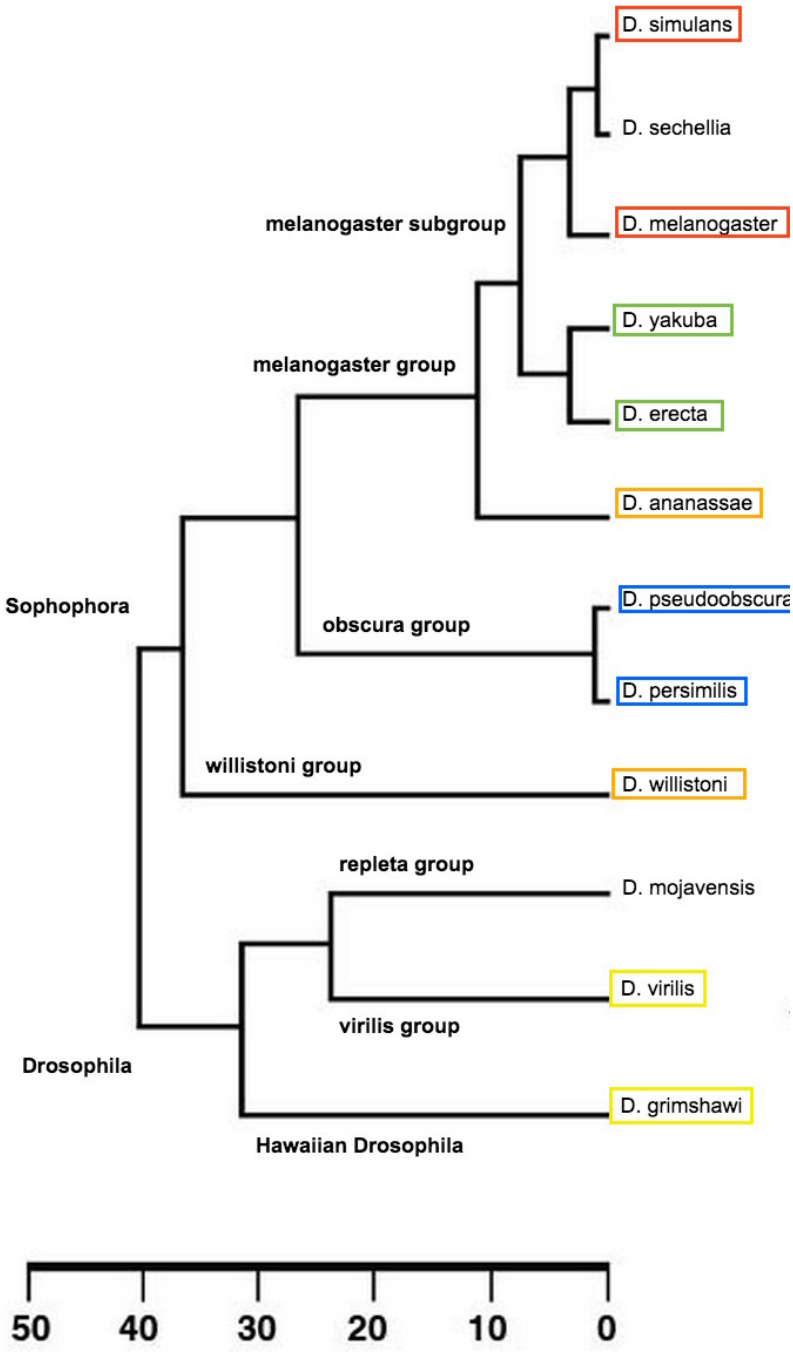


Figure 2.3: Division of 10 *Drosophila* species with different divergence times and with no sharing branch in phylogeny.

Group	Divergence Time(MY)
Dpse-Dper	1.7
Dmel-Dsim	10.8
Dyak-Dere	25.6
Dvir-Dgri	85.8
Dana-Dwil	124.4

Table 2.2: Five *Drosophila* groups with their divergence times as estimated by Tamura [2004].

2.3 Results

2.3.1 Biased distribution of SAARs between sex chromosomes and autosomes

We compared the average number of single amino acid repeats per protein coding genes located on sex chromosomes and autosomes of 34 sequenced species including 17 mammal species, 1 other mammal species (platypus), 3 *Laurasiatheria* species, 12 *Drosophila* species and 1 nematode. If we use all the single amino acid repeats which are at least 5 amino acid long and compare this ratio between sex chromosomes and autosomes, 30 out of 34 species have an enrichment of single amino acid repeats on the sex chromosomes compared to autosomes. Within this dataset, 29 out of 31 XY species have an enrichment of single amino acid repeats on X chromosome. For those three ZW species, only one out of three has a higher ratio for single amino acid repeats counts to numbers of protein coding genes. This means that we do not find evidence for the same pattern of enrichment of single amino acid repeats in ZW species as seen in XY species. However, this could due to the fact that the sample sizes of single amino acid repeats are too small for *Gallus*

gallus and *Taeniopygia guttata*. For *Gallus gallus*, only 138 single amino acid repeats were found located on the Z chromosome, and for *Taeniopygia guttata*, that number is only 60. In contrast, the number of single amino acid repeats we found located on the X chromosome of *Drosophila melanogaster* is 2146. So, the apparent absence of SAARs enrichment on Z chromosomes could be due to a small sample size.

If we use a more stringent definition of the length of single amino acid repeats to at least 7 amino acids long, we observe 32 out of 34 species with an enrichment of single amino acid repeats on sex chromosomes compared to autosomes. Within this dataset, 30 out of 31 XY species has a higher ratio of SAARs counts to protein coding genes on X chromosome than on autosomes with the only exception being the genome of Platypus. This exception could be possibly due to the fact that only 504 proteins are verified to be coded by genes located on the X chromosomes in *Platypus* and distributed on the five X chromosomes unevenly. There are 310, 17, 26, 0, 151 sequenced proteins coding genes located on chromosomes X1 through X5.

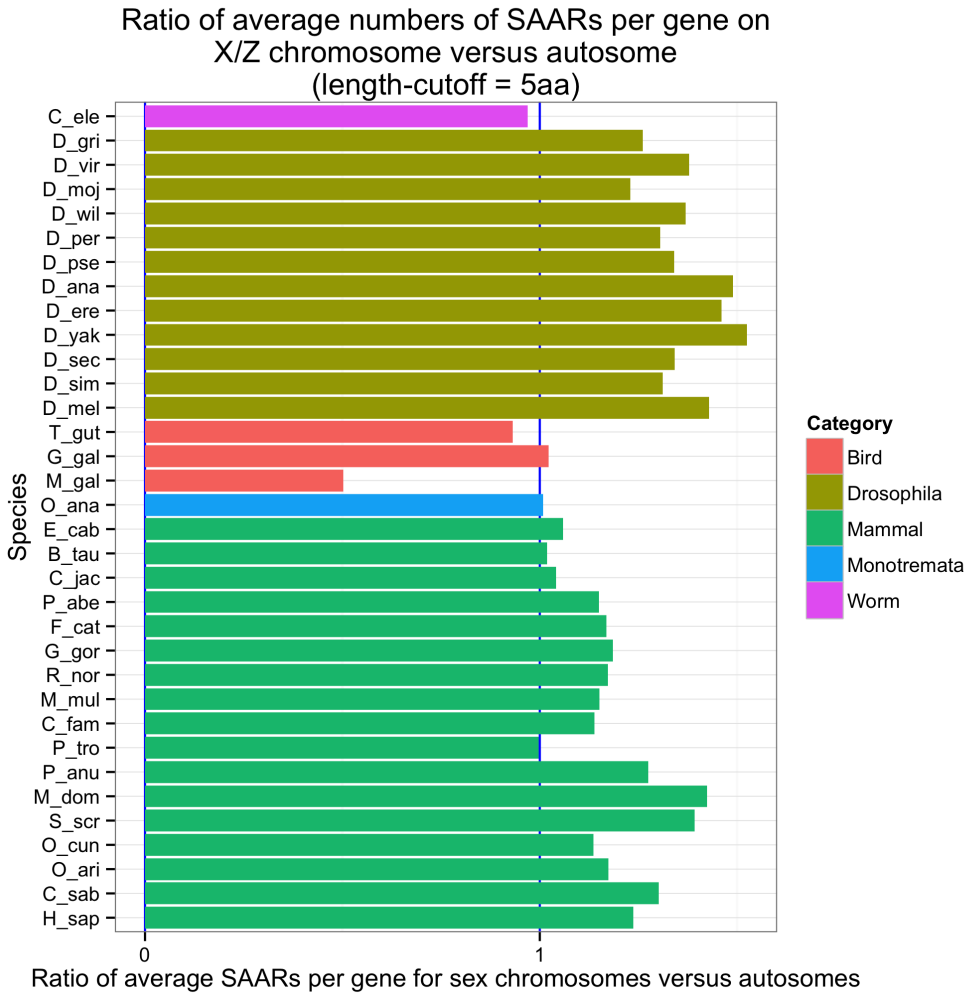


Figure 2.4: Comparison of ratio of the counts of SAARs at least 5 aa long to the counts of genes between autosomes and sex-chromosomes of 34 species.

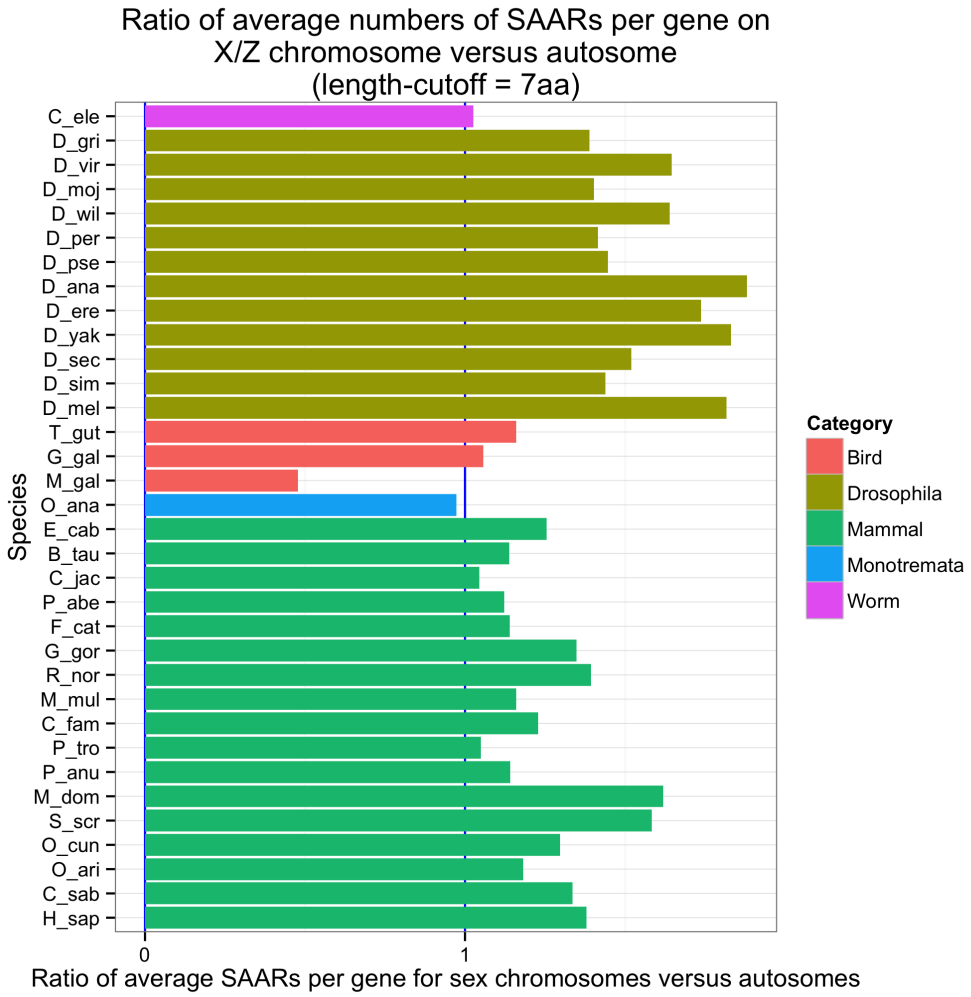


Figure 2.5: Comparison of ratio of the counts of SAARs at least 7 aa long to the counts of genes between autosomes and sex-chromosomes of 34 species.

2.3.2 Biased distribution of single amino acid repeats within protein coding genes

Figure 2.6 and Figure 2.7 show the distribution of the positions of repeats throughout the length of proteins on autosomes and X/W chromosomes. If the repeats are randomly dispersed throughout the protein coding sequences, we should expect all bins throughout the length of the genes to have almost the same number of single amino acid repeats. It has been previously noted that there is a biased distribution of repeats throughout the length of proteins in *Drosophila* species [Huntley and Clark, 2007]. In addition, studies on plant gene sequences have shown that the density of nucleotide microsatellites is higher towards the start of transcription start position. Our results are consistent with the results of Zhang et al. [2006] that there is a tendency for more amino acid repeats on the N-terminal end of proteins. However, when we compare the patterns shown in autosomes and that of X/Z chromosomes, we can see that the distribution of single amino acid repeats on X/Z chromosomes is more randomly distributed in proteins on the X chromosome than on autosomes. To evaluate whether it is true that the presence of SAARs is biased towards both ends of proteins, we divided each protein into 5 bins and recorded for each SAAR, which bin contains it. If we assume the presence of SAARs is random, the expected probability for us to observe a SAARs is present in the first bin or the last bin is 40%. We further used the observed number of SAARs presenting in first or last bin to calculate the probability using a binomial distribution for each species. We then compared the observed frequencies to the expected frequency. We found that the presence of SAARs are more general on autosome (34 out of 34 species have much higher observed frequency than expected frequency in the first and the last bins) than that of sex chromosomes (30 out of 34 species have a higher observed frequency than expected frequency on sex chromosomes in the first and the last bins)

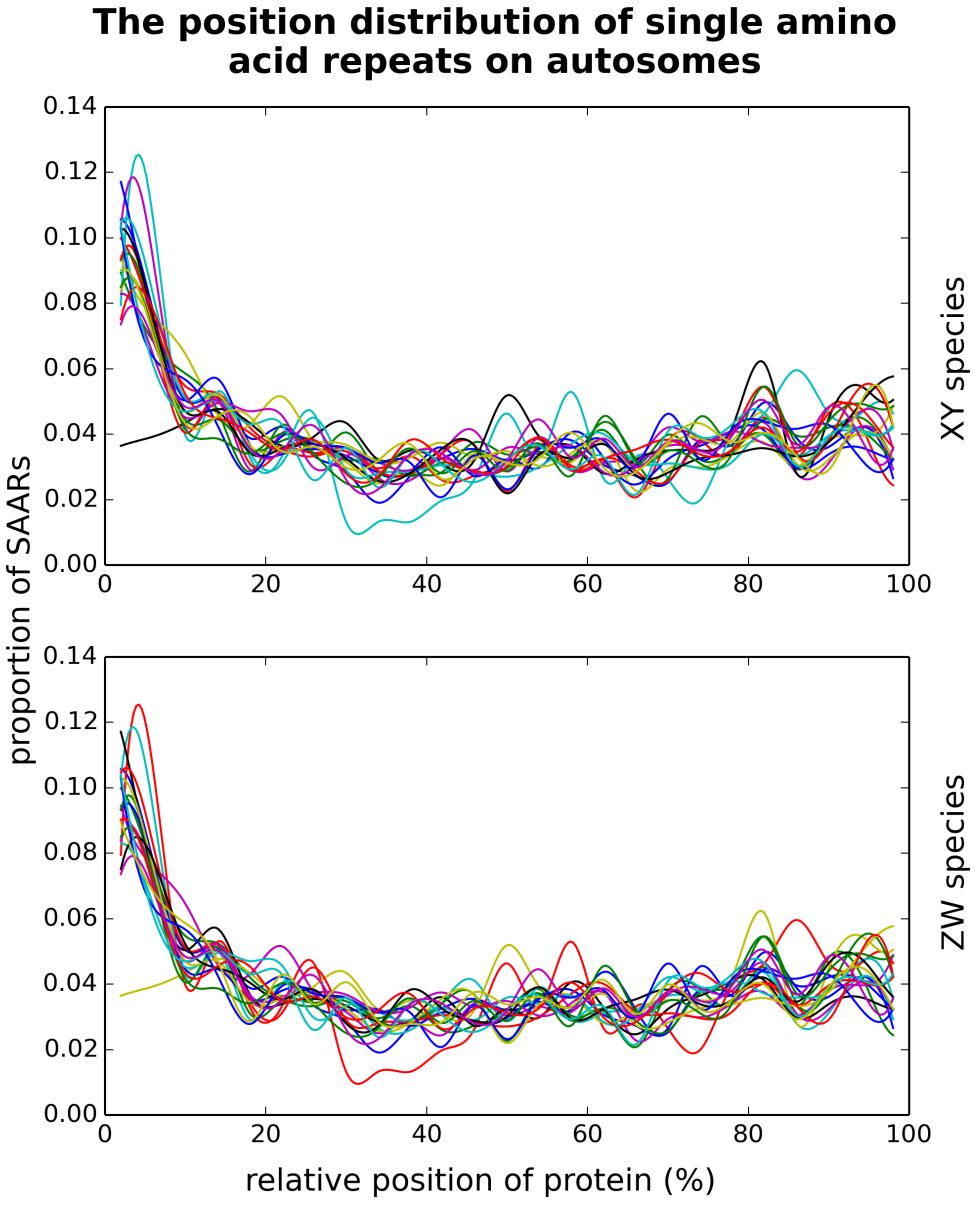


Figure 2.6: The positions of SAARs throughout the length of proteins on autosomes.

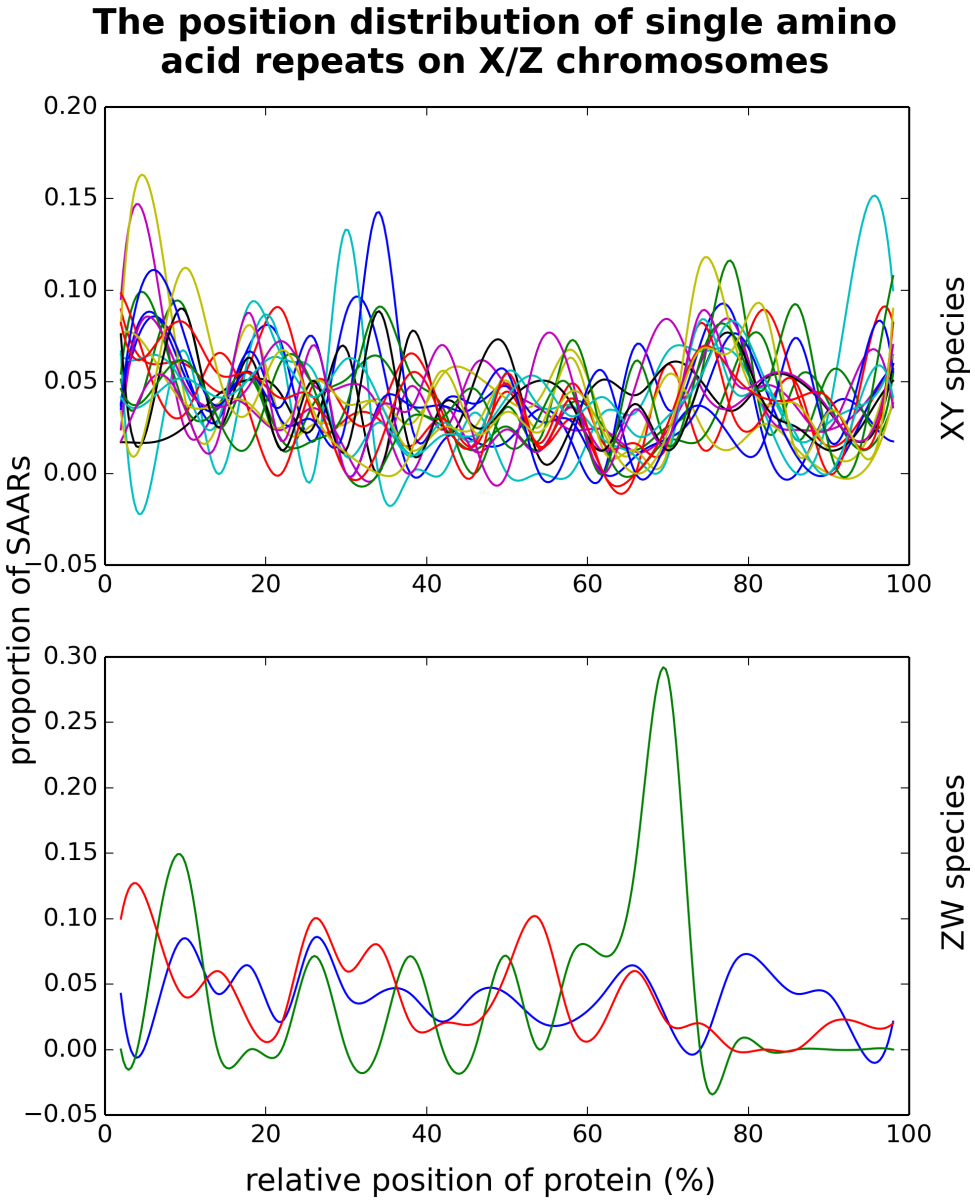


Figure 2.7: The positions of SAARs throughout the length of proteins on X chromosomes or Z chromosomes.

2.3.3 Lengths of single amino acid repeats among different species

We scaned all our genome data and searched for single amino acid repeats with lengths of at least 5 amino acids long both on sex chromosomes and on autosomes. Then we calculated all the means and variances for lengths of SAARs both on sex chromosomes and autosomes and the results are shown in Table 2.3. We can see from the table that the average length of sex chromosome linked single amino acid repeats is in general longer than that of autosome-linked SAARs and in addition (except for *Meleagris gallopavo* and *Ornithorhynchus anatinus*), the variances of lengths are also larger than that of autosomes linked to single amino acid repeats. We did one-tail t-tests with the null hypothesis that the means of lengths of sex-chromosome-linked SAARs are equal to autosome-linked SAARs. The alternative hypothesis is that the mean lengths of sex-chromosome-linked SAARs are higher than that of autosomes-linked SAARs. Since most of sex chromosomes only have 100-150 single amino acid repeats detected, we use a relatively loose significance level $\alpha = 0.1$ as cutoff and all the species without a significant difference are marked with '-' by their species names. From the table we can see that most of the species (26 out of 34 species) have significantly longer single amino acid repeats on sex chromosomes than those on autosomes. In species where this is not true, the number of sex-chromosome-linked SAARs are all less than 200 except for *Caenorhabditis elegans*, with 524 sex-chromosome-linked SAARs and 3378 autosome-linked SAARs. So it is possible that most of the species with non-significant longer sex-chromosome-linked SAARs are due to lack of sample size. In terms of 12 *Drosophila* species, all species show the pattern that SAARs on sex chromosomes, are significantly longer than those located on autosomes. In addition, all 12 *Drosophila* species have higher numbers of single amino acid repeats on sex chromosomes than other species (the average number of sex chromosome linked single amino acid repeats

is 2792 in contrast to that of autosome-linked single amino acid repeats which is 6390) of 12 *Drosophila* species.

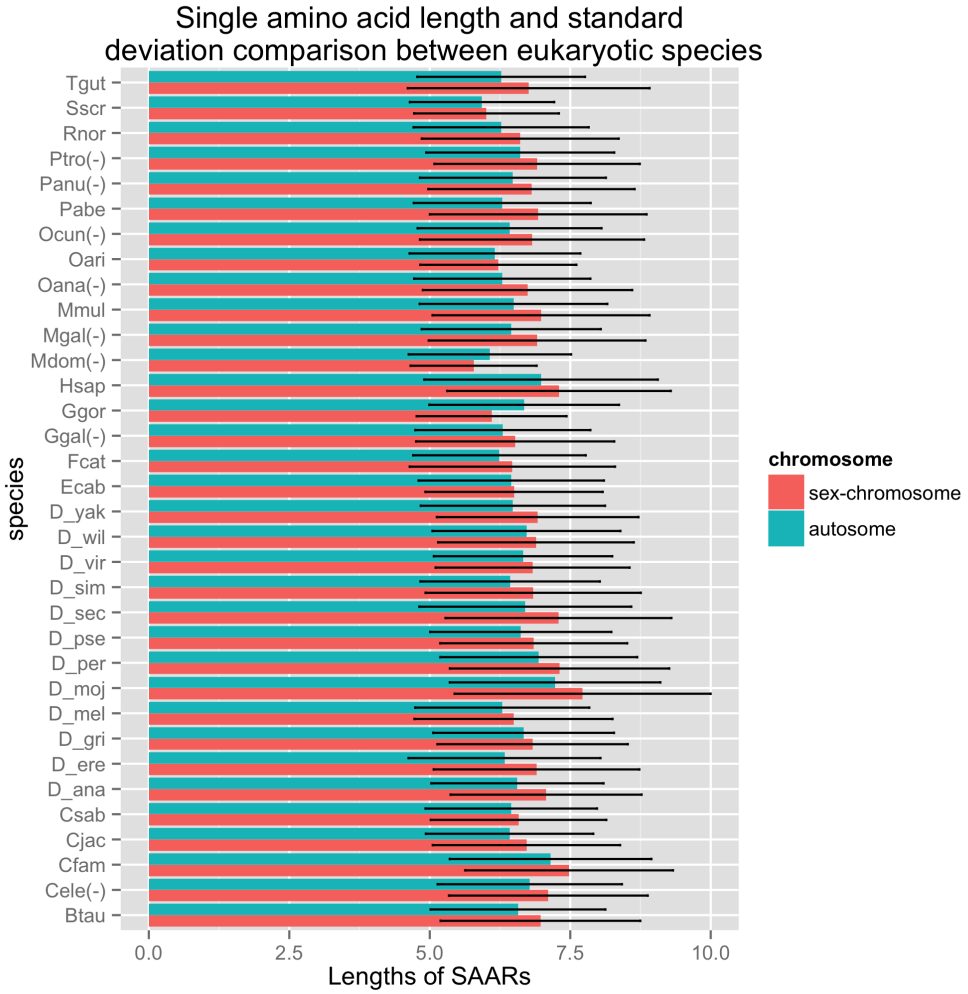


Figure 2.8: comparison between lengths of SAARs located on X/Z chromosomes and autosomes in all eukaryotic species.

Table 2.3: Means and variances of single amino acid repeats of 34 species.

species	mean-SexChr	variance-SexChr	mean-Autosome	variance-Autosome
Btau	6.76	4.63	6.27	2.23
Cele (-)	6.01	1.66	5.93	1.65
Cjac	6.61	3.07	6.27	2.43
Cfam	6.91	3.34	6.61	2.8
Csab	6.81	3.37	6.48	2.74
Ecab	6.93	3.72	6.29	2.48
Fcat	6.82	3.98	6.42	2.68
Ggal (-)	6.22	1.93	6.16	2.31
Ggor	6.74	3.47	6.29	2.46
Hsap	6.98	3.73	6.49	2.79
Mmul	6.91	3.72	6.45	2.54
Mgal (-)	5.78	1.26	6.07	2.09
Mdom (-)	7.3	3.96	6.98	4.32
Oana (-)	6.1	1.78	6.68	2.85
Ocun (-)	6.52	3.11	6.3	2.43
Oari	6.47	3.34	6.24	2.36
Ptro (-)	6.5	2.49	6.45	2.72
Panu (-)	6.49	3.11	6.29	2.4
Pabe	6.84	3.67	6.43	2.55
Rnor	7.29	4.04	6.7	3.56
Sscr	6.92	3.22	6.48	2.69
Tgut	6.9	3.35	6.33	2.92
Dana	7.07	2.89	6.56	2.35
Dere	6.85	2.77	6.62	2.6
Dgri	7.31	3.81	6.94	3.07
Dmel	6.89	3.03	6.72	2.8
Dmoj	7.72	5.19	7.23	3.51
Dper	6.83	2.98	6.66	2.52
Dpse	6.83	2.87	6.67	2.59
Dsec	6.58	2.44	6.45	2.33
Dsim	6.72	2.78	6.42	2.22
Dvir	7.48	3.42	7.15	3.22
Dwil	7.11	3.13	6.78	2.69
Dyak	6.97	3.16	6.57	2.42

2.3.4 Amino acid proportion comparison between single amino acid repeats with different lengths and between different chromosomes

We can observe from Table 2.4 that the frequencies of I, F, W, Y, V, N, C, M, in single amino acid repeats with cutoff 5 is less than that of other amino acids. Also, within these amino acids, the relative abundance is quite large within the scope of all protein sequences except for M, W and Y whose proportions are only 2.2%, 1.2% and 2.7%. Within this group of amino acids: Isoleucine, phenylalanine, valine, cysteine and methionine have hydropathy indexes greater than 0, which means that the amino acid located in that region of the protein is hydrophobic. Although the hydropathy indexes of N, W, Y are lower than zero which suggests these amino acids are hydrophilic, their abundances both in X-linked and autosome-linked proteins are low. Most of the amino acids in single amino acid repeats at least 5 amino acids long are hydrophilic (for X-linked SAARs, the proportion is 81% of hydrophilic amino acids and on autosomes, the proportion is 77%).

For single amino acid repeats of length 12 or more, the only hydrophobic amino acid that occurs is alanine. In the whole protein dataset, the proportion of alanine is 6.9% in autosome-linked proteins and 6.6% in X-linked proteins. For autosome-linked single amino acid repeats with length at least 5, 7, 12 amino acids, the proportion of alanine is 13.4%, 14.7% and 12.6%. For X-linked SAARs, the proportion of alanine is 15.7%, 16.6% and 14.1% respectively. Although alanine is hydrophobic and this seems to contradict previous research that tandem repeats of hydrophobic amino acids are not favored in proteins [Green and Wang 1994], it is still understandable based several perspectives. First of all, from Table 2.4, we can see that the hydropathy index for alanine is 1.8, the lowest among all hydrophobic

amino acids. Based on the results, we can basically answer two questions that could be interesting in terms of single amino acid repeats composition. First, since there are studies showing that hydrophobic amino acids are relatively rare compared to hydrophilic amino acids, whether specific amino acids have different proportions between SAARs versus all protein sequences should be answered. Second, since the evolution of autosomes and sex chromosomes are different, we want to know whether this could influence the amino acid composition of SAARs (leading to different proportions of amino acid can be different for X-linked SAARs and autosome-linked SAARs).

These tests were significant, even with an p-value cutoff as 1×10^{-4}, the proportion of most amino acids in SAARs are still significantly different from those of all protein sequences. The few exceptions are P for sex-chromosome-linked SAARs with length cutoff of 5 (p-value is 0.0002), T, S and H for sex-chromosome-linked SAARS with length cutoff of 7, (p-values 7.267×10^{-7}, 7.859×10^{-6} and 7.51×10^{-8} respectively) and G for sex-chromosome-linked with length cutoff of 12(p-value is 0.7242). This is strong evidence that the amino acid composition of SAARs are different from that of whole proteomes. This could support evidence from previous research that the single amino acid repeats favor hydrophilic amino acids especially E, P, Q and S since their proportions are tremendously increased for SAARs if the length cutoff is 5. In addition, two hydrophobic amino acids are increased in accord with previous research [Huntley and Clark, 2007].

To understand the abundance of amino acid composition from another perspective, we also did a repeat-based analysis. For all amino acid repeats, we counted all SAARs based on the amino acid type with minimum length both 5 and 7 amino acids. For each amino acid type, we seperately counted the number of single amino

acid repeats located on X chromosomes and autosomes.

The resulting ratios are shown in Figure 2.15 and Figure 2.16. In general, both length cutoffs are consistent. Common types of amino acid such as alanine, glutamic acid, glutamine, proline and serine also show a dominant role in repeats. The most extreme instance is tryptophan, only 8 single amino acid repeats with minimum length of 5 are recorded located on autosomes and none on sex-chromomosomes from any of 34 species recorded. Moreover, if we make the cutoff more stringent at 7 amino acids long, we can not find even one SAAR composed of tryptophan. We can compare these two amino acid proportions to their expectation. The proportion of alanines on autosomes is 6.88%, and the proportion of tryptophan is 1.17%. If the numbers of repeats composed by these two amino acids are correlated to their abundance, the ratio of their number of repeats should be 5.69 (6.88% / 1.21%). From my analysis, the total number of alanine SAARs is 22360. We should therefore expect about 3933 tryptophans SAARs. However, only 4 are observed. On sex-linked chromosomes, the ratio of these two types of amino acid is 5.93 (6.62% / 1.117%) and we found 6020 sex-chromosome linked alanine SAARs. We should expect about 1015 sex-chromosome linked tryptophan SAARs. But instead, we can not identify even one copy. These comparisons demonstrate that there must be a bias in favor of specific kinds of amino acid in SAARs independent of whether they are autosome linked or sex-chromosome linked.

The opposite extreme case is Q-repeats. From the graphs, glutamine SAARs are most frequent both for sex-chromosome linked SAARs and autosome-linked (17.27% on sex chromosomes and 27.78% on autosomes for the length cutoff of 5 amino acids long). Similarly serine SAARs are also of high frequency (11.94% on sex chromosomes and 11.86% on autosomes if the length cutoff is 5 amino acids long).

This evidence shows that the choice of amino acid in SAARs is not random.

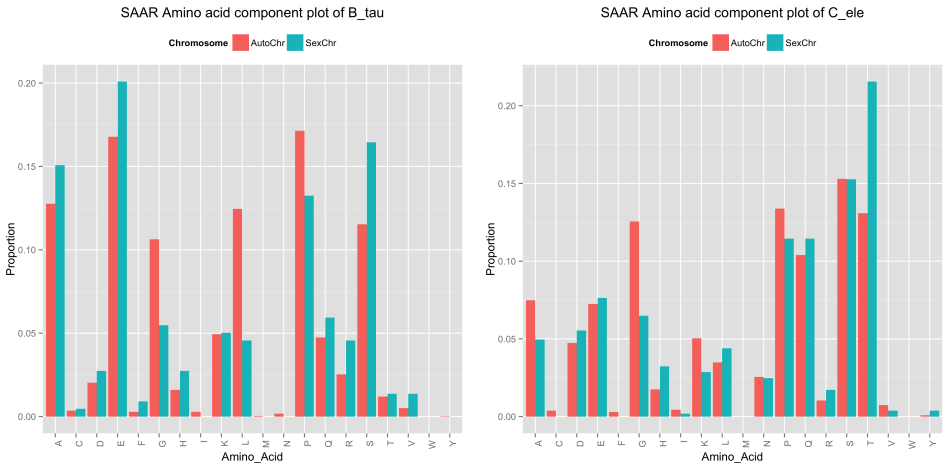
Different species may have quite different CG-content. CG-content could influence amino acid composition of SAARs. We examine the abundance of amino acid composition of each species and the result are presented in Figure 2.9 to 2.14. We observe that species with different CG-values can have slightly different amino acid composition in SAARs. For example, in *Homo sapiens*, poly-E becomes one of the most dominant SAAR type. Also, in *Caenorhabditis elegans*, poly-T becomes dominant. But there is no consistent pattern illustrating a relationship between CG content and amino acid composition of SAARs.

Table 2.4: Component of amino acids that construct to single amino acid across all species, the amino acid counts marked with (-) is not significant if we set the p-value cutoff of 1×10^{-3} when comparing the proportion of that kind of single amino acid between all protein sequences to that of SAARs. Amino acid counts under specific SAAR length cutoff with (–) shows the proportion of that kind of amino acid is not significant between that of sex chromosome and autosomes with p-value above 1×10^{-5} .

AminoAcid	HI	AutoChr-total	PropAuto	SexChr-total	PropSexChr	AutoChr-5	SexChr-5	AutoChr-7	SexChr-7	AutoChr-12	SexChr-12
A	1.8	18390757	0.0688	805333	0.0662	149488	42291	74014	23880	16153	6221
R	-4.5	14979522	0.056	659578	0.0543	16742	2520	2534	485	572	0
D	-3.5	12886629	0.0482	601947	0.0495	28729	8150	9817	3368	1315	573
N	-3.5	9896045	0.037	498854	0.041	31194	13000	14990	6985	2333	1299
C	2.5	6014999	0.0225	253516	0.0209	2130	206	466	55	25(–)	0(–)
E	-3.5	18900057	0.0707	866562	0.0713	118803	11665	53499	5036	12976	1450
Q	-3.5	12634648	0.0473	563875	0.0464	227528	87138	142443	58817	55857	25726
G	-0.4	17189880	0.0643	757072	0.0623	105841	28353	46595	14751	9341	2766(-)
H	-3.2	6921610	0.0259	308978	0.0254	22162	6166	12043(–)	3329(–)(-)	1806	487
I	4.5	12072558	0.0452	587191	0.0483	1288	145	135(–)	32(–)	43(–)	0(–)
L	3.8	26438451	0.0989	1128012	0.0928	68129	4345	16535	882	686	62
K	-3.9	15638748	0.0585	748313	0.0616	33926	2556	8851	659	679	156
M	1.9	5795078	0.0217	285442	0.0235	309(–)	103(–)	155	0	15(–)	0(–)
F	2.8	9972322	0.0373	463204	0.0381	3957	184	1778	19	490	12
P	-1.6	16368253	0.0612	735335	0.0605	125270	15829(-)	48798	6125	8444	968
S	-0.8	22227210	0.0831	1016938	0.0837	129978	29260	50293	11548(-)	13158	2999
T	-0.7	14362051	0.0537	663753	0.0546	47291	16685	19901	7416(-)	4119	1162
W	-0.9	3231792	0.0121	141906	0.0117	40(–)	0(–)	0	0	0	0
Y	-1.3	7225843	0.027	344240	0.0283	319	317	38	281(*)	0	237(*)
V	4.2	16168714	0.0605	726394	0.0598	3002	544	266	21	71(–)	0(–)
Total	*	267315167	1	12156443	1	1116126	269457	503151	143689	128083	44118

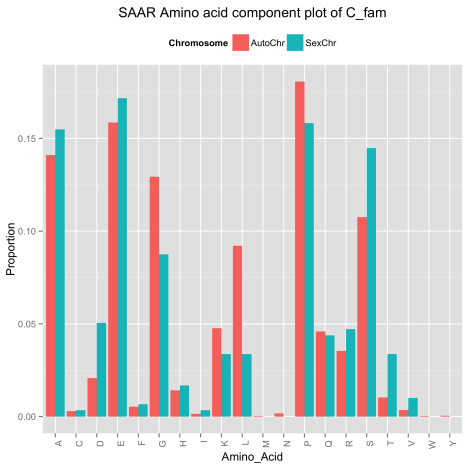
Table 2.5: Amino acid count proportion comparison of single amino acid repeats with length cutoff 5, 7, 12 amino acids long located both on X chromosomes and autosomes in all eukaryotic species.

Amino Acid	AutoProp	SexChrProp	AutoChr-5	SexChr-5	AutoChr-7	SexChr-7	AutoChr-12	SexChr-12
A	0.0688	0.0662	0.13393	0.15695	0.1471	0.16619	0.12611	0.14101
R	0.056	0.0543	0.015	0.00935	0.00504	0.00338	0.00447	0.0
D	0.0482	0.0495	0.02574	0.03025	0.01951	0.02344	0.01027	0.01299
N	0.037	0.041	0.02795	0.04825	0.02979	0.04861	0.01821	0.02944
C	0.0225	0.0209	0.00191	0.00076	0.00093	0.00038	0.0002	0.0
E	0.0707	0.0713	0.10644	0.04329	0.10633	0.03505	0.10131	0.03287
Q	0.0473	0.0464	0.20386	0.32338	0.2831	0.40934	0.4361	0.58312
G	0.0643	0.0623	0.09483	0.10522	0.09261	0.10266	0.07293	0.0627
H	0.0259	0.0254	0.01986	0.02288	0.02394	0.02317	0.0141	0.01104
I	0.0452	0.0483	0.00115	0.00054	0.00027	0.00022	0.00034	0.0
L	0.0989	0.0928	0.06104	0.01613	0.03286	0.00614	0.00536	0.00141
K	0.0585	0.0616	0.0304	0.00949	0.01759	0.00459	0.0053	0.00354
M	0.0217	0.0235	0.00028	0.00038	0.00031	0.0	0.00012	0.0
F	0.0373	0.0381	0.00355	0.00068	0.00353	0.00013	0.00383	0.00027
P	0.0612	0.0605	0.11224	0.05874	0.09698	0.04263	0.06593	0.02194
S	0.0831	0.0837	0.11645	0.10859	0.09996	0.08037	0.10273	0.06798
T	0.0537	0.0546	0.04237	0.06192	0.03955	0.05161	0.03216	0.02634
W	0.0121	0.0117	4e-05	0.0	0.0	0.0	0.0	0.0
Y	0.027	0.0283	0.00029	0.00118	8e-05	0.00196	0.0	0.00537
V	0.0605	0.0598	0.00269	0.00202	0.00053	0.00015	0.00055	0.0

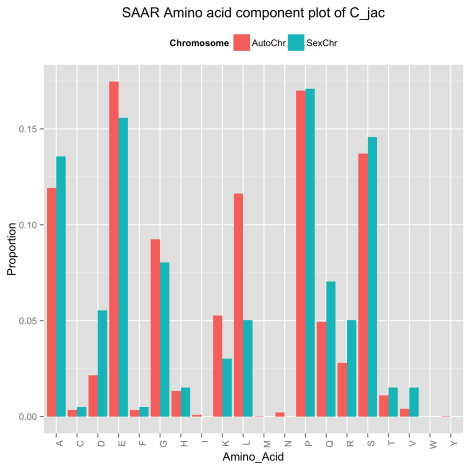


(a) *Bos taurus*

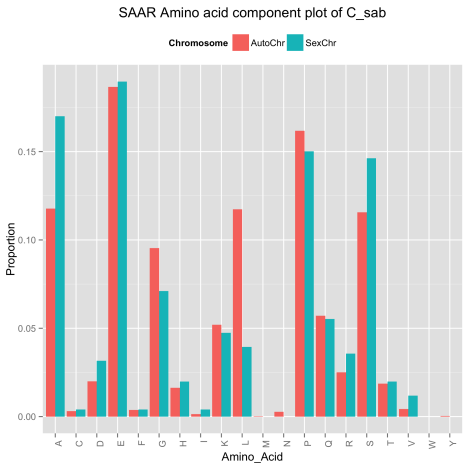
(b) *Caenorhabditil elegans*



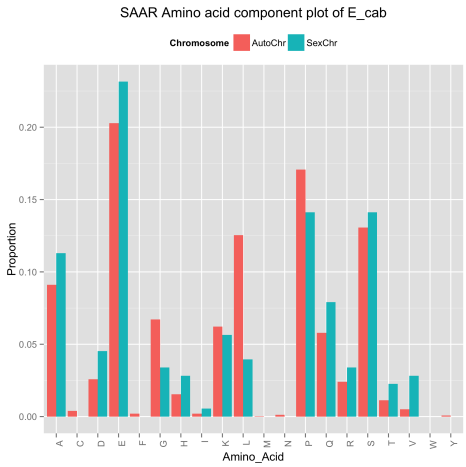
(c) *Canis familiaris*



(d) *Callithrix jacchus*

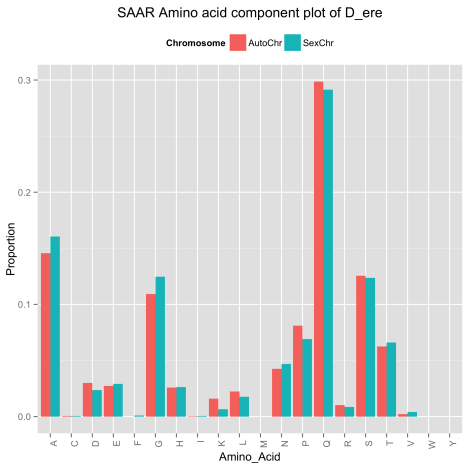


(e) *Chlorocebus sabaeus*

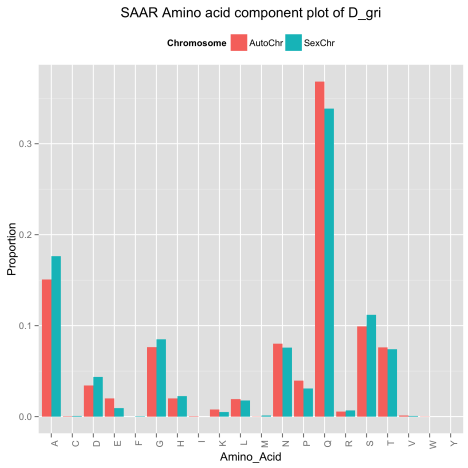


(f) *Equus caballus*

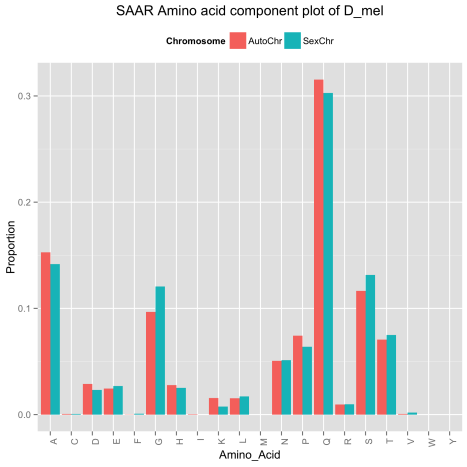
Figure 2.9: Proportions of SAARs amino acid compositions in Btau, Cele, Cfam, Cjac, Csab, Ecab.



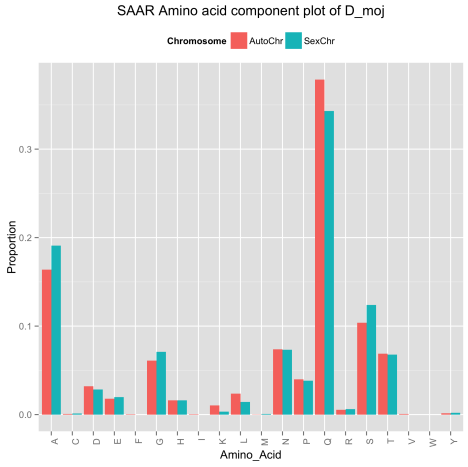
(a) *Drosophila erecta*



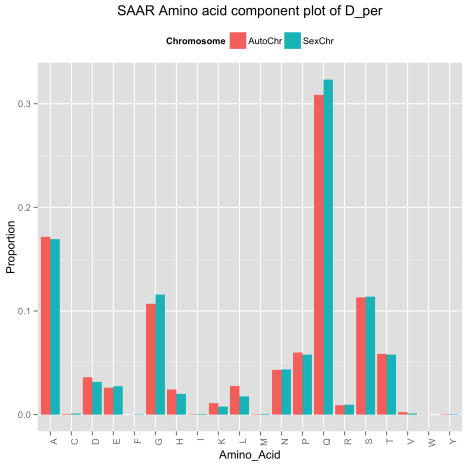
(b) *Drosophila grimshawi*



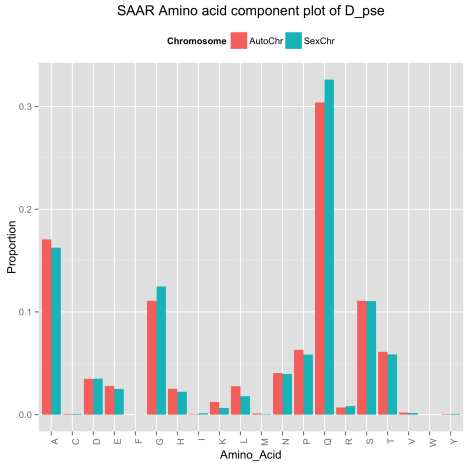
(c) *Drosophila melanogaster*



(d) *Drosophila mojavensis*

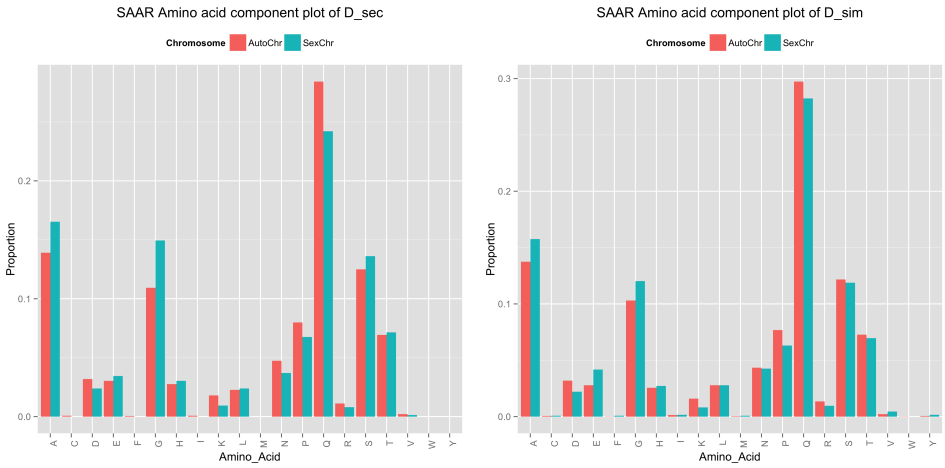


(e) *Drosophila persimilis*



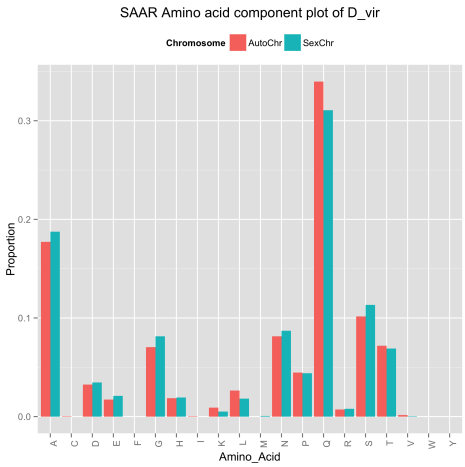
(f) *Drosophila pseudoobscura*

Figure 2.10: Proportions of SAARs amino acid compositions in Dere, Dgri, Dmel, Dmoj, Dper, Dpse.

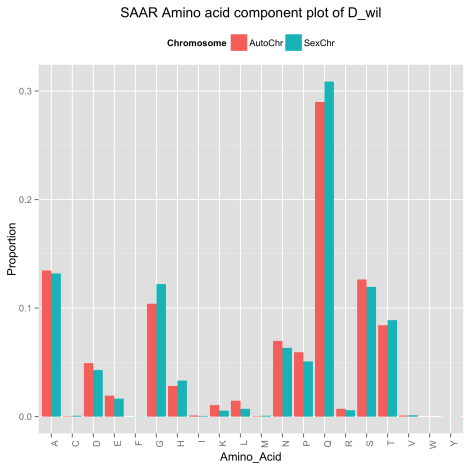


(a) *Drosophila sechellia*

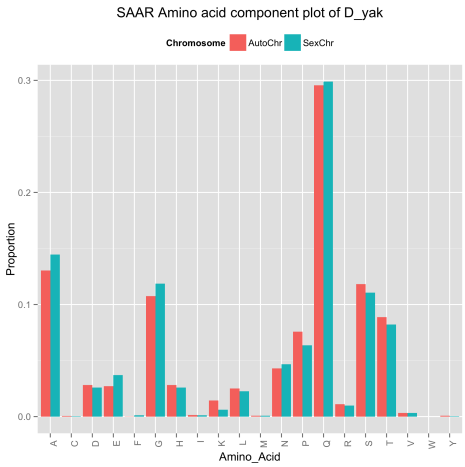
(b) *Drosophila simulans*



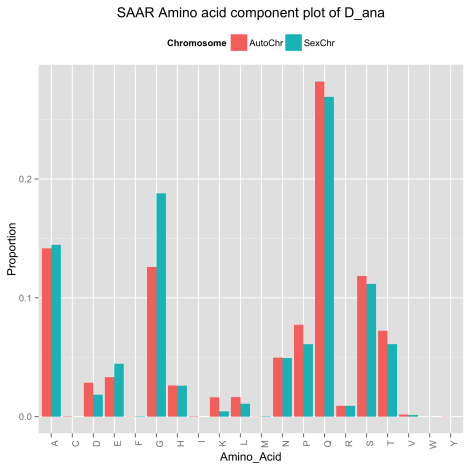
(c) *Drosophila virilis*



(d) *Drosophila willistoni*

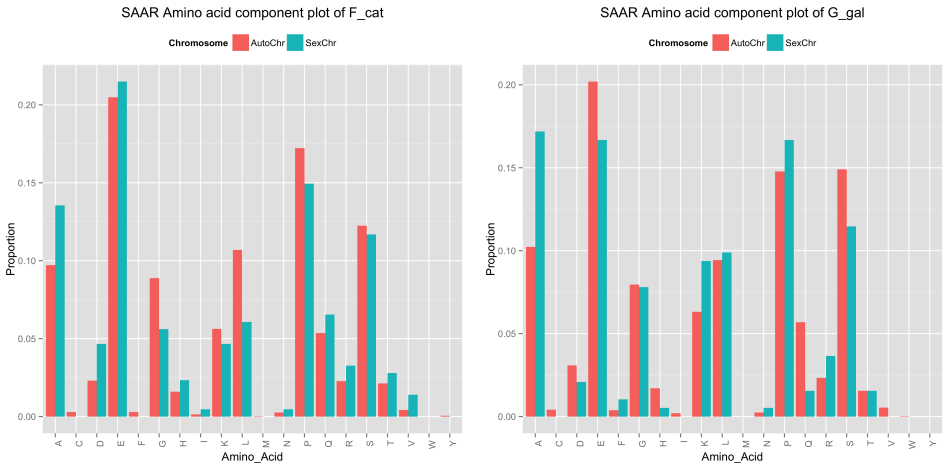


(e) *Drosophila yakuba*



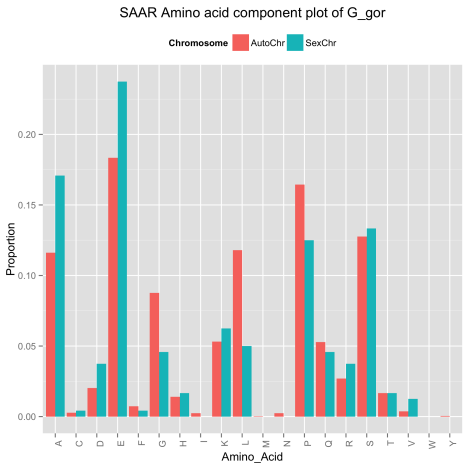
(f) *Drosophila ananassae*

Figure 2.11: Proportions of SAARs amino acid compositions in Dsec, Dsim, Dvir, Dwil, Dyak, Dana.

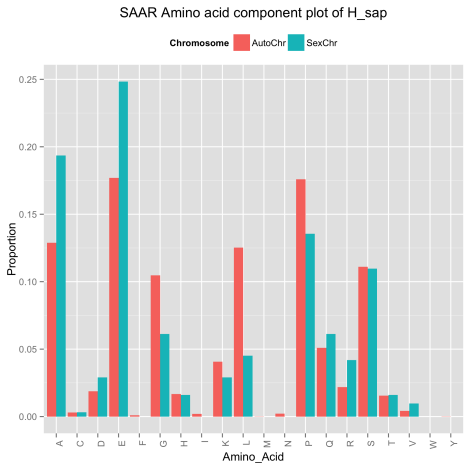


(a) *Felis catus*

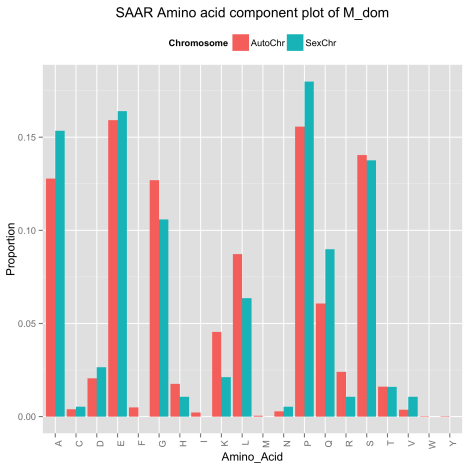
(b) *Gallus gallus*



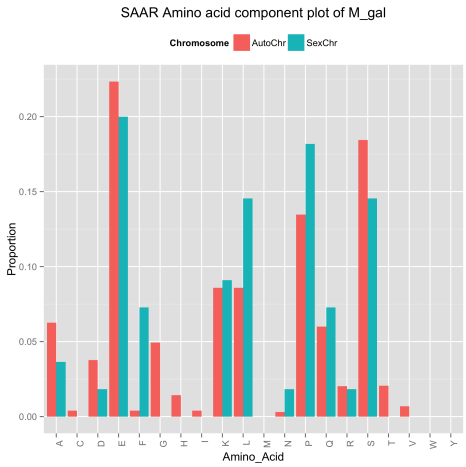
(c) *Gorilla gorilla*



(d) *Homo sapiens*

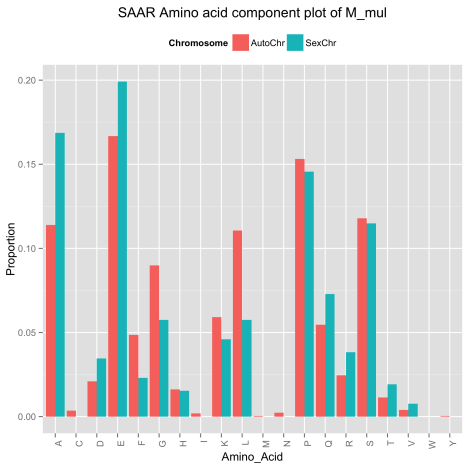


(e) *Monodelphis domestica*

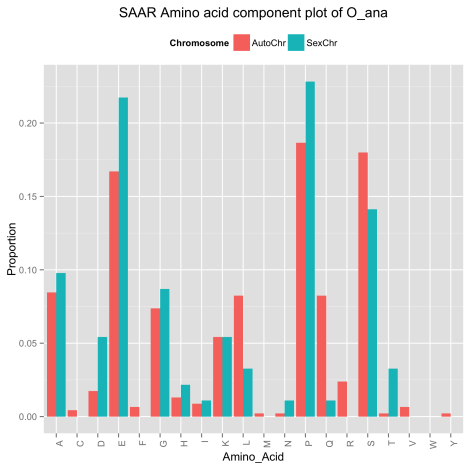


(f) *Meleagris gallopavo*

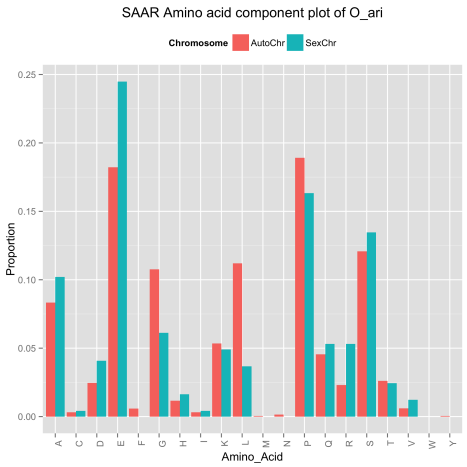
Figure 2.12: Proportions of SAARs amino acid compositions in Fcat, Ggal, Ggor, Hsap, Mdom, Mgal.



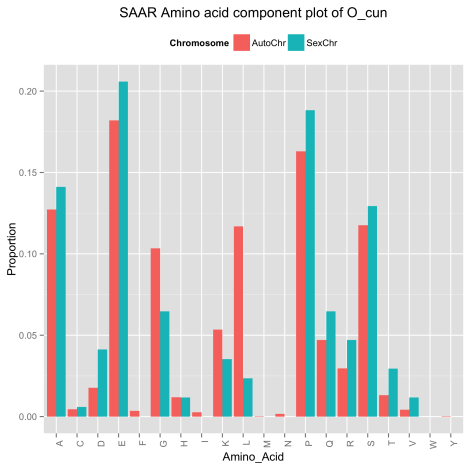
(a) *Macaca mulatta*



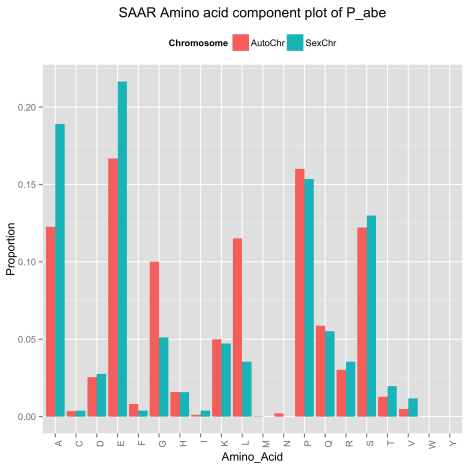
(b) *Ornithorhynchus anatinus*



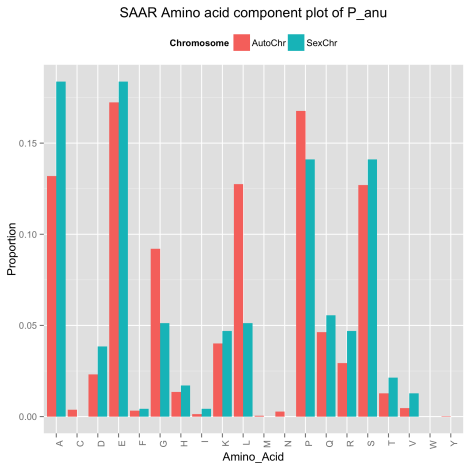
(c) *Ovis aries*



(d) *Oryctolagus cuniculus*



(e) *Pongo abelii*

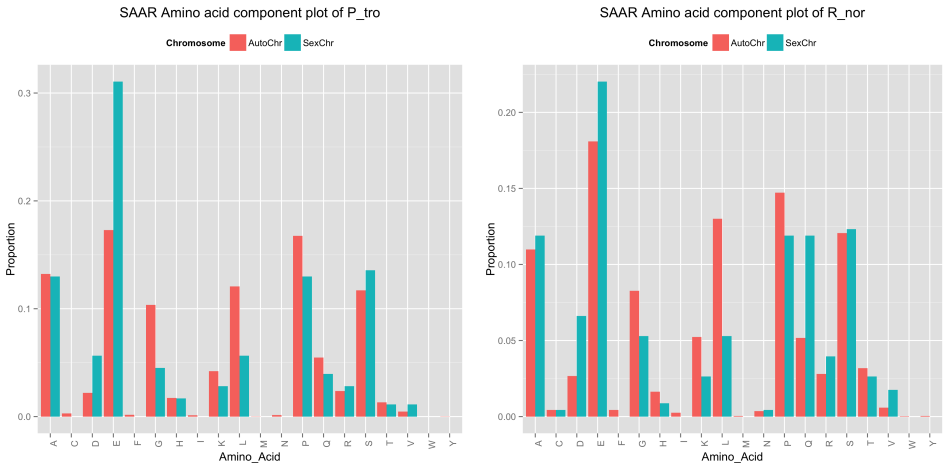


(f) *Papio anubis*

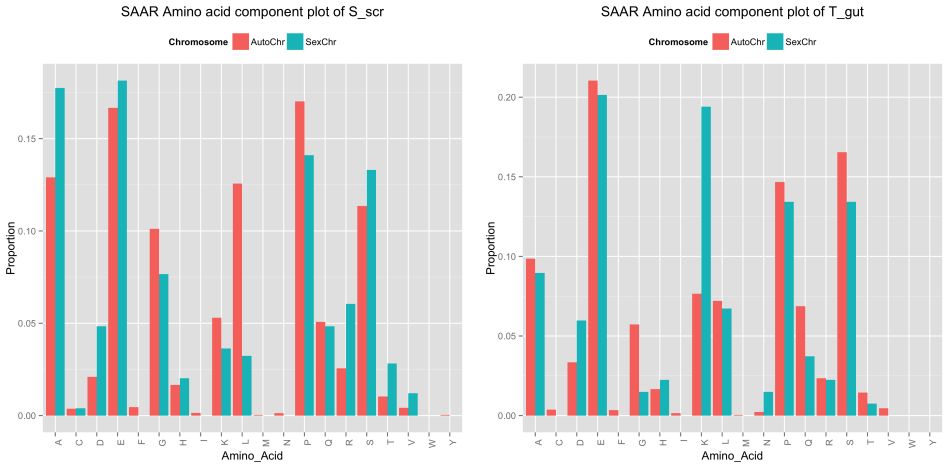
Figure 2.13: Proportions of SAARs amino acid compositions in Mmul, Oana, Oari, Ocun, Pabe, Panu.

Table 2.6: Proportions of SAARs located on X/Z chromosomes and autosomes composed by different amino acid with both length cutoff 5 and 7 amino acids long in all eukaryotic species.

Amino acids	Prop-sexchr-5	Prop-autosome-5	Prop-sexchr-7	Prop-autosome-5
A	0.13042	0.15578	0.14718	0.17097
R	0.01758	0.01159	0.00502	0.00416
D	0.02773	0.03335	0.02107	0.02573
N	0.02771	0.04927	0.03072	0.05219
C	0.00215	0.00085	0.00099	0.0004
E	0.10589	0.04637	0.10408	0.03576
Q	0.17265	0.27773	0.24768	0.36265
G	0.09541	0.10855	0.09416	0.10953
H	0.0191	0.02368	0.02487	0.02567
I	0.0014	0.00067	0.00027	0.00026
L	0.06956	0.02008	0.03746	0.00739
K	0.03409	0.0111	0.01972	0.00475
M	0.00029	0.00049	0.00036	0.0
F	0.00349	0.00088	0.00331	0.00013
P	0.11663	0.06531	0.10059	0.04764
S	0.11935	0.11862	0.09624	0.08222
T	0.04304	0.06671	0.0395	0.05556
W	5e-05	0.0	0.0	0.0
Y	0.00034	0.00039	9×10^{-5}	0.00053
V	0.00329	0.00267	0.00054	0.0002



(a) *Pan troglodytes* (b) *Rattus norvegicus*



(c) *Sus scrofa* (d) *Taeniopygia guttata*

Figure 2.14: Proportions of SAARs amino acid compositions in Ptro, Rnor, Tgut, Sscr.

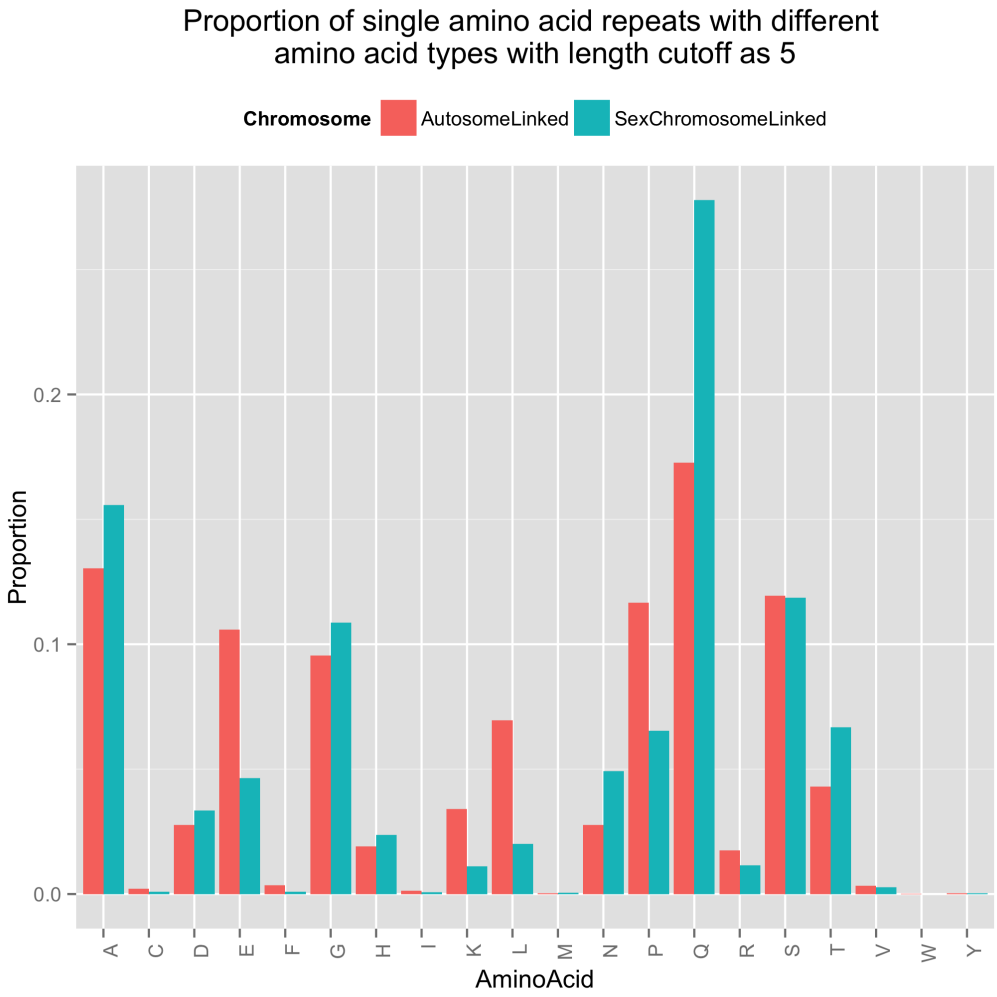


Figure 2.15: Comparison between proportions of SAARs located on X/Z chromosomes and autosomes composed by different amino acid with length cutoff 5 amino acids long in all eukaryotic species.

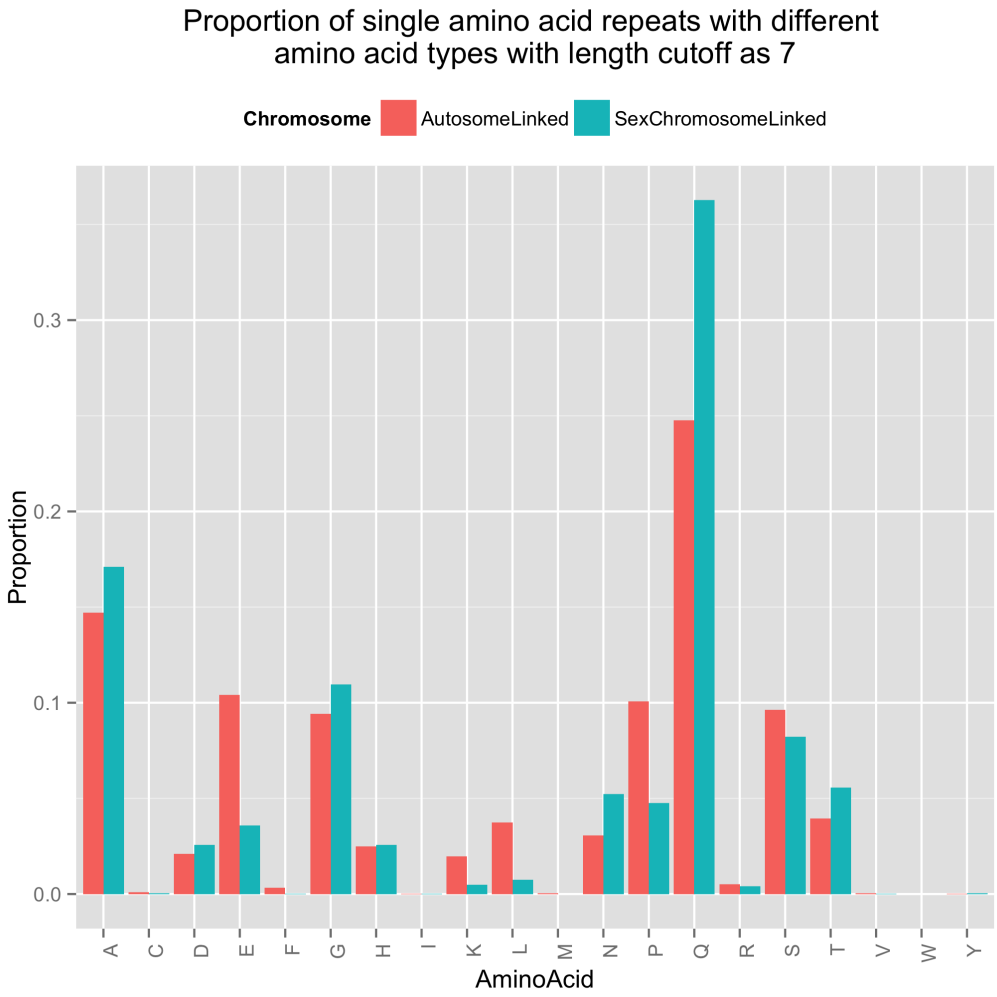


Figure 2.16: Comparison between proportions of SAARs located on X/Z chromosomes and autosomes composed by different amino acid with length cutoff 7 amino acids long in all eukaryotic species.

2.3.5 Comparison of dominant codons within SAARs in *Drosophila melanogaster*

The results of Table 2.7 and Table 2.8 are consistent with former studies that show which amino acids normally form single amino acid repeats. There are two exceptions, arginine and lysine, in *Drsophila melanogaster*, we find less than 50 instances of these two types of amino acids that contribute to X-linked SAARs at least 5 amino acids long. However, in the previous section, we find in total 2520 arginine and 2556 lysine SAARs for X-linked regions in all 34 species we analyzed. This means that we expect at least around 80 instances for each amino acid to be found in *Drosophila melanogaster* but instead we only observe less than 50 for each kind. This is despite the fact that *Drosophila* species have the most abundant SAARs on X chromosomes.

For each amino acid, we have calculated the RSCU value in 4 different categories of coding regions. These categories are X-repetitive, X-non-repetitive, autosome-repetitive and autosome-non-repetitive regions. The results are shown in Table 2.9. By comparing the dominant codon between repetitive regions and non-repetitive regions, we observe that A, D, H, P and S have different dominant codons compared to non-repetitive regions. Alanine has a positive hydropathy index and has the same dominant codon between X-linked repetitive regions and autosome-linked repetitive regions but the codon differs from non-repetitive regions. In contrast, D, H, P and S all have a negative hydropathy index and their dominant codons on X-linked repetitive regions are different from those of autosome-linked repetitive regions and non-repetitive regions. Table 2.9 shows the RSCU value for these amino acids. We can see that the RSCU values change drastically between codons prefered in repetitive regions and in non-repetitive regions. D and H rarely contribute to the constuction

of SAARs so the differences could be due to small sample random effects. The differences between dominant codons of S, A and P in repetitive regions and non-repetitive regions are more robust as they are abundant amino acids in SAARs. Based on these data, we can exclude the possibility that expression level and function might have an effect on our results because non-repetitive sequences and repetitive sequences are from the same set of genes. So patterns between them should be a net-effect of mutation rates and selection. Previous research has found that the mutation rates of tandem repeats are higher than other sequences [Gemayel et al., 2010]. Usually, the use of non-optimal codons is considered to be weakly selected against in protein coding regions since using non-optimal codons could harm the efficiency and/or accuracy of protein expression. In other theories, these codons are poorly recognized by their corresponding t-RNA which might be infrequent in cells [Rocha, 2004]. The accumulation of non-optimal codons in SAAR regions could be a sign demonstrating that selection within SAARs is relaxed compared to that of non-repetitive regions.

Table 2.7: RSCU values for X-linked repetitive/nonrepetitive regions and autosome-linked repetitive/nonrepetitive regions(elements marked as red are dominant codons that code for same amino acid but different between single amino acid repeats and non-repetitive regions, elements marked with "-" are codons with count less than 50 in total our dataset), cells with "*" are amino acids in SAARs that have significant different composition of codons compared to that of corresponding non-repetitive regions.

	x_repetitive	x_nonrepetitive	auto_repetitive	auto_nonrepetitive
A	**GCA:1.376***	GCC:1.83*	**GCA:1.446***	GCC:1.726*
C	UGC:0.0 (-)	UGC:1.516	UGC:1.5 (-)	UGC:1.436
E	GAG:1.608*	GAG:1.456*	GAG:1.511*	GAG:1.36*
D	**GAC:1.071**	GAU:1.057	GAU:1.02	GAU:1.055
G	GGC:1.597*	GGC:1.892*	GGC:1.456*	GGC:1.685*
F	UUC:0.0 (-)	UUC:1.316	UUC:0.0 (-)	UUC:1.257
I	AUC:0.0 (-)	AUC:1.596	AUC:1.6 (-)	AUC:1.422
H	CAU:1.156*	CAC:1.174*	CAC:1.048*	CAC:1.206*
K	AAG:1.556 (-)	AAG:1.511	AAG:1.441	AAG:1.427
*	UAG:0.0 (-)	UAG:1.264	UAG:0.0 (-)	UAA:1.119
M	AUG:0.0 (-)	AUG:1.0	AUG:0.0 (-)	AUG:1.0
L	CUG:2.027	CUG:2.416	CUG:2.097*	CUG:2.274*
N	AAC:1.361*	AAC:1.046*	AAC:1.316*	AAC:1.122*
Q	CAG:1.433	CAG:1.485	CAG:1.416	CAG:1.422
P	CCG:1.831*	CCG:1.415*	**CCA:1.54***	CCC:1.283*
S	**AGC:1.541**	UCG:1.555	UCC:1.787*	UCC:1.535*
R	CGC:1.474 (-)	CGC:1.765	CGC:1.344*	CGC:1.639*
T	ACC:1.827*	ACC:1.491*	ACC:1.634*	ACC:1.454*
W	UGG:0.0 (-)	UGG:1.0	UGG:0.0 (-)	UGG:1.0
V	GUG:2.133 (-)	GUG:2.002	**GUC:2.133** (-)	GUG:1.942
Y	UAC:0.0 (-)	UAC:1.247	UAC:0.0 (-)	UAC:1.271

Table 2.8: RSCU values for X-linked repetitive/nonrepetitive regions and autosome-linked repetitive/nonrepetitive regions(elements marked as red are dominant codons that code for same amino acid but different between single amino acid repeats and non-repetitive regions, elements marked with "-" are codons with count less than 50 in total our dataset), cells with "*" are amino acids for X-linked SAARs that have significant different composition of codons compared to that of corresponding autosome-linked SAAR regions.

	x_repetitive	x_nonrepetitive	auto_repetitive	auto_nonrepetitive
A	**GCA:1.376***	GCC:1.83	**GCA:1.446***	GCC:1.726
C	UGC:0.0 (-)	UGC:1.516	UGC:1.5 (-)	UGC:1.436
E	GAG:1.608	GAG:1.456	GAG:1.511	GAG:1.36
D	**GAC:1.071**	GAU:1.057	GAU:1.02	GAU:1.055
G	GGC:1.597	GGC:1.892	GGC:1.456	GGC:1.685
F	UUC:0.0 (-)	UUC:1.316	UUC:0.0 (-)	UUC:1.257
I	AUC:0.0 (-)	AUC:1.596	AUC:1.6 (-)	AUC:1.422
H	**CAU:1.156***	CAC:1.174	CAC:1.048*	CAC:1.206
K	AAG:1.556 (-)	AAG:1.511	AAG:1.441	AAG:1.427
*	UAG:0.0 (-)	UAG:1.264	UAG:0.0 (-)	UAA:1.119
M	AUG:0.0 (-)	AUG:1.0	AUG:0.0 (-)	AUG:1.0
L	CUG:2.027	CUG:2.416	CUG:2.097	CUG:2.274
N	AAC:1.361	AAC:1.046	AAC:1.316	AAC:1.122
Q	CAG:1.433	CAG:1.485	CAG:1.416	CAG:1.422
P	CCG:1.831*	CCG:1.415	**CCA:1.54***	CCC:1.283
S	**AGC:1.541***	UCG:1.555	UCC:1.787*	UCC:1.535
R	CGC:1.474 (-)	CGC:1.765	CGC:1.344	CGC:1.639
T	ACC:1.827*	ACC:1.491	ACC:1.634*	ACC:1.454
W	UGG:0.0 (-)	UGG:1.0	UGG:0.0 (-)	UGG:1.0
V	GUG:2.133 (-)	GUG:2.002	**GUC:2.133** (-)	GUG:1.942
Y	UAC:0.0 (-)	UAC:1.247	UAC:0.0 (-)	UAC:1.271

Aminoacid	Codons	X-rep	X-nonrep	Auto-rep	Auto-nonrep
A	GCA	1.376	0.676	1.446	0.729
A	GCC	0.938	1.83	0.784	1.726
A	GCU	0.852	0.611	0.842	0.747
A	GCG	0.833	0.883	0.927	0.798
H	CAC	0.844	1.174	1.048	1.206
H	CAU	1.156	0.826	0.952	0.794
S	UCU	0.374	0.391	0.441	0.495
S	AGC	1.541	1.349	1.543	1.283
S	UCG	1.241	1.555	1.118	1.36
S	UCC	1.468	1.521	1.787	1.535
S	UCA	0.916	0.533	0.654	0.61
S	AGU	0.459	0.651	0.457	0.717
D	GAU	0.929	1.057	1.02	1.055
D	GAC	1.071	0.943	0.98	0.945
P	CCU	0.437	0.366	0.71	0.497
P	CCG	1.831	1.415	1.428	1.214
P	CCA	1.451	0.974	1.54	1.006
P	CCC	0.282	1.246	0.322	1.283

Table 2.9: RSCU values comparison between different codons of amino acids that have different dominant codon for repetitive and non-repetitive regions, elements marked red are the RSCU value for optimal codons in differnet kinds of coding regions.

2.3.6 Codon usage bias within SAARs of *Drosophila melanogaster*

To quantify the ongoing selection pressures on codon usage bias in *Drosophila melanogaster*, we compared codon changes between *Drosophila melanogaster* and inferred ancestral sequence using *Drosophila melanogaster*, *Drosophila simulans* and *Drosophila yakuba*. For each pair of conserved single amino acid repeats we counted three types of changes: preferred to unpreferred change, unpreferred to preferred change and no change (either preferred to preferred change or unpreferred to unpreferred change). Whether the codon is preferred or unpreferred is determined by the RSCU value we calculated in previous section. If the RSCU value of one codon is larger than one, we would consider it as optimal codon and in contrast, codons with RSCU values smaller than one would be considered as unpreferred codons. To compare with, we also counted the numbers of different types of codon changes for non-repetitive regions. The comparison of codon changes between non-repetitive regions and single amino acid repeats coding regions are summarized in Table 2.10.

If we assume that the codon usage changes are at equilibrium, we should expect there are equal numbers of preferred codons change to non-preferred codons and non-preferred codons change to preferred codons. From Table 2.10, we can see that from ancestral sequences to sequences of *Drosophila melanogaster*, there are more preferred codons changed to unpreferred codons in non-repetitive regions. Note that our analysis relies on contrasting selected versus neutral changes assuming that changes within a codon class are neutral (i.e., preferred codon change to preferred codon and unpreferred codon change to unpreferred codon). We observe that the number of preferred codons change to unpreferred codons versus unpreferred codons to preferred codons in general is 2.55 times (84 versus 33) for X-linked SAARs, 3.92 times (11578 versus

2954) for X-linked non-repetitive regions, 1.83 times (297 versus 162) for autosome-linked SAARs and 2.98 times (66109 versus 22204) for autosome-linked non-repetitive regions. If we assume codon usage bias is at equilibrium, we find both data significantly deviate from an equal expectation (chi-square test, p-values for all four regions are less than $\alpha = 0.05$). The excess of fixations from preferred codons to unpreferred codons clearly reveals that the ongoing codon change is not at equillibrium and also, the codon changing direction is towards unpreferred codons genomewide. Begun and Aquadro [1993] had shown that populations collected from Zimbabwe had considerably greater genetic variability than populations from the USA. This suggested a population bottleneck as this species left its ancestral home [Wu et al., 1995]. The decreasing of effective population of *Drosophila melanogaster*, which can cause accumulation of recessive deleterious mutations, could be why there is an excess of codon change from preferred codons to unpreferred codons. From the result, we can also observe that for non-repetitive regions, there are less unpreferred codon changes to preferred codons compared to that of SAARs.

```
>Dmel G,5,62,62,67@G,5,62,62,67
ggaggaggcggtggt
>Dsim G,5,62,62,67@G,5,62,62,67
ggaggaggcggcggc
>Divergence
---------P2UP2U
>Polymorphism
GGA/GGA/GGC/GGT/GGT
```

Figure 2.17: Example of fixation of codon change from preferred codon to non-preferred codon, the protein ID for this protein segment is FBpp0077075 in FlyBase on 2L chromosome of *Drosophila melanogaster*.

```
>Dmel Q,5,212,212,217@Q,5,212,212,217
cagcagcagcagcag
>Dsim Q,5,212,212,217@Q,5,212,212,217
caacagcagcagcag
>Divergence
U2P----------
>Polymorphism
CAG/CAG/CAG/CAG/CAG
```

Figure 2.18: Example of fixation of codon change from unpreferred codon to preferred codon, the protein ID for this protein segment is FBpp0079183 in FlyBase on 2L chromosome of *Drosophila melanogaster*.

```
>Dmel G,5,62,62,67@G,5,62,62,67
ggaggaggcggtggt
>Dsim G,5,62,62,67@G,5,62,62,67
ggaggaggcggcggc
>Divergence
---------P2UP2U
>Polymorphism
GGA/GGA/GGC/GGT/GGT
```

Figure 2.19: Example of polymorphism of codon change from preferred codon to non-preferred codon, the protein ID for this protein segment is FBpp0073613 in FlyBase on X chromosome of *Drosophila melanogaster*.

```
>Dmel Q,5,212,212,217@Q,5,212,212,217
cagcagcagcagcag
>Dsim Q,5,212,212,217@Q,5,212,212,217
caacagcagcagcag
>Divergence
U2P------------
>Polymorphism
CAG/CAG/CAG/CAG/CAG
```

Figure 2.20: Example of polymorphism of codon change from preferred codon to non-preferred codon, the protein ID for this protein segment is FBpp0070105 in FlyBase on X chromosome of *Drosophila melanogaster*.

Table 2.10: Counts of different kinds of codon changes(No change, preferred to unpreferred and unpreferred to preferred) with state of polymorphic and fixation within single amino acid repeats coding regions of *Drosophila melanogaster*. The states of ancestral codons are inferred using *Drosophila simulans, Drosophila melanogaster* and *Drosophila yakuba*.

	No change	P to U	U to P	total
X				
SAAR	2180	84	33	2297
non-repetitive	458612	11578	2954	473144
Autosome				
SAAR	12129	297	162	12588
non-reptititve	2865559	66109	22204	2953872

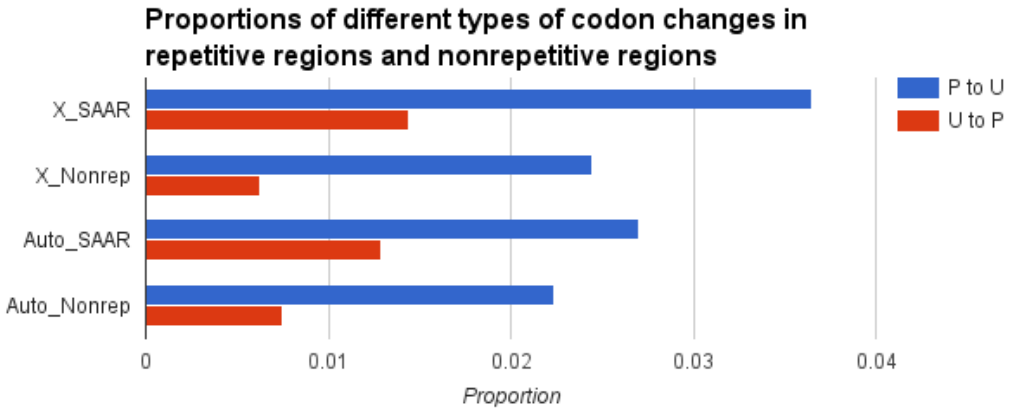


Figure 2.21: Proportions comparisons of different types codon changes within single amino acid repeats and non-repetitive regions, both on X chromosome and autosomes.

2.3.7 The conservation of single amino acid repeats within *Drosophila* species

We seperated the *Drosophila* species into 5 pairs (*D.melanogaster-D.simulans, D.pseudoobscura-D.persimilis, D.yakuba-D.erecta, D.virilis-D.grimshawi,*

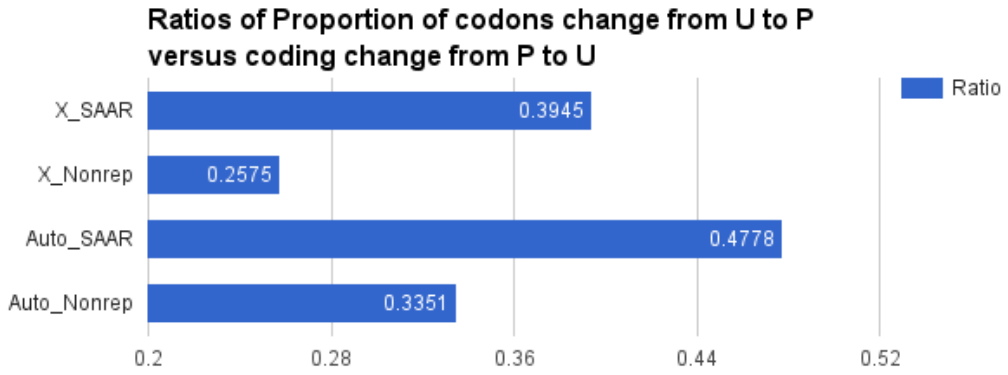


Figure 2.22: Ratios of proportion of unpreferred codons change to preferred codons versus that of preferred codons change to unpreferred codons for X-linked and autosome-linked, non-repetitive and SAAR regions.

D.ananassae-D.willistoni). These groupings were chosen to have phylogenetically independent comparisons of different evolutionary ages. All proteomes of these five groups of *Drosophila* species are scanned for homologous single amino acid repeats. Since in the phylogeny of *Drosophila*, all 5 groups have no overlaping branch, we can assume that differences in the evolution of SAARs within each group are independent from other groups. All divergence times are based on studies by Tamura [2004] using substitutions to determine the divergence time among *Drosophila* species. We then scanned each genome using a length cutoff of at least 5 amino acids searching for SAARs. The homologous SAARs are characterized by whether the two SAARs from each species have the same amino acid composition and whether they have overlapping regions in the alignment of proteins they are located in. All SAARs counts, divergence times and the number of homologous SAARs are summarized in Table 2.11.

Figure 2.23 plots the count of conserved SAARs versus divergence time. There

is a trend for decreasing conservation as divergence time increases. The only exception is the *Drosophila melanogaster-Drosophila simulans* pair. The count of conserved single amino acid repeats within this pair is lower than both *Drosophila pseudoobscura-Drosophila persimilis* which has a smaller divergence time (1.7 MY) and *Drosophila yakuba-Drosophila erecta* pair which has a higher divergence time (25.6 MY). Then we used the counts of conserved SAARs and the divergence time to construct a linear model. We used the log ratio of divergence time as a predictor to build a linear model and the result is shown in Figure 2.24. The p-value of the slope of the linear model is 0.04 and the intercept is 0.002. This model indicates a strong and significant, negative correlation between counts of conserved SAARs and divergence times. This could be explained if we assume single amino acid repeats are deleterious or slightly deleterious or neutral. Previous studies have shown that single amino acid repeats could have a negative effect on fitness (such as Huntington disease), so the presence of single amino acid repeats could be selected against and purged as is shown in *D.melanogaster*, *D.yakuba* and *D.pseudoobscura* pair. However, recently, one study found that conserved tandem repeats could be conserved through the whole eukaryotic phylogeny and in some of them, their origins could even be traced from *Homo sapiens* back to yeast [Schaper et al., 2014]. This is conservation of repeats for about one billion years (divergence time between animal kingdom and fungal lineage determined by Doolittle [1996]). Our results, contradict this hypothesis and shows that during the evolution, while species can gain new single amino acid repeats, the major trend is that these repeats are disappearing even among *Drosophila* species.

Schaper et al. [2014] observed a clear decrease of conservation in shorter tandem repeats in proteins compared to tandem repeat regions with longer units. Since SAARs are one kind of tandem repeat in proteins, it is meaningful to see whether this pattern can be shown using SAARs. Here, we used MAFFT to align all homologous

proteins. All single amino acid repeats with an overlapping region in all 12 *Drsophila* species and with the same amino acid component were characterized as "conserved". Otherwise they were characterized as "unconserved". We further seperated "conserved" and "unconserved" into X-linked and autosome-linked SAARs. For each, we calculated the mean and variance of the lengths of SAARs. The results are show in Figure 2.25. Here we can see, on both X chromosome and autosomes, conserved single amino acid repeats tend to be longer than unconserved ones (7.7 amino acids long for conserved versus 6.9 for unconserved on X chromosome; And 7.9 versus 7.2 for autosomes). Using a t-test between the lengths of conserved SAARs and that of unconserved ones both on X and autosomes, we find this pattern is significant (p-value is 8.3×10^{-5} on X chromosome and p-value $< 1 \times 10^{-10}$ on autosomes). Also, we characterized all conserved single amino acid repeats and unconserved SAARs in each *Drosophila* species pairs and calculate the mean and variance of each pair shown in Figure 2.26. Also, each pair is tested using one-side t-test with null hypothesis that the mean values of both conserved SAARs and unconserved SAARs are same within each *Drosophila* species pairs. All p-values are less than $\alpha = 0.05$, which means within each pair, conserved SAARs are significantly longer than that of unconserved SAARs. This shows that the SAARs which are conserved through *Drosophila* phylogeny are significantly longer than that of unconserved ones. Previous studies have shown that some phenotypes of SAARs are only shown when their lengths pass a threshold, (such as in Huntington's disease, the poly-Q should be longer than 34 before patients show symptoms; Usdin [2008]). This suggests that the conservation of SAARs might be related to the unknown function they have.

We also checked whether there is a difference between the amino acid composition in conserved and unconserved SAARs. We counted the presence of single amino acid repeats with different types of amino acid and then calculated the proportion of

each type of amino acids. The result is shown in Figure 2.27. We can see that the composition of amino acid in conserved SAARs is different from that of unconserved SAARs. Especially poly-Q, which is dominant in *Drosophila* species decreases a lot and poly-A increases a lot and surpasses the proportion of poly-Q in conserved single amino acid repeats. A pairwise t-test shows that conserved and unconserved SAARs have significantly different composition of amino acids (p-value = 0.009).

Species	Count of SAARs	Divergence Time(MY)	Conserved SAAR count
Dpse	5008	1.7	3993
Dper	4499		
Dmel	3870	10.8	2692
Dsim	3193		
Dyak	3753	25.6	2984
Dere	3707		
Dana	4244	85.8	2692
Dwil	6823		
Dmoj	6209	124.4	1872
Dgri	6938		

Table 2.11: Summary of counts of single amino acid repeats and conserved ones within *Drosophila* species pairs.

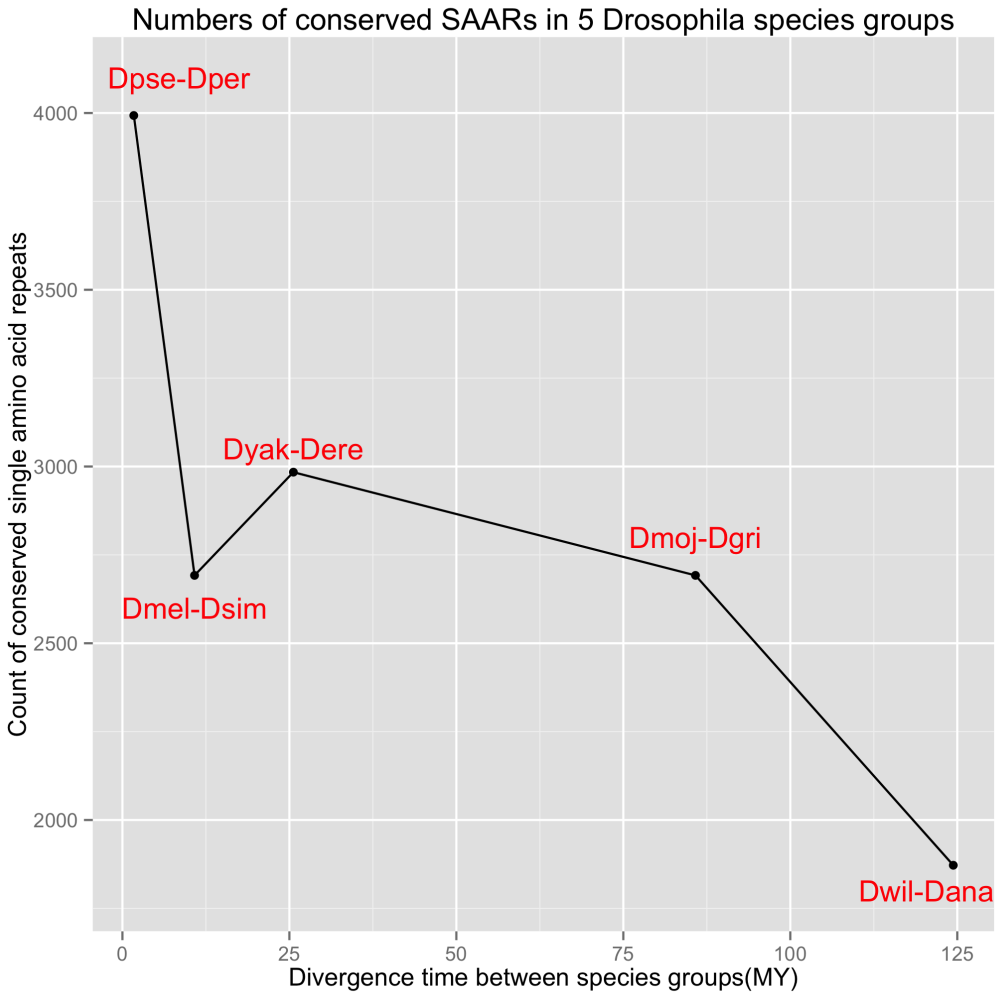


Figure 2.23: Plot of the numbers of conserved single amino acid repeats versus divergence in 5 *Drosophila* pairs.

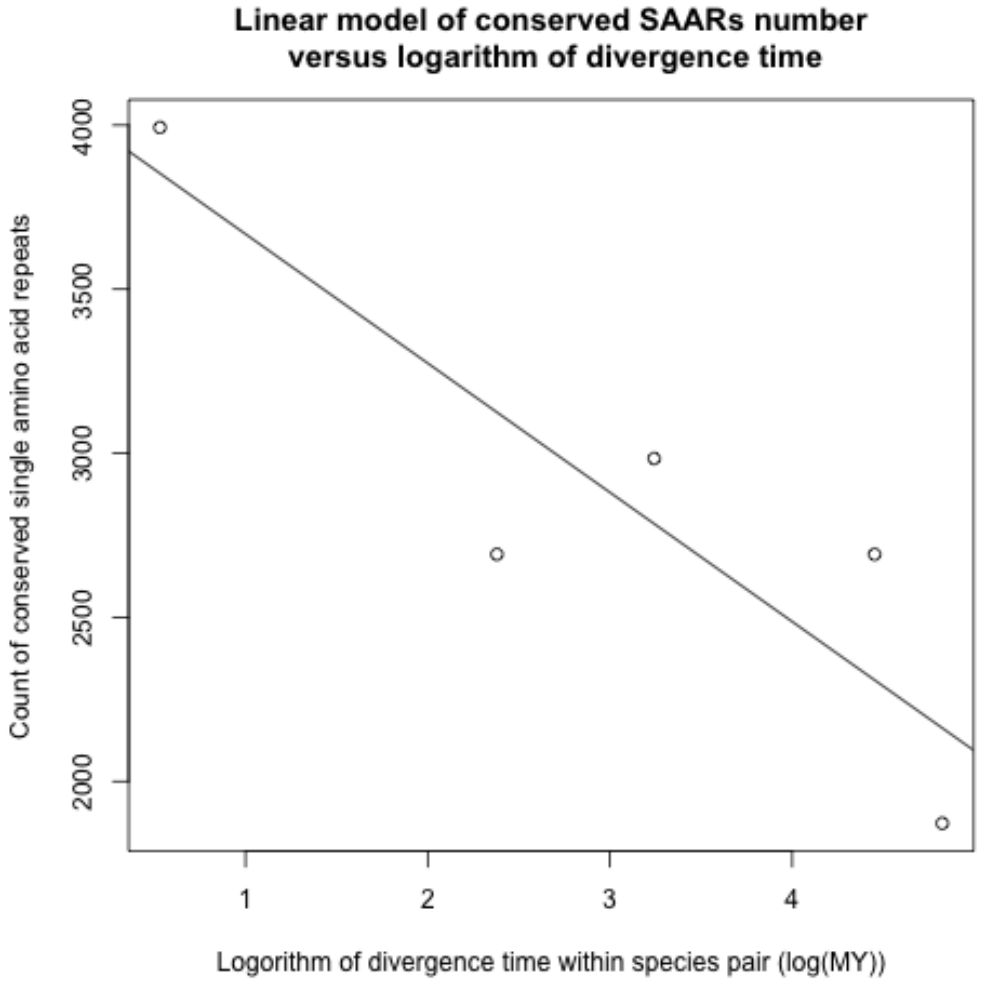


Figure 2.24: Linear model constructed using log ratio of divergence time as predictor variables and counts of conserved single amino acid repeats as responsible variables in 5 *Drosophila* pairs.

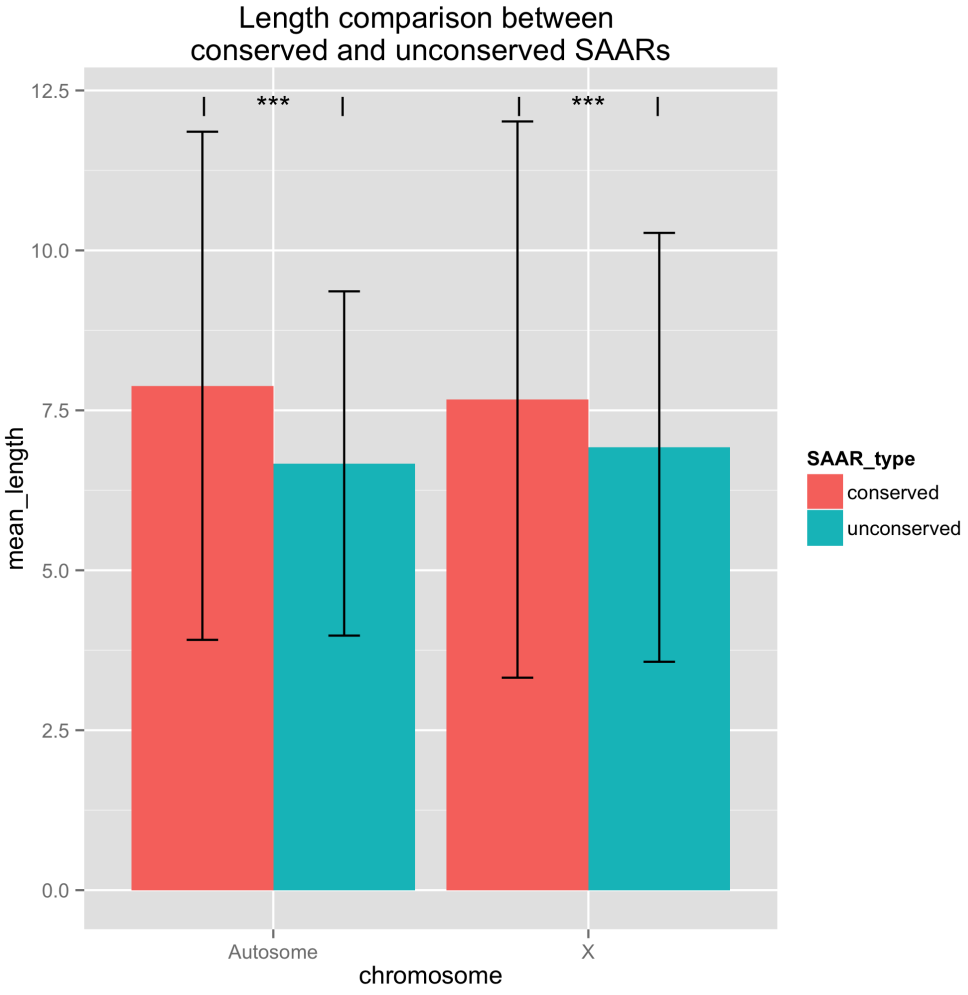


Figure 2.25: The length comparison between conserved and unconserved single amino acid repeats both on X chromosomes and autosomes from *Drosophila* species.

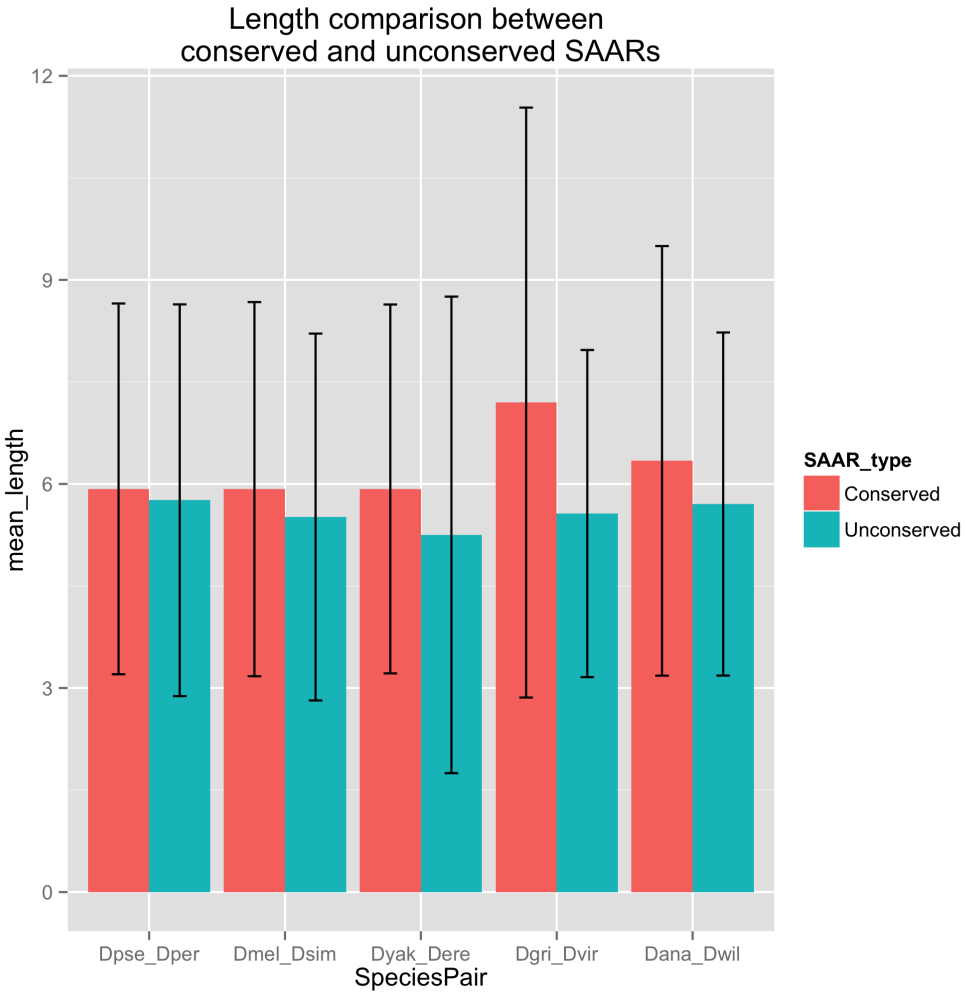


Figure 2.26: The length comparison between conserved and unconserved single amino acid repeats both on X chromosomes and autosomes for all five *Drosophila* species pairs.

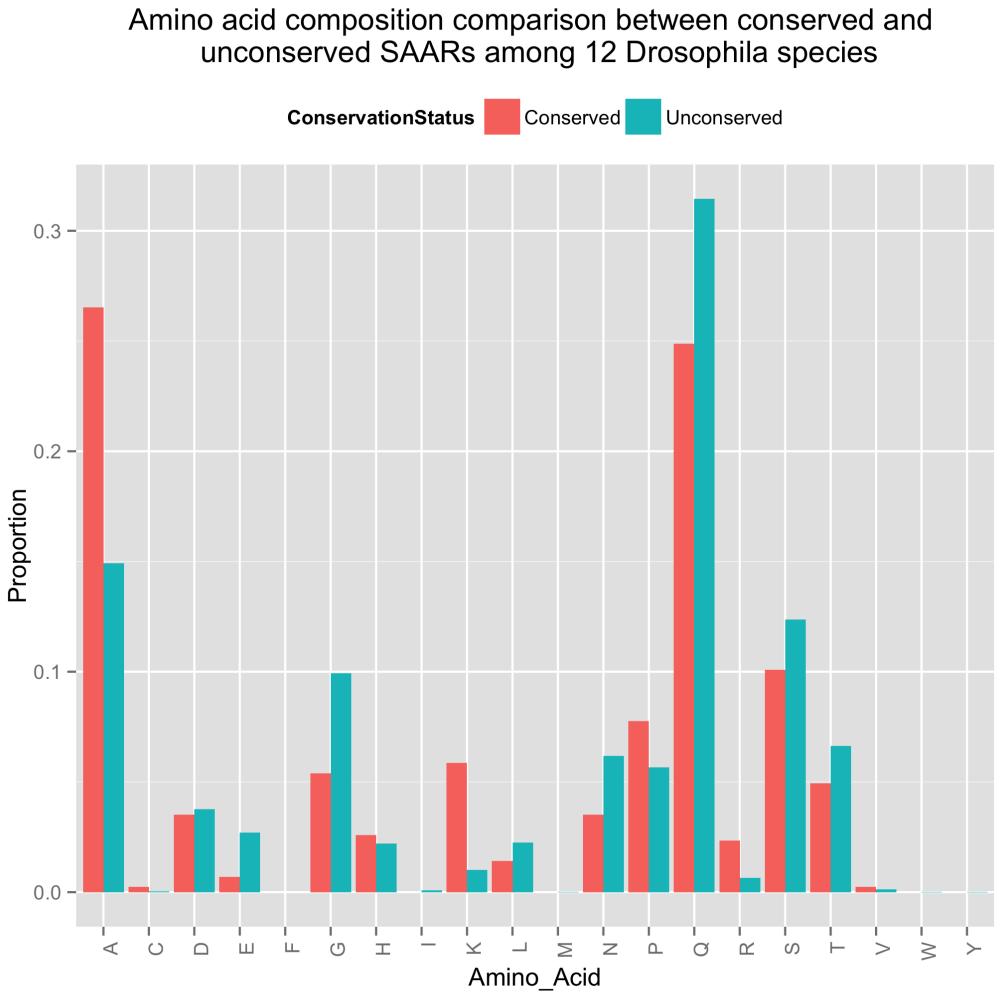


Figure 2.27: The proportion of each amino acid that construct both conserved single amino acid repeats and unconserved ones in *Drosophila* species.

2.4 Conclusion

Previous studies often considered single amino acid repeats to be evolving neutrally. In our study, we found single amino acid repeats are not randomly distributed within and between chromosomes and even within proteins. Especially, we found there is a significant tendency for an accumulation of SAARs on X chromosome among eukaryotic species compared to that of autosomes. This could suggest that the position and presence of SAARs are possible to be tuned by some underlying mechanism. Since previous studies illustrated that the expansion of some SAARs is associated with phenotypic variation, their positions might be influenced by their undetermined function. Also, we found the coding sequences for SAARs have different codon usage bias compared to that of non-repetitive regions of protein coding sequences. For example, in non-repetitive regions, the dominant codon coding for alanines is GCC on both X chromosome and autosomes. However, the dominant codon coding for alanines in SAARs is GCA on both X and autosomes. This could show us a general picture that within SAARs, the complexed effect of random genetic drift, mutation and selection on codon usage bias is significantly different from that of non-repetitive regions. To understand the ongoing codon usage change, we used *Drosophila melanogaster*, *Drosophila simulans* and *Drosophila yakuba* genomes to infer the ancestral state of codons, then further check the direction that codon changes both within SAAR coding regions and non-repetitive regions in *Drosophila melanogaster*. If we assume the codon usage bias is stable, we should expect there are equal numbers of preferred codons changing to unpreferred codons and unpreferred codons changing to preferred codons. However, based on the results, we observed that for both SAAR regions and repetitive regions, there are more preferred codons changed to unpreferred codons. Since many genomes of *Drosophila* species have been sequenced, they form a perfect dataset for comparative genetic research. We split

these species into five *Drosophila* pairs with non overlapping branch and different divergence times. We then constructed a linear model to estimate the purging pattern of SAARs. From the linear model, we found that the conservation number of single amino acid repeats is negatively correlated with the logarithm of divergence time. Also, we found that the conserved SAARs are significantly longer than those that are purged during divergence. In addition, conserved SAARs in general have different amino acid composition compared to those lost ones.

For future directions, in order to further illustrate the evolution of SAARs, more sequenced genomes with genes mapping to both autosomes and sex chromosomes could be added into this study. In addition, as another good comparative genetic database, genomes from primates could be added to research on codon usage bias and the conservation of SAARs to test the generality of what we found in this thesis. Also, what will be really interesting is the evolution of single amino acid repeats on neo-sex chromosomes of several *Drosophila* species such as *Drosophila miranda* [Zhou and Bachtrog, 2012] and *Drosophila albomicans* [Zhou et al., 2012]. These newly generated sex chromosomes could provide us perfect data to let us inference what the early evolution of SAARs looks like in primative sex chromosomes.

Chapter 3

SAAR evolution on Neo-sex chromosomes

3.1 Introduction

Generally speaking, sex chromosomes are believed to be descendents of homologous autosomes. For example, in the well studied *Drosophila melanogaster* sex chromosome system, Muller-element A is the X chromosome and the Y chromosome is a highly degenerated and pseudogenized chromosome. Although homologous sex chromosomes were commonly established millions of years ago, they still possess signs showing their evolutionary origins [Lahn et al., 2001]. Evolving pairs of sex chromosomes normally employ several evolutionary steps differentiating from each other. First of all, gene content on X and Y chromosomes could change drastically due to the suppression of recombination. Lack of recombination can lead to the degeneration of Y-linked genes by introducing repeats or early stop codons or fast fixation of deleterious mutations on Y chromosome so that lots of genes become pseudogenized [Charlesworth and Charlesworth, 2000]. In contrast,

X chromosomes are euchromatic, gene-rich and have adopted a hyperactive chromatin configuration, resulting in hyper-transcription of X-linked genes in male *Drosophila*. One hypothesis for the cause of suppression of recombination between sex chromosomes is sexually antagonistic mutations [Lahn and Page, 1999, Ross et al., 2005], which means these mutations are beneficial to one sex but detrimental to the other one. Then, in response to the gene loss on Y chromosome and employing many epigenetic modifications to silence Y-linked genes leads to the develop of a process to make X-linked genes dosage-compensated [Zhou et al., 2013]. By adopting these steps, autosomes can gradually evolve to become sex chromosomes.

Due to the lack of species with newly developed sex chromosomes, previous studies could only use highly heterochromatic Y chromosomes to infer the evolution of Y-linked genes after the establishment of sex chromosomes. The recently sequenced *Drosophila miranda* genome [Kaiser and Bachtrog, 2013], with a pair of newly formed sex chromosomes, provides perfect data for unraveling the mysteries of the evolution of sex chromosome system.

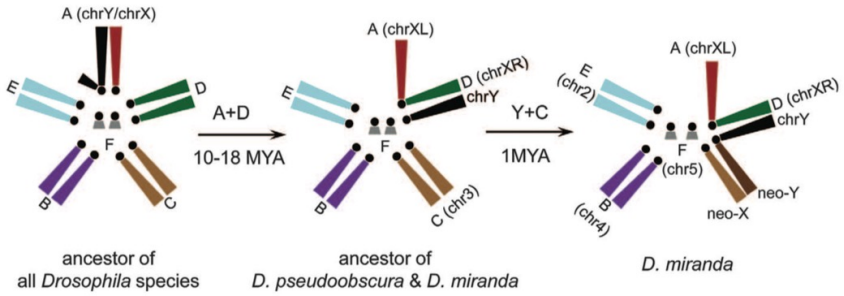


Figure 3.28: Illustration of 2 fusion events that generate new sex chromosomes in *Drosophila pseudoobscura* and *Drosophila miranda* [Kaiser and Bachtrog, 2013].

The chromosomes of *Drosophila* species can be characterized into a set of homologous chromosomal arms called "Muller elements" [Muller, 1940]. During

evolution, since the common ancestor of all *Drosophila* species, chromosomal fusions have happened between Muller-A (ancestral sex chromosome in *Drosophila* species) and autosomes to generate younger sex chromosomes. For example, Muller-A has been fused with Muller-D before the divergence of *Drosophila pseudoobscura* and *Drosophila miranda* about 10-18 million years ago [Carvalho and Clark, 2004]. This fusion made Muller-D become XR and caused the counterpart fused on Y. About 1 million years ago, another fusion happened specific to *Drosophila miranda* involving Muller-C and Y chromosome. This leads *Drosophila miranda* to have three different X chromosomes (XL, XR and Neo-X) and two Y chromosomes, one of which is the very young Y chromosome which was named "Neo-Y" [Bachtrog and Charlesworth, 2002b] and the other one is the ancestral Y. This very young sex-chromosome system is now in the process of evolving from a pair of ordinary autosomes to a pair of heteromorphic sex chromosomes. Studies employing cytogenetic methods and investigations of specific genomic regions enable us to examine the evolution of new Y chromosomes. Studies have confirmed that the recent suppression of recombination between Neo-Y and Neo-X makes the degeneration incomplete [Steinemann et al., 1993, Bachtrog et al., 2008] and further, makes the neo-X partially hemizygous.

In this chapter, we illustrate a preliminary picture of how single amino acid repeats are evolving in a young sex chromosome system. Since the lack of recombination causes the neo-Y chromosome to degenerate, including introducing fixation of deleterious mutations and accumulation of repetitive regions, lots of genes located on the neo-Y have became non-functional. Also, the neo-X could further show a picture of how genes are going to act when they are exposed to the "Faster-X" effect on Muller-C. Based on this context, if SAARs are located on neo-sex chromosomes, they could act differently compared to SAARs located on autosomes.

3.2 Materials and methods:

3.2.1 Distribution of SAARs in *Drosophila miranda*

Since the formation of neoX and neoY in *Drosophila miranda* is only 1-2 MYA, the coding regions on the neoX and neoY are very similar. Based on the genome sequence, the estimated divergence of coding regions between NeoX and NeoY is only about 1.5% [Kaiser and Bachtrog, 2013]. To analyze the single amino acid repeats distribution in *Drosophila miranda* especially for genes from neoX and neoY, we downloaded the *de novo* transcriptome assembly of *Drosophila miranda* done by the Bachtrog group [Kaiser and Bachtrog, 2013] (NCBI accession number: GALP00000000). This transcriptome assembly makes use of genomic reads mapping to the neo-sex transcripts both in males and females such that an assembly of the neo-Y transcriptome could be constructed.

The dataset in total includes 12521 transcribed RNA sequences, in which, 2141 RNAs are from neoX chromosome and 1863 are from neoY chromosome. We used TransDecoder[MacManes and Eisen, 2013] to predict the protein coding genes from these transcripts. We first used a program called TransDecoder.LongOrfs to identify open reading frames at least 100 amino acids long. Then, these open reading frames were filtered using TransDecoder, for predicted protein coding genes using proteins in *Drosophila pseudoobscura* and the pfam protein database as its searching resources. In total, 2981 and 2800 proteins from neoX and neoY chromosome were identified. Then we further filtered the proteins with several processes: First, proteins with homologous proteins in *Drosophila pseudoobscura* were gathered and then BLASTed against their *Drosophila pseudoobscura* homologous proteins and only the proteins with the highest score were kept. Second, we kept proteins with the longest putative functional open reading frames (no detectable frame-shift mutations). The number

of proteins pass the filter process for each chromosome is summarized in Table 3.12. For each chromosome, we calculated the proportions of genes with single amino acid repeats for chromosomes. The results are shown in Figure 3.30 and Figure 3.31

3.2.2 Dominant codons of neo-sex chromosome linked SAARs

Since genes linked to sex chromosomes may show different levels of functional change than autosomal genes due to different evolutionary pressures, it is meaningful to examine the codon usage within single amino acid repeats compared to that of non-repetitive regions for *Drosophila miranda*. We separated the protein coding genes into five non-overlapping types: autosome-linked, neoX-linked, neoY-linked, XL-linked and XR-linked, based on which chromosome the protein coding gene is located on. For each protein coding gene, we BLASTed it against *Drosophila pseudoobscura* genome to find the homologous gene. For each pair of genes, we used MAFFT [Katoh and Standley, 2013] to align the homologous protein coding gene from *Drosophila miranda* and *Drosophila pseudoobscura*, by setting the option to "localpair". Then, we used a customized Python script to scan both genes to find single amino acid repeats with length at least 5 amino acids long. Based on the alignments, we used two strategies to find homologous SAARs: 1. In the alignments of both homologous proteins, the position of the SAAR in the alignment should overlap at least one amino acid with the SAAR of its homologous protein. 2. For both flanking regions of SAARs (with length at least 10 amino acids long, if the flanking regions is less than 10 amino acids long, only use the flanking region on the other side), only those with at most 2 mismatches are kept (not counting gaps in the alignments). In total, for neoX proteins, 384 out of 435 SAARs were found to

be conserved and for neoY proteins, 221 out of 245 SAARs were found present in both species. On chromosome 2 and 4, we found 716 out of 804 and 502 out of 576 SAARs are present in both *Drosophila miranda* and *Drosophila pseudoobscura*. In addition, on chromosome XL and XR, 576 out of 663 and 492 out of 552 SAARs are homologous. The DNA sequences of homologous SAARs were extracted from both protein coding genes and again aligned using MAFFT. As controls, the non-repetitive regions were extracted after trimming homologous single amino acid repeats DNA off the protein coding genes and aligned using MAFFT [Katoh and Standley, 2013].

3.2.3 Codon Usage Bias within neo-sex chromosome linked SAARs

In order to check the relative abundance of synonymous codons, we calculated the RSCU (relative synonymous codon usage) value, which is defined as the ratio of the observed frequency of codons to the expected frequency given that all the synonymous codons for the same amino acids are used evenly [Sharp et al., 1986]. DAMBE was used to make this calculation [Xu, 2013]. We collected all SAARs with at least 5 amino acids in the *Drosophila miranda* genome. All SAARs coding DNA were extracted, and we then concatenated their flanking non-repetitive regions. We separated these two types of DNA based on which chromosome they come from into FASTA files. We then used these FASTA files as inputs of DAMBE and those RSCU values of different regions were then calculated.

To study changing directions of codon usage in *Drosophila miranda*, we BLASTed each protein from *Drosophila miranda* to genomes of *Drosophila pseudoobscura* and *Drosophila persimilis* to find homologous proteins coding genes. Then we used

MAFFT to align homologous proteins sequences from these three species. Based on protein alignments, we created codon alignments. Then all codon alignments are fed into FastML [Ashkenazy et al., 2012] to infer the ancestral state of each codon. We then classified codons into two different states based on the RSCU values we calculated in the previous section: Preferred codons are those codons with RSCU values larger than one, unpreferred codons are codons with RSCU values less than one. Then assuming the codon changing direction is from ancestral codon state to *Drosophila miranda* codon state, we can classify synonymous codon mutations into three different kinds: preferred codon change to unpreferred codon (P to U), unpreferred codon change to preferred codon (U to P), or codon state remains the same (unpreferred codon change to unpreferred codon and preferred codon change to preferred codon). For both non-repetitive regions and SAARs, we counted the number of each type of synonymous mutation on XL, XR, autosomes, neoX and neoY.

3.2.4 The influence of SAARs on divergence of Neo-sex chromosomes

Since the fusion between Muller-element C and ancestral Y chromosome is only about 1-2 MYA, the divergence of homologous protein coding genes on neoY and neoX in *Drosophila miranda* is only 1.5%. This means most of the protein coding genes are now in a preliminary stage of Y-degeneracy. In addition, previous studies have shown that repetitive sequences located in coding regions can accelerate the rate of evolution of their flanking coding regions. Based on these two assumptions, it is intriguing to see whether single amino acid repeats could influence the evolution or degeneracy of protein coding genes on a young Y chromosome.

In order to estimate the divergence of proteins with/without SAARs for genes from neoX and neoY, we extracted all protein coding sequences on neoX and neoY. We first constructed a protein BLAST database using all proteins from neoX. Then we used proteins from neoY as queries to BLAST against the neoX database using a length at least 200 amino acids and expect value at most 1×10^{-5} as the cutoff to find all the homologous proteins from neoX and neoY. All the proteins with incomplete open reading frames (not starting with "ATG" or possessing premature terminal codon) are discarded. In addition, due to the fact that the maximum likelihood estimates of divergence can be inaccurate for short sequences, we removed all proteins less than 100 amino acids from our dataset. We used the *Drosophila pseudoobscura* genome from FlyBase to BLAST [Altschul et al., 1990] all protein coding genes against our neoX database to find proteins that are homologous to those proteins located on neoX and neoY. Furthermore, we found all homologous proteins in the 12 sequenced *Drosophila* genomes and discarded proteins that have no homologue in one or more species of these 13 species.

For each homologous protein, we now have 14 protein sequences from *Drosophila simulans, Drosophila sechellia, Drosophila melanogaster, Drosophila yakuba, Drosophila erecta, Drosophila ananassae, Drosophila pseudoobscura, Drosophila persimilis, Drosophila willistoni, Drosophila mojavensis, Drosophila virilis, Drosophila grimshawi*, neoX and neoY from *Drosophila miranda*. We used MAFFT to align each homologous protein group. Based on the annotation information of these proteins, we extracted protein coding DNAs from their genomes. Then we used a program called PAL2NAL.pl [Suyama et al., 2006] to convert homologous protein coding DNA FASTA files to PAML format codon alignments based on their protein alignments. Then we used Bio.Phylo [Talevich et al., 2012] to write a customized python script that could automatically edit PAML [Yang, 2007] control file, run Codeml program

and read Codeml results. Based on previous studies from Clark [2007] and Gao [2007], we used the phylogenetic topology of the 14 homologous proteins given in Figure 3.29 as the topology input for PAML. For each pair of homologous coding DNA sequences, in neoX and neoY, we categorized these three situations: First, both neoY and neoX proteins have SAARs; Second, only one protein has SAARs; Third, SAARs are preserved through all 14 homologous proteins of *Drosophila* species. Based on these three situations, we analyzed the dN/dS values of neoX and neoY branches in dN/dS tree given by PAML. And we counted the number of genes with these four states: $dN/dS > 1$ with SAARs, $dN/dS < 1$ with SAARs, $dN/dS > 1$ without SAARs and $dN/dS < 1$ without SAARs. And the results are shown in Table 3.15, 3.16 and 3.17.

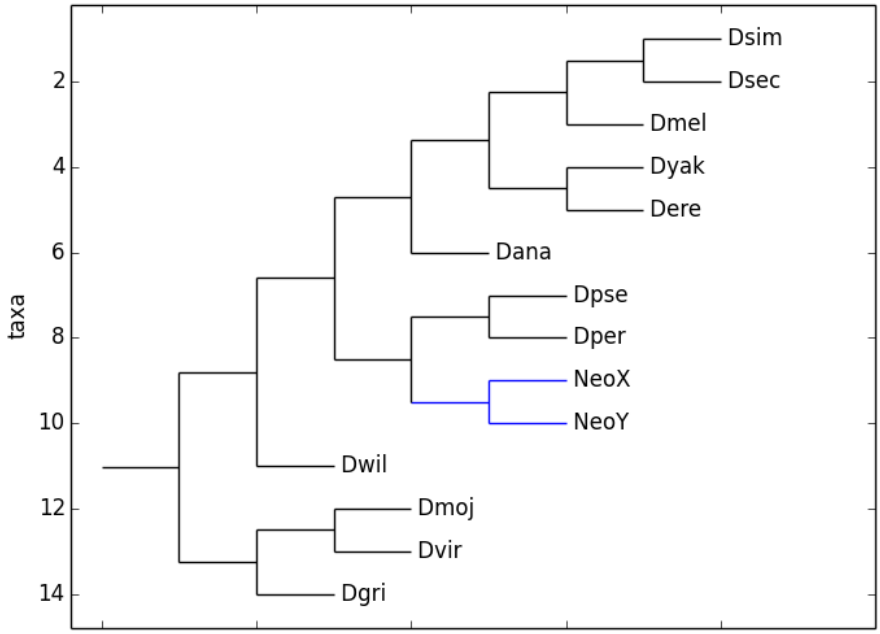


Figure 3.29: Topology of 12 sequenced sex chromosomes of*Drosophila* species and neo-sex chromosomes from *Drosophila miranda*.

3.3 Results

3.3.1 Distribution of SAARs in *Drosophila miranda*

We counted the number of single amino acids on each chromosome of *Drosophila miranda*. We filtered all proteins and only kept the longest isoform of each gene. Only single amino acid repeats with at least 5 amino acids long are counted for each chromosome. All SAARs from *Drosophila pseudoobscura* are also extracted in the same way. The results are summarized in Figure 3.30. We found that not only the counts of SAARs on neoY and neoX are much lower than for the corresponding chromosome of *Drosophila pseudoobscura* but the same phenomenon is also observed on autosomes and ancestral sex chromosomes.

In addition, we also calculated the proportion of genes with single amino acid repeats for each chromosome since in the previous chapter we found that the distribution of the proportion of genes with SAARs is biased and the proportion of X-linked genes that contain SAARs is significantly higher than that of autosomes. We used the same method like what we did for the count of SAARs and only scanned for SAARs in regions with no "X" or gap in protein sequences. Then we counted the number of genes with SAARs and divided it by the total number of proteins located on corresponding chromosome. We plotted the proportion of genes with SAARs and the results are shown in Figure 3.31.

From the Figure 3.30, we can see that there is significant difference of the counts of SAARs between *Drosophila miranda* and *Drosophila pseudoobscura* for all pairs of homologous chromosomes. This could be caused by including partial sequenced genes in our analysis. As we have shown in the previous chapter, the positions of SAARs are highest at both ends of protein sequences. In our *Drosophila*

miranda genome dataset, we have included some genes which are partially sequenced. This means that some proteins do not have their ends sequenced so that might make the proportion of genes with SAARs of each chromosome smaller compared to proportion on the homologous chromosome of *Drosophila pseudoobscura*.

Since both neoX and neoY chromosomes are derived from the same ancestral chromosome (Muller-element C), it is meaningful to see whether there is a SAAR count difference between neoX and neoY under a context of early stage of Y-chromosome effect. To evaluate the difference, we calculate the proportion of genes with SAARs for each chromosome of *Drosophila miranda*. From the result shown in Figure 3.31, we can see that the proportion of genes with SAARs on neoY chromosome are much lower than that of neoX chromosome, even though they were derived from the same chromosome around 1.5 MYA. Assuming there is no change between numbers of genes with SAARs, we can use Pearson's chi-square test to test the null hypothesis that genes with SAARs on neoX and neoY are the same. We found the p-value is 0.002 which is less than $\alpha = 0.05$. This means the numbers of SAARs of neoY and neoX are significantly different from each other. This could be caused by two reasons: First, since in previous studies, SAARs are considered by some to be evolving neutrally without negative selective constrain. Also, using 209 neo-X and neo-Y chromosome linked genes, Bachtrog [2008] found neo-Y genes which had already lost functions (early terminal codons or genes with frame shift) have twice the amino acid evolution rate compared to those with potential functions. Based on these findings, if we assume single amino acid repeats are selectively neutral, this phenomenon could be explained that since neo-Y linked SAARs have a higher amino acid evolution rate, it is highly possible that SAARs could be broken by amino acid substitutions that happened in the middle of SAARs and creating two shorter SAARs that can not pass the cutoff of 5 amino acids long. Second, if

we assume the presence of SAARs are under purifying selection, due to the fact that neo-Y chromosome lost its recombination with its homologous chromosome which is neo-X, the selection on deleterious mutations is reduced heavily [Bachtrog and Charlesworth, 2002a] so that mutations that could break down SAAR can accumulate faster on neo-Y chromosome than those on its homologous counterpart.

Chromosome	Proteins
neoX	1495
neoY	1316
2	2043
4	1490
XL	1294
XR	1393
Unknown	

Table 3.12: Summary of the homologous gene pairs we found between *Drosophila miranda* and *Drosophila pseudoobscura*.

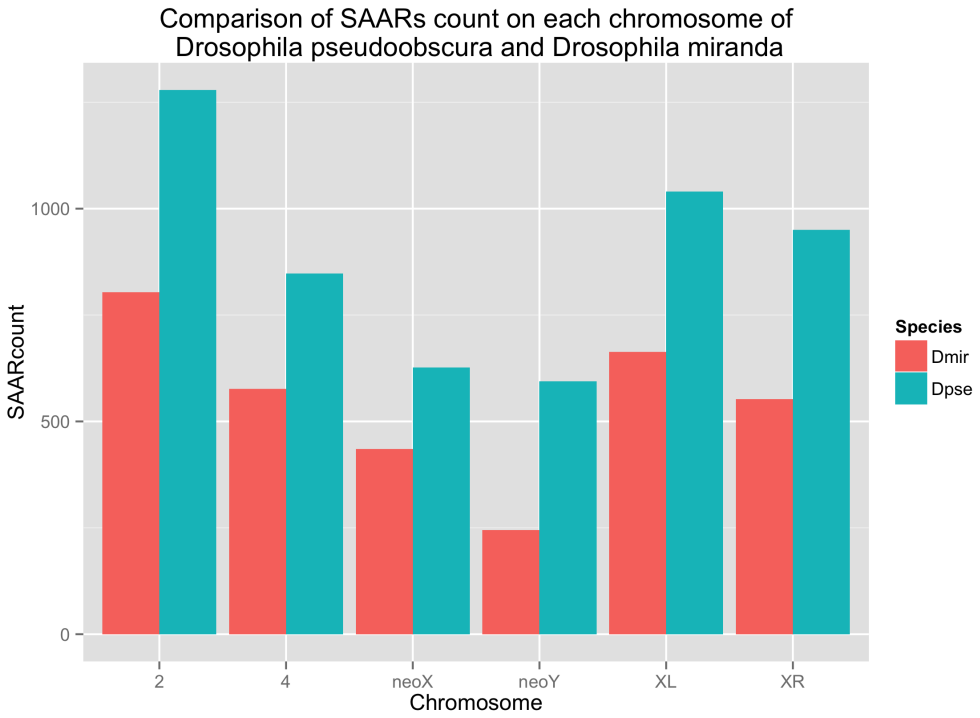


Figure 3.30: Comparison of single amino acid repeats count on each chromosome from *Drosophila miranda* and *Drosophila pseudoobscura*.

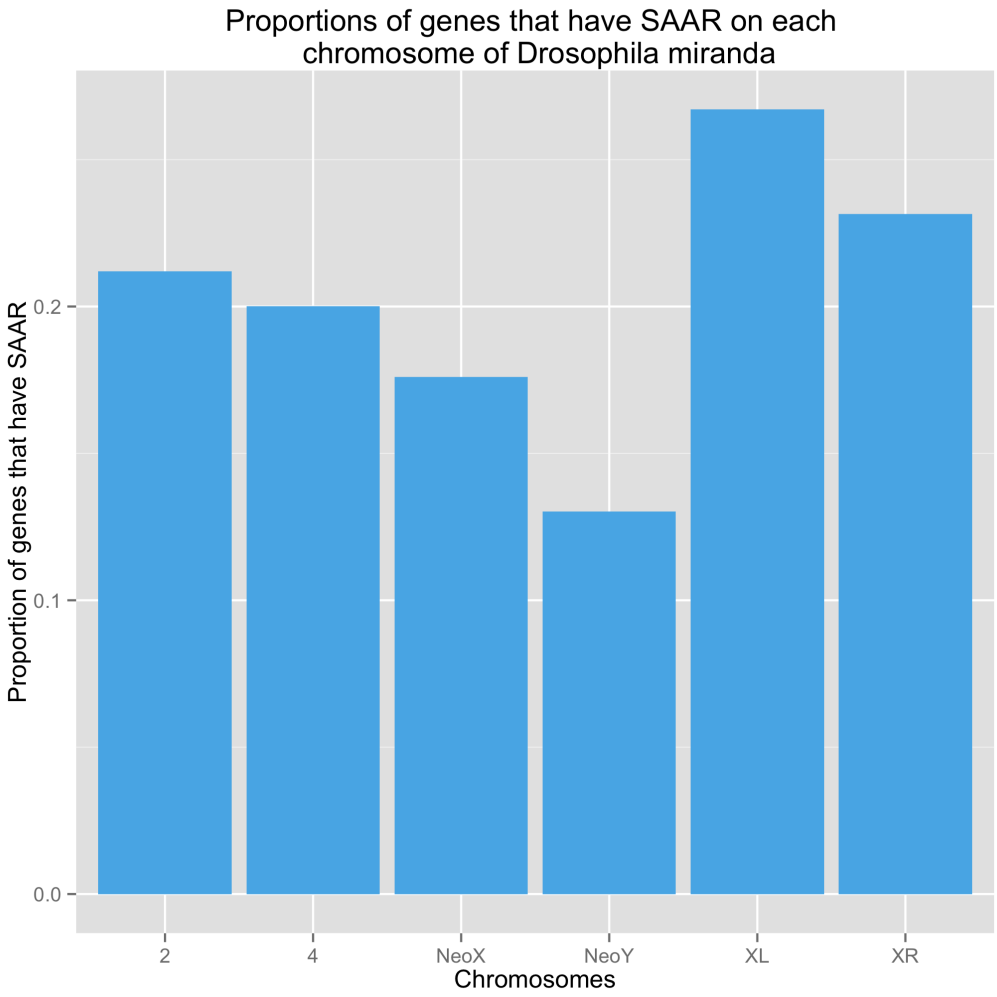


Figure 3.31: Proportions of genes with SAARs for all chromosomes of *Drosophila miranda*.

3.3.2 Dominant codons of neo-sex chromosome linked SAARs

Based on the homologous protein alignments generated by MAFFT, we extracted all protein coding DNAs both from *Drsophila miranda*, *Drosophila persimilis* and *Drosophila pseudoobscura* and aligned all codons based on their protein sequences. We use the codons predicted by FastML as ancestral codons and assume that the codon change direction is from ancestral codons to *Drosophila miranda*. We know that the codon usage is biased in functional coding regions, and unpreferred codons are weakly selected against. However, the nonrandom usage of synonymous codons is the result of combined effect of selection, mutation and random genetic drift [Akashi and Walker, 1998]. Previous studies found that the effective population size of *Drosophila miranda* is estimated only to be 1/6 of that of its sibling taxa *Drosophila pseudoobscura*, which means the random genetic drift could be much more stronger. This made us curious about how all these effects can shape the codon usage bias in SAARs coding regions and non-repetitive regions in *Drosophila miranda*.

We splitted our original data into two different types: non-repetitive sequences and SAARs coding sequences. Then, all sequences from chromosome 2 and chromosome 4 were combined as autosome data, and all sequences from XL and XR were combined as ancestral X chromosome data. All sequences that cannot be mapped to chromosomes were discarded. Then, for both SAAR coding sequences and non-repetitive regions, we have four types of data: neo-X, neo-Y, autosome and ancestral X. We use DAMBE [Xu, 2013] to calculate RSCU values for each type of data. And the results are shown in Table 3.13.

From Table 3.13, we can see that for non-repetitive regions, all optimal codons

are the same through different types of chromosomes. This means that for non-repetitive regions, although different chromosomes may have different effects of mutation, random genetic drift and selection (especially between autosomes and sex chromosomes), the overall preference towards optimal codons is almost the same within non-repetitive regions. Since the recombination between neoX and neoY chromosomes has already ceased, which means that more slightly deleterious mutations could accumulate on the neoY chromosome, and since selection doesn't favor non-optimal codons, we can expect more non-optimal codons present in non-repetitive regions on neoY chromosome. However, such expectation is not observed from our results. This could be due to the fact that the end of recombination is only 1.5 MYA, there is not enough time to accumulate non-optimal codons in non-repetitive regions.

We then compared RSCU values between non-repetitive regions and SAAR coding regions on each chromosome, except for several codons with low frequency in SAAR coding regions, such as W, Y, M, F and C. We found several non-optimal codons, like codons for A, D, H, P, R, T and V, have became dominant codons in SAAR coding regions. Among these amino acids, the dominant codons for H, P, R, T are totally changed in SAARs compared to that of non-repetitive regions. Several explainations could be used to interpret this result. First, since SAARs in previous studies are considered to be evolving neutrally, the selection pressure on certain optimal codons could be more relaxed within SAAR regions than within non-repetitive regions. This can drive the accumulation of non-optimal codons within single amino acid repeats. Second, even though we assume the selection is acting evenly on SAARs and non-repetitive regions, due to the fact that SAARs are repetitive regions, the mutational rates within single amino acid repeats can be higher compared to their flanking non-repetitive regions since studies have shown repetitive

regions could cause the mutational rates to become higher due to replication errors. Since it takes time for selection to get rid of weakly deleterious muations within the same gene, regions with higher muational rates could accumulate more non-optimal codons like in SAARs.

	Non-rep				rep			
Amino Acid	Autosome	X	neoX	neoY	Autosome	X	neoX	neoY
A	GCC	GCC	GCC	GCC	GCA	GAC	GCA	GCC
C	UGC	UGC	UGC	UGC	UGC	UGC	null	null
D	GAC	GAC	GAC	GAC	GAU	GAC	GAU	GAC
E	GAG	GAG	GAG	GAG	GAG	GAG	GAG	GAG
F	UUC	UUC	UUC	UUC	UUC	UUC	null	null
G	GGC	GGC	GGC	GGC	GGC	GGC	GGC	GGC
H	CAC	CAC	CAC	CAC	CAU	CAU	CAU	CAU
I	AUC	AUC	AUC	AUC	AUC	AUC	AUC	null
K	AAG	AAG	AAG	AAG	AAG	AAG	AAG	AAG
L	CUG	CUG	CUG	CUG	CUG	CUG	CUG	CUG
M	AUG	AUG	AUG	AUG	null	null	null	null
N	AAC	AAC	AAC	AAC	AAC	AAC	AAC	AAC
P	CCC	CCC	CCC	CCC	CCG	CCG	CCG	CCG
Q	CAG	CAG	CAG	CAG	CAG	CAG	CAG	CAG
R	CGC	CGC	CGC	CGC	CGU	AGG	AGG	CGA
S	UCC	UCC	UCC	UCC	UCC	UCC	UCC	UCC
T	ACC	ACC	ACC	ACC	ACA	ACA	ACA	ACG
V	GUG	GUG	GUG	GUG	GUG	GUU	GUC	GUC
W	UGG	UGG	UGG	UGG	null	null	null	null
Y	UAC	UAC	UAC	UAC	UAC	null	null	null

Table 3.13: Dominant codon for each amino acid evaluated using RSCU values both in SAARs and non-repetitive regions on each chromosome of *Drosophila miranda*.

3.3.3 Codon Usage Bias within neo-sex chromosome linked SAARs

We previously found that using ancestral codon states inferred by *Drsophila simulans*, *Drosophila yakuba* and *Drosophila melanogaster* to evaluate the ongoing codon changes from ancestral to *Drosophila melanogaster*, there are relatively more preferred codons change to unpreferred codons in SAAR coding regions. In addition, for SAAR regions, the ratio of preferred codons to unpreferred codons are higher than that of non-repetitive regions. In order to have a complete picture of how codons change within SAARs of newly arised sex chromosomes, we used SAARs of *Drosophila miranda* as our data. We first searched SAARs that are conserved through *Drosophila persimilis*, *Drosophila pseudoobscura* and *Drosophila miranda*. In total, we found 665 SAARs from autosomes, 471 from XL and XR, 200 from NeoX chromosome and 121 from NeoY chromosome. Based on the protein sequence alignments, we aligned all codons. Then all ancestral codons are inferred using FastML. After trimming codons from SAARs, we had 2096 non-repetitive regions from autosomes, 1527 from XL and XR, 977 from NeoX and 858 from NeoY.

We characterized synonymous codons into two different types, preferred codons and unpreferred codons, based on RSCU values calulated in previous section. Preferred codons are those codons with RSCU values larger than 1 and in contrast unpreferred codons are characterized by RSCU values smaller than 1. Further, based on the type of codons from inferred ancestral sequences and codons from sequences of *Drosophila miranda*, there are three types of codon changes: preferred codons change to unpreferred codons (P2U), unpreferred codons change to preferred codons (U2P) and preferred codons change to preferred codons and unpreferred codons change to unpreferred codons are characterized as unchange (P2P, U2U). We counted each type

on all 4 types of chromosomes: Neo-X, Neo-Y, autosomes (chromosome 2 and 4) and ancestral X chromosomes (XL and XR) and the results are summarized in Table 3.14.

We calculated the proportion of each type of codon chages within different types of coding regions and the results are shown in Figure 3.32. Firstly, to check whether there is a difference between the types of codon changes in SAARs and non-repetitive regions, we used Pearson's chi-squared test with the null hypothesis that there are the same numbers of codon changes in SAARs and in non-repetitive regions (P-values are 0.61 on NeoX, 0.30 on NeoY, 0.06 on autosomes and 0.66 on ancestral X chromosome). The insignificance of the difference between codon changes among SAARs and non-repetitive regions could be caused by their small sample size of codon changes in NeoX linked and NeoY linked SAARs. We then checked whether there are significant differences between codon changes on neoX and that on neoY. From the Figure 3.32, we can see that there are more unpreferred codons change to preferred codons compare to preferred codons change to unpreferred codons on NeoY-linked SAARs. Using Pearson's chi-square test, we got p-value equals to 0.96, which means we do not have significant power to reject the null hypothesis so that there are no difference between numbers of different types of codon changes among NeoX linked SAARs and NeoY linked SAARs.

Table 3.14: Counts of different kinds of codon changes(no change, preferred to unpreferred and unpreferred to preferred) with state of polymorphic and fixation within single amino acid repeats coding regions of *Drosophila miranda*. The ancestral states of codons are predicted using FastML using genomes of *Drosophila pseudoobscura*, *Drosophila persimilis* and *Drosophila miranda*.

	No change	P to U	U to P	total
X				
SAAR	2635	46	35	2716
non-repetitive	698193	6102	4101	708396
Autosome				
SAAR	3854	50	48	3952
non-reptititve	968405	10332	6673	985410
NeoX				
SAAR	1179	18	18	1215
non-repetitive	427932	4250	3401	435583
NeoY				
SAAR	694	9	10	713
non-reptititve	293420	3925	2466	299811

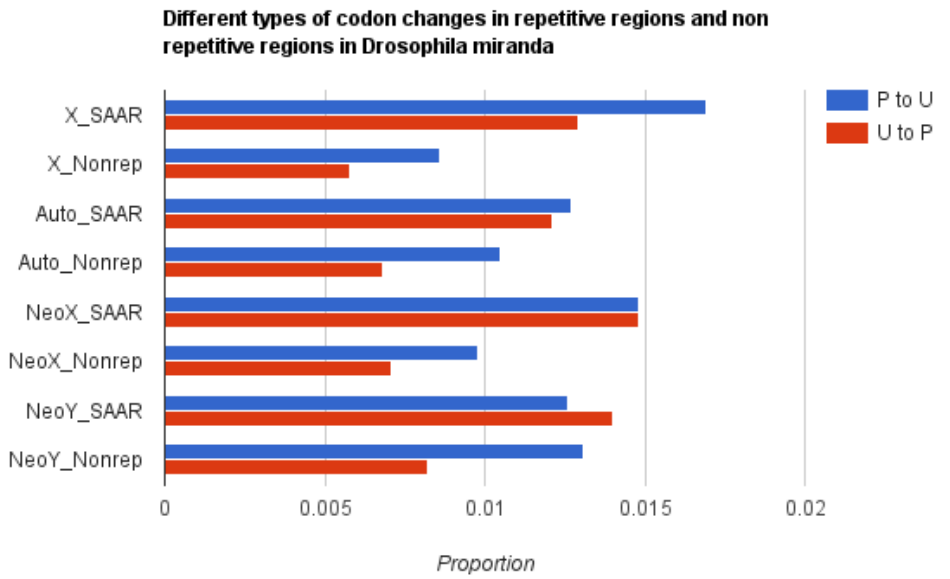


Figure 3.32: Proportions of different types of codon changes of SAARs coding regions and non-repetitive regions on each chromosome of *Drosophila miranda*.

3.3.4 The influence of SAARs on divergence of Neo-sex chromosomes

We used the free branch lengths model of codeml to estimate the dN/dS ratio for both NeoX and NeoY linked proteins. We used the BioPhylo package of BioPython to handle and process the final results of codeml. ω values on branches of the trees were extracted for NeoX-linked proteins and Neo-Y linked proteins. We filtered out the results with at least one "999" value for ω since it could be due to no synonymous differences among these sequences and is unreliable. In the end, we got results for 653 proteins.

We compared ω values to 1 for both NeoX and NeoY linked proteins. If dN/dS is equal to one, substitutions may be largely neutral. And if dN/dS is smaller than one, purifying selection might be present and in contrast, if dN/dS is larger than one, selection has caused some amino-acid substitutions. So based on this principle, we counted the number of proteins with ω larger than one or smaller than one and the results are shown in Table 3.15, 3.16 and 3.17.

We can see from these tables, that when there is a SAAR located within Neo-sex proteins, there is no difference between the number of proteins with high ω values compared to proteins with low ω values both on NeoX and NeoY in all three types of dataset. However, we did observe that when there is no SAAR within proteins, more proteins have high ω value on NeoY chromosomes than that of NeoX chromosome. The ratio of number of NeoY with high dN/dS value to that of NeoX is 1.91 times for the dataset with no SAAR, and 2.11 times for the dataset with no SAAR, and 2.21 times for dataset without SAARs in NeoX and NeoY linked proteins. We did a t test using R to examine the significance and get p=2.5×10^{-3}, p=1.02×10^{-3},

p=9×10^{-4} for these three datasets, which means that it is significant. When there is no SAAR within protein sequences, there are more proteins under positive selection on NeoY compared to NeoX. We then checked whether the presence of SAARs could influence the proportions of proteins with high and low ω values on both NeoX and NeoY from all three datasets. However, we can not find any significance using the data we have.

3.3.5 Future direction of studies for SAARs evolution on NeoSex chromosomes

In order to further illustrate the evolution of SAARs on NeoSex chromosomes, more sequenced genomes with genes mapping to NeoSex chromosomes are needed. To have a picture of the influence of sex chromosome on SAARs, we also need sex chromosome systems with different ages. *Drosophila albomicans* has an extremely young neo-sex chromosome system which was formed only about 0.12 million years ago [Chang et al., 2008, Bachtrog, 2006]. This interesting sex chromosome system is derived from the fusion of two autosomal arms, Muller-C and Muller-D. And since male flies have achiasmate meiosis, the Neo-Y cannot recombine with its homolog so it behaves like a "true" Y chromosome. Since this NeoSex system is the known youngest NeoSex system, and the NeoSex chromosomes comprise almost 40% of the genomes, these 5000 active and newly sex-linked protein-coding genes can be used to decipher the very early evolution of SAARs on sex chromosomes.

3.4 Conclusion

The genome of *Drosophila miranda* provides us excellent material to do study how SAARs could evolve under the context of a newly arised sex chromosome system. Since the formation of this sex chromosome system is recent (1-2 million years ago), protein coding genes located on neoY chromosome are still not fully degenerated [Bachtrog and Charlesworth, 2002b], we can also check how the degeneration influences the evolution SAARs on NeoY. We calculated the proportion of proteins with SAARs on NeoX and NeoY and find that the proportion of NeoY is much less than that of NeoX. Since Bachtrog [2008] found degenerated non-functional NeoY genes have a significantly higher amino acid evolution rate compared to those with functions and also, most of the SAARs are considered to be nonfunctional, we can see this phenomenon is probably caused by fast evolving of SAAR sequences caused by the newly formed Y chromosome. This phenomenon could also be limited by repetitive sequence accumulated on NeoY since it can be hard to sequence such regions. To solve this question, in the future, we can incorporate a younger NeoY with less repetitive sequences into our dataset to verify the influence of repetitive sequences on SAARs number. *Drosophila albomicans* has an extremely young neo-sex chromosome system which was formed only about 0.12 million years ago [Chang et al., 2008, Bachtrog, 2006] which is perfectly suitable for future verification but the data is not yet released. We can further check the combined forces of selection, random genetic drift and mutation on codon usage bias [Bulmer, 1991]. Firstly, we checked all RSCU values for all types of amino acids using both single amino acid repeats coding regions and non-repetitive regions. We found that many non-optimal codons of non-repetitive regions become dominant codons for some amino acids such as A, D, H, P, R, T and V in SAAR coding regions but we can not find any significant difference between sex chromosomes and autosomes. Then we tried to

study the forces forming codon usage bias within SAARs and non-repetitive regions. We cannot find differences between non-repetitive regions and SAARs which may suggest the ongoing changes of these two regions are at equillibrium. Except for the insignificance of codons changing direction of SAARs in NeoX and NeoY possibly because of small sample size, we observed there is a significant difference between the proportion of codon changes within non-repetitive regions on NeoY and NeoX. Consistent with the degeneration process of the Y chromosome which suggests deleterious mutations can preserve and accumulate on Y-linked sequences, NeoY has a higher proportion of preferred codons change to non-preferred codons compared to NeoX in non-repetitive regions. We then checked whether SAARs could influence the evolution of protein coding genes located on NeoY and NeoX. Although we cannot find significant evidence showing that SAARs could make the evolution pattern of protein coding genes on NeoX and NeoY different, we do find that proteins without SAARs have twice as much positive selection on NeoY compared to NeoX.

	NeoX-ω>1	NeoY-ω>1	NeoX-ω<1	NeoX-ω<1
Have SAAR	0	0	12	12
Don't hava SAAR	32	61	609	580

Table 3.15: dN/dS values for 653 proteins with or without SAARs that conserved through all 13 *Drosophila* species linked to NeoX and NeoY chromosomes of *Drosophila miranda*.

	NeoX-ω>1	NeoY-ω>1	NeoX-ω<1	NeoX-ω<1
Have SAAR	5	4	66	67
Don't hava SAAR	27	57	555	525

Table 3.16: dN/dS values for 653 Neo-sex chromomosome linked proteins with or without SAARs that present both on NeoX and NeoY of *Drosophila miranda*.

	NeoX-ω>1	NeoY-ω>1	NeoX-ω<1	NeoX-ω<1
Have SAAR	8	8	127	127
Don't hava SAAR	24	53	494	465

Table 3.17: dN/dS values for 653 Neo-sex chromomosome linked proteins with or without SAARs that present at least on NeoX or NeoY of *Drosophila miranda*.

Bibliography

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

M. Mar Alba and R. Guigo. Comparative analysis of amino acid repeats in rodents and humans. *Genome Research*, 14:549–554, 2004.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

H. Ashkenazy, O. Penn, A. Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*, pages 580–584, 2012.

D. Bachtrog. The speciation history of the *Drosophila* nasuta complex. *Genome Research*, 88:13–26, 2006.

D. Bachtrog and B. Charlesworth. Reduced adaptation of a non-recombining neo-Y chromosome. *Nature*, 416:323–326, 2002a.

D. Bachtrog and B. Charlesworth. Reduced adaptation of a non-recombining neo-Y chromosome. *Nature*, 416:323–326, 2002b.

D. Bachtrog, E. Hom, K. Wong, X. Maside, and P. Jong. Genomic degradation of a young Y chromosome in *Drosophila miranda*. *Genome Biology*, 9, 2008.

J. Baines and B. Harr. Reduced X-linked diversity in derived populations of house mice. *Genetics*, 175:1911–1921, 2007.

R. Bannen, C. Bingman, and G. Phillips Jr. Effect of low-complexity regions on protein structure determination. *Journal of Structural and Functional Genomics*, 8:217–226, 2008.

D. Begun. Population genomics: Whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *Plos Biology*, 5:2534–2559, 2007.

A. Bhutkar, S. Russo, T. F. Smith, and W. M. Gelbart. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics*, 17:152–161, 2011.

M. Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129:897–907, 1991.

G. Butler, M. D. Rasmussen, M. F. Lin, M. Santos, S. Sakthikumar, C. A. Munro, E. Rheinbay, M. Grabherr, A. Forche, J. L. Reedy, I. Agrafioti, and M. B. Arnaud. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, 459:657–662, 2009.

A. Carvalho and A. Clark. Y chromosome of *Drosophila pseudoobscura* is not homologous to the ancestral *Drosophila* Y. *Science*, 307:108–110, 2004.

T. Chang, T. Tsai, and H. Chang. Fusions of muller's elements during the chromosome evolution of *Drosophila albomicans*. *Zoological Science*, 47:574–584, 2008.

B. Charlesworth. The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist*, 130:113–146, 1987.

B. Charlesworth and D. Charlesworth. The degeneration of Y chromosomes. *Philosophical Transactions of the Royal Society of London*, 355:1563–1572, 2000.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25:14221423, 2009.

B. Counterman and M. Noor. Using comparative genomic data to test for fast-X evolution. *Evolution*, 58:656–660, 2004.

T. K. Darlington. Closing the circadian loop: CLOCK-induced transcription of its own inhibitors per and tim. *Science*, 280:1599–1603, 1998.

K. Dunker. The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, 9, 2008.

J. Fondon and H. Garner. Molecular origins of rapid and continuous morphological evolution. *PNAS*, 101:18058–18063, 2004.

R. Gemayel, M. Vinces, M. Legendre, and K. Verstrepen. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annual Review of Genetics*, 44:445–477, 2010.

B. Golding. Simple sequence is abundant in eukaryotic proteins. *Protein Science*, 8: 1358–1361, 1999.

P. Goodfellow. The human Y chromosome. *Journal of Medical Genetics*, 5:329–344, 1985.

J. A. Graves. Sex chromosome specialization and degeneration in mammals. *Cell*, 124:901–914, 2006.

W. Haerty and B. Golding. Genome-wide evidence for selection acting on single amino acid repeats. *Genome Research*, 20:755–760, 2010.

W. Haerty and B Golding. Increased polymorphism near low-complexity sequences across the genomes of *Plasmodium falciparum* isolates. *Genome Biology annd Evolution*, 3:539–550, 2011.

M. Huntley and A. Clark. Evolutionary analysis of amino acid repeats across the genomes of 12 *Drosophila* species. *Molecular Biology and Evolution*, 24:2598–2609, 2007.

M. Huntley and B. Golding. Evolution of simple sequence in proteins. *Journal of Molecular Biology*, 51:131–140, 2000.

M. Huntley and B. Golding. Simple sequences are rare in the protein data bank. proteins: Structure, function, and genetics. *Proteins: Structure, Function, and Bioinformatics*, 48:134–140, 2002.

M. Huntley and G. Golding. Selection and slippage creating serine homopolymers. *Molecular Biology and Evolution*, 23:2017–2025, 2006.

V. Kaiser and D. Bachtrog. *De novo* transcriptome assembly reveals sex-specific selection acting on evolving neo-sex chromosomes in *Drosophila miranda*. *BMC Genomics*, 15:241, 2013.

S. Karlina, L. Brocchieria, J. Trentb, B. Blaisdella, and J. Mrzeka. Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes. *Theor Popul Biol*, 61:367–390, 2002.

K. Katoh and K. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30: 772–780, 2013.

P. Khaitovich. Parallel patterns of evolution in the genomes and transcriptomes of Humans and Chimpanzees. *Science*, 309:1850–1854, 2005.

D. P. King. Positional cloning of the mouse circadian clock gene. *Cell*, 89:641–653, 1997.

B. Lahn and D. Page. Four evolutionary strata on the human X chromosome. *Science*, 286:964–967, 1999.

B. Lahn, N. Pearson, and K. Jegalian. The human Y chromosome, in the light of evolution. *Nature Reviews Genetics*, 2:207–216, 2001.

S. Lovell. Are non-functional, unfolded proteins (junk proteins) common in the genome? *FEBS Letters*, 554:237–239, 2003.

M. D. MacManes and M. B. Eisen. Improving transcriptome assembly through error correction of high-throughput sequence reads. *PeerJ Computer Science*, e113, 2013.

J. Mank, E. Axelsson, and H. Ellegren. Fast-X on the Z: Rapid evolution of sex-linked genes in birds. *Genome Research*, 17:618–624, 2007.

E. M. Marcotte, M. Pellegrini, H. Ng, D. W. Rice, T. O. Yeates, and D. Eisenberg. Detecting protein function and protein-protein interactions from genome sequences. *Science*, 285:751–753, 1999.

M. Marcy. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72:971–983, 1993.

Hermann Joseph Muller. *Bearings of the Drosophila Work on Systematics*. The Clarendon Press, 1940.

S. Nygaard, A. Braunstein, and D. Jeffares. Long- and short-term selective forces on malaria parasite genomes. *PLoS Genetics*, 6, 2010.

E. Pizzi and C. Frontali. Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Research*, 11:218–219, 2001.

S. Pulst, A. Nechiporuk, T. Nechiporuk, S. Gispert, XN Chen, I. Cendes, S. Pearlman, S. Starkman, G. Diaz, A. Lunkes, P. DeJong, G. A. Rouleau, G. Auburger, J. R.

Korenberg, C. Figueroa, and S Sahba. Moderate expansion of a normally biallelic trinucleotide repeat in spinocerebellar ataxia type 2. *Nature Genetics*, 14:269–276, 1996.

R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, 2008.

E. Rocha. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 11:22792286, 2004.

M. T. Ross, , J. L. Ashurst, R. S. Fulton, R. Sudbrak, G. Wen, M. C. Jones, and D. R. Bentley. The DNA sequence of the human X chromosome. *Nature*, 434: 325–337, 2005.

E. Schaper, O. Gascuel, and M. Anisimova. Deep conservation of human protein tandem repeats within the eukaryotes. *Molecular Biology and Evolution*, 31:1132–1148, 2014.

The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*, 437:69–87, 2005.

P. Sharp, T. Tuohy, and K. Mosursk. Codon usage in yeast: cluster analysis dearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, 14:5125–5143, 1986.

M. Simon and J. M. Hancock. Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins. *Genome Biology*, 10:R59.1–R59.16, 2009.

A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A.

Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. Ruby, J. Brennecke, Harvard FlyBase curators, Berkeley *Drosophila* Genome Project, E. Hodges, A. Hinrichs, A. Caspi, B. Paten, S. Park, and M. Han. Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature*, 450:219–232, 2007.

M. Steinemann, S. Steinemann, and F. Lottspeich. How Y chromosomes become genetically inert. *Genetics*, 90:5737–5741, 1993.

M. Suyama, D. Torrents, and P. Bork. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34:609–612, 2006.

E. Talevich, B. M. Invergo, P. Cock, and B. A. Chapman. Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, 13:209, 2012.

K. Thornton, D. Bachtrog, and P. Andolfatto. X chromosomes and autosomes evolve at similar rates in *Drosophila*: no evidence for faster-X protein evolution. *Genome Research*, 16:498–504, 2006.

D. Torgerson and R. Singh. Sex-linked mammalian sperm proteins evolve faster than autosomal ones. *Molecular Biology and Evolution*, 20:1705–1709, 2003.

D. G. Torgerson and R. S. Singh. Enhanced adaptive evolution of sperm-expressed genes on the mammalian X chromosome. *Heredity*, 96:39–44, 2006.

N. Trilla, K. Arat, C. Pegueroles, A. Raya, S. Luna, and M. Alb. Key role of amino acid repeat expansions in the functional diversification of duplicated transcription factors. *Molecular Biology and Evolution*, 32, 2015.

K. Usdin. The biological effects of simple tandem repeats: Lessons from the repeat expansion diseases. *Genome Research*, 18:1011–1019, 2008.

K. J. Verstrepen, A. Jansen, F. Lewitter, and G. R. Fink. Intragenic tandem repeats generate functional variability. *Nature Genetics*, 37:986–990, 2005.

B. Vicoso and B. Charlesworth. A multispecies approach for comparing sequence evolution of X-linked and autosomal sites in *Drosophila*. *Genome Research*, 90: 421–431, 2008.

B. Vicoso and B. Charlesworth. Effective population size and the faster-X effect: An extended model. *Evolution*, 63:2413–2426, 2009.

F. Walker. Huntington's disease. *The Lancet*, 369:218–228, 2007.

S. Walt, C. Colbert, and G. Varoquaux. The NumPy array: A structure for efficient numerical computation. *Computing in Science & Engineering*, 13:22–30, 2011.

X. Xu. DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. *Molecular Biology and Evolution*, 30:1720–1728, 2013.

Z. Yang. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24:1586–1592, 2007.

F. Yu, P. C. Sabeti, P. Hardenbol, Q. Fu, B. Fry, X. Lu, S. Ghose, R. Vega, A. Perez, S. Pasternak, S. M. Leal, T. D. Willis, D. L. Nelson, J. Belmont, and R. A. Gibbs. Positive selection of a pre-expansion CAG repeat of the human SCA2 gene. *PLOS genetics*, 1:404–412, 2005.

Q. Zhou and D. Bachtrog. Sex-specific adaptation drives early sex chromosome evolution in *Drosophila*. *Science*, 337:341–345, 2012.

Q. Zhou, H. Zhu, Q. Huang, L. Zhao, G. Zhang, S. W. Roy, B. Vicoso, Z. Xuan, J Ruan, Y Zhang, R. Zhao, C. Ye, X. Zhang, J. Wang, W. Wang, and D. Bachtrog.

Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*, 13:109, 2012.

Q. Zhou, C. E. Ellison, V. B. Kaiser, A. A. Alekseyenko, A. A. Gorchakov, and D. Bachtrog. The epigenome of evolving *Drosophila* neo-sex chromosomes: Dosage compensation and heterochromatin formation. *Plos Biology*, 2013.