FASTER EVOLUTION OF X-LINKED AND MALE SEX GENES IN *DROSOPHILA*

A GENOMIC COMPARISON OF *DROSOPHILA MELANOGASTER* AND *DROSOPHILA PSEUDOOBSCRUA*: EVIDENCE FOR FASTER-X, FASTER-SEX AND FASTER-MALE EVOLUTION

By

HEIDI MUSTERS, 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

MASTER OF SCIENCE (2004) McMASTER UNIVERSITY
(Biology) Hamilton, Ontario

TITLE: A Genomic Comparison of *Drosophila melanogaster* and *Drosophila pseudoobscura*: Evidence for Faster-X, Faster-sex and Faster-male Evolution

AUTHOR: Heidi Musters, B.Sc. (Graceland University)

SUPERVISOR: Professor R.S. Singh

NUMBER OF PAGES: ix; 56

ABSTRACT

A genomic comparison of *Drosophila melanogaster* and *Drosophila pseudoobscura* provides a unique opportunity to investigate factors involved in sequence divergence. The chromosomal arrangements of these species include an autosomal segment in *D. melanogaster* which is homologous to part of the X chromosome in *D. pseudoobscura*. Using orthologs to calculate sequence divergence and ratios of non-synonymous to synonymous substitutions, we found sequences on the X chromosome significantly more diverged than sequences on the autosomes. Mean divergence for sequences having sex-related functions are higher than sequences without reproductive function and even higher for sequences with male-specific reproductive functions. These estimates of divergence for sex-related sequences are most likely underestimates, as the very rapidly evolving sex-genes would tend to loose homology sooner and thus not be included in the comparison of orthologs. These results suggest that the role of sexual selection in genomic evolution is more pervasive than imagined from the interplay of female choice and secondary sexual male traits.

ACKNOWLEDGEMENTS

There are so many people to thank, that it's hard to express my gratitude to all those who have helped me along this path. I would like to begin by thanking my supervisor Dr. Rama Singh for giving me the opportunity to be a part of his lab and for his kind words and suggestions throughout my time at McMaster. I would also like to thank Dr. Richard Morton for his many suggestions, which helped to make my project more effective. Thanks also to Dr. JP Xu for his willingness to be on my committee and for his helpful suggestions as I was preparing my manuscript and thesis.

Without the help of Melanie Huntley, this project would not have been possible. Not only did she help with her bioinformatics skills, she helped by providing me with useful suggestions and above all support and friendship. I'm so grateful for all her help. Thanks also to Brett Whitty for sharing his bioinformatics skills and to Adrian Musters, my brother, who coached me through the statistical analysis. Adrian was also a great emotional support every time I needed it.

I would also like to express my appreciation to the members of the lab for their help and support. Thanks go to Dara Togerson for her wisdom and helpful suggestions, but most especially for her friendship. I'm grateful for her confidence in me, which always gave me the motivational boost I needed. I'm also grateful to Santosh Jagadeeshan for his kind remarks and smiles as well as for help in the Flyroom, learning all about Drosophila. Without Aaron Thompson and his expertise in ordering materials,

I'm sure I would have been lost. Thanks to Manhal Younes and Carla Gibson for always bringing a smile into the lab. To Wilfried Haerty and Abha Ahuja, although our time together has been short, your smiles and comments have helped more than you know.

Lastly, I would like to thank my family for their support and confidence during this time in my life. Their confidence helped me to know I could accomplish anything I set my mind to. Thanks!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

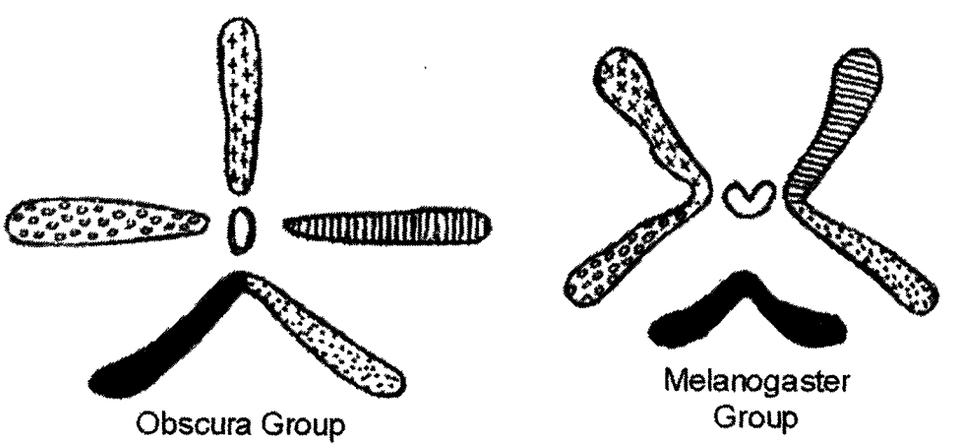
Genomic comparisons between species provide a powerful means for measuring the historical effects of selection and rates of evolution on various classes of genes. These comparisons are truly panoramic as they provide a retrospective view of the accumulated effects of evolutionary forces acting along different branches leading to the current taxa. The observed rates of evolution among genes are determined by the rates of mutation, their functional constraints and potentially their physical locations in the genome. For example, small scale genomic comparisons have suggested that genes located on the X chromosome (Thornton and Long 2002; Counterman et al. 2004) as well as genes related to sex and reproduction (Civetta and Singh 1995), particularly those with male reproductive functions (Stevison et al. 2004), are evolving at a faster rate than other genes.

Drosophila melanogaster and Drosophila pseudoobscura Homology

Drosophila melanogaster and *D. pseudoobscura* are estimated to have diverged 55 million years ago (Tamura et al. 2004) and their complete genome sequences provide a unique opportunity for comparing rates of evolution. Six chromosomal segments are homologous between these species and among these, one chromosomal segment has remained as an X chromosome in both species (see Figure 1). A second segment which is autosomal in *D. melanogaster* (the left arm of chromosome 3 denoted 3L) is homologous

Figure 1

Chromosome Homology between *D. melanogaster* and *D. pseudoobscura*. Homologous chromosome elements are indicated by similar markings. The solid black chromosome indicates the X chromosome in *D. melanogaster* and the left arm of the X chromosome in *D. pseudoobscura*. (From Dobzhansky 1970).


Obscura Group
Melanogaster
Group

to the right arm of the X chromosome (XR) in *D. pseudoobscura* (hereafter 3L-XR). The remaining 4 segments are autosomal in both species (Strutevant and Novitski 1941).

The Faster-X Theory

The Faster-X theory has been proposed and studied several times yet a clear answer to whether or not the X chromosome and the genes that reside there evolve at a faster rate has yet to be answered. It has been proposed that X-linked genes should evolve faster than the same gene if it were autosomal due to selection on hemizygous genes (Charlesworth et al. 1987). In a given population, the population of X chromosomes is smaller than that of autosomes since half of the individuals in the population will only have one X chromosome. At the sequence level, this will allow for a more rapid fixation of non-synonymous substitutions than the autosomes (Counterman et al 2004). Therefore, a greater number of mutations will be fixed on the X chromosome (assuming they are beneficial) (Charlesworth et al. 1987). Also, since a hemizygous chromosome will express the mutations which are recessive, this gives natural selection a greater opportunity to act on the hemizygous individual. However, it has also been proposed that if mutations are deleterious, and on the hemizygous chromosome, they will be more quickly eliminated from the population and will therefore show less non-synonymous mutation compared to the autosomes. Unlike the X chromosome, a mutation on the autosomes that is deleterious and recessive in a diploid species, may be maintained because it can be hidden in the genome (Charlesworth *et al.* 1987).

In *Drosophila*, the X chromosome is hemizygously expressed in males, therefore, beneficial mutations on the X chromosome can be fixed more rapidly than those on the autosomes and may allow for rapid divergence of genes on the X chromosome (Charlesworth et al. 1987). Duplicated genes in *D. melanogaster* show higher ratios of non-synonymous to synonymous substitutions (d_N/d_S) in genes located on the X chromosome compared to those observed when both copies are located on the autosomes (Thornton and Long 2002). Mammalian sperm proteins located on the X chromosome also appear to evolving faster than those on the autosomes (Torgerson and Singh 2003).

A recent study compared the divergence of homologous genes between *D. melanogaster* and *D. simulans* (3L) to *D. pseudoobscura* and *D. miranda* (XR) to test for faster-X evolution (Counterman et al. 2004). The authors compared genes on the 3L and XR chromosome to evaluate whether different ratios of d_N/d_S are observed for particular genes when they are X-linked rather than autosomal. They found that genes on the right arm of X chromosome (XR) in *D. pseudoobscura* had higher d_N/d_S values than their orthologs which are autosomal (3L) in *D. melanogaster*.

Rapid Evolution of Reproductive Proteins

Betancourt and colleagues (2002) suggested that the "faster-X" effect is further enhanced for genes with sex specific functions. In numerous studies (Coulthart and Singh 1988; Civetta and Singh 1998; Swanson et al. 2001; Torgerson et al. 2002), a select group of proteins involved in reproduction, particularly male reproduction, have been identified as having high divergence rates. The rapid evolution of these proteins may have several

explanations, including lack of functional constraint, adaptive evolution driven by sperm competition, sexual selection and sexual conflict (Wu and Davis 1993; Rice 1996; Swanson and Vacquier 2002). The latter mechanisms affecting males could account for the faster-male evolution hypothesis proposed to explain the faster evolution of hybrid male sterility in X/Y heterogametic species (Wu and Davis 1993). Using two-dimensional electrophoresis, proteins in both the testis and ovaries were found to be twice as diverged when compared to proteins in other tissues (Civetta and Singh 1995). In a study of male reproductive genes in *Drosophila*, accessory gland (Acp) expressed sequence tags (ESTs) were shown to be highly diverged (Swanson et al. 2001).

Objectives of the Study

Although many smaller scale studies have suggested a possible faster evolution of candidate genes on the X chromosome and genes involved in reproduction, comprehensive genome wide studies have not been done to see if these results hold on a long term evolutionary time scale. In this study a genomic comparison of *D. melanogaster* and *D. pseudoobscura* is used to test three hypotheses: (1) that X-linked genes evolve faster than autosomal genes (faster-X evolution), (2) that sex and reproduction related genes evolve faster than other genes (faster-sex evolution), (3) and that sex genes affecting male functions evolve faster than other genes (faster-male evolution).

MATERIALS AND METHODS

Identification of Orthologs

All available *D. melanogaster* protein sequences (a total of 18,746) were retrieved from the National Center for Biotechnology Information (NCBI) database as well as all available *D. pseudoobscura* predicted protein sequences (a total of 18,331) from the Baylor College of Medicine, Human Genome Sequencing Center (BCM-HGSC) website (http://www.hgsc.bcm.tmc.edu/projects/drosophila/). A total of 10,881 putative orthologs between the two species were identified by implementing a reciprocal smallest distance algorithm (Wall et al. 2003). This method has been shown to find many putative orthologs that are missed by the reciprocal best BLAST hit method. This is particularly important when there is the presence of a close paralog. Koski and Golding (2001) have shown that the best BLAST hit is often not the closest phylogenetic neighbor, while the reciprocal smallest distance method uses global alignment and evolutionary distance estimation to recover the nearest neighbor without requiring it to be the best BLAST hit.

Calculating Sequence Divergence

Sequences were first screened using BLAST (as part of the reciprocal smallest distance program) and only those with expect values $<10^{-3}$ were further individually aligned in ClustalW. If the alignable region was at least 80% of the alignments total length, PAML was then used to calculate a maximum likelihood estimate of the number

of amino acid substitutions per site (divergence), using an empirical amino acid substitution rate matrix. Thus for each pair of putative orthologs that detected, there was a quantitative measure of the divergence (substitutions per site) between the two sequences.

Identification of Chromosomal Location

Chromosomal location for the orthologs were determined using the known location of the *Drosophila melanogaster* genes. This assumes that orthologs have remained on the homologous chromosome throughout the divergence of *Drosophila* melanogaster and *Drosophila pseudoobscura*, yet understanding that there is a possibility that some genes between the species may have moved to a non-homologous chromosomes. This is a relatively safe assumption knowing that transpositions in the genus *Drosophila* are scarce (Gonzalez et al 2002; Ranz et al. 2003). In a study using clones to identify transposition events, only 2 out of 328 clones indicated a possible transposition event (Gonzalez et al. 2002).

Chromosome Comparison Using Sequence Divergence

A t-test was performed to detect significant differences between the means of divergence on each of the chromosomes. Each of the autosomes was compared to the X chromosome and significance was identified where it existed. Because the data did not fit a normal distribution, as it contained many outliers, an exploratory analysis was performed and the tests were re-run while excluding data beyond two standard deviations

of the mean for each category. This ensured that the results were not different than running the analysis with the complete set of data.

Calculating Ratios of Non-synonymous to Synonymous Substitutions

Using the same set of orthologs, ratios of non-synonymous to synonymous substitutions (d_N/d_S) for each protein sequence was determined. To do this, nucleotide sequences were aligned according to the protein alignment using RevTrans. Following the alignment, SNAP (Synonymous Non-synonymous Analysis Program), available at www.hiv.lanl.gov, was used to calculate the value of d_N/d_S (Korber 2000). This program, written by Bette Korber (2000), calculates the synonymous vs. non-synonymous substitutions according to the method described in Nei and Gojobori (1986) for all pairwise comparisons of sequences in an alignment. The number of each type of substitution is calculated as well as the number of potential changes. The SNAP program also incorporates a Jukes-Cantor transformation. This transformation could not be performed for sequences where more than 75% of the possible synonymous or non-synonymous sites were considered changed as these sequences had reached a level of saturation. A total of 1508 orthologous pairs, out of 10,881 pairs previously determined by the reciprocal smallest distance program, which the SNAP program indicated a saturation of substitutions. These sequences were omitted from further data analysis, as there was no value of d_N/d_S generated yielding 9373 orthologous pairs for further analysis.

Chromosome Comparisons Using Ratios of d_N/d_S

Chromosome elements were compared using ratios of d_N/d_S as was done with values of sequence divergence. The distribution of the data was tested using a parametric, one-sample Kolmogorov-Smirnov Z test which indicated the data was not normally distributed. However, due to the high numbers of sequences on each chromosome it is still appropriate to run an analysis of variance test (Lumley et al. 2002), which is only thought to be appropriate for normally distributed data. Lumley et al. (2002) have indicated that simulation studies show that 100 data points is the minimum. Using ratios of d_N/d_S, chromosomes were compared using a t-test, one-way ANOVA and both Tukey and Bonferroni post-hoc tests. Because the Tukey and Bonferroni tests yielded identical results, only the results from the Tukey test are presented.

Classification of Sequences

Orthologous sequences were classified into the following categories using a keyword search on the Flybase website (http://flybase.net/genes/) in order to make comparisons: non-reproductive, reproductive, male-reproductive, and male–specific reproductive. The reproductive sequence classification includes any gene functionally related to aspects of mating behaviour, spermatogenesis, oogenesis or fertilization or that is expressed in reproductive tissues. The terms used in the key word search to identify reproductive genes were the following: accessory gland, courtship, ejaculatory, genital, gamete, meiosis, nebenkern, oogenesis, ovary, ovariole, ovulation, oviduct, oocyte, reproduction, seminal, sperm-, testi-, testes, uterus, vagina, vulva, and vas deferens. The

reproductive class was further classified into male reproductive sequences, including any of the reproductive sequences affecting the male, regardless of function or expression in the female. Sequences were even further classified into a male-specific reproductive category if the sequence was shown to be expressed or have function in the male and not shown to have expression in the female. Non-reproductive sequences included any sequence without known reproductive function according to the Flybase website.

Since there is currently no annotation available for the *Drosophila pseudoobscura* genome, it is assumed that the genes with known function in *D. melanogaster* have the same function in *D. pseudoobscura*. Therefore, a sequence with known reproductive function in *D. melanogaster* which matches to a sequence in *D. pseudoobscura* is considered a reproductive sequence even though the function in *D. pseudoobscura* may not be known. Although the function in the two species may be different, this assumption is necessary in order to make comparisons between reproductive classes between the two species.

Sequences involved in female reproduction were not identified due to a bias in the number of known female reproductive proteins compared to the number known to be involved in male reproduction. There are approximately 50 genes currently known to be involved in female reproduction on each of the chromosome arms, compared to approximately 300 genes involved in male reproduction. Also, few female specific reproductive genes are known. The majority of the genes which are involved in female reproduction (according to the Flybase website) also have a wide variety of other functions and therefore may have a selective constraint due to the multiple functions of

the gene. This would also present a bias if comparing male and female reproductive genes. There have been several studies identifying sequences of unknown function to be involved in male reproduction. Of these studies, one has identified nearly 1600 testis EST's (Andrews et al 2000) and two others have identified just under 100 EST's expressed in accessory glands of *Drosophila* male flies (Swanson et al. 2001, Wolfner 2002). Large scale studies identifying sequences involved in female reproduction have not been performed to date, and would be necessary in order to make meaningful comparisons between groups as have been done with the male reproductive sequence comparisons.

Comparison of Reproductive Classes

Using the classifications previously described, reproductive groups were compared to the non-reproductive group within a chromosome using both d_N/d_S and divergence data the same way the entire chromosomes were compared. Student's t-tests were performed and results were verified using a one-way ANOVA and post-hoc tests; Tukey and Bonferroni.

RESULTS

Faster Evolution of X-Linked Genes

Gene divergence between protein sequences was calculated for each protein sequence match between *Drosophila melanogaster* and *Drosophila pseudoobscura*. Ratios of non-synonymous to synonymous substitutions (d_N/d_S) were also then calculated to verify the results obtained from the divergence data. For each chromosome arm, the distributions of d_N/d_S for the sequences are shown in Figure 2. The data is clearly skewed towards zero and is not normally distributed, however, it is still appropriate to use tests which assume normality such as a t-test and ANOVA (Lumley et al. 2002). Lumley et al. (2002) have shown that even if a large data set does not fit a normal distribution, accurate results can still be obtained using statistical tests that have been thought to only be appropriate when comparing normally distributed data.

Table 1 presents the mean values of d_N/d_S and divergence for each of the chromosome elements being compared. A quick glance at the values of d_N/d_S for each chromosome makes it clear that the X-XL chromosome element has higher values of both divergence and d_N/d_S than any of the other chromosomes. Using a t-test to compare each of the chromosome elements to X-XL, it was determined that X-XL had significantly ($p<0.001$) higher values of both mean d_N/d_S and mean divergence (0.1305 and 0.4708 respectively) than any of the other chromosome elements.

Figure 2

Distribution of ratios of d_N/d_S for sequences on each of the chromosome elements. All distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. The chromosome the sequences are located on is indicated to the left of the distributions.

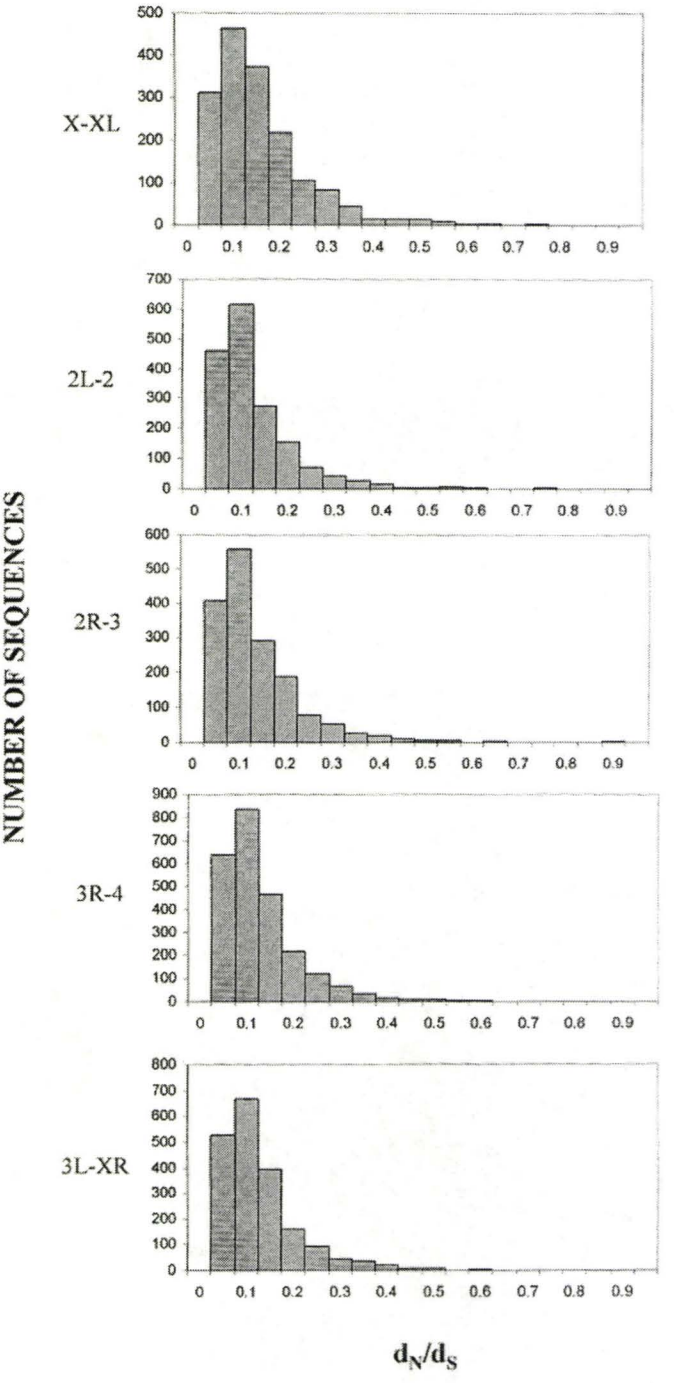

X-XL
2L-2
2R-3
3R-4
3L-XR
NUMBER OF SEQUENCES
d_N/d_S

Table 1

Mean d_N/d_S and Mean divergence (substitutions per site) of all identified protein sequence matches on the chromosomes of *D. melanogaster* and *D. pseudoobscura* and their comparisons between chromosomes. Significantly lower ratios of d_N/d_S from the X-XL chromosome, determined using a Student's t-test, are indicated by an asterix (* $p< 0.05$, ** $p< 0.01$, *** $p< 0.001$). The number of genes in each group are shown in parentheses. Tukey post-hoc test indicates chromosome X-XL is significantly different from all other chromosome elements ($p < 0.001$).

NOTE: *D.m. – D. melanogaster, D.p.- D. pseudoobscura.*

Chromosome *(D.m.)*	Chromosome *(D.p.)*	Mean d_N/d_S	Mean divergence (substitutions per site)
X	XL	0.1305 (1659) N/A	.4708 (1848) N/A
2L	2	0.1020 (1679) ***	.4124 (2001) **
2R	3	0.1112 (1642) ***	.3975 (1968) ***
3R	4	0.1050 (2430) ***	.3775 (2798) ***
3L	XR	0.1053 (1963) ***	.3726 (2266) ***

Ensuring that accurate results were obtained using a t-test, additional tests were used to confirm the findings. A one-way ANOVA tested the chromosome elements together and indicated that there was a difference in the group ($p<0.001$). Post-hoc tests, Tukey and Bonferroni, were used to determine the nature of the differences. Both tests yielded identical results to the t-tests, indicating that the X-XL chromosome is significantly more diverged than the autosomal elements ($p <0.001$).

The mean d_S was not significantly different on chromosome X-XL compared to any of the other chromosomal segments indicating that the increase of non-synonymous substitutions, are causing the increase in the ratio of d_N/d_S.

Similar to the X-XL chromosome, the 3L-XR chromosome element was tested to see if differences existed between this chromosome and the other autosomal elements. Contrary to an initial hypothesis, a Tukey test showed no differences between this chromosome and those that are autosomal in both species (see Table 2). The 3L-XR chromosome was different from the X-XL chromosome as indicated previously, but the mean values of d_N/d_S and divergence were remarkably similar to those values on the other autosomal elements (see Table 1).

Faster Evolution of Sex-Related Genes

Table 3 presents the values of divergence (substitutions per site) for reproductive protein and non-reproductive protein sequences compared within each of the homologous chromosomes. In every comparison sequences involved in reproduction show higher

Table 2

Post hoc test: Tukey, testing if chromosome 3L-XR is different from other autosomes elements. In every comparison, the p value indicates 3L-XR is not significantly ($\alpha = 0.05$) different from any of the other autosomal elements (2L-2, 2R-3 and 3R-4) as the mean values of each of the autosomes are nearly identical. 3L-XR is clearly different from X-XL.

Chromosome I x Chromosome J			Mean difference | I-J |	SE	p
3L-XR	x	X-XL	0.02545	0.0029	< 0.001
		2L-2	0.00297	0.0029	0.848
		2R-3	0.00622	0.0028	0.214
		3R-4	0.00031	0.0029	1.000

Table 3

Mean divergence (substitutions per site) of protein sequences with reproductive vs. non-reproductive function within each chromosome arm. Significantly different values from the non-reproductive class are indicated by an asterix (* $p< 0.05$, **$p< 0.01$, *** $p< 0.001$) and determined using a t-test. The number of genes are shown in parentheses.

Functions	X-XL	2L-2	2R-3	3R-4	3L-XR
Non-reproductive	0.4645 (1644)	0.3843 (1718)	0.3956 (1713)	0.3623 (2446)	0.3655 (1988)
Reproductive	0.5216 (204)	0.5834 (283) ***	0.4101 (255)	0.4835 (352) ***	0.4234 (278) *
Male reproductive	0.6141 (148) **	0.6294 (228) ***	0.4616 (200)	0.5265 (285) ***	0.4572 (230) **
Male-specific reproductive	0.7055 (117) ***	0.6699 (207) ***	0.4805 (177)*	0.5480 (249) ***	0.4637 (213) **

divergence than those without reproductive function (see Figure 3) and in almost every comparison, the means are significantly different. Interestingly, when the data is limited to the male reproductive sequences, the results were even more pronounced than when the more broad reproductive classification was used for the comparison. This result was even more apparent when the male-specific reproductive proteins were used for comparison to the non-reproductive sequences, showing high significance for every comparison (see Table 3, Figure 3). In order to verify the results of the t-tests, an exploratory analysis was done excluding all data beyond two standard deviations, since many outliers were present in the data set. Results indicated that outliers were not significantly altering the data.

Using values of d_N/d_S for each of the sequences, the same comparisons were made within chromosomes as was done with the values of divergence. On each chromosome, the distribution of the values of d_N/d_S for each sequence category (see figures 4-8) shows the data is skewed towards zero in every case, and was not normally distributed according to a Kolmogorov-Smirnov test. However, as was the case with the distribution of all sequences on each chromosome, it is still appropriate to use further statistical tests, which assume normality of data. The proportion of synonymous substitutions for reproductive groups was not higher than the non-reproductive classification indicating that increased ratios of d_N/d_S. are caused by an increased proportions of non-synonymous substitutions rather than an increase in synonymous mutation in these genes.

Figure 3

Divergence of male-specific reproductive, male reproductive, reproductive, and non-reproductive protein sequences on each chromosome. The histogram variables for each chromosomes from left to right are: male-specific reproductive, male-reproductive, reproductive and non-reproductive. On each chromosome, protein sequences involved in reproduction have divergence values higher than non-reproductive sequences. Data is significant on all chromosomes when male-specific is compared to non-reproductive and significant for almost all other comparisons. Error bars indicate standard error. Data from table 3.

NOTE: *D.m. – D. melanogaster, D.p.- D. pseudoobscura*

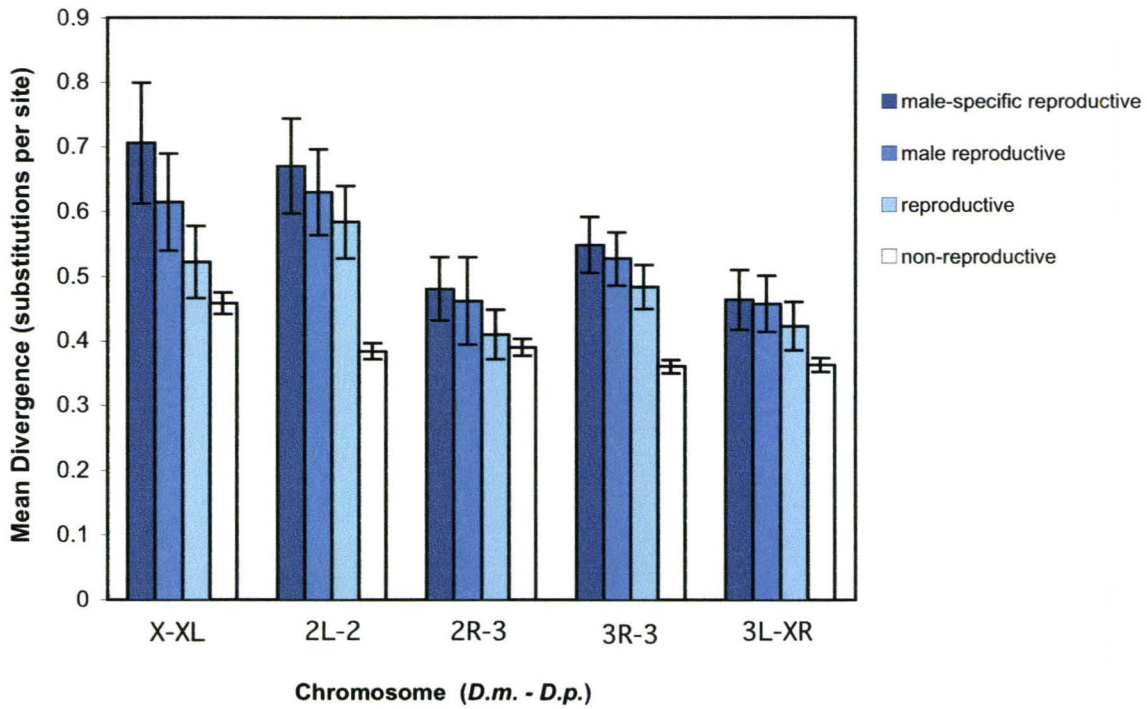

0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Mean Divergence (substitutions per site)
X-XL
2L-2
2R-3
3R-3
3L-XR
Chromosome (D.m. - D.p.)
male-specific reproductive
male reproductive
reproductive
non-reproductive

Figure 4

Distribution of ratios of d_N/d_S for sequences in each of the reproductive classes on chromosome X-XL. Distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. Sequence classification is indicated on the right of the figure.

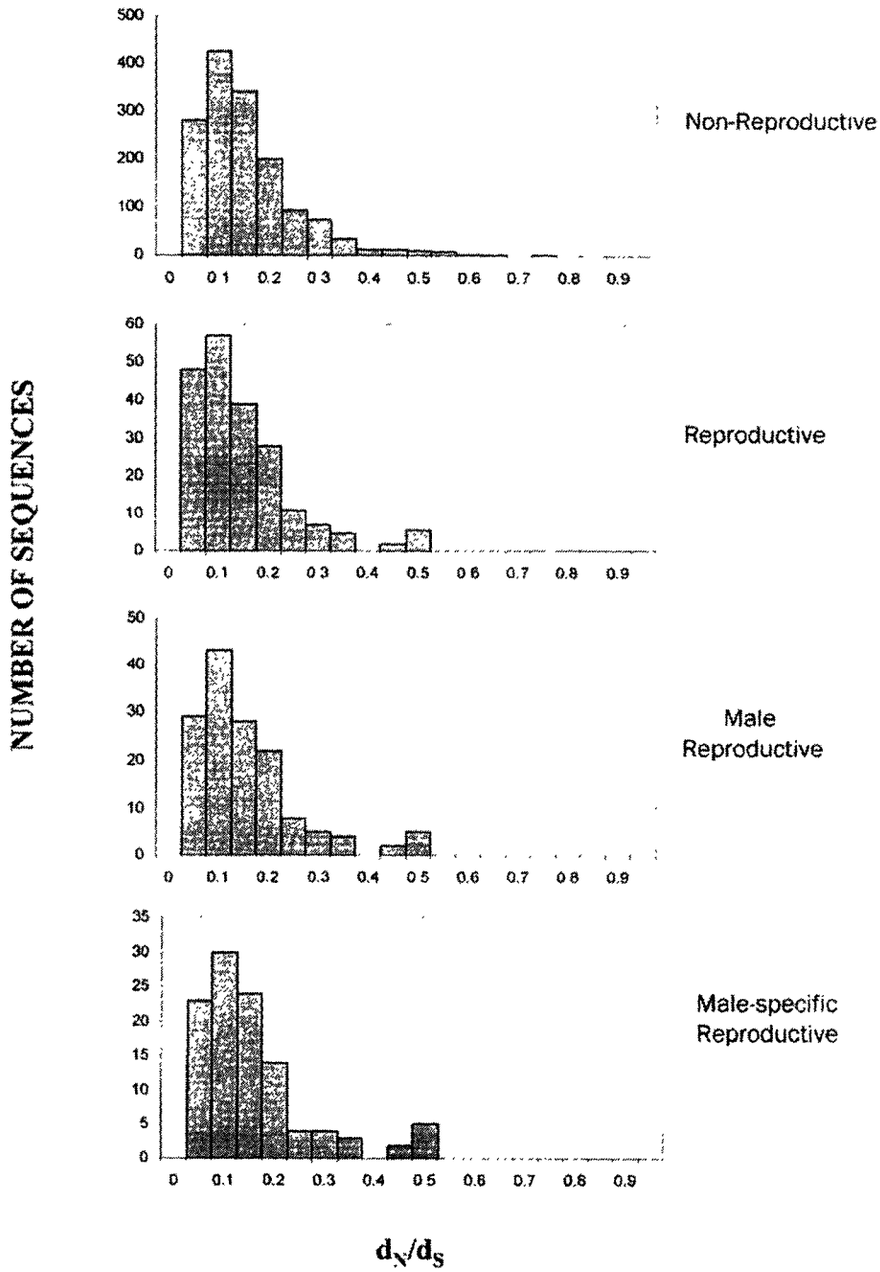

NUMBER OF SEQUENCES
Non-Reproductive
Reproductive
Male
Reproductive
Male-specific
Reproductive
d_N/d_S

Figure 5

Distribution of ratios of d_N/d_S for sequences in each of the reproductive classes on chromosome 2L-2. Distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. Sequence classification is indicated on the right of the figure.

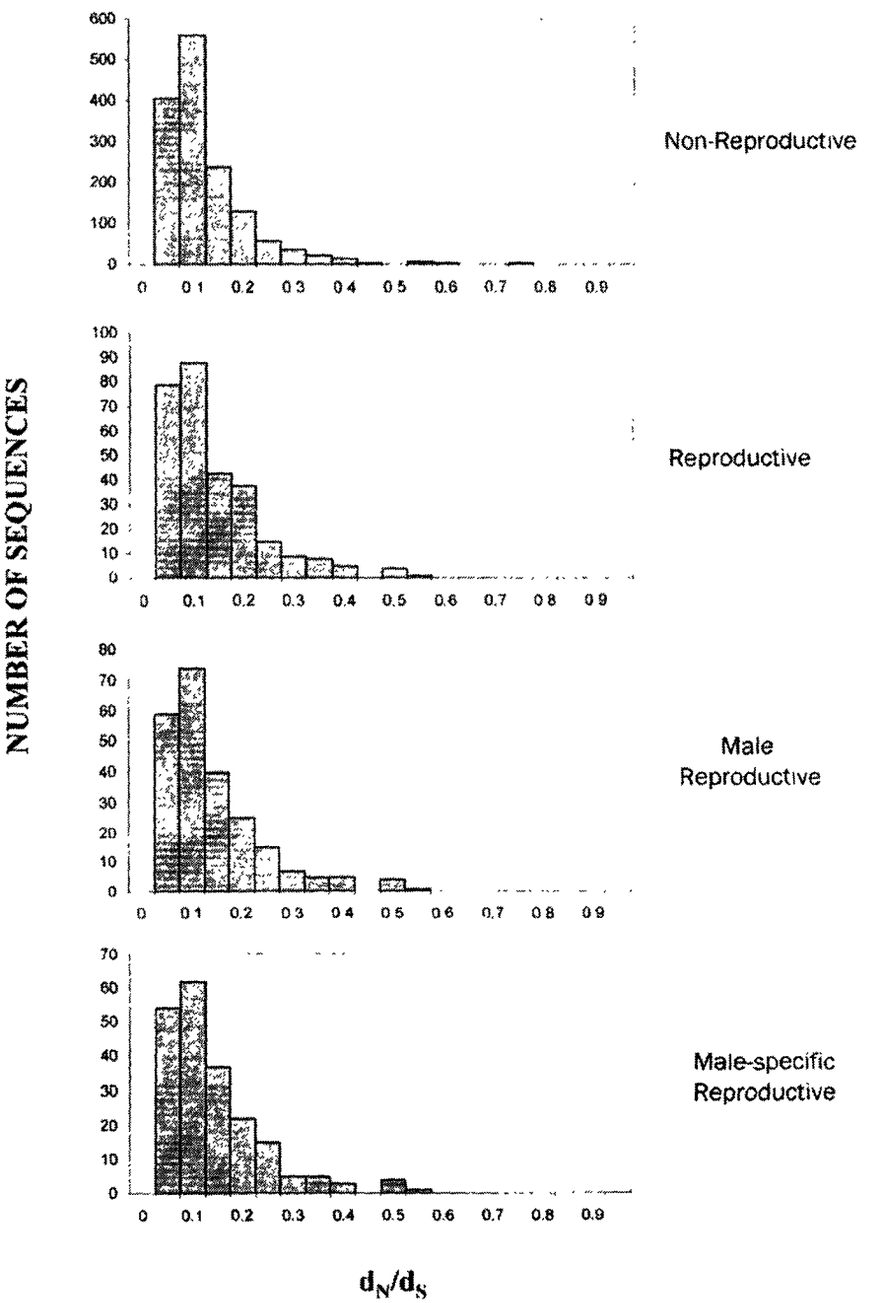

NUMBER OF SEQUENCES
Non-Reproductive
Reproductive
Male
Reproductive
Male-specific
Reproductive
d_N/d_S

Figure 6

Distribution of ratios of d_N/d_S for sequences in each of the reproductive classes on chromosome 2R-3. Distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. Sequence classification is indicated on the right of the figure.

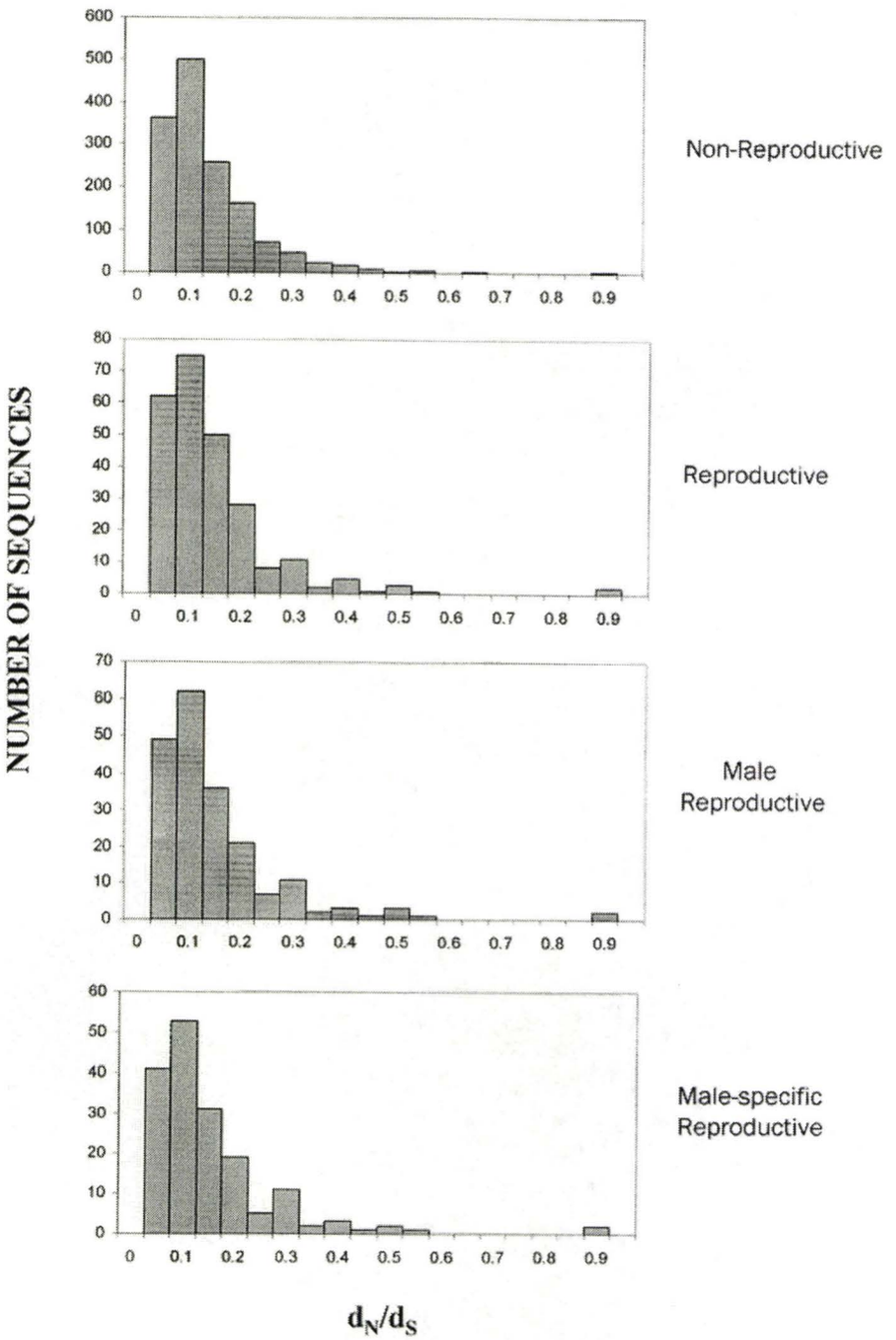

NUMBER OF SEQUENCES
Non-Reproductive
Reproductive
Male
Reproductive
Male-specific
Reproductive
d_N/d_S

Figure 7

Distribution of ratios of d_N/d_S for sequences in each of the reproductive classes on chromosome 3R-4. Distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. Sequence classification is indicated on the right of the figure.

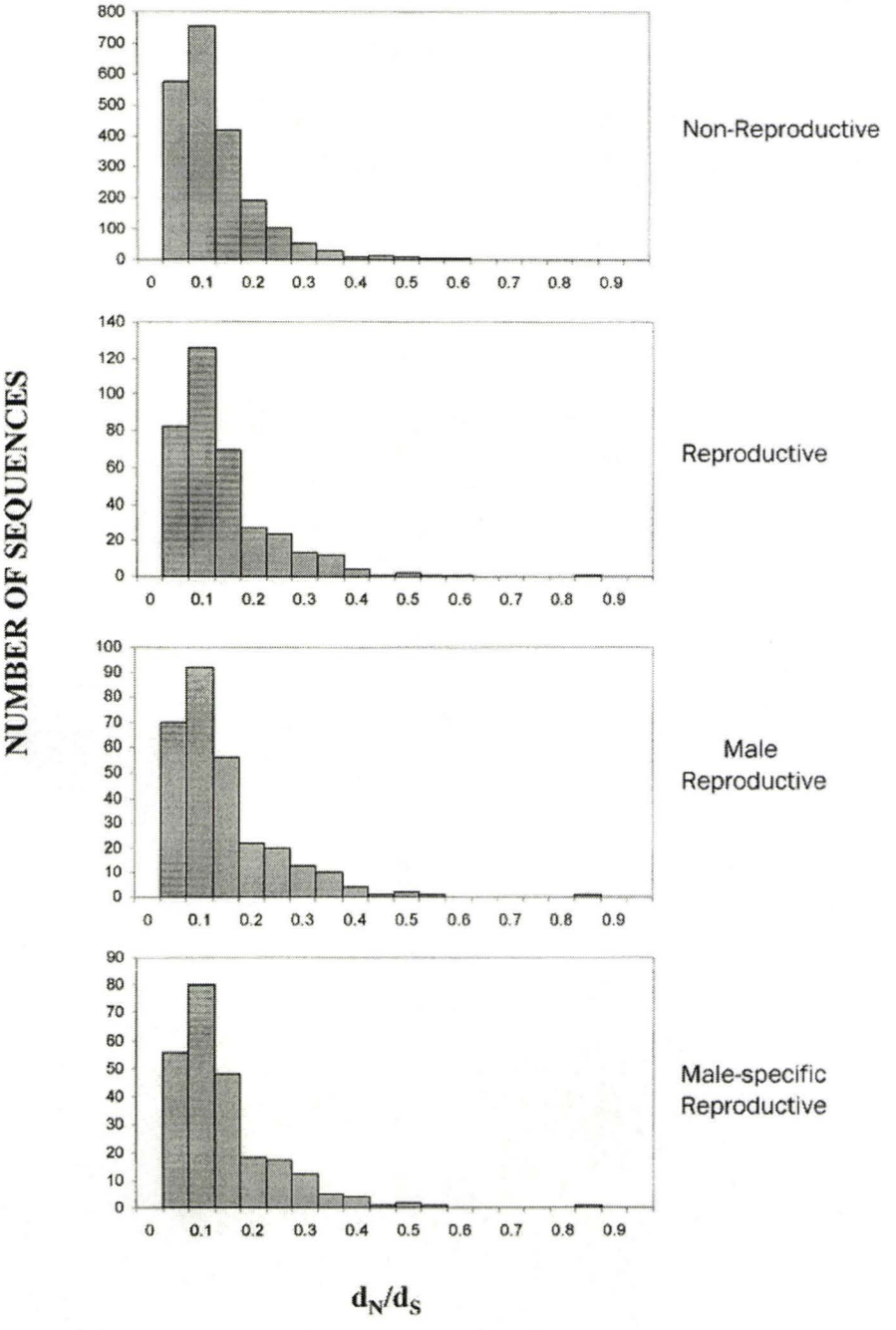

NUMBER OF SEQUENCES
Non-Reproductive
Reproductive
Male Reproductive
Male-specific Reproductive
d_N/d_S

Figure 8

Distribution of ratios of d_N/d_S for sequences in each of the reproductive classes on chromosome 3L-XR Distributions are not normally distributed according to a Kolmogorov-Smirnov Z test. Sequence classification is indicated on the right of the figure.

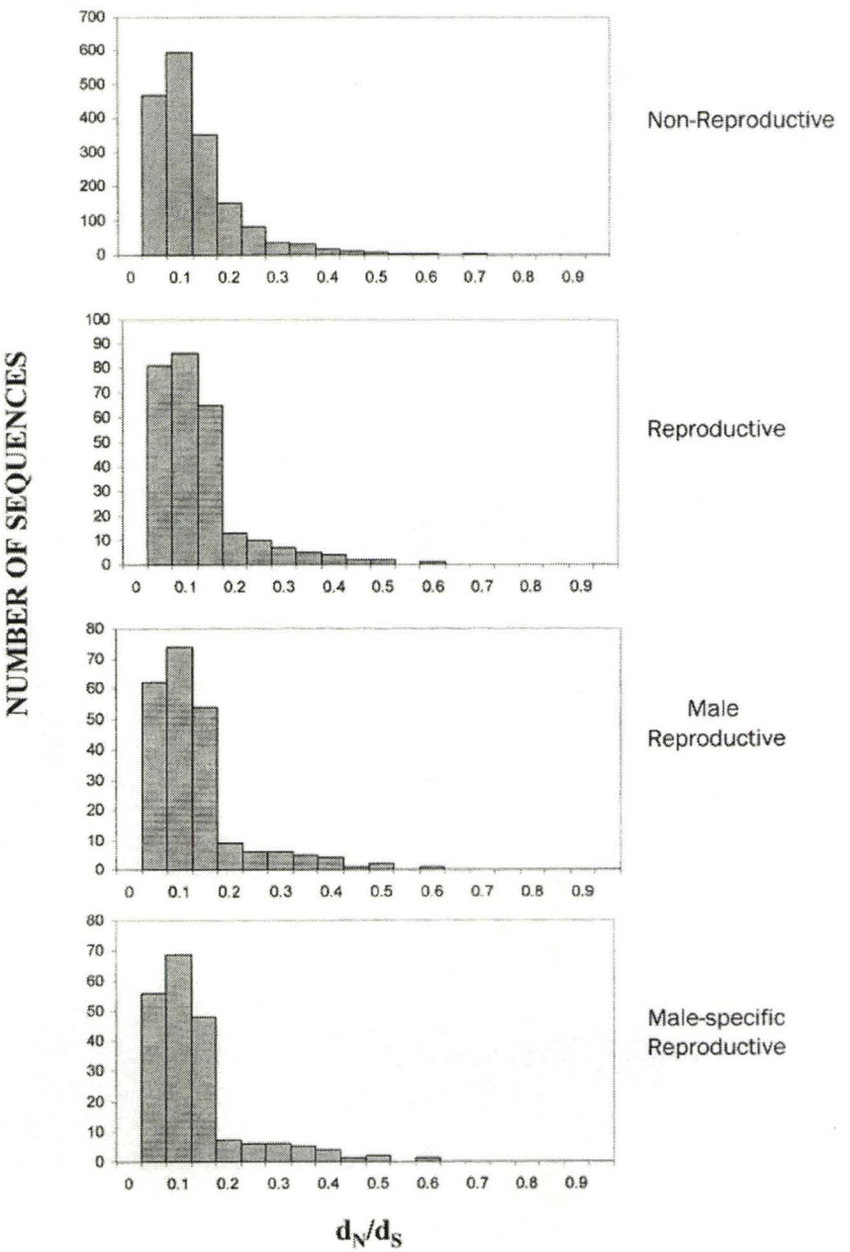

NUMBER OF SEQUENCES
Non-Reproductive
Reproductive
Male
Reproductive
Male-specific
Reproductive
d_N/d_S

Table 4

Mean ratios of d_N/d_S for protein sequences with reproductive vs. non-reproductive function within each chromosome arm. Significantly different values from the non-reproductive class, determined using a t-test, are indicated by an asterix (* $p< 0.05$, ** $p< 0.01$). The number of genes are shown in parentheses. Significant differences are only present on chromosomes 2L-2, 2R-3, and 3R-4.

Functions	X-XL	2L-2	2R-3	3R-4	3L-XR
Non-reproductive	0.1319 (1507)	0.0998 (1482)	0.1110 (1465)	0.1032 (2163)	0.1045 (1752)
Reproductive	0.1228 (203)	0.1142 (290) **	0.1204 (248)	0.1178 (364) **	0.1030 (276)
Male reproductive	0.1303 (146)	0.1169 (235) **	0.1250 (198)	0.1202 (292) **	0.1047 (224)
Male-specific reproductive	0.1337 (109)	0.1166 (208) **	0.1284 (171)*	0.1203 (245) **	0.1070 (205)

Ratios of d_N/d_S for sequence classes on each of the chromosomes were first compared using a Student's t-test (see Table 4) followed by a one-way ANOVA (1-ANOVA) and then a Tukey test to identify differences between the groups (see Table 5).

Sex-related evolution on chromosome X-XL

Using a t-test to compare mean ratios of d_N/d_S for each of the reproductive groups to the non-reproductive class (see table 4), it was determined that none of the groups were significantly different from one another. Using a one-way ANOVA (1-ANOVA) test to verify the results of the t-test, a value of $p = 0.752$ was returned indicating no differences between these groups. Even though differences in the groups were not indicated from 1-ANOVA, a Tukey post-hoc test was performed to see the individual differences between groups (table 5). As expected, the mean difference was so small between the different groups that any difference present was not even close to significant. In Figure 9 the mean d_N/d_S for each of the sequence classifications on this chromosome are shown. Although it appears that mean value for the reproductive class are much smaller than the other classes, the spread of values along the y-axis covers such a limited range, that the values do not statistically differ at all.

Table 5

One-way ANOVA (1-ANOVA) and Tukey test results indicating the presence of differences between groups on each chromosome and the nature of those differences. The significance to which differences exist within a chromosome is indicated under the 1-ANOVA. The value of significance each is from the non-reproductive class is indicated under the Tukey test, and shows the nature of the differences observed (if any) from the 1-ANOVA test. P values indicating significance in the Tukey test indicate that the reproductive class is significantly different from the non-reproductive class on that chromosome. Chromosomes X-XL and 3L-XR show no differences between the different classifications.

Chromosome (*D.m –D.p.*)	1-ANOVA p	Tukey p		
		Reproductive	Male-Reproductive	Male-specific Reproductive
X-XL	0.752	0.711	0.998	0.999
2L-2	0.002 *	0.066	0.038 *	0.062
2R-3	0.055 *	0.546	0.276	0.157
3R-4	0.001 *	0.025 *	0.015 *	0.028 *
3L-XR	0.970	0.994	1.000	0.980

Sex-related evolution on chromosome 2L-2

Like the X-XL chromosome, sequence classifications on chromosome2L-2 were compared using a t-test. Results indicate that all of the reproductive classification groups had significantly higher values of d_N/d_S than the non-reproductive category they were compared to (see Table 4). Using a 1-ANOVA test, a value of $p = 0.002$ was returned indicating significant differences between the groups being compared. A Tukeyt post-hoc test indicated that the male reproductive group of sequences is significantly different than the non-reproductive group and the reproductive and male-specific reproductive groups are just slightly past being significant (see Table 5). Figure 9 clearly shows the trend of higher d_N/d_S values as you go from non-reproductive to reproductive to male reproductive on this chromosome.

Sex-related evolution on chromosome 2R-3

On chromosome 2R-3, it was determined using a t-test that only the male-specific reproductive class of sequences was significantly different from the non-reproductive class. All other reproductive groups have higher mean ratios of d_N/d_S, however these values are not significantly different. A value of $p = 0.055$ was returned from a one-way ANOVA test indicating differences between the groups is nearly significant, but not quite. Further analysis was done using a Tukey post-hoc test (table 5). Even though there seemed to be a difference in the groups, no two groups compared were significantly different from one another according to a Tukey test. However, when looking at Figure 9,

in the comparison of groups on this chromosome, there is a trend that the ratios of d_N/d_S increase as you go from non-reproductive to reproductive to male-reproductive.

Sex-related evolution on chromosome 3R-4

Mean ratios of d_N/d_S for all reproductive classes were determined to be significantly different from the non-reproductive class according to a student's t-test ($p<0.001$) on the 3R-4 chromosome element. When testing these using a one-way ANOVA test a value of $p< 0.001$ was given. This indicates there are significant differences between the groups. A Tukey test yielded identical results to the student's t-test, indicating significant difference between the non-reproductive class and all of the reproductive classes on this chromosome (see Table 5). Figure 9 also shows the same trend for this chromosome element that was seen on chromosomes 2L-2 and 2R-3, which shows that sequences are more diverged when they belong to a reproductive class then when they have no involvement with reproduction.

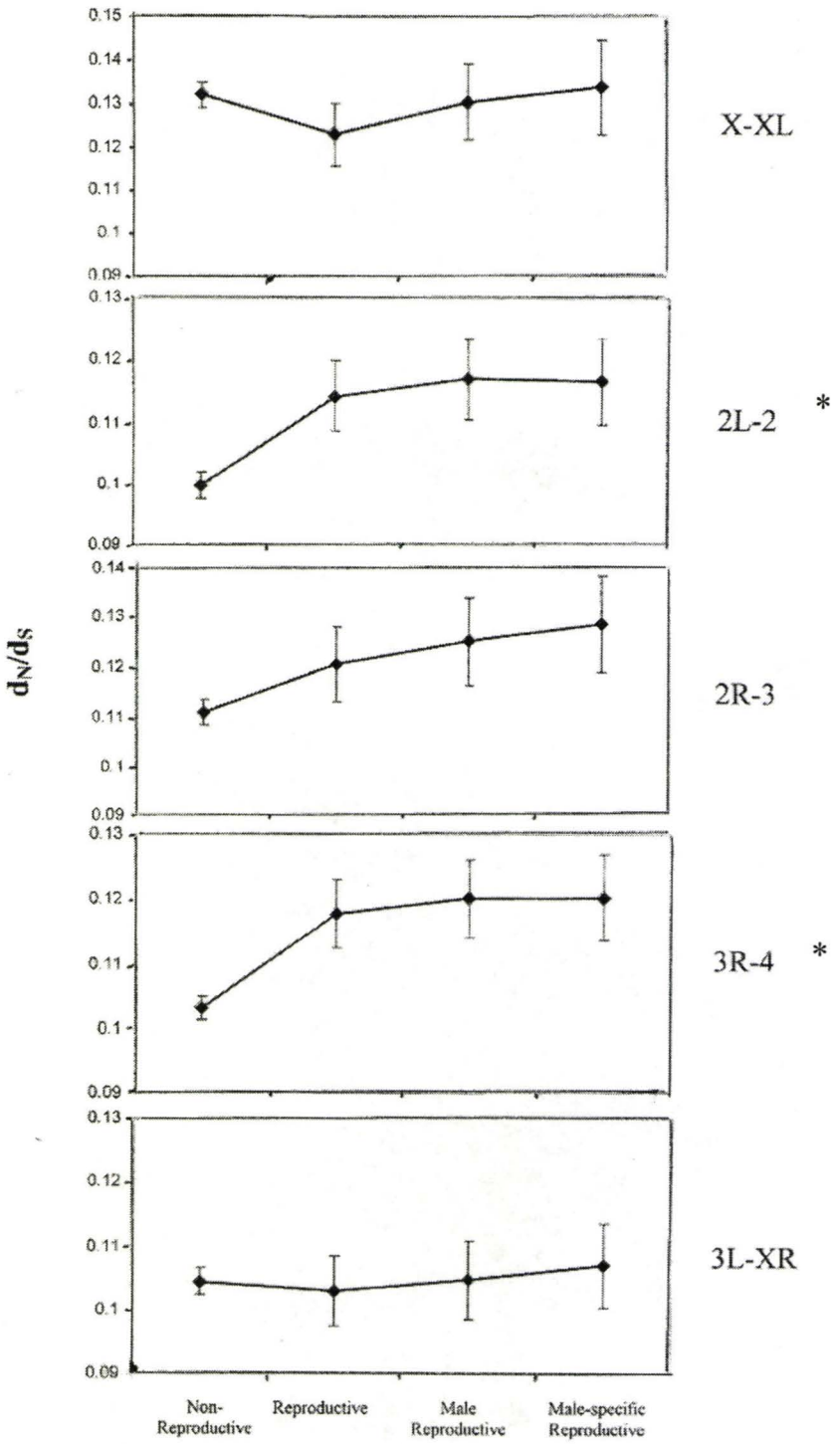
Sex-related evolution on chromosome 3L-XR

On the 3L-XR chromosome, t-test's compared each of the reproductive groups to the non-reproductive class and no two groups were found to be significantly different from one another. One-way ANOVA test yielded a value of $p = 0.970$ indicating no differences between the groups being compared. Even though differences in the groups were not indicated from the ANOVA, a Tukey test was performed to see the individual

differences between groups (see Table 5). As expected, the mean difference was so small between the different groups that any difference present was not even close to significance. In Figure 9, it appears that mean ratios of d_N/d_S for the male-specific reproductive class still follows the trend of higher values, however, the range on the y-axis is so small that these values could be considered nearly the same number and therefore do not statistically differ.

Figure 9

Mean d_N/d_S for each reproductive classification on the chromosome elements, X-XL, 2L-2, 2R-3, 3R-4 and 3L-XR. Differences between groups in the chromosomes were determined significant or non-significant using a One-way ANOVA and a Tukey Post Hoc test at $p < 0.05$. Chromosomes showing significant differences are indicated with and asterix. Error bars indicate Standard Error. Data from Table 1.


d_N/d_S
X-XL
2L-2 *
2R-3
3R-4 *
3L-XR
Non-Reproductive
Reproductive
Male Reproductive
Male-specific Reproductive

DISCUSSION

Faster-X Evolution

The results suggest that genes on the X-XL chromosome are indeed evolving at a faster rate compared to those on the autosomes. This is indicated from chromosome comparisons using values for both mean divergence and d_N/d_S. For both of these analyses, results were extremely significant ($p < 0.001$) indicating more rapid evolution of genes on the X-XL chromosome. Although, there is conflicting evidence in the literature, the results are consistent with the hypothesis of faster evolution of the X chromosome (Charlesworth et al. 1987; Thornton and Long 2002). Betancourt *et al.* (2002) found no evidence for faster X evolution in a data set of male-specific genes although they also indicate that this could be because they eliminated many of the rapidly evolving genes (Accessory Gland proteins). They also mentioned that since their data set was small, and only included coding sequences, they may have missed some relevant loci. However, duplicated genes on the X chromosome were shown to be more diverged compared to gene copies on the autosomes (Thornton and Long 2002). The present results are also consistent with a mammalian study, which found sperm genes on the X chromosome to be evolving faster than sperm genes on the autosomes (Torgerson and Singh 2003).

Faster-X evolution makes sense since mutations on the X chromosome will have more opportunities to be expressed compared to the autosomes due to the hemizygous

expression of the this chromosome. In any given population of *Drosophila*, the population of the X chromosomes is smaller than that of the autosomes since males only have one X chromosome. Therefore, beneficial and neutral mutations will be more quickly incorporated into the population if they reside on this chromosome (Counterman et al 2004). If this is the case, genes on the X chromosome will show more sequence divergence than genes residing on the autosomes.

Evolution of Orthologs Located on X chromosome vs. Autosome

Chromosome 3L in *D. melanogaster* is the homologous chromosome segment to the XR segment in *D. pseudoobscura* (3L-XR). In the comparison of these species, one might expect genes on the 3L-XR chromosome to have divergence values somewhere in between those on the X-XL and those on the autosomes since it is a sex chromosome in one species and an autosome in the other. A recent study compared genes on 3L in D. melanogaster and D. simulans to XR in D. pseudoobscura and D. miranda and found genes on XR to be more diverged than those on 3L (Counterman et al. 2004). The data presented here does not support this trend, as 3L-XR was not significantly different from any of the autosomes (see Table 2) and in fact appeared to be the least diverged of all of the chromosomes, having a mean divergence value of 0.3726 substitutions per site (see Table 1). Assuming that genes on the X chromosome evolve at a faster rate than autosomal genes due to hemizygous expression in males, one can expect that chromosome 3L-XR would have mean divergence values similar to the X chromosome.

The right arm of the X chromosome in *D. pseudoobscura* should act the same as the left arm and this was seen in the results of Counterman *et al.* (2004). However, these authors compared ratios of d_N/d_S for genes between the *Drosophila* sibling species pair *melanogaster/simulans* vs. *pseudoobscura/miranda* (3L vs. XR). The left arm of chromosome 3 in *D. melanogaster* should be similar to the other autosomes in *D. melanogaster.* Therefore, when comparing these two more distantly related species, the divergence of this chromosome (3L-XR) should fall somewhere in between the X chromosome and the autosomes. Despite this, the results presented here indicate that the 3L-XR chromosome arm does not have a higher divergence values as seen with the X-XL chromosome.

Although Counterman *et al.* (2004) found significantly faster rates of sequence evolution in XR in *D. Pseudoobscura/D. miranda* than its ortholog (3L) in *D. melanogaster/D. simulans,* this result was not seen in the present *D. melanogaster/D. pseudoobscura* genomic comparison. The differences between the two studies that could be responsible for this are: a longer evolutionary time and a larger and presumably more random selection of genes in the present study compared to the Counterman *et al.* (2004) study. It is possible that the longer evolutionary time makes all chromosomes show significant differences although this was not true of XL chromosome (see Table 1). Another factor could be that comparisons of divergence between closely related species, as done in the Counterman et al. (2004) study, are more likely to find differences if sex genes evolve rapidly during the early stages of speciation- a hypothesis that has been

considered before (Civetta and Singh 1998) and is currently being tested (Jagadeeshan and Singh, unpublished).

Faster-Sex / Faster-Male Evolution

Using a measure of divergence, protein sequences involved in reproduction were found to have much higher values than proteins without reproductive function (see Table 3, Figure 3). This result is consistent with various studies identifying groups of reproductive genes as rapidly evolving (Coulthart and Singh 1988; Civetta and Singh 1998; Swanson et al 2001). When comparing ratios of d_N/d_S for each sequence, the same clear result, as was seen with values of divergence, was not as apparent. However, when looking closely at the data, although the numbers are not all significant, they still present the same trend that sequences involved in reproduction have higher ratios of d_N/d_S (see Figure 9).

Since *D. melanogaster* and *D. pseudoobscura* have been diverged for approximately 55 million years (Tamura et al. 2004), many of the highly diverged genes between the two species can no longer be identified as orthologs. When initially searching for orthologs, the reciprocal smallest distance program could not find matches for all of the protein sequences. Only 55% of the genes were identified as orthologs between the species. This is either because the genes were too far diverged between the species, or the genes are lost in one of the species. Also, once orthologs were identified, and values of d_N/d_S were calculated using SNAP, many of the most diverged genes were

not able to give a value of d_N/d_S due to a saturation of substitutions (see Methods and Materials). Because no value was given, these sequences were no longer used. Therefore, many of the most highly diverged sequences between these species were not used in the comparisons giving an underestimate of the differences seen between the groups.

Of the 45% of genes where no ortholog between the species could be determined, 16% of those genes were reproductive genes, compared to the 14% of reproductive genes present in the full 100%. This is a first indication that more reproductive genes were lost than other types of genes. As a test, a list of known highly diverged genes *in Drosophila melanogaster*, identified from various studies, were searched within the data set to see if they were among the genes with an ortholog match in *D. pseudoobscura*. The majority of the rapidly evolving genes identified did not have a homologous sequence match in *D. pseudoobscura* (see Table 6) identified in the present study. This supports the conclusion that the results are an underestimate of the differences because many of the rapidly evolving genes listed, having no ortholog, are involved in reproduction.

The results indicating faster-sex evolution are even more evident when you look specifically at proteins involved in male reproduction. This too is consistent with previous studies, which have shown accessory gland proteins to be one of the most rapidly evolving classes of genes in *Drosophila* (Stevison et al. 2004; Swanson et al. 2001; Tsaur et al. 2001). Although this trend was not as apparent with values of d_N/d_S, as mentioned previously, if values of d_N/d_S for the most highly diverged genes were available, I suspect the same trend would be seen.

Table 6

Highly Diverged Genes. 20 Highly diverged genes are indicated in the table from various studies. Only 4 of the 20 had an ortholog identified in *Drosophila pseudoobscura* using the method in this study. Genes in which orthologs were identified all had higher than average values of divergence. This shows that many of the known rapidly evolving genes were missed in the present study.

Note: *D.p. – Drosophila pseudoobscura*

Gene Name	Ortholog in *D.p* ?	Classification	Study identifying rapid evolution
Andropin	no	Male-specific reproductive	Civetta and Singh 1998
Transformer	no	Male reproductive	Civetta and Singh 1998; McAllister and McVean 2000
Exuperantia	yes	Male reproductive	Civetta and Singh 1998
Period	yes	Male reproductive	Civetta and Singh 1998
Janus A	no	Reproductive	Parsch et al. 2001
Janus B	no	Male-specific reproductive	Parsch et al. 2001
ocnus	no	Male specific reproductive	Parsch et al. 2001
Hybrid male rescue	no	Non-reproductive	Barbash et al. 2003
Acp26Ab	no	Male-specific reproductive	Aguade 1999
Acp26Aa	no	Male-specific reproductive	Civetta and Singh 1998; Tsaur and Wu 1997; Begun et al. 2000
Acp36DE	no	Male-specific reproductive	Tsaur and Wu 1997; Begun et al. 2000
Acp29AB	no	Male-specific reproductive	Begun et al. 2000
Acp32CD	yes	Male-specific reproductive	Begun et al. 2000
Acp33A	no	Male-specific reproductive	Begun et al. 2000
Acp53Ea	yes	Male-specific reproductive	Begun et al. 2000
Acp62F	no	Male-specific reproductive	Begun et al. 2000
Acp63F	no	Male-specific reproductive	Begun et al. 2000
Acp76A	no	Male-specific reproductive	Begun et al. 2000
Acp95EF	no	Male-specific reproductive	Begun et al. 2000
Acp98AB	no	Male-specific reproductive	Begun et al. 2000

Several theories have been proposed to explain the faster evolution of reproductive proteins however many of them are still in the process of being tested (Swanson and Vacquier 2002). Hemizygously expressed genes have a selective advantage as changes on the X chromosome are immediately exposed to sexual selection (Charlesworth et al. 1987). allowing for beneficial mutations to be immediately expressed and quickly fixed in the population. Sexual selection includes processes such as female choice, sperm competition, sexual conflict and co-evolution.

Female choice can drive rapid evolution of reproductive genes as females are competing for males with ideal characteristics. Sperm competition, sexual conflict and co-evolution drive adaptive evolution of reproductive proteins, as these proteins must change in order to maintain function with other rapidly evolving proteins in which they interact. For example, if egg membrane proteins are rapidly evolving to prevent polyspermy, sperm, in order to continue to be able to fertilize, must co-evolve and therefore evolve rapidly (Swanson and Vacquier 2002). Other theories proposed include relaxed constraint, reinforcement and gene duplication (Swanson and Vacquier 2002). Although many theories have been proposed, adaptive evolution through sexual selection (Torgerson and Singh 2002; Rice 1996; Swanson and Vacquier 2002) is the most likely explanation for the faster evolution of reproductive proteins in the male.

CONCLUSION

The results indicate that the mean divergence of sequences on the X chromosome are significantly higher than those found on the autosomes. Although the 3L-XR chromosome arm shows no evidence of faster-X evolution, our results clearly indicate faster evolution on the X-XL chromosome compared to the autosomes. The results also show, for every chromosome, that mean divergence of protein sequences involved in reproduction, particularly male-reproduction, are higher than protein sequences with no reproductive function. These results suggest that the rapid evolution of sex and reproduction related genes is driven by sexual selection through a variety of physiological processes including female-choice (classical sexual selection), sperm competition, sexual conflict, and co-evolution (Rice 1996; Civetta and Singh 1999; Carson 2003).

LITERATURE CITED

Aguade M. 1999. Positive Selection drives the evolution of the *Acp29AB* accessory gland protein in Drosophila. Genetics. **152**: 543-551.

Andrews J., G.C. Bouffard, C. Cheadle, J. Lu, K.G. Becker and B. Oliver. 2000. Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Research. **10:** 2030-2043.

Barbash, D.A., D.F. Siino, A.M. Tarone and J. Roote, 2003. A rapidly evolving MYB-related protein causes species isolation in *Drosophila*. Proceeding of the National Academy of Science. **100**: 5302-5307.

Begun, D.J., P. Whitley, B.L. Todd, H.M. Waldrip-Dail and A.G. Clark. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics. **156:** 1879-1888.

Betancourt, A.J., D.C. Presgraves, and W.J. Swanson. 2002. A test for faster X evolution in *Drosophila*. Molecular Biology and Evolution **19:** 1816-1819.

Carson, H. L. 2003. Mate choice theory and the mode of selection in sexual populations. Proceeding of the National Academy of Science **100:** 6584-7.

Charlesworth, B., J.A. Coyne, and N.H. Barton. 1987. The relative rates of evolution of sex chromosomes and autosomes. American Naturalist **130:** 113-146.

Civetta, A., and R.S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. Journal of Molecular Evolution **41:** 1085-1095.

Civetta, A. and R.S. Singh. 1998. Sex-related genes, directional selection, and speciation. Molecular Biology and Evolution **15:** 901-909.

Civetta, A., and R. S. Singh. 1999. Broad-sense sexual selection, sex gene pool evolution, and speciation. Genome **42:** 1033-1041.

Coulthart, M.B., and R.S. Singh. 1988. High level of divergence of male-reproductive-tract proteins, between *Drosophila melanogaster* and its sibling species, *D. simulans*. Molecular Biology and Evolution **5:** 182-91.

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

Dobzhansky T. 1970. Genetics of the Evolutionary Process. New York and London: Columbia University Press.

Gonzalez, J., J.M. Ranz, and A. Ruiz. 2002. Chromosomal Elements Evolve at Different Rates in the *Drosophila* Genome. Genetics **161:** 1137-1154.

Korber, B. 2000. HIV Signature and Sequence Variation Analysis. In: Learn GH (eds) Computational Analysis of HIV Molecular Sequences. Kluwer Academic Publishers, Dordrecht, Netherlands. Pp 55-72.

Koski, L.B., and G.B. Golding. 2001. The closest BLAST hit is often not the nearest neighbor. Journal of Molecular Evolution **52:** 540-542.

Lumley, T., P. Diehr, S. Emerson, and L. Chen. 2002. The Importance of the Normality Assumption in Large Public Health Data Sets. Annual Review of Public Health **23:** 151-169.

McAllister, B.F. and G.V. McVean. 2002. Neutral Evolution of the sex-determining gene *transformer* in *Drosophila*. Genetics. **154:** 1711-1720.

Nei, M., and T. Gojobori. 1986. Simple methods for estimating the numbers of synonymous and non-synonymous nucleotide substitutions. Molecular Biology and Evolution **3:**418-426.

Parsch J., C.D. Meiklejohn, E. Hauschteck-Jungen, P. Hunziker and D.L. Hartl. 2001. Molecular Evolution of the *ocnus* and *janus* genes in the *Drosophila melanogaster* species subgroup. Molecular Biology and Evolution. **18:** 801-811

Ranz, J.M., J. Gonzalez, F. Casals, and A. Ruiz. 2003. Low occurrence of transposition events during the evolution of the Genus *Drosophila*. Evolution **57:** 1325-1335.

Rice, W.R. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature **381:** 232-234.

Stevison, L.S., B.A. Counterman, and M.A.F. Noor. 2004. Molecular evolution of X-linked accessory Gland Proteins in *Drosophila pseudoobscura*. Journal of Heredity **95:** 114-118.

Strutevant, A.H., and E. Novitski. 1941. The Homologies of the chromosome elements in the Genus *Drosophila*. Genetics **26:** 517-541.

Swanson, W.J., A.G. Clark, H.M. Waldrip-Dail, M.F. Wolfner, and C.F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proceeding of the National Academy of Science **98:** 7375-7379.

Swanson, W.J., and V.D. Vacquier. 2002. Rapid evolution of reproductive proteins. Nat. Rev. Genet. 3:137-144.

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

Thornton, K., and M.J .Long. 2002. Rapid divergence of gene duplicates on the *Drosophila melanogaster* X Chromosome. Molecular Biology and Evolution **19:** 918-925.

Torgerson, D.G., R.J. Kulathinal, and R.S. Singh. 2002. Mammalian sperm proteins are rapidly evolving: evidence of positive selection in functionally diverse genes. Molecular Biology and Evolution **19:** 1973-1980.

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

Tsaur, S-C. and C-I. Wu. 1997. Positive Selection and the molecular evolution of a gene male reproduction, *Acp26Aa* of *Drosophila*. Molecular Biology and Evolution. **14:** 544-549.

Tsaur, S-C., C-T. Ting, and C-I. Wu. 2001. Sex in *Drosophila mauritiana*: A very high level of Amino Acid polymorphism in a male reproductive protein gene, Acp26Aa. Molecular Biology and Evolution **18:** 22-26.

Wall, D.P., H.B. Fraser, and A.E. Hirsh. 2003. Detecting putative orthologs. Bioinformatics **19:** 1710-1711.

Wolfner, M.F., 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity **88:** 85-93.

Wu, C.-I., and A.W. Davis. 1993. Evolution of postmating reproductive isolation: The composite nature of Haldane's rule and its genetic basis. American Naturalist **142:** 187-212.