Stony Brook University

Testes-Specific Genes are Evolving Faster than Their Broadly Expressed Paralogs due to Reduced Expression Level and Relaxed Functional Constraint

A Thesis Presented

by

Seiji Kumagai

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Arts

in

Biological Sciences

Stony Brook University

August 2007

Stony Brook University

The Graduate School

Seiji Kumagai

We, the thesis committee for the above candidate for the
Master of Arts degree, hereby recommend
acceptance of this thesis.

Walter F. Eanes — Thesis Adviser
Professor, Department of Ecology and Evolution

Lev R. Ginzburg — Chairperson of Defense
Professor, Department of Ecology and Evolution

R. Geeta
Associate Professor, Department of Ecology and Evolution

This thesis is accepted by the Graduate School

Lawrence Martin
Dean of the Graduate School

Abstract of the Thesis

Testes-Specific Genes are Evolving Faster than Their Broadly Expressed Paralogs due to Reduced Expression Level and Relaxed Functional Constraint

by
Seiji Kumagai

Master of Arts
in
Biological Sciences
Stony Brook University
2007

Despite reports have repeatedly shown that expression level is a major determinant of evolutionary rate of proteins, the relationship between rapid evolution of testes-specific genes and their expression level has not been widely addressed. Here we report cases supporting that expression level affects the rate of protein evolution and that the testes genes may have acquired novel functions. Pairs of duplicates genes, of which one was expressed in the testes and the other with broad expression, was tested using comparative framework based on twelve genomes of *Drosophila* species. We show that the genes with testes-biased expression evolve faster than their broad-expression paralogs. Additionally, by using amino acid sequence alignments of *Drosophila*, *Homo sapiens*, and *Saccharomyces cerevisiae*, more substitutions were found in residues that were generally well conserved across eukaryote in the testes genes than their paralogs. The observed patterns of evolution is consistent with a prediction that expression level plays a major role in evolution of the testes-specific genes. The faster rate of evolution in the testes genes was also consistently detected in relatively recent branches. This suggests that the testes genes have been evolving faster than their paralogs throughout the phylogeny of *Drosophila*. A subset of analyses was performed on the head-specific genes and their broad-expression paralogs, accelerate rate of evolution in the head-specific genes was not found. These results strongly indicate that the rapid evolution of the testes-specific genes is not simply due to positive Darwinian selection around the time of gene duplication or due to the fact that the testes-specific genes are expressed in a reduced number of tissues.

Contents

List of Tables

List of Figures

Acknowledgement

First, I would like to thank my parents, Mitsuko and Yoshio Kumagai, to let me do what I want to do for such a long time. I would also like to thank members of Eanes' lab for their continuous encouragement throughout this project. Especially, I am in great debt of Jon Flowers. He provided invaluable advices since the earliest stage of this project. Thomas Merritt guided me into the world of *Drosophila* genetics when I started working in the lab as an undergraduate student. I would not have considered working on this project without his guidance. Walt Eanes has always been patient enough to have me in his lab since my undergraduate period. From him, I learnt the meaning of being a good researcher. R Geeta and Lev Ginzburg are kind enough to read my thesis with a short notice. Joe Lachance's enthusiasm kept me motivated during the period that the progress of the project was stalled.

Introduction

Genes expressed in the male reproductive system of animals have drawn the attention of evolutionary biologists. It has been shown that many of the genes expressed in the testes evolve under positive selection (Nurminsky *et al.*, 1998; Betrán and Long, 2003). In fact, this class of genes is one of only a few where researchers consistently find signatures of positive selection. Accessory gland proteins (*Acp*) in *Drosophila* are well known examples of rapidly evolving genes (Cirera and Aguadé, 1997; Begun *et al.*, 2000). Similar observations have been reported in other organisms (Lee *et al.*, 1995; Swanson and Vacquier, 1995; Sanwon and Vacquier, 1998; Rooney and Zhang, 1999; Torgerson *et al.*, 2002).

In *Drosophila*, evolutionary rates of genes with testes-biased expression are accelerated relative to the genome-wide average (Richards *et al.*, 2005; Jagadeeshan and Singh, 2005). These findings were attributed to selection related to reproductive activities such as sexual conflict and sperm competition (Zhang and Li, 2004), but other hypotheses such as the influence of gene expression on evolutionary rate have not been explicitly considered. This is despite the fact that gene expression has repeatedly been reported as the major determinant of evolutionary rates of proteins (Duret and Mouchiroud, 2000; Pál *et al.*, 2001; Herbeck *et al.*, 2003; Rocha and Danchin, 2004; Subramanian and Kumar, 2004; Lemos *et al.*, 2005). Other explanations for the accelerated rate of testes-specific genes such as a global relaxation of selective constraint on these genes have also not been widely considered.

In the present study, we use a comparative approach to better understand why genes with testes-biased expression evolve rapidly. We consider the evolution of paralogous gene pairs; one with testes-specific and the other with a broad expression profile. We compared the patterns of molecular evolution between 50 genes specific to the testes and their paralogs with broad expression based on the comparative annotation freeze 1 (CAF1) genomes of twelve *Drosophila* species. Additional analyses were also conducted on 31 genes with head-biased expression and their broadly expressed paralogs to test if the patterns we observed in the testes were specific to this tissue or were common to the genes specifically expressed in a restricted set of tissues. We hypothesized that more testes-specific genes have experienced positive Darwinian selection than the head-specific genes or the broadly expressed paralogs if positive selection drove the observed escalation of evolutionary rate in the testes. On the other hand, we predicted that the testes-specific genes were expressed lower than their paralogs with broad expression if reduced expression level in the testes caused the rapid evolution. Finally, if functional divergence causes the marked increase of evolutionary rate in the testes, we expected to observe more substitutions accumulated in the sites that were highly conserved across wide range of taxa due to the functional importance of those sites. The results indicate that genes expressed exclusively in testes tend to evolve faster than their broadly expressed paralogs, while genes specific to the head did not show the same pattern. We also found indirect evidence that genes with testes expression have expressed in lower quantity than their broadly expressed paralogs. This pattern could explain the rapid evolution of the testes-specific paralogs, and should be considered along with adaptive explanations for the rapid

evolution of genes expressed in the testes.

Materials and Methods

Expression Pattern of Genes

The expression breadth of a gene is defined as the number of tissues, where the gene is expressed. This information was obtained using Expressed Sequence Tags (ESTs) derived from tissue-specific EST libraries available from UniGene release 45 (http://www.ncbi.nlm.nih/UniGene/). There are eight tissue types in this database: hemocyte, nervous system, embryonic tissue, fat body, salivary gland, testes, ovary, and head. The number of observed ESTs was normalized by total number of ESTs in each tissue-specific library. If the proportion of ESTs from a single location, after the normalization, was at least 0.9, genes were considered to have biased expression in that tissue. We focused on genes, whose expression was biased toward the testes or head, as the two largest tissue-specific EST libraries were available for these tissues.

Identification of Paralogs

Genes with testes- or head-specific expression often belong to protein families that contain more than two genes in *D. melanogaster*. Our approach was designed for a pair of paralogs with high similarity to facilitate unambiguous alignment and close evolutionary relatedness. In order to determine which of the widely expressed genes was most similar to the gene that was expressed in the testes or head, we chose a paralog that had the highest amino acid identity with the tissue-specific gene, and that belonged to the same protein family. We determined the family membership of the specific genes based using EnsEMBL release 40 (Hubbard *et al.*, 2007). If this family contained more than two *D. melanogaster* sequences, the one with the highest identity, also from EnsEMBL, to the gene with the tissue-biased expression was selected. If the identity of the most similar protein was not more than 10% of the others, we excluded the pair from further analyses. We subsequently filtered out alignments for which the shorter paralog was less than 70% of the other. The final dataset consisted of 50 pairs for the testes and 31 pairs for the head. Information for each of the genes, including the estimated evolutionary rates, molecular and biological functions, and homology to human genes, can be found in supplementary tables 1 to 4. Briefly, none of the *Acp* genes reported in Mueller *et al.* (2005) were present in our dataset, and biological functions of the genes spanned from metabolism, including some of the core metabolic enzymes and transcription factors to signal transduction related proteins and other miscellaneous functions.

DNA Alignments

Aligned coding DNA sequences of the 12 *Drosophila* species generated by Chatterji and Pachter (2006) were downloaded from http://bio.math.berkeley.edu/genemapper/CAF1_genes_v0.2/. The species are *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. mojavensis* *D. virilis*, and *D. grimshawi* (*D. erecta*, *D. ananassae*, *D. mojavensis*, *D. virilis* and *D. grimshawi* genomes were sequenced by Agencourt BioScience Corporation. *D. simulans* and *D. yakuba* genomes were

sequenced by Genome Sequencing Center at Washington University. *D. sechellia* and *D. persimilis* were sequenced by Broad Institute at Massachusetts Institute of Technology and Harvard University. *D. willistoni* was sequenced by The Institute of Genomic Research, now a part of J. Craig Venter Institute. *D. melanogaster* was sequenced by the Berkeley *Drosophila* Genome Project and Celera. *D. pseudoobscura* was sequenced by Human Genome Sequencing Center at Baylor College of Medicine). Sequences with premature stop codons or those that were shorter than 90% of the *D. melanogaster* sequence were excluded from the alignment. The number of species in each alignment spans from 2 to 12 with mean 9.8, median 11, and standard deviation 2.6. In many cases, gene duplications appear to predate the common ancestor of the twelve *Drosophila* species. For the alignment of the two paralogs, a pairwise amino acid alignment of *D. melanogaster* sequences was created using MUSCLE version 3.6 (Edgar, 2004). The alignment was back-translated into DNA. Then, the sequences of the rest of species were added to the alignment.

Phylogenetic Analysis of Molecular Evolution

We conducted codon-based analyses of molecular evolution using the codeml program of PAML package version 3.15 (Yang, 1997). All the models discussed in this section are implemented in codeml. Throughout this part of analyses, we used the species tree topology reported by Powell (1997). We used two approaches to test for rate differences between paralogs. In one approach, we tested whether evolutionary rates differ between a pair of duplicates in a maximum likelihood framework. An alignment, containing a pair of genes, was used to obtain estimates of evolutionary rates based on the nearly neutral model (M1a) and clade model C in codeml (Yang, 1997). M1a partitions all sites into two classes, of which a class of sites is constrained to have $0 < d_N/d_S < 1$, and the other is set to be $d_N/d_S = 1$. The clade model C has three rate categories. In addition to the two classes equivalent to those in M1a, two lineages (i.e., the two paralogous genes, or clades) have a third class of sites in which the ratio of nonsynonymous to synonymous substitutions (d_N/d_S) is unconstrained and estimated from the data for each lineage (for details of the clade model C, see Bielawski and Yang (2004)). Because M1a is nested within the clade model C, a Likelihood Ratio Test (LRT) was conducted by comparing $-2\Delta l$, where Δl is the difference of log likelihood values between M1a and the clade model C, with χ^2 with 3 degrees of freedom (Yang, 1997). In the second approach, we estimated the nonsynonymous substitution rate (d_N) and the synonymous substitution rate(d_S) from a separate alignment of each paralog using the one ratio model (M0), which assigns a single rate to over the entire tree. Then, distributions of the rate for the tissue-specific genes and their paralogs were compared using nonparametric paired tests.

Estimating Expression Level by Codon Bias as a Proxy

In order to test whether expression level explains a large proportion of variation in evolutionary rates of the tissue-specific and their paralogs, we needed to have a reliable estimate of expression level. This is difficult to estimate, however, because gene expression could fluctuate temporally and spatially not to mention depen-

dency on environments. Moreover, there is no method available to compare expression level in one tissue with that of another in any evolutionary meaningful manner. Because of the difficulties in comparing absolute expression level, we decided to use codon bias as a proxy. Codon bias is negatively associated with expression level (Powell and Moriyama, 1994) through weak selection on translational accuracy (Akashi, 1994). Using codonW version 1.4.4 (http://codonw.sourceforge.net/), we calculated ENC, which is an index of how evenly alternative codons encoding the same amino acid are used in a gene (Wright, 1990). The range of ENC spans from 20 to 61. The low value of ENC indicates that a subset of codons are preferentially used in a gene, and highly expressed genes have low ENC as the use of preferred codons reduces mistranslation.

Assessing Functional Divergence through Comparison of Amino Acid Sequences with Human and Yeast

The degree of functional constraint has a huge impact on the evolutionary rate of a protein. We were interested if duplicates evolve at different rates due to differences in their functions. Some of the amino acid residues in a protein are well conserved across a wide range of taxa because these residues are essential for proper functioning of the protein. If substitutions are found in those conserved positions, it could indicate that the protein has lost the ancestral function and has acquired a new one. To determine if amino acids that are highly conserved among eukaryote homologs have substituted in either the tissue-specific or broadly expressed *Drosophila* paralogs, we compared amino acid sequences of *Homo sapiens* (human), *Saccharomyces cerevisiae* (yeast) and each of the *D. melanogaster* sequences. First, aligned human, yeast, and fly amino acid sequences of a protein family were downloaded from EnsEMBL release 43 (Hubbard *et al.*, 2007). Then, the positions that were completely conserved between human and yeast protein families were identified. Amino acids corresponding to those conserved positions were extracted from this alignment to create a new alignment consisting only of sites perfectly conserved between human and yeast. Any position with gaps were eliminated. An example of one alignment is in fig. 1. The fractions of mismatched amino acids between *D. melanogaster* and the human/yeast sequences were computed separately for each *D. melanogaster* gene. Similarly, we obtained pairwise amino acid distances between each *D. melanogaster* sequence and the human/yeast sequences to evaluate the extent of protein divergence of each paralog at this conserved set of sites. The amino acid distances based on JTT substitution matrix of Jones *et al.* (1992) were computed by PROTDIST program from PHYLIP package version 3.66 of Felsenstein (2005).

Results and Discussion

A primary goal of our study was to establish if genes with testes-biased expression have faster rates of evolution than their broadly expressed paralogs and, if this is the case, to identify the causes. For comparative purposes, we also tested for rate differences in a second set of duplicate gene pairs — paralogs with head-biased and broad expression. We conducted maximum likelihood based tests of evolutionary rates within the paralogous gene pairs. LRTs of clade model C with the nearly neutral model (M1a) were significant with $p < 0.001$ in all testes and head pairs suggesting that there is a class of sites evolving at different rates in the two genes. Inspection of the d_N/d_S for this class of sites indicated that these sites were evolving faster in the genes with testes-biased expression in 42 of 50 pairs. In the head, only 13 of 31 pairs showed a faster rate in the head-specific paralog.

Similar results were also found when we compared evolutionary rates estimated using the one ratio model (M0) for each gene separately. The mean d_N of the tissue-specific genes was 2.10 fold higher in the testes and effectively the same in the head as their corresponding paralogs with broad expression. Indeed, we found a significant skew in the distributions of d_N for testes-specific and broadly expressed paralogs, with the testes genes evolving significantly faster ($p < 0.001$, paired one-tailed Wilcoxon signed rank test). In addition, a similar skew in the distributions of d_N/d_S in the testes comparisons was observed (fig. 2, $p < 0.001$, paired one-tailed Wilcoxon signed rank test). The distributions of the head-specific genes and their paralogs did not differ in d_N or in d_N/d_S. Additionally, we tested if distributions of d_S differed between the paralogous gene pairs. The distributions were skewed toward higher d_S in the tissue specific genes ($p < 0.025$ for the testes, and $p < 0.05$ for the head, paired one-tailed Wilcoxon signed rank test). The higher d_S makes d_N/d_S lower, so the increased evolutionary rate found in the testes could not be explained by the observed escalation of d_S.

We were interested if the accelerated rate observed in the testes genes was due to a higher rate of evolution in one or a few branches in the phylogeny as might be expected if the rate difference was due to a burst of amino acid substitutions following an ancient duplication. To evaluate this, we computed pairwise estimates of d_N and d_S for *D. melanogaster*/*D. sechellia*, *D. yakuba*/*D. erecta*, and *D. mojavensis*/*D. virilis* for a subset of gene pairs. In the three independent pairs of lineages, all three estimates were significantly higher in the testes-specific genes ($p < 0.05$ for d_S and d_N/d_S and $p < 0.01$ for d_N, paired one-tailed Wilcoxon ranked sum test), whereas none were significantly higher in the head-biased genes (fig. 2). This indicates that the rapid rate of evolution in the testes is consistently found throughout the phylogeny, and is not limited to one or a few branches. An important discovery emerging from this work, therefore, is that testes genes, which are retained, evolve at faster rates for millions of years after duplication. Thus, whatever mechanism is responsible for the accelerated rate cannot be attributed to rapid evolution immediately following gene duplication.

In order to identify the factors causing these patterns, we conducted additional analyses focusing on two alternative explanations: positive selection and relaxation of selective constraint. We first tested for evidence of positive selection. If the accelerated rate in the testes is driven by positive selection, we expected to observe signatures of positive selection more often in the testes-specific genes.

To detect these signatures, we performed LRTs on M1a and the positive selection model (M2a), and the beta distribution model (M7) and the corresponding positive selection model (M8) in codeml (Yang *et al.*, 2000). While M2a did not fit the data better than M1a in any pair, there were a few pairs where M8 fit significantly better than M7 (i.e., consistent with positive selection) and at least a single amino acid was inferred to be under positive selection with probability 0.95 or higher by a Bayes empirical Bayes procedure (Yang *et al.*, 2005). The testes genes and their paralogs contained three genes each that showed evidence of adaptive evolution. In the head, these signatures were found in two of the specific genes and one of the broadly expressed paralogs. Our prediction that there are more positively selected genes with testes-biased expression than the broadly expressed genes or the head-specific genes was not supported. Therefore, we found no support for the hypothesis that positive selection is the cause of rapid evolution in the testes. However, it is plausible that more genes have undergone positive selection than our estimate because the test may be too conservative to detect most of positively selected genes as $d_N/d_S > 1$ is attained only under extremely strong adaptive evolution (Kreitman and Akashi, 1995).

Several researchers have argued that a high proportion of fixed amino acid differences between closely related species are due to positive selection. This has been reported in *Drosophila* (Akashi, 1999; Fay *et al.*, 2002; Smith and Eyre-Walker, 2002; Sawyer *et al.*, 2003; Bierne and Eyre-Walker, 2004; Shapiro *et al.*, 2007), in humans (Fay *et al.*, 2001; Gojobiri *et al.*, 2007), and in bacteria (Charlesworth and Eyre-Walker, 2006). Their claims suggest that positive selection may have considerable influence on the overall rate of protein evolution. However, if positive selection explains the elevated rate in testes-specific genes, these genes must be evolving adaptively in three independent lineages millions of years after duplication. Although the recurrent positive selection in the three lineage is one possible explanation of accelerated evolution found in the testes, it is more parsimonious to infer that a parameter that globally affects the rates of protein evolution, such as reduction of gene expression level or reduced functional constraint in the testes, has changed after duplication.

As noted above, gene expression level has been argued to be a primary factor that determines the rate of protein evolution; highly expressed genes evolve slower than genes with low expression. Therefore, it is possible that the pattern we observed is due to reduced expression level in the testes. By using ENC as a proxy of expression level, we found significantly higher ENC (i.e., lower codon bias) in the genes with testes-biased expression indicating that lower codon bias was present in the set (fig. 3, $p < 0.001$, paired one-tailed Wilcoxon ranked sum test). ENC of the head-specific genes did not differ from their broadly expressed paralogs. This suggests that the expression levels of the testes-specific genes are lower than that of their paralogs, while the relationship is not found in the head. This observation is consistent with the hypothesis that expression level can explain rapid evolution in genes with testes-biased expression.

To determine whether our finding about lower expression level of the testes-specific genes can be applied in broader context of the rapid evolution of testes-biased genes, we also calculated ENC of testes-specific genes and broadly expressed genes at genome-wide scale. The average ENCs of our data set were 49.3 and 43.8 for the testes-biased and their paralogous genes ($p < 10^{-5}$, paired one-tailed

Wilcoxon ranked sum test). However, when we computed the mean ENC for all testes-biased genes in *D. melanogaster* genome, we found that the mean was 52.5, while all genes with broad expression had a mean of 51.3. Although this difference was highly significant ($p < 10^{-10}$, one-tailed Wilcoxon ranked sum test), the difference was much smaller than that in analysis of the testes-biased genes and the paralogs. The difference in codon bias may explain a part of reasons that testes-specific genes evolved faster than other genes, but we expect that explanatory power of expression level, inferred from codon bias, may be relatively low in genome scale dataset such as the one studied by RICHARDS *et al.* (2005), where testes-specific genes were shown to evolve faster than other genes. This indicates that there are other factors that contribute to the accelerated rate of evolution of testes-specific genes. Nevertheless, our results suggest that gene expression could be a major determinant of the rate difference observed in our dataset.

Finally, we tested if the accelerated evolution of the genes in the testes was due to functional divergence (*subfunctionalization* and/or *neofunctionalization*) (FORCE *et al.*, 1999) and subsequent relaxation of functional constraint. One prediction of functional divergence is that amino acid residues which are strongly conserved in orthologs throughout evolution could be substituted upon change in function, because a new function may require a different set of functionally important residues. To test this prediction, we compared amino acid sequences of the *Drosophila* paralogs against sequences of human and yeast. We found that the fraction of sites that are substituted in the testes-specific genes are significantly higher than their paralogs (fig. 4 A, $p < 0.01$, paired one-tailed Wilcoxon ranked sign test). A similar pattern was observed in the pairwise amino acid distances (fig. 4 B, $p < 0.01$) where the yeast/human–testes paralog distances were significantly greater than the yeast/human–broad paralog distances at this set of evolutionarily conserved sites. Our rationale is that, if amino acids are conserved between the two distantly related species, there is a high chance that those sites are conserved because of functional importance. Then, if a gene in *D. melanogaster* has substitutions within those residues, it could indicate that the ancestral function has been disrupted. This suggests that the genes with testes restricted expression tend to lose their ancestral functions. Together with the fact that most of the genes analysed have persisted over millions of years so that they are likely to be functional in *Drosophila* (PETROV *et al.*, 1996; PETROV and HARTL, 1998; PETROV *et al.*, 1998), it could be that the testes-specific genes have gained novel functions.

Moreover, we predicted that if the functions of the two genes are similar after gene duplication, their rate of evolution would also be similar, because they would experience similar functional constraints. To test this possibility, we examined correlations between d_N/d_S for the broadly expressed and the tissue-specific genes for both the testes and head. Interestingly, we found a positive correlation in the head paralogs ($\tau = 0.376$, $p < 0.05$, Kendall's rank correlation), but not in the testes, despite having greater statistical power in the latter analysis (fig. 2). This pattern is consistent with conservation of function between head paralogs as evident in the conservation of evolutionary rate. An alternative explanation is that gene expression level is conserved between paralogs, and the observed correlations of d_N/d_S are merely artifacts of expression level. However, we observed a significant correlation between ENC of the testes genes and their paralogs ($\tau = 0.230$, $p < 0.05$, Kendall's rank correlation) but not for the pairs of paralogs in the head

(fig. 3). Although the correlation of ENC found in the testes paralogs is puzzling, the lack of association in the head does not fit with the explanation that association of d_N/d_S in the head comparison is due to expression level.

In this study, we found an increased rate of evolution in the genes with testes-biased expression compared with their broadly expressed paralogs. The rate acceleration of protein evolution in the testes has previously been attributed to adaptive evolution. However, we find that the observed pattern could also be caused by reduced expression and reduced functional constraints. Moreover, the genes with testes-biased expression evolve significantly faster in three independent lineages relatively recently indicates that change in a genome-wide parameter is more likely to explain the accelerate evolution of testes-specific genes than positive selection in our dataset. This suggests that the model of gene evolution in testes may require revision to include neutral, non-adaptive explanations.

References

Akashi, H., 1994 Synonymous codon usage in drosophila melanogaster: Natural selection and translational accuracy. Genetics **136**: 927–935.

Akashi, H., 1999 Inferring the fitness effects of dna mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151**: 221–238.

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.

Betrán, E., and M. Long, 2003 *Dntf-2r*, a young Drosophila retroposed gene with specific male expression under positive darwinian selection. Genetics **164**: 977–988.

Bielawski, J. P., and Z. Yang, 2004 A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene familyevolution. J. Mol. Evol. **59**: 121–132.

Bierne, N., and A. Eyre-Walker, 2004 The genome rate of adaptive amino acid substitution in *Drosophila*. Mol. Evol. Biol. **21**: 1350–1360.

Charlesworth, J., and A. Eyre-Walker, 2006 The rate of adaptive evolution in enteric bacteria. Mol. Evol. Biol. **23**: 1348–1356.

Chatterji, S., and L. Pachter, 2006 Reference based annotation with GeneMapper. Genome Biol. **7**: R29.

Cirera, S., and M. Aguadé, 1997 Evolutionary history of the sex-peptide (Acp70A) gene region in *Drosophila melanogaster*. Genetics **147**: 189–197.

Duret, L., and D. Mouchiroud, 2000 Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. Mol. Evol. Biol. **17**: 68–74.

Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic. Acids. Res. **32**: 1792–1797.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu, 2001 Positive and negative selection on the human genome. Genetics **158**: 1227–1234.

Fay, J. C., G. J. Wyckoff, and C.-I. Wu, 2002 Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. Nature **415**: 1024–1026.

Felsenstein, J., 2005 PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, University of Washinton, Seattle.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y.-l. Yan, *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151**: 1531–1545.

Gojobiri, J., H. Tang, J. M. Akey, and C.-I. Wu, 2007 Adaptive evolution in humans revealed by the negative correlation between the polymorphism and fixation phases of evolution. Proc. Natl. Acad. Sci. USA **104**: 3907–3912.

Herbeck, J. T., D. P. Wall, and J. J. Wernegreen, 2003 Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology **149**: 2585–2596.

Hubbard, T. J. P., B. L. Aken, K. Beal, B. Ballester, M. Caccamo, *et al.*, 2007 Ensembl 2007. Nucleic. Acids. Res. **35**: 610–617.

Jagadeeshan, S., and R. S. Singh, 2005 Rapidly evolving genes of *Drosophila*: Differing levels of selective pressure in testis, ovary, and head tissues between sibling species. Mol. Evol. Biol. **22**: 1793–1801.

Jones, D. T., W. R. Taylor, and T. J. M, 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8**: 275–282.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26**: 403–422.

Lee, Y.-H., T. Ota, and V. D. Vacquier, 1995 Positive selection is a general phenomenon in the evolution of Abalone sperm Lysin. Mol. Evol. Biol. **12**: 231–238.

Lemos, B., B. R. Bettencourt, C. D. Meiklejohn, and D. L. Hartl, 2005 Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mrna abundance, protein length, and number of protein-protein interactions. Mol. Evol. Biol. **22**: 1345–1354.

Mueller, J. L., K. Ravi Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, *et al.*, 2005 Cross-species comparison of drosophila male accessory gland protein genes. Genetics **171**: 131–143.

Nurminsky, D. I., M. V. Nurminskaya, D. De Aquiar, and D. L. Hartl, 1998 Selective sweep of evolved sperm-specific gene in *Drosophila*. Nature **396**: 572–575.

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

Petrov, D. A., Y.-C. Chao, E. C. Stephenson, and D. L. Hartl, 1998 Pseudogene evolution in *Drosophila* suggests a high rate of DNA loss. Mol. Evol. Biol. **15**: 1562–1567.

Petrov, D. A., and D. L. Hartl, 1998 High rate of DNA loss in *Drosophila melanogaster* and *Drosophila virilis* species groups. Mol. Evol. Biol. **15**: 293–302. Similar to Petrov98?

Petrov, D. A., E. R. Lozovskaya, and D. L. Hartl, 1996 High intrinsic rate of DNA loss in *Drosophila*. Nature **384**: 346–349. Why only a few pseudogenes are present in drosophila.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford Univ Press, New York.

Powell, J. R., and E. S. Moriyama, 1994 Evolution of codon usage bias in *Drosophila*. Proc. Natl. Acad. Sci. USA **94**: 7784–7790.

Richards, S., Y. Liu, B. R. Bettencourt, P. Hradecky, S. Letovsky, *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. Genome Res. **15**: 1–18.

Rocha, E. P., and A. Danchin, 2004 An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol. Evol. Biol. **21**: 108–116.

Rooney, A. P., and J. Zhang, 1999 Rapid evolution of a primate sperm proteins: Relaxation of functional constraint or positive Darwinian selection? Mol. Biol. Evol. **16**: 706–710.

Sanwon, W. J., and V. D. Vacquier, 1998 Concerted evolution in an egg for a rapidly evolving Abalone sperm protein. Science **281**: 710–712.

Sawyer, S. A., R. J. Kulathinal, C. D. Bustamante, and D. L. Hartl, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. Mol. Evol. Biol. **57**: S154–S164.

Shapiro, J. A., W. Huang, C. Zhang, M. J. Hubisz, J. Lu, *et al.*, 2007 Adaptive genic evolution in the *Drosophila* genomes. Proc. Natl. Acad. Sci. USA **104**: 2271–2276.

Smith, N. G., and A. Eyre-Walker, 2002 Adaptive protein evolution in *Drosophila*. Nature **415**: 1022–1024.

Subramanian, S., and S. Kumar, 2004 Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168**: 373–381.

Swanson, W. J., and V. D. Vacquier, 1995 Extraordinary divergence and positive Darwinian selection in a fusagenic protein coding the acrosomal process of abalone spermatozoa. Proc. Natl. Acad. Sci. USA **92**: 4957–4961.

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. Mol. Evol. Biol. **19**: 1973–1980.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87**: 23–29.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**: 555–556.

Yang, Z., W. J. Swanson, and V. D. Vacquier, 2000 Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Evol. Biol. **17**: 1446–1455.

YANG, Z., W. S. WONG, and R. NIELSEN, 2005 Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Evol. Biol. **22**: 1107–1118.

ZHANG, L., and W.-H. LI, 2004 Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol. Evol. Biol. **21**: 236–239.

Figures

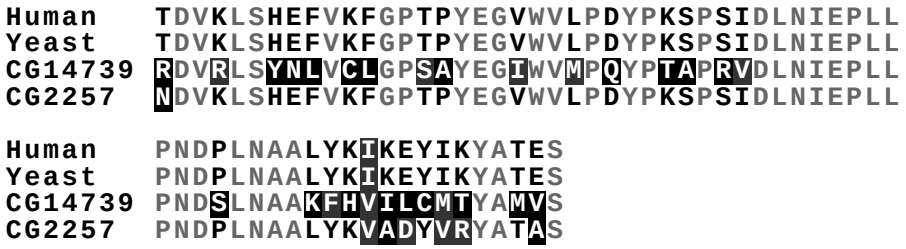


Figure 1: An amino acid alignment consisting of perfectly conserved sites between human and yeast. *CG14739* is a testes-specific gene in *D. melanogaster*, and *CG2257* (*Ubc-E2H*) is its paralog with broad expression. The percentage of mismatched sites between a *Drosophila* gene and a human/yeast homolog for each paralog was 45% in the testes-specific gene and 11% for its paralog. The protein distances of *CG14739* was 0.689, and that of *CG2257* was 0.125.

Figure 2: Comparison of d_N/d_S between the tissue-specific genes and their paralogs with broad expression. Broadly expressed genes and tissue-biased expression paralogs are plotted as x- and y-axis. The values are partitioned based on tissue types of the biased expression genes. The sample sizes are 31 and 50 for head and testes, respectively. The dashed line has an intercept 0 and slope 1, and indicates equivalent rate of evolution. The head genes and their paralogs are correlated ($p < 0.005$), but the testes and their paralogs are not.

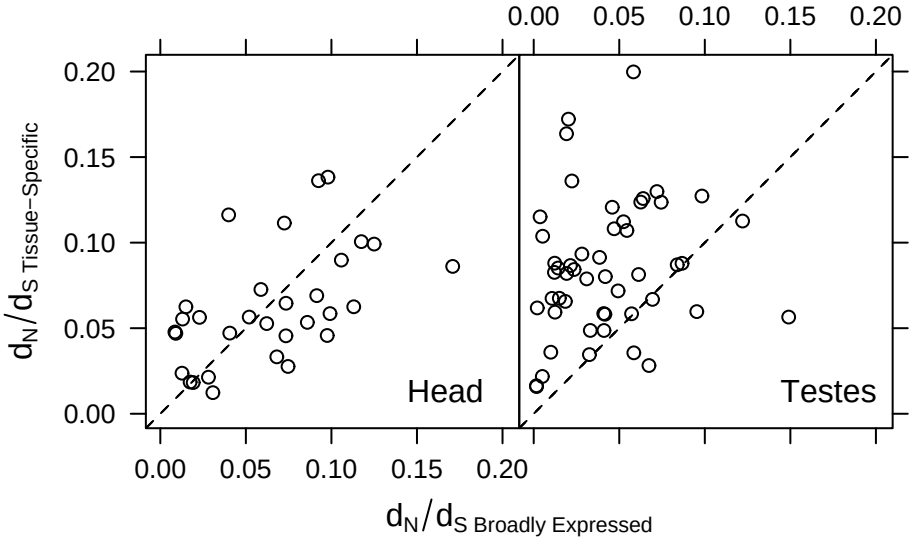


Figure 3: Comparison of ENC between the the genes with tissue-biased expression with their broadly expressed paralogs. The effective number of codons (ENC) of WRIGHT (1990) was computed for the *D. melanogaster* sequences by codonw 1.4.4 (http://codonw.sourceforge.net/). ENC is an index of codon bias; the higher the bias is, the smaller is ENC. The dashed line has an intercept 0 and slope 1, and indicates points where paralog pairs locate if they have the same ENC. The testes genes and thier paralogs are correlated ($p < 0.05$), but the head-specific genes and their paralogs are not.

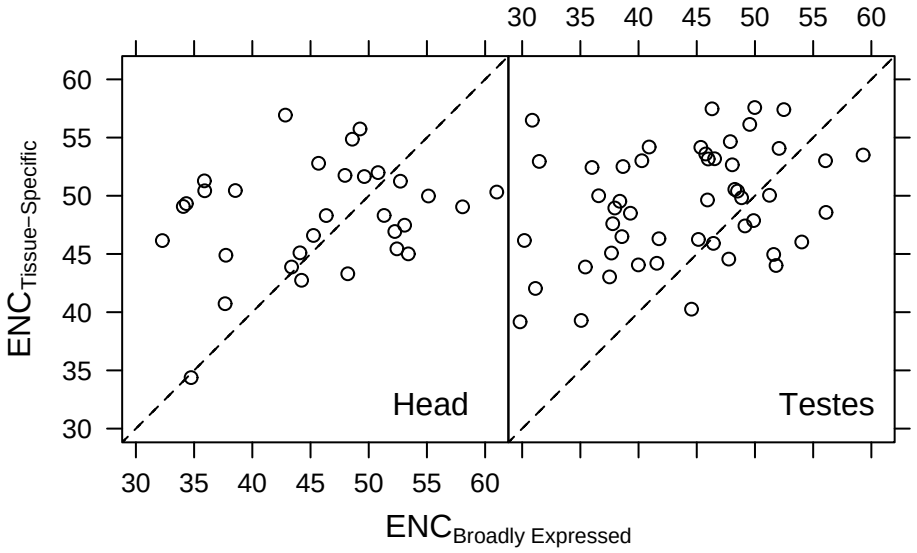
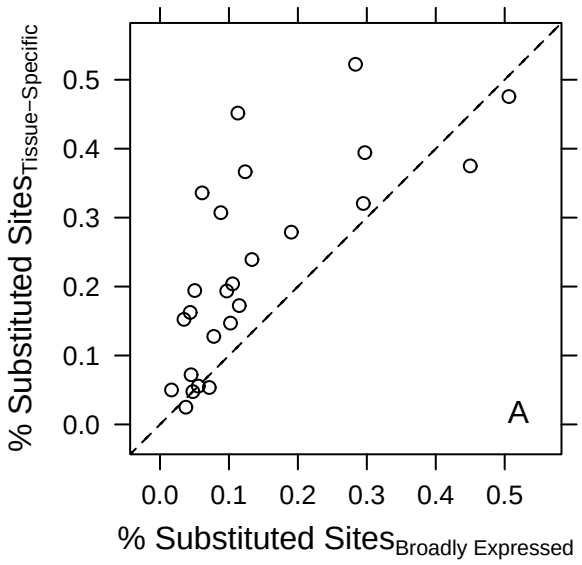
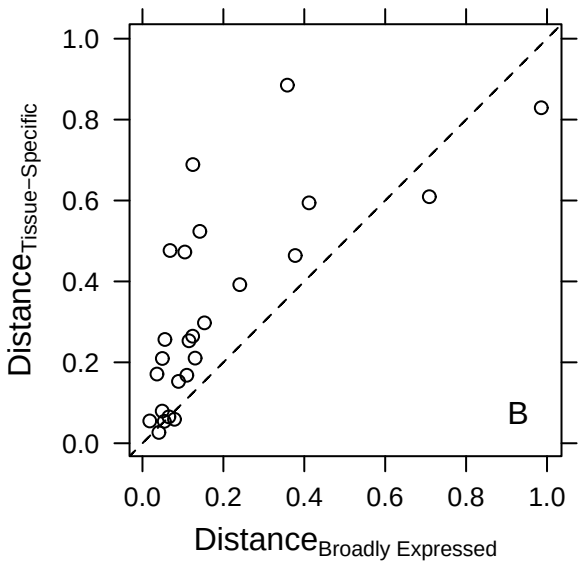


Figure 4: Comparison of amino acid conservation in the testes-specific genes and their paralogs in *D. melanogaster* where human and yeast have the same residues. (A) Fractions of residues that are identical in human and yeast but differs in *D. melanogaster*. (B) The distance between *D. melanogaster* and other two species based on JTT matrix. Both are given over the residues that are perfectly conserved between human and yeast. Sample size is 25. The dashed line has an intercept 0 and slope 1, and indicates points where paralog pairs locate if they diverge from human/yeast at the same degree.





Tables

Table 1: Means and standard deviations of d_N, d_S, d_N/d_S, ENC and number of sequences in the each tissue-specific classes and their paralogs. Sample mean are reported followed by standard deviation in parenthesis.

	Broad		**Biased**		p[a]
Testes	$N = 50$ gene pairs				
d_N	0.243	(0.237)	0.511	(0.303)	1.42×10^{-6}
d_S	5.21	(1.98)	6.21	(2.72)	0.0153
d_N/d_S	0.0422	(0.0327)	0.0845	(0.0395)	1.91×10^{-7}
ENC	43.8	(7.39)	49.3	(4.86)	8.90×10^{-6}
no. seqs	10.62	(1.84)	8.88	(2.90)	
Head	$N = 31$ gene pairs				
d_N	0.319	(0.275)	0.356	(0.274)	0.360
d_S	4.73	(2.05)	5.57	(2.54)	0.0471
d_N/d_S	0.0623	(0.0336)	0.0642	(0.0414)	0.632
ENC	45.7	(7.72)	48.3	(4.63)	0.0650
no. seqs	9.86	(2.89)	10.10	(2.36)	

[a] p values are from one-tailed Wilcoxon signed rank test with a null hypothesis that the estimates of tissue-biased expression genes are lower or equal to those of broadly expressed genes

Table 2: Means and standard deviations of pairwise d_N, d_S, and d_N/d_S in the three independent branches of *D. melanogaster*. The reported values are means followed by standard deviation in parenthesis. The estimates were obtained for the following three pairs of species per gene: *D. melanogaster* and *D. sechellia*; *D. yakuba* and *D. erecta*; and *D. mojavensis* and *D. virilis*.

	Broad		**Biased**		p[a]
Testes	$N = 18$ gene pairs				
D. melanogaster* and *D. sechellia					
d_N	0.00882	(0.0138)	0.0163	(0.0130)	9.18×10^{-3}
d_S	0.126	(0.0437)	0.155	(0.0437)	0.0274
d_N/d_S	0.0772	(0.120)	0.112	(0.0887)	0.0277
D. yakuba* and *D. erecta					
d_N	0.0267	(0.0300)	0.0177	(0.0168)	3.74×10^{-3}
d_S	0.185	(0.0943)	0.283	(0.0881)	4.72×10^{-4}
d_N/d_S	0.0766	(0.0961)	0.0925	(0.0491)	0.0483
D. mojavensis* and *D. virilis					
d_N	0.0457	(0.0480)	0.110	(0.0720)	1.71×10^{-3}
d_S	0.860	(0.374)	1.18	(0.355)	0.0274
d_N/d_S	0.0424	(0.0382)	0.0911	(0.0545)	2.31×10^{-3}
Head	$N = 15$ gene pairs				
D. melanogaster* and *D. sechellia					
d_N	0.00662	(0.00675)	0.00991	(0.00815)	0.0884
d_S	0.115	(0.0456)	0.132	(0.0338)	0.0631
d_N/d_S	0.0647	(0.0678)	0.0717	(0.0513)	0.281
D. yakuba* and *D. erecta					
d_N	0.0130	(0.0144)	0.0163	(0.0164)	0.165
d_S	0.206	(0.0771)	0.234	(0.0802)	0.10
d_N/d_S	0.0629	(0.0615)	0.0815	(0.0580)	0.165
D. mojavensis* and *D. virilis					
d_N	0.0683	(0.0671)	0.0815	(0.0648)	0.288
d_S	0.957	(0.400)	1.36	(1.56)	0.475
d_N/d_S	0.0668	(0.0539)	0.0682	(0.0470)	0.455

[a] p values are from one-tailed Wilcoxon signed rank test with a null hypothesis that the estimates of tissue-biased expression genes are lower or equal to those of broadly expressed genes