J Mol Evol (2005) 60:774–785
DOI: 10.1007/s00239-004-0241-2

JOURNAL OF MOLECULAR EVOLUTION

Protein Evolution in the Context of *Drosophila* Development

Jerel C. Davis,[1]* Onn Brandman,[2]* Dmitri A. Petrov[1]

[1] Department of Biological Science, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA
[2] Department of Molecular Pharmacology, W200 Clark, 318 Campus Drive, Stanford University Medical School, Stanford, CA 94305, USA

Received: 1 August 2004 / Accepted: 16 January 2005 [*Reviewing Editor:* Dr. Manyuan Long]

Abstract. The tempo at which a protein evolves depends not only on the rate at which mutations arise but also on the selective effects that those mutations have at the organismal level. It is intuitive that proteins functioning during different stages of development may be predisposed to having mutations of different selective effects. For example, it has been hypothesized that changes to proteins expressed during early development should have larger phenotypic consequences because later stages depend on them. Conversely, changes to proteins expressed much later in development should have smaller consequences at the organismal level. Here we assess whether proteins expressed at different times during *Drosophila* development vary systematically in their rates of evolution. We find that proteins expressed early in development and particularly during mid–late embryonic development evolve unusually slowly. In addition, proteins expressed in adult males show an elevated evolutionary rate. These two trends are independent of each other and cannot be explained by peculiar rates of mutation or levels of codon bias. Moreover, the observed patterns appear to hold across several functional classes of genes, although the exact developmental time of the slowest protein evolution differs among each class. We discuss our results in connection with data on the evolution of development.

Key words: Expression profile — Developmental Constraint — Phylotypic stage — Nonsynonymous rate — Synonymous rate

Introduction

The evolutionary rate of a protein depends not only on the rate of mutation but on the likelihood that any particular mutation is deleterious, beneficial, or has such a small effect that it is selectively neutral. The selective effects of mutations in turn depend on how they affect gene function throughout development and ultimately affect the phenotype at the organismal level. Thus, a deep understanding of molecular evolution in multicellular organisms requires integrating information about the patterns and rates of gene evolution with knowledge about when and how these genes function during development (Arthur 2000).

Some authors have proposed that mutations occurring in genes expressed early in development will on average have more deleterious fitness consequences than mutations occurring in genes expressed later on (Arthur 1988; Arthur 1997; Castillo-Davis and Hartl 2002; Powell et al. 1993; Rasmussen 1987; Riedl 1978). Changes to early-expressed genes may have cascading consequences for the later steps in development and thus may lead to large consequences for the adult phenotype. Such phenotypic jumps are unlikely to produce "hopeful monsters" and will usually be strongly deleterious (Fisher 1930; Orr 2000). By contrast, mutations in late-acting genes

*These authors contributed equally to this work.
Correspondence to: Onn Brandman; *email:* Onn@stanford.edu

do not affect the early steps in development and therefore may have a greater chance of leading to small or no phenotypic effect. If these intuitions about the selective effects of mutations are true, then genes functioning in early development should be under stronger purifying selection and thus evolve more slowly than genes which function later on.

Proteins expressed during particular developmental stages may possess unusually fast or slow rates of evolution regardless of how early or late these stages occur. For example, some developmental phases might be highly robust (canalized) (Waddington 1966), to genetic changes (see von Dassow et al. 2000; Von Dassow and Odell 2002) such that alterations to the sequence of genes expressed during these periods will have especially small effects on the phenotype and should pass through the sieve of natural selection at an elevated rate. On the other hand, some developmental stages may be particularly sensitive to genetic change so that an exceptionally high proportion of sequence changes to stage-specific proteins will have profound (and therefore usually deleterious) fitness effects. Such proteins are expected to be under strong purifying selection. Finally, some developmental stages may underlie phenotypic traits that are subject to persistent positive selection. Genes participating in the generation of these traits are likely to have elevated rates of evolution. For example, some male-specific adult genes exhibit especially high rates of sequence evolution (Civetta and Singh 1998; Coulthart and Singh 1988; Swanson et al. 2001; Swanson and Vacquier 2002) likely due to sexual selection acting on male sexual traits.

Investigating the relationship between when a gene is expressed and its rate of evolution may help us understand key questions regarding the evolution of development itself. Taxonomically diverse studies have revealed that the evolutionary conservation of development at the morphologic level has an "hourglass" shape (Raff 1996); that is, both the latest and the earliest periods of development appear more morphologically flexible, while an intermediate developmental stage is often the most conserved (Galis and Metz 2001; Goldstein et al. 1998; Wagner and Misof 1993). It is not clear, however, why the hourglass pattern of morphological conservation exists. One possibility is that the slow phenotypic evolution of particular developmental time points is due to the generally heightened sensitivity of these stages to genetic perturbation. If this hypothesis is true, many genes expressed predominantly during the most conserved periods of development are expected to show reduced rates of sequence evolution. On the other hand, the conservation of an intermediate developmental period may be due to its robustness to genetic perturbation such that changes to stage-specific genes have little phenotypic effect. As a result, a relaxation of purifying selection may be expected for proteins functioning at morphologically conserved developmental stages. Yet a third possibility is that no general relationship exists between evolutionary rate and the time of genes expression. This may be true if the morphological conservation of developmental time points depends on a small number of developmentally important genes (Arthur 1997; Raff 1996). Yet another possibility is that the evolution of development might depend primarily on the rate of gain and loss of genes (Castillo-Davis and Hartl 2002; Domazet-Loso and Tautz 2003) or on the changes in patterns of gene expression (Gellon and McGinnis 1998) rather than on protein sequence evolution.

Previous studies addressing the relationship between the timing of expression and the rates of sequence evolution have largely been limited by experimental design and available data. One study using DNA–DNA hybridizations suggested that genes expressed early in embryonic development may evolve slower than those expressed later on (Powell et al. 1993). The methodology of the study was limited, however, in that it could not discriminate between synonymous and replacement differences. In addition, sequence divergence was subject to saturation at synonymous sites. A more recent study in *Caenorhabditis elegans* failed to find any difference in the rate of protein sequence evolution between genes expressed at different times during development (Castillo-Davis and Hartl 2002). However, the study did provide evidence for a lower rate of duplication for genes expressed during embryogenesis. The failure of the study to find trends regarding the rate of protein evolution may partly be due to the paucity of available data. A more recent study in *C. elegans* using a larger set of expression and evolutionary rate data has revealed that genes expressed during embryogenesis in *C. elegans* appear to evolve at a significantly slower rate than proteins expressed later on (Cutter 2004, personal communication).

To further understand how the rate of gene evolution depends on the developmental context, here we investigate the relationship between the developmental timing of gene expression, as determined by microarray analyses (Arbeitman et al. 2002), and the rate of protein evolution for a large number of *Drosophila* genes. We find a highly significant relationship between the developmental timing of expression of genes and their nonsynonymous rate of evolution. Genes that are expressed earlier in development are likely to have a slower rate of evolution at the protein level than those expressed at later developmental stages. In particular, late embryonic stage genes tend to evolve at the slowest rate, with the exact time of maximum conservation varying across functional protein category. The location of this conserved stage, between early and late development, is reminiscent of the hourglass shape of

the rate of evolution of development itself. In addition, we find that proteins expressed in adults, especially males, have a striking acceleration of molecular evolution possibly due to the action of positive selection on genes involving reproduction.

Methods

Finding Orthologs and Obtaining Measurements of K_S, K_A *and FOP*

For each of the 4028 *D. melanogaster* genes in the expression dataset (Arbeitman et al. 2002), we attempted to find a *D. pseudoobscura* ortholog using alignments of the two genome sequences generated by the Berkeley Genome Pipeline—information found at http://pipeline.lbl.gov/pseudo. This provides an alignment of the whole *D. melanogaster* genome with the scaffolds of the *D. pseudoobscura* genome (which were not completely assembled at the time). For each known or predicted *D. melanogaster* gene, we checked to see if (1) a *D. pseudoobscura* scaffold was aligned with the region and (2) if there was a significant alignment score for the predicted coding region of the gene. This method thus uses information both about synteny and sequence homology to identify orthologs.

To generate sequence alignments we used TBLASTX to align the protein sequence of each predicted exon of the *D. melanogaster* gene with the orthologous region of *D. pseudoobscura*. We removed gaps in these alignments by trimming back from both ends of each gap until an anchor pair was found (following Conery and Lynch 2001) and then replaced the amino acid alignment with the respective nucleotide sequence. We then concatenated the alignment from all exons of a particular gene and used the PAML software package (Yang 1997) to estimate the number of synonymous and nonsynonymous substitutions per site. Whole gene alignments that were fewer than 150 nucleotides in length were removed from the dataset because estimates based on such short alignments may be erroneous.

While our methodology was successful in finding many orthologs, it may have been ineffective in discovering others. At the time we undertook this study, the alignment between the two fly genomes was incomplete and therefore orthologs could not be identified in regions in *D. melanogaster* that were not associated with a consensus sequence in *D. pseudoobscura*. Furthermore, the genome alignments used may be problematic in regions that have undergone complex structural evolution (e.g., segmental duplication). Finally, orthology of especially fast-evolving genes may be difficult to establish because of insufficient sequence similarity. This last assertion is supported by the fact that the average expression profile of the *D. melanogaster* genes for which we found no orthologs is similar to that of faster-evolving genes (supplementary Fig. 1).

We obtained measures of FOP for each gene in *D. melanogaster* using the program CodonW (written by John Peden and available from ftp://molbiol.ox.ac.uk/Win95.codonW.zip). The default *D. melanogaster* preferred codon table included in the package was used.

Expression Data

We used expression data from experiments by Arbeitman et al (Arbeitman et al. 2002). cDNA microarrays were used to analyze the RNA expression levels of 4028 genes in wild-type flies. Samples were collected during 66 sequential time periods beginning at fertilization and spanning the embryonic, larval, and pupal periods and the first 30 days of adulthood, during which males and females were sampled separately. For all of our figures, the time of sample collection for embryonic, larval, and pupal stages is provided in hours, with samples for the embryonic and larval stages collected over a single time course. In addition, embryonic samples were collected over hour intervals; for embryonic time points our *x*-axis denotes the start of these intervals. Each sample was compared to a common reference sample made from pooled mRNA representing all stages of the life cycle. For each time point, the average ratio of each gene's expression to the reference sample was normalized to 1. This normalization step builds on the hypothesis that each time point is likely to have a similar level of expression complexity (Levy and Manning 1981).

Correlation Coefficients

Spearman rank correlation coefficients were used for all of the correlations used in this paper. Because the number of data points used to assess the correlation coefficients and partial correlation coefficients was much greater than 30, we used the following *t*-statistics to calculate significance. $t_s = r\sqrt{\frac{(n-2)}{(1-r^2)}}$ and $t_s = r\sqrt{\frac{(n-2-m)}{(1-r^2)}}$, respectively, where n is the sample size, m is the number of variables held constant, and r is the rank correlation coefficient (Sokal and Rohlf 1995).

Greedy Algorithm to Identify Trend-Causing Genes

Trend strength for a given time interval was calculated as the absolute value of the mean Spearman coefficient between $K_A$and expression at each relevant time point. Initial trend strength was measured for the interval of interest using all genes. Next, a single gene was removed and trend strength measured again. If removing the gene decreased the trend strength, then that gene was classified as a trend-causing gene. This process was repeated for each gene and all trend-causing genes were identified.

Gene Ontology Annotations

Gene Otology information was downloaded from the GO Web sites at http://www.geneontology.org/. From this site we obtained all available functional annotations of the *D. melanogaster* genes in our dataset as well as the functional annotation DAG (directed acyclic graph). Functional annotations for each gene correspond to a particular node in this DAG. In order to analyze the functional composition of a set of genes we wanted to classify them into high-level functional classifications to obtain a substantial number of genes in each category. We obtained each gene's annotation at a specific level in the DAG by using a computer program to trace its annotation provided by the GO Web site up to that level. In some cases a gene can have several, say n, high-level functional classifications, since nodes in the DAG often have multiple parents. In such cases a gene contributed a count of $1/n$ to each category.

Results

Evolutionary rate data

For 2905 of the 4028 *Drosophila melanogaster* genes used in the expression experiments by Arbeitman et al. (2002), we identified orthologs in *D. pseudoobscura* (see Methods) using the full genome alignment made available via the Vista browser (http://pipeline.lbl.gov/pseudo/). We were able to generate suitably long alignments and to obtain estimates of both nonsynonymous substitutions per nonsynonymous

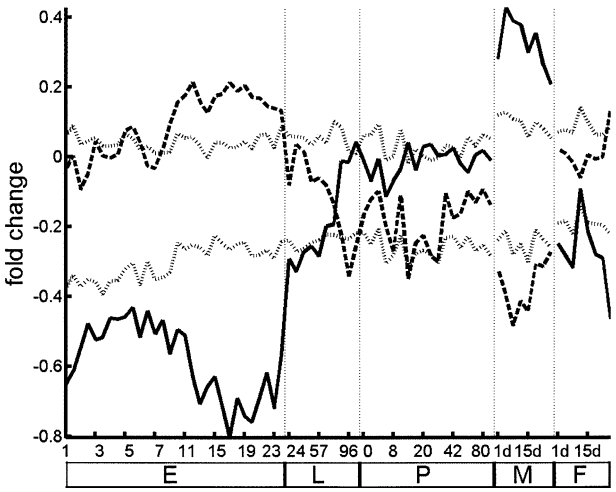

Fig. 1. Mean expression profiles for genes with the highest and lowest K_A values. The solid line was generated from genes with the 150 highest K_A values and the dashed line from the 150 lowest. Dotted lines mark two standard deviations around the mean for a random sample of 150 genes. The Spearman correlation between the two average expression profiles is −0.91. The *x*-axis describes the times at which samples were taken for gene expression analysis. Time points for adult stages were measured in days, whereas hours are used for the other stages (see Methods for details).

site (K_A) and synonymous substitutions per synonymous site (K_S) for 2,617 of these genes (see Methods). For 3% of these alignments one or more exons predicted in *D. melanogaster* were not used because they could not be aligned with the sequence in *D. pseudoobscura*. Different alternative splicing between the two fly species, as noted in previous studies (Clyne et al, 2000; Robertson et al. 2003), may be responsible for a lack of sufficient sequence homology for some exons. The mean number of nonsynonymous substitution per nonsynonymous site (K_A) is 0.107 and the variance in K_A is 0.0073.

The Timing of Protein Expression and Evolutionary Rate

As a first step in determining whether or not there is a relationship between evolutionary rate and time of expression, we plotted the average expression profile for the 150 fastest and the 150 slowest-evolving genes (Fig. 1). The slowest-evolving genes are expressed at significantly high levels during the late embryonic stage (the "late embryo" profile) and have their lowest expression in males. The fastest-evolving genes exhibit the opposite trend, with a low expression during the embryonic phase and high expression during the adult male stage (the "male" profile). The inverse correlation between the expression profiles of the fastest-and the slowest-evolving genes is striking, with a Spearman Rank correlation coefficient of −0.91.

To examine the expression profiles for the entire distribution of K_A values, we grouped genes based on their rate of evolution (150 genes in each bin) and then plotted the average expression profile for the genes in each bin (Fig. 2A). We quantified this relationship further by calculating the Spearman rank correlation coefficient between expression profile and evolutionary rate for each developmental time point (Fig. 2B). A negative correlation of $r_s = -0.10$ to−0.21 exists between expression and K_A during late embryonic stages and a positive correlation, $r_s = 0.11$, during later life stages, with the strongest correlation in adult males. The correlation coefficients for the late embryonic and male stages were significant at the 0.01 level, as assessed by a *t*-statistic (see Methods). Moreover, the nonparametric Spearman rank correlations are most likely conservative. While we use the nonparametric Spearman rank correlations throughout our analysis, it should be noted that a parametric Pearson correlation shows an even more pronounced negative ($r_s = -0.245$) correlation in the late embryonic stage and a stronger positive correlation ($r_s = 0.145$) in the adult male stage.

Figure 2A suggests that the genes with the most extreme evolutionary rate values are not solely responsible for the nonindependence between expression profile and evolutionary rate. We quantitatively verified this result by removing proteins with the most extreme expression profiles and recomputing the correlation coefficients. Even after removing the 400 proteins with the most extreme K_A values (the 200 highest and 200 lowest values) and the 800 proteins with the most extreme rates (the 400 highest and 400 lowest values) correlation coefficients remained significant (supplementary Fig. 2)

In Fig. 2C we plot the average expression of the only three segment polarity and Hox genes found in the expression dataset, *wingless* (*wg*), *antennapedia* (*antp*), and *abdominal-A* (*abdA*). A comparison of this profile with the correlation plot (Fig. 2B), reveals that the strongest negative relationship between expression and evolutionary rate takes place somewhat after the main burst of expression of the segment polarity and Hox genes in embryonic development.

Controlling for K_S *and Codon Bias*

The association between expression and evolutionary rate may indicate that the time of gene expressing directly affects its rate of evolution. On the other hand, this relationship could be mediated by some other properties of genes that are correlated with both expression profile and K_A. One possibility is that genes expressed at particular times in development experience similar rates of mutation and this is the reason for their correlated rates of molecular evolution. This

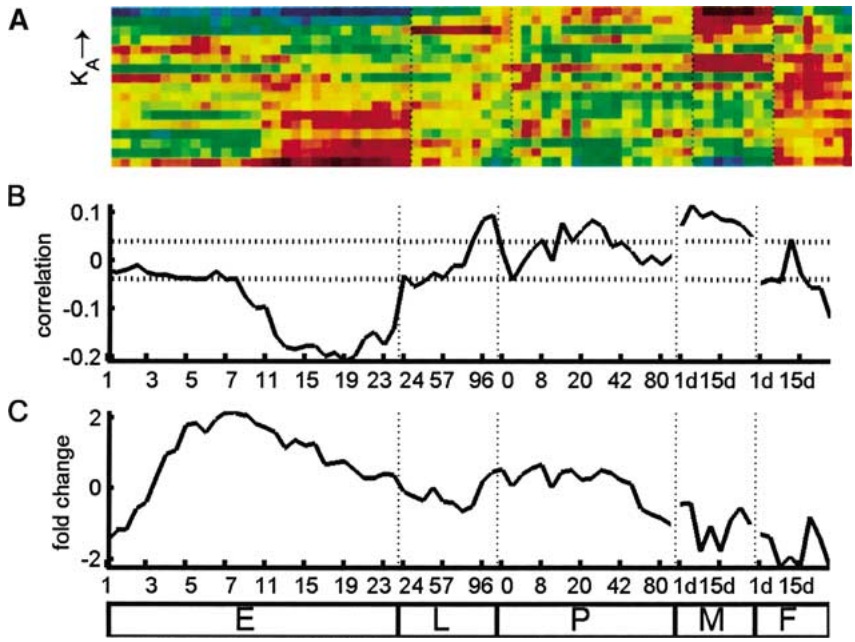

Fig. 2. **A**. The average expression across development is shown for genes of various K_A classes. Genes were sorted by K_A and then partitioned into buckets of size 150. The mean relative expression value for each bucket at each time point was computed and is depicted in the figure via color coding, with blue corresponding to the lowest level of expression, red corresponding to the highest, and a linear scale for colors between. Cutoffs for buckets are provided in supplementary Table 1. **B** The Spearman correlation between relative expression and K_A for each timer point is shown. Dotted lines represent the 95% confidence interval as determined using randomization. **C** Mean expression profiles for three Hox and segment polarity genes, *wingless, antennapedia, and abdominal*-A, found in our data.

scenario is plausible given that genes with similar expression profiles are often found very close to each other on chromosomes (Boutanaev et al. 2002; Spellman and Rubin 2002) and different chromosomal regions experience different rates of mutation. Another possibility is that genes expressed at particular times may have correlated levels of synonymous codon bias, Given that genes with higher codon bias tend to evolve more slowly (Akashi 2001; Pal et al. 2001), codon bias may mediate a spurious relationship between patterns of expression and rates of protein evolution.

To determine whether our results are robust to these parameters, we estimated levels of mutation and codon bias and determined whether a significant trend remains once these factors are corrected for. To estimate the rate of mutation, we measured the rate of synonymous evolution (measured by the number of synonymous substitutions per synonymous site K_S which is known to reflect regional mutation rates (Hartl and Clark 1997). We measured codon bias for each gene using the frequency of optimal codons (FOP) (Ikemura 1981). We then corrected for these factors by calculating the Spearman rank partial correlation profile between K_A and expression at each time-point while holding K_S and FOP constant. The results in Fig. 3 show that codon bias and K_S do not explain the existence of the "late embryo" and "male" patterns and these

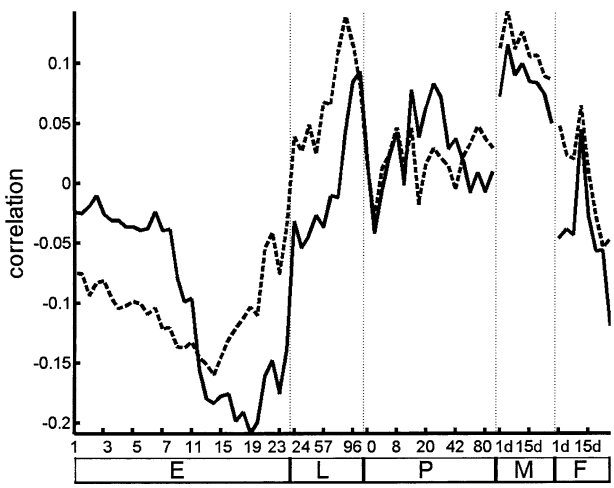

Fig. 3. The direct and partial correlations between expression and K_A. The solid line shows the direct correlation between relative expression and K_A; the dashed line shows the partial correlation between these variables with FOP and K_S held constant.

patterns remain strongly significant. It is worth noting, however, that controlling for these factors does shift the phase of slowest development to a slightly earlier time point during embryonic development and also reveals a negative correlation between expression and the rate of evolution during the earliest developmental stages.

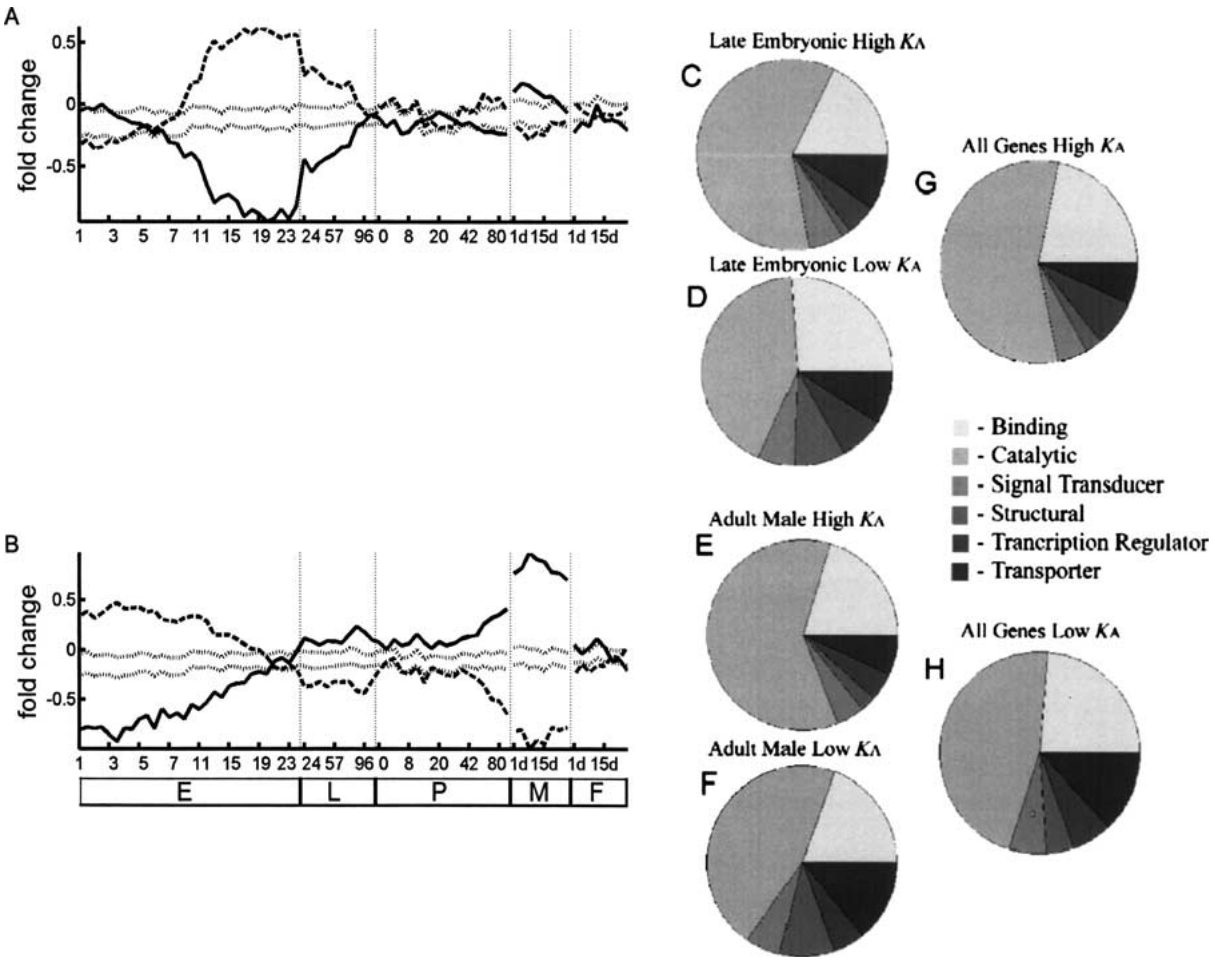

Fig. 4. A greedy algorithm was used to select the genes most responsible for the correlations between expression and evolutionary rate during the late embryonic and the adult male stages. The genes causing each trend are dived into two groups; those below and above the median K_A. The expression profiles for these two groups are shown for the late embryonic trend (**A**) and the male trend (**B**). The proportion of genes in each of these sets that belong to various functional classes is shown in **C**–**F**. **G** and **H** depict the proportion of genes in the entire set that belong to various functional classes.

The Independence of Trends

It is possible that the set of genes responsible, for the ''late embryo'' trend is also responsible for the ''adult male'' trend, perhaps due to a negative correlation in the level of expression of these genes during the late embryonic and adult male stages. To investigate this possibility we used a greedy algorithm (see Methods) to independently identify those genes most responsible for causing the late embryonic trend (from time point 15 to 30) and those most responsible for the male trend (from 58 to 66). In this way, we identified the 1200 and 1250 genes most responsible for generating the ''late embryo'' and the ''male'' trends. The number of genes that are shared by both sets is 600. Based on these numbers, the hypothesis of independence of the two trends cannot be rejected (G-test, $p \sim 0.9$), suggesting that different sets of genes are responsible for the male and the late embryo trends.

To further investigate the genes causing the male and late embryo trends, we divided each set into those genes with K_A values above and below the global median and plotted the average expression profiles (Figs. 4A and B). Figure 4A reveals that the genes causing the late embryonic trend do not systematically have similar expression profiles during any other time in development. Thus, the rate of evolution of genes in this set appears to be determined by their expression during this stage of development and no other. By contrast, the expression profile of genes causing the adult male trend (Fig. 4B) reveals nonrandom expression at other times during development. Specifically, fast-evolving genes also tend to be expressed at moderate levels during the larval and pupal stages of development and expressed at very low levels during the earliest embryonic stages. Because the larvae and pupae used for this experiment included both male and females, it is possible that the high rate of evolution for genes expressed in these stages could

Table 1. The number of genes and rates of evolution for various GO functional classes

Functional group	Mean K_A	Median K_A	No. of genes
Structural	0.065	0.035	60
Transporter	0.071	0.057	159
Binding	0.087	0.071	402
Signal transducer	0.090	0.068	109
Catalytic	0.094	0.074	703
Transcriptional regulator	0.097	0.085	135
All annotated	0.090	0.071	1290
Not annotated	0.12	0.10	1327
All	0.11	0.084	2617

reflect the expression of male specific genes at pre-adult stages. Importantly, it appears that the "late embryo" and "male" trends are independent from each other, whereas the "male" and the weaker "early embryo" trends are due largely to the same genes.

The consistency of Patterns Across Functional Classes

To determine if a small number of functional classes are responsible for the trends we observe or whether the trends hold across all functional classes, we took two approaches. For both approaches we identified proteins in our set belonging to the six largest Gene Ontology (Ashburner et al. 2000) functional classes: *catalytic*, *binding*, *transcription regulator*, *signal transducer*, *structural*, and *transporter* (see Methods). First, we have compared the functional composition of the fast and slowly evolving proteins responsible for the "late embryo" and "male" trends, which we identified in the previous section, to the fast-and slow-evolving proteins from the entire set (Figs. 4C–H). If the relationship between the timing of expression and evolutionary rate is confined to particular functional classes, proteins belonging to these classes should be overrepresented in the set of trend causing genes we identified. Our analysis reveals that this is not the case. The fast and slow-evolving proteins causing both trends (Figs. 4C–F) appear to span all functional classes. The differences in functional constitution of gene sets with high and low K_A values in Fig. 4 (C vs D and E vs F) are expected given that different functional classes have different mean rates of evolution (and the mean K_A of each functional class is provided in Table 1). Moreover, the proportion of proteins causing each of the trends is roughly proportional to the functional composition of the entire set of genes (Figs. 4C and E vs G, Figs. 4D and F vs H). It therefore appears that the late embryonic and the adult male trends do not merely reflect a relationship between expression and rate of evolution for a few developmental proteins belonging to particular functional classes but reflect a relationship that exists across proteins of all function.

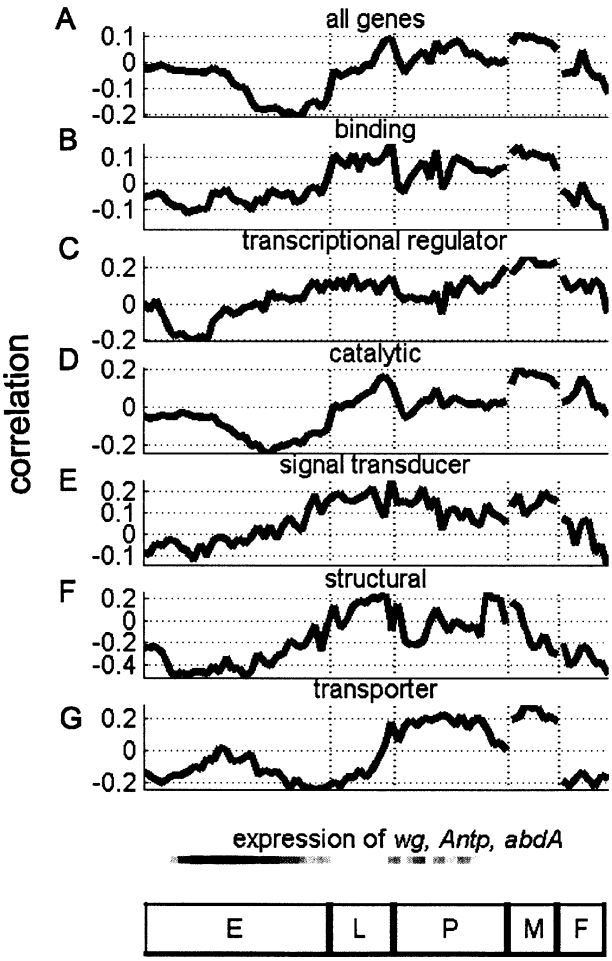

Fig. 5. Spearman correlations between K_A and expression at each time point are shown for all genes and genes that belong to the six most common Gene Ontology functional classes. The shaded bar below shows the average expression profile of three Hox and segment polarity genes, with darker shading indicating higher expression.

As a second approach to assessing the generality of our findings for all functional classes, we analyzed the Spearman rank correlation coefficients between expression and rate of evolution for genes in each functional class independently (Figs. 5A–G). Each of these plots reveals the same general trend—genes expressed at some intermediate stage in the embryo tend to evolve slowly, whereas genes expressed in the later developmental stages evolve more rapidly. Interestingly, the period with the strongest negative correlation does not come at the same developmental time point for each of the functional classes. The *transcription regulators*, the class that contains many of the classic developmental program genes, have their strongest negative correlation during time points 8-12, which corresponds exactly to the stage of development in which Hox and segment polarity genes are expressed and which is most evolutionarily conserved at the morphologic level (Campos-Ortega and Hartenstein 1997; Lawrence 1992). Similarly, the *binding,*

signal transduction, and *structural* classes appear to have the strongest negative correlation between expression and evolutionary rate at approximately the same time-point and are constrained at similar levels for the rest of embryonic development. Maximum constraint occurs somewhat later during development for the *catalytic* and *transporter* functional classes.

Importantly, separating genes into functional categories in some cases leads to higher rather than lower correlation coefficients. For example, there is a very strong negative correlation (Spearman rank correlation, $r_S = -0.5$) between expression of structural proteins in the late embryonic stages and evolutionary rate. This implies that all functional classes of genes exhibit a similar relationship between expression and evolutionary rate, and moreover, that analyzing genes of all functional classes together obscures the true relationships between expression and evolutionary rate.

Adult Trends: Fast Evolution of Male-Specific Proteins and Slow Evolution of Maternal Genes

In addition to the "late embryo" trend we also identified an independent positive correlation between the rate of evolution and the level of gene expression in adult males. In contrast, the adult female stage reveals little correlation between these two variables during early female stages and a negative correlation for the latest female stages. Several explanations might work together to account for the observed adult trends. First, the fast evolution of male-specific genes (Civetta and Singh 1998; Coulthart and Singh 1988; Swanson et al. 2001) may explain the positive correlation between expression and rate of evolution in males. Second, the high expression of maternal proteins (especially in older, actively egg-laying females) may account for the negative correlation between expression and the rate of evolution in the late adult female stage. Indeed, maternally derived genes are likely to have a slow rate of evolution as do the majority of genes which are specific to the early embryonic stage. Finally, the fast evolution of adult-specific genes (expressed in both sexes) may, at least partially, account for the positive correlation between expression and evolutionary rate in adult males. To test each of these explanations we attempted first to identify maternal genes and then, after filtering out these genes, to compare rates of evolution for male-specific, female-specific, and adult-specific genes.

On the assumption that older females express proportionately more maternal genes compared to younger females, we separated genes into those expressed in young (day 1) and old (day 30) females. Consistent with this assumption, there is a strong positive correlation (Spearman rank correlation, $r_S = 0.49$) for gene expression in the old females and early embryos (1 h), but no such correlation for the young females and early embryos (Spearman rank correlation, $r_S = -0.04$).

We can use the difference between young and old females for the purpose of filtering out, albeit crudely, many of the maternal genes. We classify maternal genes as those expressed at lower than median levels in males and with increased expression in older females compared to younger females. As expected, such genes evolve slowly ($K_A = 0.092$), and significantly so ($p \sim 10^{-6}$, randomization test). We can use similar logic to identify adult-specific genes (genes expressed at higher than median levels in males and young females and that have a higher expression in young females compared to old females), male-specific genes (higher than median male expression, lower than median expression in young and old females), and female-specinc genes (lower than median expression in males, higher than median expression in young females, and higher expression in young females compared to old females). The average K_A is slightly higher for adult-specific and female-specific genes ($K_A = 0.110$ and 0.114, respectively), however, neither of these values is significant ($p = 0.21$, $p = 0.26$, randomization test). The male-specific genes, on the other hand, evolve much faster as a group, with an average K_A of 0.126. This difference is strongly significant ($p = 4 \cdot 10^{-6}$, randomization test) suggesting that male-specific proteins experience an increased rate of evolution.

Discussion

Timing of Gene Expression and the Rate of Protein Evolution

In this study we investigate the rate of molecular evolution of proteins in relation to the timing of their expression in *Drosophila* development. Previous theoretical and empirical work has suggested at least two, not mutually exclusive, reasons why the timing of a protein's expression might affect its rate of evolution. The first possibility is that since later stages of development depend on the successful execution of earlier ones, proteins expressed earlier in development should evolve slower than those expressed later (Arthur 1988, 1997; Riedl 1978). A second hypothesis argues that some periods of development may be more or less sensitive to perturbation (mutation) than others because of different stage-specific architectures (Goodwin et al. 1993; Raff 1996; von Dassow et al. 2000). The varying sensitivity to perturbation may be reflected in the rate of sequence evolution of proteins primarily expressed during those stages.

We compared rates of protein evolution and developmental expression profiles using a large set of 2617 proteins in *Drosophila* for which clear orthologs between *D. melanogaster* and *D. pseudoobscura* could be identified and time course expression data were available (Arbeitman et al. 2002). The expression profiles generated by Arbeitman et al. provide an estimate of relative expression differences for any particular gene at different times of development. Although we do not have information about the absolute abundance of transcripts or proteins, these data do provide us with a good way to assess the variation of protein abundance across development (see methods). Our analysis reveals a highly significant relationship between expression profile and evolutionary rate (Figs. 2A, B). This result cannot be explained by differences in codon bias or mutation rates among proteins (Fig. 3) and remains significant even when genes with the most extreme evolutionary rates are removed (supplementary Fig. 2).

When evaluating the relationship between expression and evolutionary rate within each of the six major functional classes of genes (Fig. 5), similar trends emerge. Individual functional classes often show stronger correlations than the total set of proteins. This strength is likely due to the fact that separating genes into various functional classes removes "noise" from the relationship between timing of expression and rate of evolution. Because some functional classes have particularly fast or slow rates of evolution as a whole, as evidenced by the different functional constituencies in Figs. 4C–H and previous work (Hartl and Clark 1997), the relationship between the rate of evolution and the timing of expression is obscured when all functional classes are analyzed together.

The finding that proteins expressed during embryogenesis have slower rates of evolution than those expressed later in development roughly agrees with long-standing speculation (Riedl 1978) and supports preliminary conclusions from previous empirical work (Powell et al. 1993). Our data also suggest that there is something peculiar about mid–late embryonic development that causes genes expressed at this stage to evolve more slowly than those expressed at slightly earlier and later stages. In addition to trends in early development, our data reveal that proteins expressed in adult males have unusually fast rates of evolution.

Rate of Protein Evolution and the Evolution of Drosophila *Development*

Taxonomically diverse studies of ontogeny have revealed that at a morphological level the conservation of development appears to have an "hourglass" shape (Raff 1996); early and late stages of development appear to be somewhat variable among taxa, whereas an intermediate developmental stage is often the most conserved (Galis and Metz 2001; Goldstein et al. 1998; Wagner and Misof 1993). Two theories have been proposed to explain the conservation of this intermediate stage at the morphologic level (Galis et al. 2002). One explanation is that this stage is robust to genetic perturbation so that even if mutational changes occur they do not impact the phenotype of this stage. A second explanation is that the intermediate stage is subject to strong stabilizing selection. That is, mutations that alter the phenotype of this stage are strongly deleterious and thus removed by natural selection. Importantly, these mechanisms are not mutually exclusive; both may work together to generate the morphological conservation of the intermediate developmental stage.

The relative degree to which these mechanisms work to preserve the most conserved stage of development may be revealed by evolutionary divergence of genetic networks and the sequences of genes specific to that stage. Assuming that most mutations are slightly deleterious, an elevated robustness of the intermediate stage should lead to an increased tolerance to mutations which alter the function of stage-specific genes; these genes may therefore be likely to accumulate nonsynonymous mutations or even be lost over the course of evolution. On the other hand, if stabilizing selection alone acts to preserve the morphology of the intermediate time point, then the genetics are more likely to be conserved both in terms of gene identity and sequence. Finally, if both mechanisms work together, then genes expressed at this stage may diverge more or less rapidly than genes at other stages depending on the relative effects of robustness and stabilizing selection.

Studies of the conservation of developmental genes and genetic networks provide some evidence that stabilizing selection is acting more strongly at this intermediate stage of development. In the case of *Drosophila*, for example, genetic networks acting in the earliest stages of development evolve surprisingly quickly; *bicoid*, one of the first genes acting in *Drosophila* development, appears to be specific to flies (Wilkins 1997). By contrast, networks functioning at an intermediate stage in development such as the segment polarity genes (e.g., *engrailed* and *wingless*) and Hox genes are a universal feature of Arthropoda (Wilkins 1997). Thus it appears that genes specific to the morphologically conserved, intermediate stage of development are maintained across organisms presumably because of their indispensability.

In addition to the preservation of genes, a slow rate of sequence evolution of genes functioning at a particular time point may reflect a higher degree of stage-specific stabilizing selection. Our analysis here

reveals that proteins expressed after the main burst of expression of segment polarity and Hox genes evolve at particularly slow rates (Figs. 2B and C). Thus our findings provide further evidence that genes expressed during the most morphologically conserved stage are under a high degree of stabilizing selection. Moreover, our analysis reveals that evidence for stabilizing selection at the genetic level is not confined to only a small number of genes, but is revealed when all genes are taken together and is robust to removing genes with the most extreme evolutionary rates (supplementary Fig. 2). Separating genes into functional classes provides even an even more detailed picture of stabilizing selection at the level of sequence evolution (Fig. 5). The most constrained transcription factors and signal transducers, the functional group that contains many developmentally key genes, are expressed precisely during the time of expression of the segment polarity and Hox genes. For other functional categories, the strongest correlation between evolutionary rate and expression occurs at a slightly later stage in embryonic development.

Although heightened stabilizing selection during the most morphologically conserved stage of development is the most straightforward interpretation of our data, one alternative hypothesis deserves mentioning. It is possible that the slow evolutionary rate of genes expressed during the segment polarity stage is not due to strong stabilizing selection on their function during this developmental time point but results from selection on their function at some other stage when they are expressed. Our analysis reveals, however, that expression at the intermediate stage of development is not correlated with expression and any other time point (Fig. 4A), and thus it is difficult to imagine how this alternative hypothesis can explain the strong correlation between rate of evolution and expression and the late embryonic stage.

It should be noted that our data can neither confirm nor refute the hypothesis that robustness is also partially responsible for the conservation of this intermediate stage. Robustness and stabilizing selection may be working together. If this is the case, the strength of stabilizing selection must be considerable since its signal can be observed over the molecular signal of robustness which is likely to militate against the slow rate of evolution. Future evidence suggesting high levels of robustness of the late embryonic stages will render our conclusions about strong stabilizing selection acting at those times conservative.

The Adult Stages; Male and Female-Specific Trends

In adult females, the latest stages (20–30 days) exhibit a significant negative correlation between expression and evolution. Our analysis reveals that the expression of maternal genes is the most likely explanation for this relationship. Such genes are both slowly evolving and preferentially expressed in the late females stages. As discussed above, we believe that an increase in purifying selection is responsible for the slow rate of evolution of proteins expressed in the mid–late embryo. For similar reasons, we believe that an increase in purifying selection is the most likely explanation for the slow evolutionary pace of maternal genes.

The adult male stage exhibits an even more consistent relationship between level of expression and rate of evolution. This stage exhibits the strongest positive relationship between these two variables in our data, and moreover, the genes that cause the trend are independent from the genes responsible for the trend in embryonic development (Figs. 4A, B). In order to determine whether male-specific or adult-specific genes cause the trend, we attempted to identify genes in both of these categories. Our analysis reveals that male-specific genes, rather than adult-specific genes, have an elevated evolutionary rate. One explanation for this pattern is that a relaxation of purifying selection acting on male specific genes is responsible for their increased rate of evolution. A second explanation is that the elevated evolutionary rate of male-specific proteins is due to the action of positive selection. Indeed, models of sperm competition (Clark et al. 1999; Wyckoff et al. 2000) and sexual conflict (Gavrilets 2000; Rice 1996) predict, and empirical results have suggested (e.g., Civetta and Singh 1998; Coulthart and Singh 1988; Swanson et al. 2001), that many male-specific proteins involved in reproduction experience strong adaptive evolution at the sequence level. Using current data, we cannot determine which of these explanations accounts for the positive correlation between male-specific expression and evolutionary rate. In the future, polymorphism data, which can distinguish between the action of positive selection and a relaxation of purifying selection, should help resolve this question.

Conclusions

Our results reveal a systematic relationship between the timing of gene expression and the rate of coding sequence evolution. Genes expressed early on, and during the late embryonic stage, in particular, tend to have the slowest rates of evolution, while those expressed at later stages and in adult males have considerably faster rates of evolution. This relationship appears to be consistent across many genes and to be robust to systematic differences in rates of mutation and codon bias. Moreover, the relationship holds independently for various functional classes. While a connection between evolution at the molecular level

and developmental processes has been advocated for some time (Arthur 2000; Riedl 1978), and while several previous studies have attempted to find a relationship between timing of expression and rate of evolution (Castillo-Davis and Haitl 2002; Powell et al. 1993), this is the first study to demonstrate such a relationship on the genomic scale.

Our findings add to the suite of known "rules" governing the rate at which proteins evolve. Recent genome-scale studies have shown that a protein's level of expression and number of interactions are negatively correlated with its rate of evolution in yeast (Fraser et al. 2002; Pal et al. 2001). In addition, other studies have shown that protein dispensability in bacteria, yeast, and nematode is correlated with the rate of protein evolution (Cutter et al. 2003; Hirsh and Fraser 2001; Jordan et al. 2002). Our observations suggest that the timing of expression during development is another significant, and possibly independent, factor that affects the evolutionary rate of proteins.

Acknowledgments. We would like to thank Mark Siegal, Michelle Arbeitman, Bruce Baker, Cristian Castillo-Davis, Asher Cutter, Tobias Meyer, and Nadia Singh for helpful advice and discussions regarding the manuscript and Michael Brudno for technical advice. This work was funded by NSF predoctoral Fellowships awarded to J.C.D. and O.B. and the Sloan Research Fellowship awarded to D.A.P.

References

Akashi H (2001) Gene expression and molecular evolution. Curr Opin Genet Dev 11:660–666

Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP (2002) Gene expression during the life cycle of *Drosophila melanogaster*. Science 297:2270–2275

Arthur W (1988) A theory of the evolution of development. Wiley Chichester, UK

Arthur W (1997) The origin of animal body plans: A study in evolutionary developmental biology. Cambridge University Press, Cambridge

Arthur W (2000) The concept of developmental reprogramming and the quest for an inclusive theory of evolutionary mechanisms. Evol Dev 2:49–57

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI (2002) Large clusters of co-expressed genes in the Drosophila genome. Nature 420:666–669

Campos-Ortega JA, Hartenstein V (1997) The embryonic development of *Drosophila melanogaster*. Springer, Berlin

Castillo-Davis CI, Hartl DL (2002) Genome evolution and developmental constraint in *Caenorhabditis elegans*. Mol Biol Evol 19:728–735

Civetta A, Singh RS (1998) Sex-related genes, directional sexual selection, and speciation. Mol Biol Evol 15:901–909

Clark AG, Begun DJ, Prout T (1999) Female × male interactions in Drosophila sperm competition. Science 283:217–220

Clyne PJ, Warr CG, Carlson JR (2000) Candidate taste receptors in Drosophila. Science 287:1830–1834

Conery JS, Lynch M (2001) Nucleotide substitutions and the evolution of duplicate genes. Pac Symp Biocomput pp 167–178

Coulthart MB, Singh RS (1988) High level of divergence of male-reproductive tract proteins, between *Drosophila melanogaster* and its sibling species, D. simulans. Mol Biol Evol 5:182–191

Cutter AD, Payseur BA, Salcedo T, Estes AM, Good JM, Wood E, Hartl T, Maughan H, Strempel J, Wang B, Bryan AC, Dellos M (2003) Molecular correlates of genes exhibiting RNAi phenotypes in *Caenorhabditis elegans*. Genome Res 13: 2651–2657

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. Genome Res 13:2213–2219

Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296:750–752

Galies F, Metz JA (2001) Testing the vulnerability of the phylotypic stage: On modularity and evolutionary conservation. J Exp Zool 291:195–204

Galis F, van Dooren TJ, Metz JA (2002) Conservation of the segmented germband stage: robustness or pleiotropy? Trends Genet 18:504–509

Gavrilets S (2000) Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889

Gellon G, McGinnis W (1998) Shaping animal body plans in development and evolution by modulation of Hox expression patterns. Bioessays 20:116–125

Goldstein B, Frisse LM, Thomas WK (1998) Embryonic axis specification in nematodes: evolution of the first step in development. Curr Biol 8:157–160

Goodwill BC, Kauffman S, Murray JD (1993) Is morphogenesis an intrinsically robust process? J Theor Biol 163:135–144

Hartl DL, Clark AG (1997) Principles of population genetics. Sinauer Associates, Sunderland, MA

Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. Nature 411:1046–1049

Ikemura T (1981) Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J Mol Biol 151:389–409

Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–968

Lawrence PA (1992) The making of a fly: The genetics of animal design. Blackwell Science Oxford

Levy LS, Manning JE (1981) Messenger RNA sequence complexity and homology in developmental stages of Drosophila. Dev Biol 85:141–149

Orr HA (2000) Adaptation and the cost of complexity. Evolution Int J Org Evolution 54:13–20

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

Powell JR, Caccone A, Gleason JM, Nigro L (1993) Rates of DNA evolution in Drosophila depend on function and developmental stage of expression. Genetics 133:291–298

Raff RA (1996) The shape of life: Genes, development, and the evolution of animal form. University of Chicago Press, Chicago

Rasmussen N (1996) A new model of developmental constraints as applied to the Drosophila system. J Theor Biol 127:271–299

Rice WR (1996) Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. Nature 381:232–234

Riedl R (1978) Order in living organisms: A systems analysis of evolution. Wiley, New York

Robertson HM, Warr CG, Carlson JR (2003) Molecular evolution of the insect chemoreceptor gene superfamily in *Drosophila melanogaster*. Proc Natl Acad Sci USA 100(Suppl 2):14537–14542

Sokal RR, Rohlf FJ (1995) Biometry: The principles and practice of statistics in biological research. WH Freeman, New York

Spellman PT, Rubin GM (2002) Evidence for large domains of similarly expressed genes in the Drosophila genome. J Biol 1:5

Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. Nat Rev Genet 3:137–144

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci 98:7375–7379

Von Dassow G , Odell GM (2002) Design and constraints of the Drosophila segment polarity module: robust spatial patterning emerges from intertwined cell state switches. J Exp Zool 294:179–215

von Dassow G , Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. Nature 406:188–192

Waddington CH (1966) Principles of development and differentiation. Macmillan, New York

Wagner GP, Misof BY (1993) How can a character be developmentally constrained despite variation in developmental pathways? JEvolut Biol 6:449–455

Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309

Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556