Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes

Andreas Heger and Chris P. Ponting

Genome Res. 2007 17: 1837-1849 originally published online November 7, 2007
Access the most recent version at doi:10.1101/gr.6249707

Evolutionary rate analyses of orthologs and paralogs from 12 *Drosophila* genomes

Andreas Heger[1] and Chris P. Ponting

MRC Functional Genetics Unit, University of Oxford, Department of Physiology, Anatomy and Genetics, Oxford OX1 3QX, United Kingdom

The newly sequenced genome sequences of 11 *Drosophila* species provide the first opportunity to investigate variations in evolutionary rates across a clade of closely related species. Protein-coding genes were predicted using established *Drosophila melanogaster* genes as templates, with recovery rates ranging from 81%–97% depending on species divergence and on genome assembly quality. Orthology and paralogy assignments were shown to be self-consistent among the different *Drosophila* species and to be consistent with regions of conserved gene order (synteny blocks). Next, we investigated the rates of diversification among these species' gene repertoires with respect to amino acid substitutions and to gene duplications. Constraints on amino acid sequences appear to have been most pronounced on *D. ananassae* and least pronounced on *D. simulans* and *D. erecta* terminal lineages. Codons predicted to have been subject to positive selection were found to be significantly over-represented among genes with roles in immune response and RNA metabolism, with the latter category including each subunit of the Dicer-2/r2d2 heterodimer. The vast majority of gene duplications (96.5%) and synteny rearrangements were found to occur, as expected, within single Müller elements. We show that the rate of ancient gene duplications was relatively uniform. However, gene duplications in terminal lineages are strongly skewed toward very recent events, consistent with either a rapid-birth and rapid-death model or the presence of large proportions of copy number variable genes in these *Drosophila* populations. Duplications were significantly more frequent among trypsin-like proteases and DM8 putative lipid-binding domain proteins.

[Supplemental material is available online at www.genome.org. Multiple alignments, species trees, and orthologous groups can be found at http://genserv.anat.ox.ac.uk/clades/flies.]

Of all species, the fruit fly *Drosophila melanogaster* has perhaps best illuminated the conserved biology of animals. Not only is *Drosophila* an organism of choice in evolutionary genetics, population genetics, and ecology (Rubin and Lewis 2000), it is also fast becoming one in comparative genomics. To add to the accurate, comprehensive, and well-annotated euchromatic genome of *D. melanogaster* (Ashburner and Bergman 2005), there are now 11 other *Drosophila* genomes that recently have been sequenced and assembled (Richards et al. 2005; *Drosophila* 12 Genomes Consortium 2007). These species sample different branches of the *Drosophila* phylogeny. Relative to *D. melanogaster*, four (*D. willistoni, D. grimshawi, D. virilis*, and *D. mojavensis*) are divergent species, two (*D. pseudoobscura* and *D. persimilis*) are from the obscura group, and five close relatives (*D. simulans, D. sechellia, D. yakuba, D. erecta*, and *D. ananassae*) are from the *melanogaster* subgroup.

This broad span of species presents unprecedented opportunities to investigate the evolution, not of a pair, or a few, species as hitherto, but of a diverse clade of species, each associated with very different habitats, morphologies, and behaviors. These species' genome sequences are expected to assist the functional annotation of the *D. melanogaster* genome and to inform on evolutionary issues such as speciation. However, the progression from analyzing a pair of genome sequences to analyzing a dozen presents substantial challenges, owing to the quadratic increase in the number of sequence comparisons. Previously simple inferences, such as ortholog assignment between a species pair, suddenly necessitate fully phylogenetic approaches when several genomes are considered. Indeed, methodological advances stemming from the sequencing and analysis of these dozen fruit fly genomes are expected, in time, to directly benefit analyses of multiple mammalian genomes (http://flybase.net/data/docs/CommunityWhitePapers/GenomesWP2003.html).

The challenges of the multiple fruit fly genome sequencing project are manifold. These genomes have been sequenced by different centers and often they have been assembled using different algorithms; their statistical coverage of sequencing varies from 3- to 12-fold (Table 1), which results in different degrees of incompleteness and error; and their divergences range from slight to substantial. Nevertheless, to provide objective comparisons of these genomes and their genes it is essential that single annotation and analytical approaches ("pipelines") are applied equally to them all to avoid methodological biases.

We were interested in extending our approaches, previously applied only to pairs of genomes (Waterston et al. 2002; Gibbs et al. 2004; International Human Genome Sequencing Center 2004; Goodstadt and Ponting 2006; Goodstadt et al. 2007), for predicting genes, orthologs, and paralogs of these dozen fruit fly genomes and inferring from them differences in selective constraints on genes and on their proteins' amino acids. We first needed to construct a novel gene prediction pipeline to apply to each genome in turn because our usual source of such predictions, Ensembl (Birney et al. 2006), was not a contributor to this project. Then, we needed to extend from two genomes, to many, our previously described phylogenetic approach (PhyOP, Goodstadt and Ponting 2006) to inferring orthology, paralogy, and conserved synteny. Subsequently, we made these predictions available via the World Wide Web (http://wwwfgu.anat.ox.ac.uk/flies/).

[1]**Corresponding author.**
E-mail Andreas.Heger@dpag.ox.ac.uk; fax 44-1865-285862.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6249707. Freely available online through the *Genome Research* Open Access option.

Table 1. Gene prediction for 11 *Drosophila* species' genomes

	Genomes						Recall of *D. melanogaster* templates[a]						Number of predicted genes		
Species	Contigs	Size/ Mb	Coverage	No. of gaps	Nucleotides in gaps/Mb		*D. melanogaster* genes		*D. melanogaster* transcripts		Median identity[b]	Median coverage[b]	Genes	Pseudogenes	Total
D. simulans	17	142	~3-fold + 6-fold[c]	23,640	15	11%	13,228	96%	14,735	92%	97	100	9,717	3,108	12,825
D. sechellia	14,730	167	~3-fold	6,695	9	6%	13,448	97%	15,155	95%	97	100	11,261	3,773	15,034
D. yakuba	20	169	~6-fold	13,502	6	4%	13,305	96%	14,984	94%	95	100	12,646	1,871	14,517
D. erecta	5,124	153	~12-fold	2,486	8	5%	13,229	96%	14,911	93%	95	100	12,190	1,466	13,656
D. ananassae	13,749	231	~8-fold	6,783	17	7%	12,535	91%	14,038	88%	85	100	11,841	1,838	13,679
D. pseudoobscura	4,896	153	~7-fold	8,729	7	4%	11,951	86%	13,389	84%	81	99	10,726	1,086	11,812
D. persimilis	12,838	188	~4-fold	13,975	13	7%	12,067	87%	13,484	84%	80	99	8,968	3,382	12,350
D. willistonis	14,927	237	~6-fold	5,716	12	5%	11,213	81%	12,572	79%	77	99	10,536	1,501	12,037
D. virilis	13,530	206	~9-fold	4,852	17	8%	11,661	84%	13,042	82%	77	99	10,082	1,336	11,418
D. mojavensis	6,841	194	~8-fold	5,033	14	7%	11,541	83%	12,891	81%	76	99	9,806	1,321	11,127
D. grimshawi	17,440	200	~8-fold	6,717	14	7%	11,469	83%	12,849	80%	76	99	10,058	1,222	11,280

Analysis is based on 19,369 *D. melanogaster* transcripts from 13,836 *D. melanogaster* genes. Only genes with conserved gene structure are considered, where predicted genes with conserved gene structure contain at least two exons with conserved exon boundaries or are single-exon predictions stemming from single-exon templates. Pseudogenes are predictions with disruptions that contain at least one in-frame stop codon or frameshift.

[a]Transcripts/gene in *D. melanogaster* with matches in target genome.

[b]Between template and best prediction in percent.

[c]Approximately threefold whole genome shotgun (WGS) of w501 strain, onefold coverage of six other strains.

The principal advantage afforded by the 12 *Drosophila* genomes is that evolutionary analyses, previously necessarily confined to small data sets, are now comprehensive. From our predicted sets of genes, orthologs, and paralogs, we sought to understand the divergences and the topology of the *Drosophila* species' phylogeny, using the estimated number of synonymous substitutions at silent sites as a molecular clock. In a companion paper, we discuss the evolution of codon bias in this clade (Heger and Ponting 2007). Here, using the species phylogeny, we consider how selective pressures vary among different fruit flies, and among their chromosomes, genes, and codons.

Results

Gene prediction

Recovery of template transcripts

The majority of the 19,369 transcripts and 13,836 genes from *D. melanogaster* aligned with high coverage and percent identity to each of the 11 other genomes (Table 1). The recovery rate per species was dependent on the evolutionary distance between its genome and that of *D. melanogaster*. The highest recovery rates, up to 97% for genes and up to 92% for transcripts, were achieved for the most closely related species *D. simulans* and *D. sechellia*. The recovery rate dropped to 81% for genes and 79% for transcripts among the species furthest diverged from *D. melanogaster*.

The majority of predictions spanned more than 90% of the template sequence (Fig. 1A), but occasionally coverage dropped to as low as 20%. Sequence identity between a template and its best prediction peaked at high percent identities for species closely related to *D. melanogaster*. For further diverged species, however, the distribution was broader and peaked at 80%–95% identity, with a sizable number of predictions at low percent identity but high coverage. There were no predictions of less than 30% identity, so we expected our procedure to fail for the most rapidly evolving genes.

Quality control

We used three measures to assess whether predictions were accurate: (1) the presence of frameshifts and/or stop codons; (2) the coverage of the template sequence, i.e., how many residues of the template sequence can be aligned to a predicted gene; and (3) the conservation of exon boundaries. On the basis of these three properties, we grouped predictions into a set of 15 categories. The categories were ranked from putative ortholog predictions with conserved gene structure down to pseudogenes and fragments (Supplemental Table S1). The quality of a gene prediction was determined by its highest ranking transcript.

We predicted 8968 to 12,579 genes in each genome assembly (Fig, 1B; Table 1) that contain no frameshifts or in-frame stop codons, that align to at least 80% of the template sequence, and that have partially or fully conserved exon structures. The proportion of genes with fully conserved exon structure was dependent on the divergence between template and target genome and dropped from at least 80% for species closely related to *D. melanogaster* to 53% for the more distantly related *D. grimshawi, D. willistoni*, and *D. mojavensis* species. We estimated that at least one quarter of apparent changes in gene structure are due to assembly or prediction artifacts (see Supplemental materials).

As expected, the quality of a genome assembly directly affected the quality of its predicted transcripts. We observed a relatively low number of conserved and partially conserved genes in *D. simulans, D. sechellia*, and *D. persimilis,* which were balanced by a corresponding increase in the number of predicted pseudogenes with conserved or partially conserved exon structure. These three genomes differed from other assemblies in sequence coverage (that for *D. sechellia* was threefold and that for *D. persimilis* was fourfold) or in assembly process (*D. simulans* is a mosaic assembly from multiple strains). Many of these predicted pseudogenes will thus prove to be full-length genes when these assemblies are more accurately known.

Differences in gene structure between template and target were mostly due to dubious exon predictions (see below), deleted

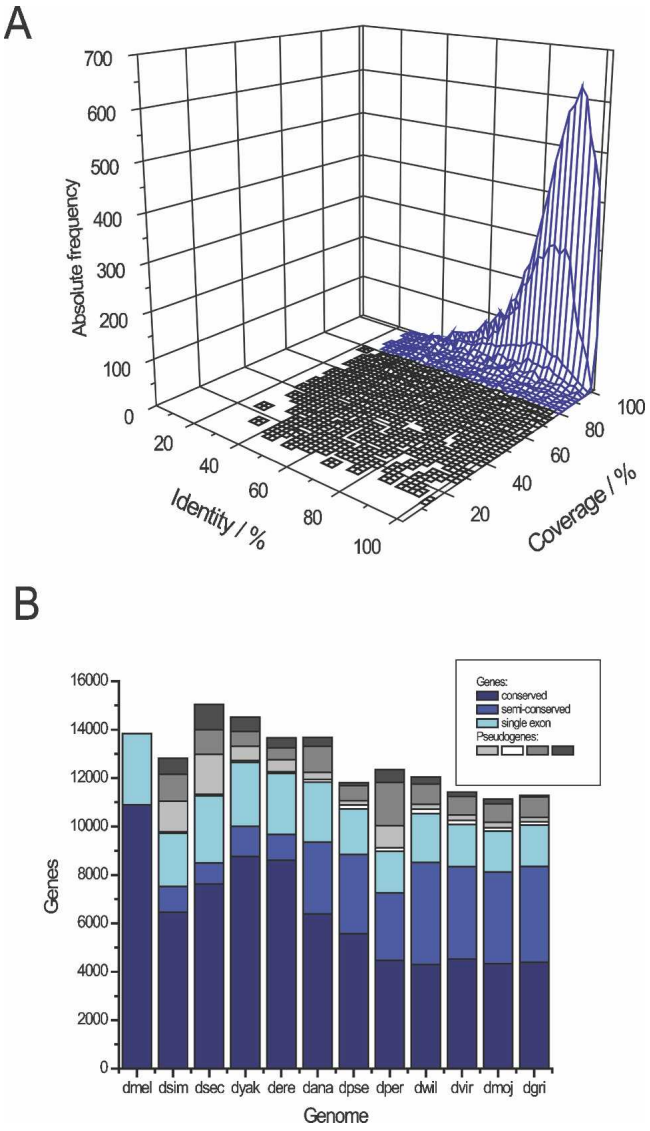

Figure 1. Gene prediction results. (*A*) Two-dimensional histogram of percentage identity and alignment coverage of *D. melanogaster* transcripts to their best matching predictions in *D. pseudoobscura*. Transcripts predicted with conserved gene structure and >80% coverage were retained for further analysis, the remainder were removed. (*B*) Numbers of predicted genes in all fly genomes. Genes with conserved or partially conserved gene structure are shown in blue shades, pseudogenes are shown in gray shades indicating conservation of gene structure: conserved (light), partially conserved (medium), single exon (dark), and retrotransposed (white). Species names have been abbreviated.

introns, and missed terminal exons. Between 6% and 11% of predictions missed a terminal exon, where N-terminal exons are more likely to be absent than C-terminal exons. Internal exons were never entirely absent in predictions since Exonerate tends to produce alignments accommodating all exons even if orthologous exons are not present in the assembly (i.e., are absent because of an assembly gap). In all species, a predicted transcript was twice as likely to contain an inserted intron as a deleted intron when compared with its template transcript.

Dubious exons are those exhibiting low sequence identity to the template compared with the other exons in the predicted transcript. The presence of dubious exons is an indicator that the alignment in this region is likely to contain errors. Between 6% and 8% of all predictions, with conserved gene structure, in a species closely related to *D. melanogaster* contained such dubious exons. This proportion rose to 26% for predictions in species further diverged from *D. melanogaster*.

Overall, we concluded that gene prediction by homology yields high-quality gene predictions for the additional 11 *Drosophila* genomes. Although predicted transcripts showed some variation in gene structure compared with their templates, transcripts in other species will need to be validated experimentally before conclusions concerning gene structure evolution can be drawn. The Supplemental material contains more extensive discussions of these gene prediction results.

Orthology assignment

Orthologs to D. melanogaster *genes among the other* Drosophila *genomes*

The orthology assignment process predicted orthologs for between 73% and 93% of *D. melanogaster* genes depending on the evolutionary distance of the target genome to *D. melanogaster* (Table 2; Fig. 2A). The numbers of orthologs decreased with increasing distance to *D. melanogaster*, while the number and proportion of orphans and degenerate orthologs (orthologs in one-to-many or many-to-many relationships) increased. *D. yakuba* contains an extraordinarily large number of genes that are apparent duplications (1:2 orthologs), but these appear to represent artifacts of its genome's assembly (see Fig. 2C).

Orphans

Orphans are transcripts in *D. melanogaster* without a predicted ortholog in another genome. Apart from lineage-specific deletions, the failure to detect an ortholog can result from various methodological artifacts: (1) the *D. melanogaster* transcript is a spurious or nonprotein-coding gene; (2) the ortholog is not recognized as such by the orthology prediction method; or (3) the ortholog is not detected by the gene prediction method. The latter can be due to several reasons. For example, some genes may have become nonfunctional and have then diverged beyond recognition or genes are located in a gapped or misassembled region in the genome.

The number of genes in *D. melanogaster* without orthologs was minimal for *D. yakuba* (915 genes) whereas it was maximal for *D. willistoni* (3801), which perhaps reflects its greater divergence from *D. melanogaster*. Most of the orphaned genes failed to generate predictions in many species or else were those whose predictions are discarded in the quality-control step on the basis of their fragmentation or disruption. Orphans might represent noncoding rather than open reading frame sequence in the *D. melanogaster* genome. For example, there were 94 gene predictions in *D. melanogaster* that were orphaned in all other species. All these genes encode repeats or low-complexity regions and were thus masked. These predictions should now be targeted for experimental verification, or otherwise, of their protein coding capacity. Issues related to these are explored in greater depth elsewhere (see *Drosophila* 12 Genomes Consortium 2007).

Validation of orthology assignments

The true orthology and paralogy relationships between homologous predicted genes in the *Drosophila* species are unknown a priori, and appropriate benchmark sets thus do not yet exist. Therefore, we considered three expectations against the predicted orthology assignments: (1) sequence similarities between out-paralogs should be larger than those between orthologs and in-paralogs; (2) orthology assignments are consistent among several genome pairs; and (3) orthologs are present in syntenic order.

We observed little overlap between sequence similarities in terms of normalized bitscores between out-paralogous gene pairs and orthologous gene pairs (Supplemental Fig. S3). This is not a trivial result, as the phylogeny-based orthology assignment by PhyOP does not impose a fixed threshold on the basis of sequence dissimilarity but instead assigns orthology based on tree topology alone.

Consistency

If all orthologs and in-paralogs have been predicted correctly among all 12 species, graph clustering by connected components ought to aggregate them into orthologous groups. To test this, we computed all possible orthology triplets and found, indeed, that 98% of all these were self consistent. Degenerate orthologs are here counted as a single orthology assignment.

We note that this high number is still likely to be a lower estimate as duplications that are not lineage-specific will give rise to inconsistencies. This is because in-paralogs grouped into two separate orthologous pairs will naturally be inconsistent when joined with a common ortholog in another species. Indeed, we observed that the number of inconsistent triplets was largest if they involved the sibling species *D. simulans* and *D. sechellia,* or *D. pseudoobscura* and *D. persimilis*.

Conserved Synteny

Rearrangements of chromosomes are rare events and tend to happen in a block-wise fashion that mainly preserves the local order of genes on the chromosome. Thus, even after long periods of divergence between species, synteny blocks, defined as conserved runs of consecutive orthologous genes, remain discernible. We computed synteny blocks (as previously, Richards et al. 2005) as runs of ortholog gene pairs, discounting local duplications and allowing for local rearrangements. Orthologs on unplaced contigs in *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. pseudoobscura* genome assemblies were ignored.

As expected, we observed high rates of rearrangements within Müller elements (Ranz et al. 2001), with an increasing number of rearrangements with increasing evolutionary distance between genome pairs (Fig. 2C,D). The size of synteny blocks, however, was of course dependent on the assembly status. For example, the median synteny block lengths for the sibling species *D. simulans* and *D. sechellia* were relatively low and quite different (742 kb and 416 kb, respectively), reflecting the high number of contigs and low median contig size in these two ge-

Table 2. Orthology assignment results for 11 *Drosophila* species' genomes

	Orthology assignment											
	Bitscore-based orthologs					d_S-based orthologs					Lineage-specific duplications	
Species	1:1	Dmel 1:m/m:1	Target 1:m/m:1	Dmel	Target	1:1	Dmel 1:m/m:1	Target 1:m/m:1	Dmel	Target	Genes and pseudogenes	Genes only[a]
D. simulans	11,236	331	457	84%	91%	11,306	272	375	84%	91%	103	104
D. sechellia	12,442	368	810	93%	88%	12,513	321	678	93%	88%	99	63
D. yakuba	12,246	675	1,220	93%	93%	12,305	655	1,122	94%	92%	329	200
D. erecta	12,579	304	372	93%	95%	12,640	288	306	93%	95%	173	133
D. ananassae	11,217	464	575	84%	86%	10,384	405	449	78%	79%	174	129
D. pseudoobscura	10,224	605	914	78%	94%	8,941	476	685	68%	81%	NA	NA
D. persimilis	9,897	577	1,071	76%	89%	8,644	439	853	66%	77%	NA	NA
D. willistonis	9,455	580	1,039	73%	87%	5,991	342	581	46%	55%	NA	NA
D. virilis	9,848	551	617	75%	92%	7,048	303	381	53%	65%	NA	NA
D. mojavensis	9,621	542	599	73%	92%	6,439	288	300	49%	61%	NA	NA
D. grimshawi	9,323	825	1,226	73%	94%	6,813	507	740	53%	67%	NA	NA

Orthology assignment was performed using normalized bit scores and later verified using d_S values. Lineage-specific duplications have only been recorded in the *D. melanogaster* subgroup.
[a]Tree reconciliation was performed without considering pseudogenes. Because tree reconciliation only considers tree topology, the number of lineage-specific duplications can rise in one lineage, when an orthologous pseudogene in a second lineage is removed.
NA, Not applicable.

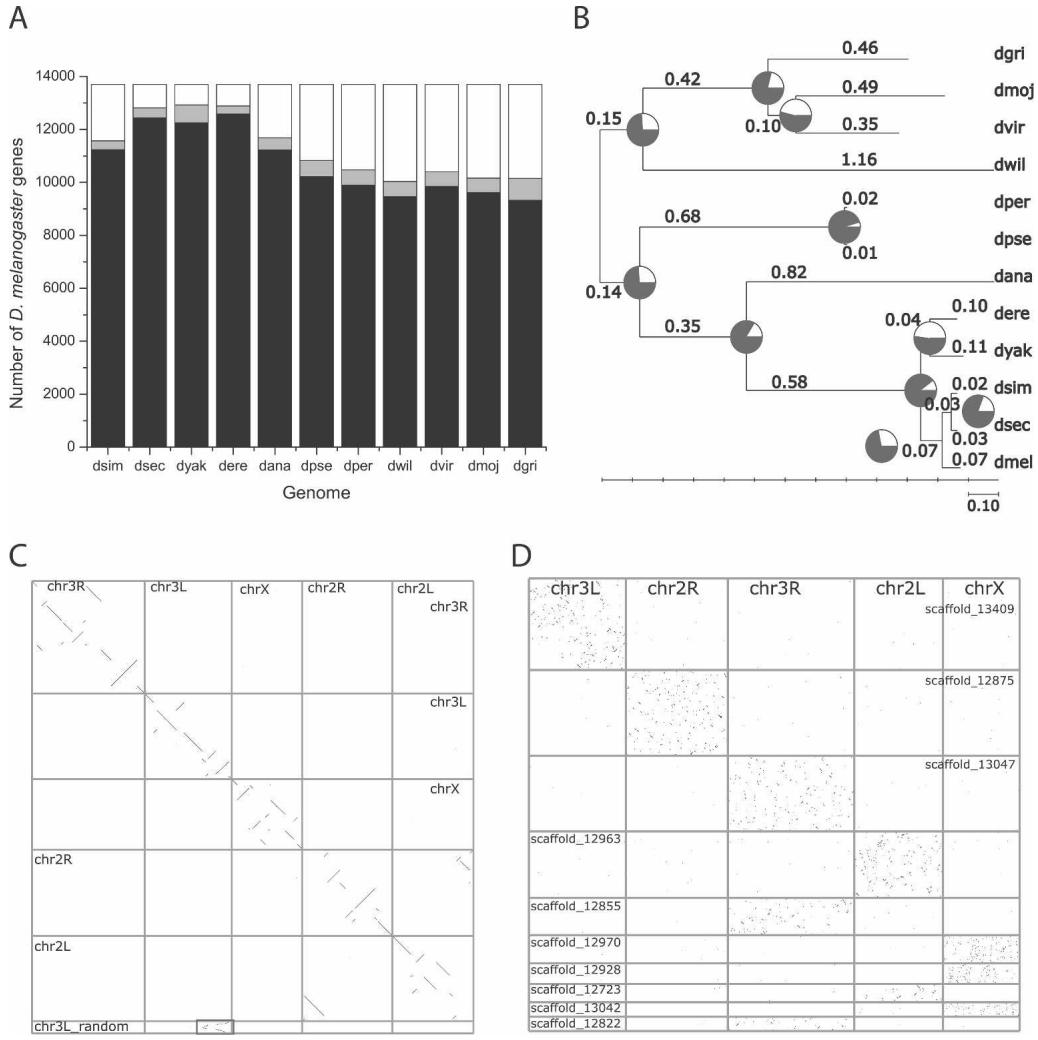

Figure 2. Orthology assignment. (*A*) Numbers of *D. melanogaster* genes with orthologs in other *Drosophila* species. These ortholog counts increase with increasing statistical coverage of genome sequence and decrease with increasing species divergence. The numbers of sequences in 1:1 orthology assignments are shown in black, while the numbers of 1:many orthologs are shown in gray. (*B*) The inferred phylogeny of *Drosophila* species based on median d_s values among orthologs. The tree was computed using the FITCH program of the PHYLIP package (Felsenstein 1989). Branch lengths are given in d_s. Branch support values, computed as percentage of gene phylogenies that are consistent with the species phylogeny, are shown as red pie slices. (*C*) Gene-based synteny plot between *D. melanogaster* (*X*-axis) and *D. yakuba* (*Y*-axis). Genes are sorted by physical locations on the chromosomes. The box marks an artefactual duplication between chromosome 3L and chromosome 3L_random in *D. yakuba* that explains the excess of 1:2 orthologs in this assembly. (*D*) Gene-based synteny plot between the more divergent species pair *D. melanogaster* (*X*-axis) and *D. virilis* (*Y*-axis). Species names have been abbreviated.

nomes. In *D. pseudoobscura* the average length of a synteny block was 100 kb, spanning on average 9.3 genes, which corresponds well to the average block length (83 kb) and gene count (10.7) per synteny block reported previously (Richards et al. 2005).

We took advantage of these predicted conserved synteny blocks to identify dubious orthology assignments since these are more likely to be inconsistent in their placement within a synteny block. The false assignment rate rose to a maximum of 7% for the pair of *D. willistoni* and *D. simulans* assemblies. Although this represents an upper bound on the misprediction of orthology, the true proportion will be influenced by assembly quality. For example, on removing unplaced contigs the number of non-syntenic ortholog assignments dropped from 629 to 66 for the pair of *D. melanogaster* and *D. simulans* assemblies. We assume that many of these misplaced orthologs resulted from assembly error, although, because the prevalence of dubious orthologs increases with increased divergence, this indicates that accurate orthology assignment is progressively more difficult with increasing evolutionary distance.

Orthologous groups

We have shown that pairwise orthology assignments are highly consistent between species. Thus, it seems feasible to build orthology groups between multiple genomes based on pairwise orthology assignments using a simple clustering procedure, as long as this accommodates an occasional spurious orthology assignment. Thus, perfect gene prediction and orthology assignment should result in crisp clusters of all 12 species. In reality, we expected three types of orthology groups: (1) those containing all species (each with one or more genes); (2) those containing only

a subset of species and that are monophyletic (i.e., involving lineage-specific deletions and rapid evolution); and (3) those where absences have led to nonmonophyly (i.e., absences due to gene or orthology prediction errors, or assembly incompleteness).

Using a graph clustering approach (see Methods) we found 14,258 orthologous clusters from 13,836 *D. melanogaster* genes. A total of 6647 (45%) clusters contained all twelve species and 2623 (18%) clusters lacked a single species. A further 2106 (15%) clusters lacked more than one species, but the pattern of absences was consistent with the species phylogeny. Thus, a total of 11,376 (78%) clusters reflected a pattern that is phylogenetically consistent. Given the unfinished status of the genomes and the multiplication of errors during the prediction process, a considerable number of incomplete clusters was to be expected.

The clustering method we used is agnostic of the species to which a gene belongs. It was thus reassuring that 12,924 (91%) clusters contained a single *D. melanogaster* gene indicating that out-paralogous sequences have not been wrongly grouped together. An additional 989 (7%) clusters contained no *D. melanogaster* gene, leaving only 345 clusters (2%) with more than one *D. melanogaster* gene, which might well represent gene duplications in the lineages leading to *D. melanogaster*.

The presence of additional genomes assists, albeit only marginally, in orthology assignment. We found that sensitivity, defined here as the number of assigned orthologs for *D. melanogaster*, increased by up to 1.8% using 12, as opposed to only two, species' information (Supplemental Table S3).

Evolutionary rate analysis

Phylogenetic trees for 6375 clusters of orthologous transcripts, using synonymous substitution rates as a distance metric, recapitulated the established *Drosophila* phylogeny (Russo et al. 1995; Ko et al. 2003) (Fig. 2B). Support for the established phylogeny from gene trees was, in general, high. The most difficult groups to resolve were the *D. mojavensis–D. virilis–D. grimshawi* clade, most likely because of their large divergences, and the split between *D. yakuba, D. erecta*, and *D. melanogaster*, which is suggested to be subject to lineage sorting effects (Pollard et al. 2006). Analyses of noncoding sequence and the application of coalescence models should allow further insights into this issue (Hobolth et al. 2007).

Branch-specific d_N/d_S

We used 6375 clusters of 1:1 orthologs to assess the variation of selective strength for different branches of the tree. In 20 iterations, we sampled 200 alignments from the full set of alignments and compared two models using PAML: (1) with a single d_N/d_S ratio estimated for all branches and (2) with branch-specific d_N/d_S ratios for all possible clades. The two models were compared in a log-likelihood ratio test. The species *D. willistoni, D. sechellia*, and *D. persimilis* were not included in this analysis, the former because of its different G + C content (see Heger and Ponting 2007), the latter two because they are closely related sibling species to *D. simulans* and to *D. pseudoobscura*, respectively.

The multiple branch model passed the likelihood ratio test ($P < 0.05$) in all instances. Substitution rates among the 20 samples were reproducible (Fig. 3), and d_S values and d_N/d_S ratios were not significantly linearly correlated ($P = 0.14$). Among the terminal branches, *D. simulans* and *D. erecta* had an elevated d_N/d_S ratio, *D. ananassae* showed a decreased d_N/d_S ratio, while the remainder had intermediate values. In a pairwise t-test among all combinations, only *D. ananassae* had a significantly different d_N/d_S ratio from the rest ($P < 0.05$, with Bonferroni correction). We repeated the analysis with a set of 2662 genes with low codon usage bias with little change in the results. Similar results were obtained when the analysis was restricted to the *D. melanogaster* subgroup together with *D. pseudoobscura*.

Analyses of paralogous gene families

Our analysis was based on 13,126 maximum likelihood trees from the *D. melanogaster* subgroup, using sequences from *D. pseudoobscura* and *D. persimilis* as outgroups, because synonymous rate estimation suffers less from saturation in this group, and because the presence of an outgroup facilitates tree rooting. Of these trees, 11,122 included at least one outgroup, and 8198 contained a full species complement as well as an outgroup. Although the height of these trees (distance to root) varied considerably (Supplemental Figure S7), branch tips within a tree were largely contemporaneous with ~10% variation, and there were no discernible differences between species (Supplemental Figure S8). A large variation in tree height has implications when comparing the dates of duplication events among gene families because trees then need to be scaled to better approximate linear time. We employed two scaling methods: (1) scaling by median tree height and (2) scaling by branch length (see Methods in Supplemental material).

We observed 1935 events of (pseudo)gene duplication in 836 families. The majority, 1389 events, were terminal lineage-specific duplications, while the remaining 546 events were more

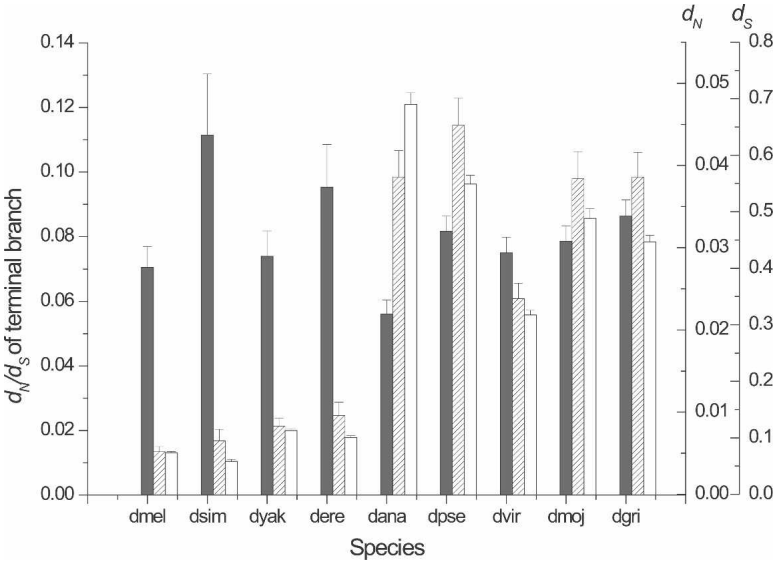

Figure 3. Branch-specific terminal d_N/d_S, d_N, and d_S values in solid, hatched, and open bars, respectively. The error bars indicate the standard deviation from 20 replicates. Species names have been abbreviated.

ancient duplications internal to the tree. A large proportion of duplications (33%–40%) were due to predicted genes with in-frame stop codons or frameshifts. There was a marked discrepancy between the numbers of lineage-specific duplications per species (Table 1). An elevated number of terminal lineage-specific duplications in *D. ananassae* could partly be explained by the longer branch length on which duplications have taken place, but this fails to explain the excess of terminal lineage-specific duplications in *D. sechellia* compared with *D. simulans*. A possible explanation might be differences in the assembly method: While the *D. simulans* genome sequence has been assembled using *D. melanogaster* as a blueprint, the *D. sechellia* assembly was assembled ab initio.

We found a strong excess of very recent duplications with a long tail of older duplication events (Fig. 4). The duplication rate was more uniform for times that were prior to this burst of near-contemporaneous duplications. Identical patterns of duplications were also observed for predicted intact genes and for likely pseudogenes (Supplemental Figure S11). These distributions are consistent with models of rapid birth-and-death of genes or copy number variation of *Drosophila* genes (see Discussion).

We found that lineage-specific duplications have almost always occurred within the same Müller element (Fig. 5). In the three species whose sequence has been assembled into chromosomes (*D. melanogaster*, *D. simulans*, and *D. yakuba*), we found 97.6% (285 out of 292) of duplications to be confined to a single Müller element. Of seven events of duplication between Müller elements, only four appear to have given rise to full-length genes. These include one in *D. yakuba*, from chromosome X to 3R, which has recently been proposed to be associated with a gene's avoidance of X chromosomal inactivation during early spermatogenesis (Betran et al. 2006). By contrast, a transposition has occurred in the opposite direction from chromosome 3R to X, involving *D. melanogaster CG33213* and *CG33221* genes.

Another transposition in the *D. yakuba* lineage is of a gene encoding a DM8 domain (Ponting et al. 2001), which has copies on four Müller elements and thus accounts for three transpositions, and the fourth, this time in the *D. simulans* lineage, is of a gene encoding an ML (MD-2-related lipid-recognition) lipid-binding domain (Inohara and Nunez 2002). Although these two domain families were previously thought to be evolutionarily distinct, we found that they are homologous, with the DM8 family representing simply an arthropod-specific expansion of the ML domain family. A PSI-BLAST (Altschul et al. 1997) search of current protein sequences found, after four iterations, significant similarity ($E = 10^{-3}$) between a DM8 domain query sequence (CG14455-PA) and a ML domain protein (rat GM2 activator protein). Not only have two of these genes been translocated onto the X chromosome, which might indicate a selective advantage of their presence there, many others have been duplicated locally on independent *Drosophila* lineages. The numbers of such genes were lowest for *D. melanogaster* (26) and highest for *D. pseudoobscura* and *D. persimilis* (55 and 73, respectively). DM8 domain proteins are expressed in chemosensory sensilla and are thought to be involved in male-specific chemosensation (Xu et al. 2002). The diversification of DM8 domain-encoding genes and their sequences would thus be compatible with frequent episodes of adaptation to varied odorant-detection needs and thus may have been driven by sexual selection. The DM8 domain family bears many resemblances to rodent pheromone carriers (Emes et al. 2004). They are secreted molecules whose sequences have diversified and duplicated rapidly, and they bind small lipids. Our finding that they represent an offshoot of the ML lipid-binding domain family would thus be consistent with a pheromone-binding function.

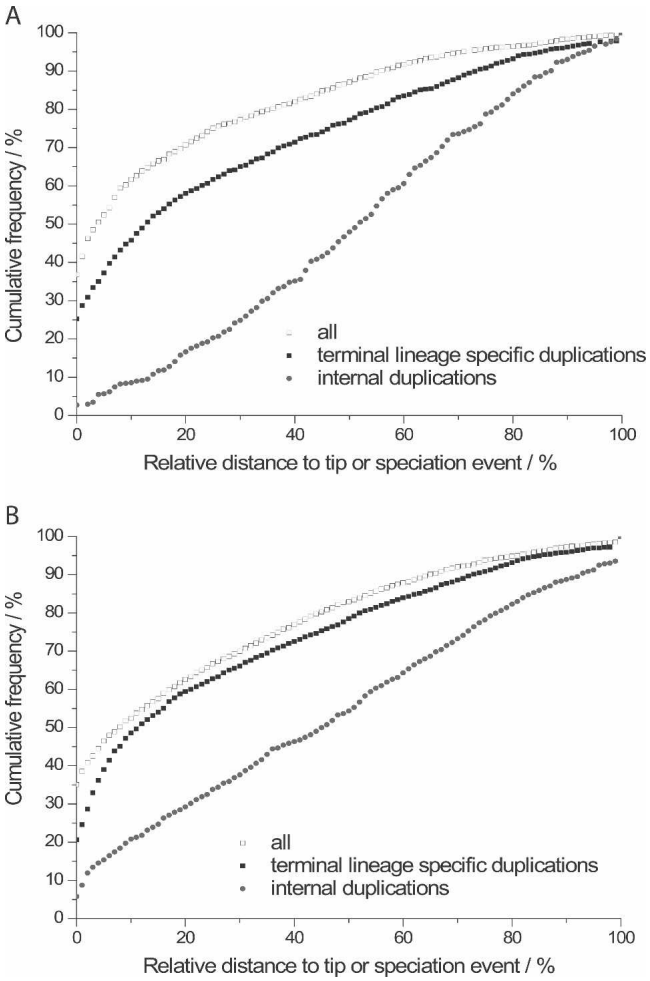

Figure 4. Approximately 20%–30% of all gene duplications have been very recent, whereas duplications inferred to have been more ancient occurred less frequently and more uniformly. Duplication events have been dated by the synonymous substitution rate d_S and normalized by the overall height of each gene tree (open squares) or by the corresponding branch length (solid squares) of the species tree after reconciliation with the gene tree. The latter have been aggregated over all internal or terminal lineage branches, respectively. A similar picture emerges when considering each branch separately (Supplemental Figs. S9 and S10). Duplications from *D. melanogaster* subgroup only rooted with *D. pseudoobscura* and/or *D. persimilis* sequences (*A*) or all 12 species (*B*). Among 13,132 clusters, 5851 had the full species complement and there were 1853 clusters with 1305 internal and 3794 lineage-specific duplications.

Analyses of fast evolving proteins

Extant species have successfully adapted to fluctuating selective pressures. The effects of molecular adaptation are, among others, (1) fixation of beneficial amino acids substitutions and (2) changes to the proteome by gene duplication and subsequent specialization (Prince and Pickett 2002; Emes et al. 2003). In the following section we characterize the fastest evolving genes with respect to (1) site substitution in 1:1 orthologs, (2) subgroup-specific genes, and (3) genes arising from duplications.

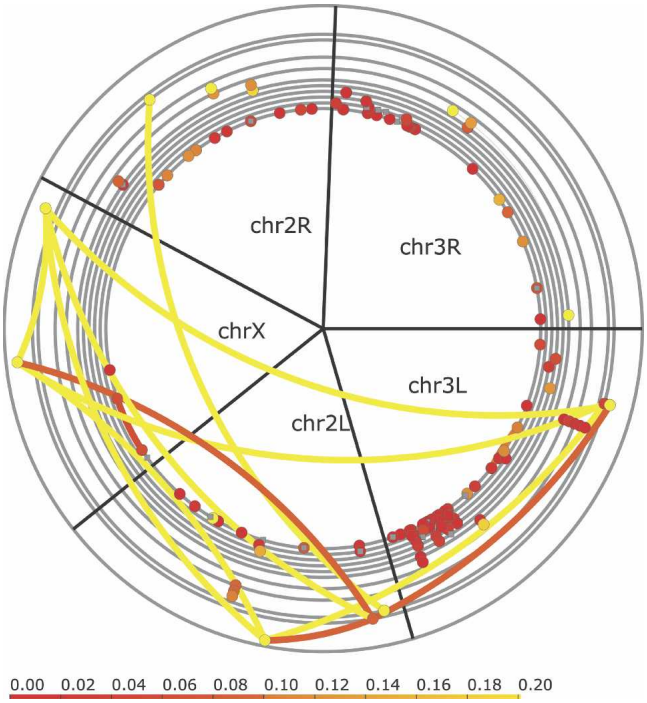

Figure 5. Most gene duplications are closely linked and are recent (d_S < 0.04). Shown here are lineage-specific duplications in *D. yakuba* for the five large chromosomal arms, but similar results are seen for other species (Supplemental Figs. S12 and S13). Each duplication is represented by two dots connected by an arc. These are colored by their divergence (d_S value, see scale). Pseudogenes are shown in gray. Genes are placed on the chromosomal arms according to their physical location. Most duplications are local such that only a single dot is visible. Overlapping or very close duplications are stacked on *top* of each other. Multiple duplications within the same gene family are stacked on *top* of each other in the outer rings whose increased radius reflects the family size. Each member of a multigene family is connected to all other members resulting in a connected path of arcs within a family. Translocations involving three families of likely transposable elements have not been shown to simplify the image.

Site-specific analysis

With the availability of several completely sequenced genomes it is now possible to present a global overview of adaptive evolution across multiple whole genomes. Apart from technical issues surrounding the accuracy of gene and orthology predictions, the problem of multiple hypotheses testing now gains greater prominence. While the number of tests performed within a single gene family was already considered large and multiple testing procedures were thus adopted (Massingham and Goldman 2005), this number is now dwarfed by the number of tests employed in genome-wide studies.

Here, we adopted a heuristic and conservative approach to obtain a list of proteins containing positively selected sites (see Methods). We found 2869 sites within 1618 multiple alignments (out of a total of 6375 multiple alignments of 1:1 orthologous transcripts analyzed) predicted by the SLR method (Massingham and Goldman 2005) to have been subject to positive selection (P < 0.05). Of these, 121 alignments with a total of 553 sites remain that contain more positively selected sites than expected (binomial distribution test, P < 0.01, Supplemental Table S4). According to an analysis of Gene Ontology (GO) terms (Ashburner et al. 2000), these 121 multiple alignments are significantly enriched for proteins involved in two functional categories: the epigenetic regulation of gene expression (five genes) and carbohydrate binding (six genes) (hypergeometric distribution test, P < 0.05, Fig. 6).

Three positively selected genes, Dicer-2 (*Dcr-2*), *r2d2*, and *spn-E*, that belong to the first category are involved in the RNA-initiated silencing complex (RISC) pathway. Antiviral RNAi genes, among them *Dcr-2* (CG6493) and *r2d2* (CG7138), have previously been predicted to be under positive selection based on their high d_N/d_S ratio and nonsynonymous divergence (Obbard et al. 2006). Dicer-2 and r2d2, which were predicted to have nine and three positively selected sites, respectively, form a heterodimer, required for the loading of small interfering RNA (siRNA) onto the RISC (Liu et al. 2003).

In the second category, five of the six identified carbohydrate-binding genes (*CG6497*, *CG7298*, *CG14247*, *CG14880*, and *CG13075*) contain one or more copies of the chitin-binding peritrophin A domain (PFAM identifier: PF01607). Proteins with this domain are located in the peritrophic matrix that lines the gut of most insects (Tellam et al. 1999) and that, among other functions, impedes the invasion of pathogens. Positive sites were found to cluster spatially within a known structure of this domain, consistent with positive selection having acted upon a binding function. The sixth gene (*CG9134*), a C-type lectin, is expressed during eye development (Michaut et al. 2003) but is also thought to act as a peptidoglycan recognition protein in innate immunity (Ao et al. 2007).

The list of genes predicted to be under positive selection contains further components of the innate immune response of *D. melanogaster*, including the thiol-ester containing macroglobulins *TepII* (CG7052) and *TepIV* (CG10363) (Lagueux et al. 2000), and *baz* (CG5505), a peptidase that inhibits the growth of the intracellular pathogen *Listeria monocytogenes* (Cheng et al. 2005).

Analysis of D. melanogaster *subgroup-specific genes*

The fastest evolving genes will have diverged to such extent that groupings into 1:1 orthology clades over the entire *D. melanogaster* subgroup will be disfavored. To identify such rapidly evolving genes, we examined monophyletic subgroups within the *D. melanogaster* subgroup. In particular, we considered groups that contain orthologs from (1) *D. melanogaster* and *D. simulans* and/or *D. sechellia* or from (2) each of these three species, together with *D. yakuba* and/or *D. erecta,* but no orthologs from any of the remaining seven species.

We found 795 *D. melanogaster* genes that are specific to the subgroup. According to an analysis of GO terms (hypergeometric distribution test, P < 0.05, Fig. 6), this set is significantly enriched for proteins involved in behavior (n = 16 genes), response to biotic stimulus (n = 12), and symbiosis and parasite response (n = 2). The latter category contains the two drosomycins *dro2* (CG32279) and *dro5* (CG10812), peptides involved in the antifungal response but not believed to co-evolve with pathogens (Jiggins and Kim 2005). The set of behavioral genes contains accessory gland proteins (n = 11) and ejaculatory peptides (n = 3), previously shown to be fast evolving (Swanson and Vacquier 2002). The set representing the response to biotic stimuli contains further drosomycins (*dro2, dro3* [CG32283], *dro5, dro6* [CG32268], *Drs-l* [CG32274]), cecropins (*Anp* [CG1361], *CecA1* [CG1365], *CecA2* [CG1367]), and proteins of the hemolymph (*Dox-A3* [CG2952]) and hemocytes (*He* [CG31770]).

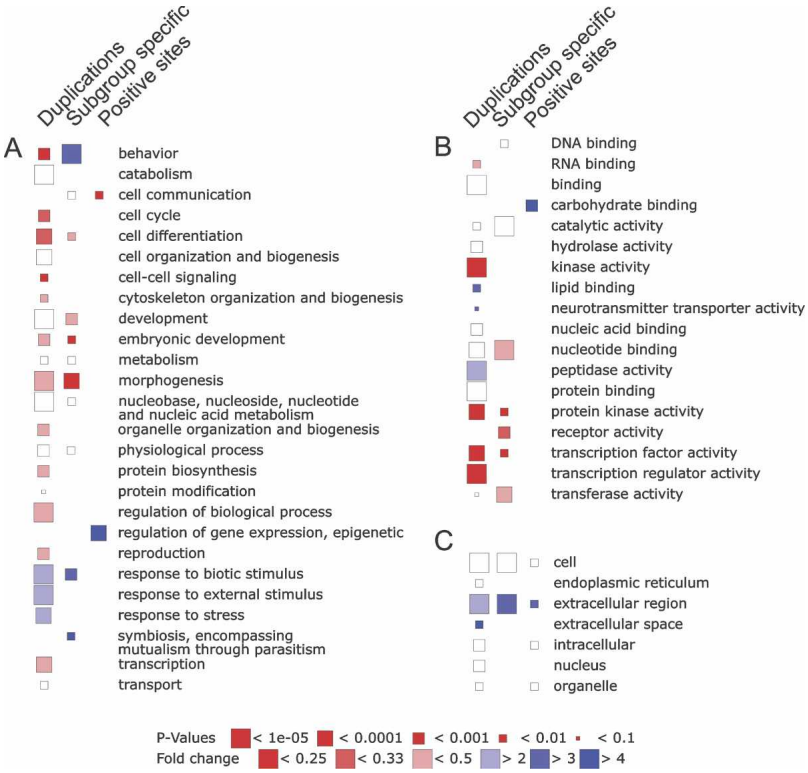

Figure 6. Different classes of rapidly evolving genes. Duplicated genes are often involved in adaptive functions such as responses to external stimuli, whereas they are under-represented in transcription factors and regulatory genes. Shown are over-/under-represented GOSlim categories of *D. melanogaster* genes present in clusters containing gene duplications ($n = 1126$) (*A*), without detectable orthologs in species further diverged than *D. yakuba* and *D. erecta* ($n = 795$) (*B*), and with sites predicted to have been subject to positive selection ($n = 121$) (*C*). The size of the box represents the *P* value of the over/under-representation while the fold over-/under-presentation is indicated by the color of the box (see scale at *bottom*). False-positive predictions arising from the application of multiple tests were controlled using a false discovery rate of 0.05.

Analysis of paralog families

GO term analysis of gene families exhibiting duplications indicated that this set is enriched in genes involved in various stimulus responses ($n = 54$) and genes with peptidase activity ($n = 75$). Among the former are odorant receptors ($n = 15$), heat shock proteins ($n = 6$), and lysozymes ($n = 4$). We found nine genes encoding trypsin-like peptidases, whose known roles in Drosophilids include digestion, development, and immune responses (Ross et al. 2003). Furthermore, we found five DM8 domain proteins, which may participate in odorant and pheromone detection (for a full list, see Supplemental Table S5).

In conclusion, the set of fast evolving genes, defined in terms of substitutions or duplications, is greatly enriched in genes involved in immunity and reproduction, as previously observed among mammals (Emes et al. 2003). Nevertheless, only 59% of the fastest evolving proteins currently contribute to either a GO biological process or a GO molecular function. Future functional characterization of the unannotated sequences is thus likely to reveal further genes involved in immunity and reproduction. A more detailed analysis of the evolution of immunity genes in *Drosophila* species can be found elsewhere (T.B. Sackton, B.P. Lazarro, T.A. Schlenke, J.D. Evans, D. Hultmark, and A.G. Clark, in prep.).

Discussion

We have shown how protein-coding gene sets can be predicted successfully for multiple closely related *Drosophila* species by comparison with a well-annotated reference gene set. The orthology and paralogy relationships among these *Drosophila* species' genes can, as for mammals (Goodstadt and Ponting 2006), be inferred successfully using rates of synonymous substitutions as a molecular clock. The accuracies of gene, ortholog, and paralog predictions, and indeed the accuracies of these genomes' sequences and assemblies, are reflected both by the great majority of genes possessing orthology relationships and by the high conservation of gene order ("synteny") between closely related genomes.

The application of d_S estimates for inferring phylogeny and our evolutionary rate analyses appear to have been successful despite inaccuracies arising from codon usage biases, nonequilibrium mutational biases, and saturation of substitutions at synonymous sites. Elsewhere (Heger and Ponting 2007), we show that codon usage bias and G + C content have remained largely constant within the *D. melanogaster* subgroup, and that *D. willistoni* is distinguished by its exceedingly low G + C content. The similar unimodal d_S distributions of orthologous genes between *D. melanogaster* and even the most diverged *Drosophila* species (Supplemental Fig. S5) indicate that these analyses have been appropriate. Nevertheless, for comparative purposes we also applied a normalized bitscore as a divergence metric to predict orthology relationships.

It had been argued previously that substitutions at silent sites are saturated for species pairs of the distance *D. melanogaster–D. pseudoobscura* and beyond (Bergman et al. 2002). We do, indeed, observe saturation of transitions at fourfold degenerate sites between such species pairs. However, we find that transversions still retain sufficient information to allow maximum likelihood methods to infer rates at an appropriate level of accuracy. In simulations, synonymous substitution rates can be recovered accurately up to at least 2.5, although error bars widen at larger distances (L. Goodstadt, pers. comm.). Forty-five percent of orthologs to *D. melanogaster* possess d_S values less than 2.5 in *D. mojavensis*, *D. virilis*, and *D. grimshawi* and consequently these remain amenable for evolutionary rate analyses. For example, when we extended our gene duplication analysis from only the *D. melanogaster* subgroup species to all 12 species, we obtained similar results in spite of the increased uncertainties in rate estimation due to saturation and changes in codon usage bias (Fig. 4B).

Each lineage-specific d_N/d_S value represents the degree of purifying selection that has prevailed since divergence from a

closely related species. If this strength of selection has remained constant to the present day, then these values might be correlated with either these species' extant effective population sizes, or the strengths of their extant codon biases. However, we find that the strengths of selection on codon bias for these species (Heger and Ponting 2007) are not significantly correlated with their lineage-specific d_N/d_S values. Moreover, those species whose effective population sizes are currently high (e.g., *D. virilis* or *D. simulans*) (Aquadro et al. 1988; Akashi 1996; McVean and Vieira 2001), relative to *D. melanogaster*, appear not to show appreciable decreases in lineage-specific d_N/d_S values, as might be expected from population theory arguments (Ohta 1973). Consequently, it appears likely that effective population sizes and selection acting on codon bias usage have often varied rapidly, relative to the divergence time, as has been proposed by others (Akashi et al. 2006).

We find that extant gene duplicates apparent in each of the 12 *Drosophila* species' genomes arose considerably more frequently in very recent evolution than they arose, for example, prior to the last common ancestors of any species' pair (Fig. 4). Such observations for many species have been attributed, by Lynch and Conery (2000), to a rapid-birth and rapid-death model of gene duplication. In this model, gene duplicates arise frequently, but only the minority of genes remain functional after a short period of time; for *D. melanogaster*, this has been estimated as 2.9 million years (Lynch and Conery 2000). Our data imply that this half-life of *Drosophila* gene duplicates is shorter still since, proportionately, we observe gene duplicate pairs more frequently with vanishingly small divergence ($d_S < 0.01$), than did Lynch and Conery (2000), perhaps because they discounted large multigene families in their analysis.

This relatively high abundance of low-divergence, virtually identical, gene pairs suggests a second model, different from that of Lynch and Conery (2000). This is a rapid-birth but infrequent-fixation model, which hypothesizes that although gene duplications do indeed occur rapidly, very few duplicates are fixed. Instead, the numbers of essentially identical gene pairs we observe are copy number variable, and most will, over time, be lost by genetic drift. This model predicts that gene duplicates are frequently lost by drift, not through degenerative mutations producing nonfunctional pseudogenes, and therefore that the frequency spectrum, as a function of sequence divergence, of pseudogene duplications should match the frequency spectrum of genes. Indeed, this is what we observe (Supplemental Figure S11). Alternatively, the duplication rate profile can also be explained by the well-established rapid deletion rate of *Drosophila* DNA (Petrov et al. 1996): At a constant duplication rate but high deletion rate, ancient duplications are unlikely to persist, which thus creates a bias toward the observation of recent duplications.

The excess of virtually identical genes might also be accounted for by episodes of persistent gene conversion. However, documented examples of gene conversion among *Drosophila* genes, such as *Hsp70* genes (Bettencourt and Feder 2002), are rare, and this process is not applicable to the translocated duplications seen, for example, among the DM8 domain-encoding genes. Whether these are ultimately fixed or not, the high number (45) of gene duplications that have occurred in the *D. melanogaster* lineage since its split with the *D. simulans* lineage ~2.3 million years ago (Russo et al. 1995) attests to an exceedingly rapid turnover of genes.

Many of the genes that have conferred selective advantage from translocations, duplications, and nonsynonymous substitutions within the *Drosophila* clade are involved in immunity and reproduction. *Drosophila* genes encoding RNA-binding proteins have rarely been duplicated (Fig. 6) but appear frequently to have acquired amino acid substitutions by positive selection (Obbard et al. 2006; this work). We note that 41% of the fastest evolving genes have no GO biological process or molecular function assignment, and thus they are interesting targets for experimental characterization.

We have presented initial evolutionary rate analyses among these *Drosophila* genomes and their genes. Our approach for prediction of positively selected sites has been to apply well-established gene family approaches across this clade and to apply relatively simple corrections for multiple testing. In the future, it will be important to test for selection, while accounting for false positive predictions, across all codons from all ortholog sets genome-wide. Moreover, we sought evidence for positively selected sites only among the five species of the *D. melanogaster* subgroup whose genome sequences are currently known, whereas a larger number of such closely related species would likely have provided greater predictive power. In this regard, we note that when this site-specific analysis was performed on these five genomes, together with that of the more-distantly related species *D. pseudoobscura*, this resulted in a substantial reduction in predicted site count.

These analyses are complementary to those described in the primary publication of these genomes (*Drosophila* 12 Genomes Consortium 2007). We have focused more on providing gene, alternative transcript, and pseudogene sets, and orthology and paralogy predictions, among these dozen genomes that should, in the future, prove beneficial to *Drosophila* evolutionary biologists. These predictions are available in full at http://wwwfgu.anat.ox.ac.uk/flies/.

Methods

Data sets

Chromosomes, transcripts, and translations for *D. melanogaster* were acquired from Ensembl release 37 (Birney et al. 2006). The sequence data are based on BDGP assembly release 4 and annotations are based on FlyBase release 4.2.1 (Grumbling and Strelets 2006). This set contained 19,369 transcripts from 13,836 genes.

Assembled genomic sequences for *D. simulans*, *D. sechellia*, *D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. grimshawi, D. virilis,* and *D. mojavensis* were obtained from the community server for the assembly/alignment/annotation project (http://rana.lbl.gov/drosophila/wiki/index.php/Main_Page), release comparative analysis freeze 1 (caf1).

Transcript prediction and orthology assignment

We predicted the transcripts for the 11 additional *Drosophila* species on the basis of homology with known transcripts from *D. melanogaster*. It was our objective to predict transcripts rather than genes with maximal sensitivity. Assignment of predicted transcripts to specific genes was deferred to a final quality-control step. The transcript prediction pipeline is centered on the Exonerate implementation (Slater and Birney 2005, version 0.9.0) of the GeneWise model (Birney et al. 2004). This model allows alignment of an amino acid sequence directly to all six reading frames of a genomic sequence while accounting for DNA frameshifts, in-frame stop codons, and introns. Orthology assignment was performed in two stages. First, pairwise orthology was com-

puted in pairwise species comparisons using PhyOP (Goodstadt and Ponting 2006). Then, multiple orthology assignments involving more than two species were inferred from clusters derived from the graph of pairwise orthology relationships.

The Supplemental material contains a complete description of transcript prediction and orthology assignment.

Rate measurements

Synonymous and nonsynonymous substitution rates were estimated using CodonML from the PAML package (Yang and Nielsen 2002). In all measurements, codon frequencies were estimated from nucleotide frequencies at each codon position (model F3x4). No correlation among sites was assumed, and the transition/transversion ratio was allowed to vary.

Rates were measured in two sets of multiple alignments. The first set contained 6375 multiple alignments of 1:1 orthologous transcripts where each of the 12 species was represented. This set was used to establish the species phylogeny, to measure branch-specific d_N, d_S, d_N/d_S, values, and to identify sites under positive selection. The second set contained 13,126 multiple alignments of ortholog and in-paralog transcripts of transcripts within the *melanogaster* subgroup, *D. pseudoobscura*, and *D. persimilis*. The second set of multiple alignments was used for the analysis of the duplication rate for the GO analysis of subgroup-specific families and families with duplications. More details are provided in the Supplemental materials.

Acknowledgments

We thank the *Drosophila* genomes' sequencing consortium, in particular the various sequencing centers, for the data, and Mike Eisen for hosting the AAA web site. We thank Leo Goodstadt for advice and assistance in implementing PhyOP, Gerton Lunter for advice on GOSlim false discovery rate estimation, and Caleb Webber for the gene prediction prototype. This work was funded by the UK Medical Research Council.

References

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

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

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25:** 3389–3402.

Ao, J., Ling, E., and Yu, X. 2007. *Drosophila* C-type lectins enhance cellular encapsulation. *Mol. Immunol.* **44:** 2541–2548.

Aquadro, C.F., Lado, K.M., and Noon, W.A. 1988. The rosy region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation and divergence. *Genetics* **119:** 875–888.

Ashburner, M. and Bergman, C.M. 2005. *Drosophila melanogaster*: A case study of a model genomic sequence and its consequences. *Genome Res.* **15:** 1661–1667.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Bergman C.M., Pfeiffer B.D., Rincon-Limas D.E., Hoskins R.A., Gnirke A., Mungall C.J., Wang A.M., Kronmiller B., Pacleb J., Park S., et al. 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome. *Genome Biol.* **3:** RESEARCH0086.1–0086.20. doi: 10.1186/gb-2002-3-12-research0086.

Betran, E., Bai, Y., and Motiwale, M. 2006. Fast protein evolution and germ line expression of a *Drosophila* parental gene and its young retroposed paralog. *Mol. Biol. Evol.* **23:** 2191–2202.

Bettencourt, B.R. and Feder, M.E. 2002. Rapid concerted evolution via gene conversion at the *Drosophila hsp70* genes. *J. Mol. Evol.* **54:** 569–586.

Birney, E., Clamp, M., and Durbin, R. 2004. GeneWise and Genomewise. *Genome Res.* **14:** 988–995.

Birney, E., Andrews, D., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cox, T., Cunningham, F., Curwen, V., Cutts, T., et al. 2006. Ensembl 2006. *Nucleic Acids Res.* **34:** D556–D561.

Cheng, L.W., Viala, J.P.M., Stuurman, N., Wiedemann, U., Vale, R.D., and Portnoy, D.A. 2005. Use of RNA interference in *Drosophila* S2 cells to identify host pathways controlling compartmentalization of an intracellular pathogen. *Proc. Natl. Acad. Sci.* **102:** 13646–13651.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* (in press) doi: 10.1038/nature06341.

Emes, R.D., Goodstadt, L., Winter, E.E., and Ponting, C.P. 2003. Comparison of the genomes of human and mouse lays the foundation of genome zoology. *Hum. Mol. Genet.* **12:** 701–709.

Emes, R.D., Beatson, S.A., Ponting, C.P., and Goodstadt, L. 2004. Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. *Genome Res.* **14:** 591–602.

Felsenstein, J. 1989. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5:** 164–166.

Gibbs, R.A., Weinstock, G.M., Metzker, M.L., Muzny, D.M., Sodergren, E.J., Scherer, S., Scott, G., Steffen, D., Worley, K.C., Burch, P.E., et al. 2004. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428:** 493–521.

Goodstadt, L. and Ponting, C.P. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput. Biol.* **2:** e133. doi: 10.1371/journal.pcbi.0020133.

Goodstadt, L., Heger, A., Webber, C., and Ponting, C. 2007. An analysis of the gene complement of a marsupial, *Monodelphis domestica*: Evolution of lineage-specific genes and giant chromosomes. *Genome Res.* doi: 10.1101/gr.6093907.

Grumbling, G. and Strelets, V. 2006. FlyBase: Anatomical data, images and queries. *Nucleic Acids Res.* **34:** D484–D488.

Heger, A. and Ponting, C. 2007. Variable strength of translational selection among twelve *Drosophila* species. *Genetics* (in press) doi: 10.1534/genetics.107.070466.

Hobolth, A., Christensen, O.F., Mailund, T., and Schierup, M.H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS Genet.* **3:** e7. doi: 10.1371/journal.pgen.0030007.

Inohara, N. and Nunez, G. 2002. ML—A conserved domain involved in innate immunity and lipid metabolism. *Trends Biochem. Sci.* **27:** 219–221.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. *Nature* **431:** 931–945.

Jiggins, F.M. and Kim, K. 2005. The evolution of antifungal peptides in *Drosophila*. *Genetics* **171:** 1847–1859.

Ko, W., David, R.M., and Akashi, H. 2003. Molecular phylogeny of the *Drosophila melanogaster* species subgroup. *J. Mol. Evol.* **57:** 562–573.

Lagueux, M., Perrodou, E., Levashina, E.A., Capovilla, M., and Hoffmann, J.A. 2000. Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of *Drosophila*. *Proc. Natl. Acad. Sci.* **97:** 11427–11432.

Liu, Q., Rand, T.A., Kalidas, S., Du, F., Kim, H., Smith, D.P., and Wang, X. 2003. R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science* **301:** 1921–1925.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. *Science* **290:** 1151–1155.

Massingham, T. and Goldman, N. 2005. Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169:** 1753–1762.

McVean, G.A. and Vieira, J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157:** 245–257.

Michaut, L., Flister, S., Neeb, M., White, K.P., Certa, U., and Gehring, W.J. 2003. Analysis of the eye developmental pathway in *Drosophila* using DNA microarrays. *Proc. Natl. Acad. Sci.* **100:** 4024–4029.

Obbard, D.J., Jiggins, F.M., Halligan, D.L., and Little, T.J. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.* **16:** 580–585.

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

Petrov, D.A., Lozovskaya, E.R., and Hartl, D.L. 1996. High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384:** 346–349.

Pollard, D.A., Iyer, V.N., Moses, A.M., and Eisen, M.B. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. *PLoS Genet.* **2:** e173. doi: 10.1371/journal.pgen.0020173.

Ponting, C.P., Mott, R., Bork, P., and Copley, R.R. 2001. Novel protein domains and repeats in *Drosophila melanogaster*: Insights into structure, function, and evolution. *Genome Res.* **11:** 1996–2008.

Prince, V.E. and Pickett, F.B. 2002. Splitting pairs: The diverging fates of duplicated genes. *Nat. Rev. Genet.* **3:** 827–837.

Ranz, J.M., Casals, F., and Ruiz, A. 2001. How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila*. *Genome Res.* **11:** 230–239.

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

Ross, J., Jiang, H., Kanost, M.R., and Wang, Y. 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: An initial analysis of sequence conservation and phylogenetic relationships. *Gene* **304:** 117–131.

Rubin, G.M. and Lewis, E.B. 2000. A brief history of *Drosophila*'s contribution to genome research. *Science* **287:** 2216–2218.

Russo, C.A., Takezaki, N., and Nei, M. 1995. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12:** 391–404.

Slater, G.S.C. and Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.

Swanson, W.J. and Vacquier, V.D. 2002. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3:** 137–144.

Tellam, R.L., Wijffels, G., and Willadsen, P. 1999. Peritrophic matrix proteins. *Insect Biochem. Mol. Biol.* **29:** 87–101.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Xu, A., Park, S., D'Mello, S., Kim, E., Wang, Q., and Pikielny, C.W. 2002. Novel genes expressed in subsets of chemosensory sensilla on the front legs of male *Drosophila melanogaster*. *Cell Tissue Res.* **307:** 381–392.

Yang, Z. and Nielsen, R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19:** 908–917.

Received December 30, 2006; accepted in revised form March 26, 2007.