SUPPLEMENTARY INFORMATION

1. Sequence data

1.1 Genomic DNA and mRNA preparation

We obtained DNA for nine of the ten newly sequenced species from single lines that had been inbred for 8-14 generations. A single isofemale line of *D. grimshawi* was used because this species is prohibitively difficult to inbreed. Genomic DNA for preparation of plasmid, fosmid and BAC libraries was prepared from the following strains and biological materials (equal mixtures of males and females, except as noted). The sequence reads generated for each of the new projects are summarized in Supplemental Table 1. For strains where DNA was prepared from embryos, embryos were collected and DNA was isolated from nuclei using standard procedures; for the remaining strains, genomic DNA was isolated from mixed sex adults. RNA was isolated from embryos using the Trizol reagent and normalized cDNA libraries were prepared by standard methods. Stocks used for sequencing the new species presented here are described below. All sequenced strains are available from the Tucson Stock Center (http://stockcenter.arl.arizona.edu/).

***D. ananassae* AABBg1**. We used embryos of Tucson stock center strain number 14024-0371.13. This line was established by Y. Tobari and K. Kojima in 1967 from a single female and was maintained as a small mass culture for 35 years and then inbred by single pair mating for two generations in 2003. Pooled RNA from mixed total embryonic stages and total adult RNA from a white-eyed mutant strain from M. Matsuda at Kyorin University was used to make a normalized cDNA library for EST sequencing.

D. erecta. We used embryos of Tucson stock center strain number 14021-0224.01. This strain came from the wild type strain 14021-0224.00 inbred for eight generations by S. Castrezana in the Stock Center between 2003-2004. Pooled RNA from mixed total embryonic stages and total adult RNA from the same strain was used to make a normalized cDNA library for EST sequencing.

***D. grimshawi* G1**. We used embryos of Tucson stock center number 15287-2541.00. This is a wild-type strain collected in Maui, Hawaii. Pooled RNA from mixed total embryonic stages and total adult RNA from the same strain were used to make a normalized cDNA library for EST sequencing.

***D. mojavensis* CI 12 IB-4 g8**. We used embryos of Tucson stock center strain number 15081-1352.22. This is a wild type strain, inbred eight generations through single brother-sister pairs by L. Reed between the years of 2003-2004. Pooled RNA from mixed total embryonic stages and total adult RNA from Tucson stock center strain number 15081-1352.05 was used to make a normalized cDNA library for EST sequencing.

***D. sechellia* C**. We used embryos of Tucson stock center strain number 14021-0248.25. This line was started from a "Robertson" strain female obtained from J. Coyne. It was inbred three generations by C. Jones (UNC) and six generations by S. Castrezana at the Stock Center (2004).

***D. persimilis* MSH3**. We used embryos of Tucson stock center strain number 14011-0111.49. This strain was originally collected from Mount St. Helena (near Calistoga, CA), and identified by M. Noor (Duke University) using polytene chromosomes. This stock was inbred 15 generations by C. Machado (University of Arizona) (2003).

***D. simulans* 4**. We used adult flies of Tucson stock center strain number 14021-0251.216. This line was established by 10 generations of sib mating from a single inseminated female collected by D. Begun in the Wolfskill orchard, Winters, CA, (Summer 1995).

***D. simulans* 6.** We used adult flies of Tucson stock center strain number 14021-0251.194. This line was established by 10 generations of sib mating from a single inseminated female collected by D. Begun (University of California, Davis) in the Wolfskill orchard, Winters, CA (Summer 1995).

***D. simulans* w501**. We used embryos from the Tucson stock center strain number 14021-0251.195. This strain carries a white (eye color) mutation and has been in culture since the mid 20th century. It was likely descended from a female collected in N. America. The strain used for sequencing was sib mated for nine generations by D. Barbash at UC-Davis. Libraries for sequencing were prepared from DNA isolated from embryos.

***D. simulans* MD106TS**. We used adult flies of Tucson stock center strain number 14021-0251.196. This line is descended from a single inseminated female collected by J. W. O. Ballard in Ansirabe, Madagascar on 19 March 1998. It carries a siII mitochondrial genotype, and was cured of *Wolbachia* by tetracycline. This line was sib mated for five generations in the Ballard lab, followed by an additional five generations of sib mating by D. Begun.

***D. simulans* MD199S.** We used adult flies of Tucson stock center strain number 14021-0251.197 (females only). This line was descended from a single inseminated female collected by J. W. O. Ballard in Joffreville, Madagascar on 28 March 1998. This line has a siIII mitochondrial genotype, and has probably lost any Wolbachia infection. It was sib mated for five generations in the Ballard lab, followed by an additional five generations of sib mating by D. Begun. All-female DNA was made to assist in assembly of the Y chromosome by comparison to mixed-sex libraries of other lines.

***D. simulans* NC48S**. We used adult flies of Tucson stock center strain number 14021-0251.198. This line descended from a collection by F. Baba-Aissa in Noumea, New Caledonia in 1991. This line has a siI mitochondrial genotype. It was sib mated for five generations in the Ballard lab, followed by an additional five generations of sib mating by D. Begun.

***D. simulans* C167.4**. We used adult flies of Tucson stock center strain number 14021-0251.199. Descended from a collection in Nanyuki, Kenya. This strain is unusual in that can produce fertile females when hybridized to *D. melanogaster*. The line used for the genome project was obtained from the Ashburner laboratory via D. Barbash, and was

subjected to a total of 13 generations of sib mating.

D. virilis. We used eggs and embryos of Tucson stock center strain number 15010-1051.87. This stock was inbred for more than 14 generations from mutant stock # 15010-1051.46 by B. McAllister (and B. Charlesworth), then for two more generations in the Stock Center by S. Castrezana (2004). This stock is a multiple marked strain apparently created by The Institute for Developmental Biology in Moscow before it was donated to holdings currently maintained at the TSC. Pooled RNA from mixed total embryonic stages and total adult RNA from Tucson stock center strain number 15010-1051.45 was used to make a normalized cDNA library for EST sequencing.

***D. willistoni* Gd-H4-1**. We used eggs and embryos of Tucson stock center strain number 14030-0811.24. J. Powell inbred this stock for four generations in 2003 (Yale University), then it was inbred for five generations by S. Castrezana at the Stock Center (7/23/2004). Pooled RNA from mixed total embryonic stages and total adult RNA from Tucson stock center strain number 14030-0811.33 were used to make a normalized cDNA library for EST sequencing.

***D. yakuba* Tai18E2**. We used eggs and embryos of Tucson stock center strain number 14021-0261.01. This line derives from a single inseminated female captured in 1983 by D. Lachaise (CNRS, Gif-sur-Yvette) in the Taï rainforest, on the border of Liberia and Ivory Coast. This line was sib mated for 10 generations by A. Llopart and J. Coyne (University of Chicago). Inspection of 21 salivary gland polytene chromosomes showed no chromosomal rearrangements segregating within the strain. Therefore, Tai18E2 appears homokaryotypic for the standard arrangement in all chromosome arms, save 2R, which is homokaryotypic for 2Rn.

1.2 Heterozygosity testing

We used an adaptive heterozygosity testing procedure in order to verify the homozygosity of the stocks used for sequencing. We selected random loci for resequencing in 8-10 individuals of the stock for each species and assessed the data for the presence of any signs of heterozygous loci. The following species stocks were tested as described: *D. sechellia* (15 loci; 8 individuals), *D. persimilis* (35 loci; 8 individuals), *D. virilis* (5 loci; 8 individuals), *D. ananassae* (7 loci; 10 individuals), *D. mojavensis* (7 loci; 10 individuals), *D. erecta* (10 loci; 10 individuals), and *D. grimshawi* (10 loci, 10 individuals). No heterozygous bases were detected for *D. sechellia*, *D. virilis*, *D. ananassae*, *D. mojavensis*, *D. erecta*, and *D. grimshawi*; two heterozygous bases were detected in *D. persimilis*. The two heterozygous bases detected in the *D. persimilis* samples were heterozygous in all individuals, leading to the conclusion that all of these species were homogenously inbred at the loci tested (note that heterozygous-seeming bases that appear across all samples may also derive from PCR resequencing of loci that are not unique in the genome).

2. Sequence assembly

***D. persimilis* assisted assembly.** A new draft assembly of *Drosophila pseudoobscura* was generated with ARACHNE, to be used as a reference for the assembly of *Drosophila persimilis*. The summary data for that assembly are shown in Supplemental Table 17.

***D. yakuba* assembly.** To create the *D. yakuba* chromosomal FASTA files, we began by aligning the *D. yakuba* WGS assembly data against the *D. melanogaster* genome. *D. yakuba* supercontigs were artificially broken into 1000 bp fragments and aligned against the *D. melanogaster* genome using BLAT[1]. An alignment was defined as "unique" if its best scoring match had a score at least twice that of its next best scoring alignment. The *D. yakuba* contigs were initially ordered by the positions of their unique alignments along the assigned *D. melanogaster* chromosomes. Because there are rearrangements in *D. yakuba* as compared to *D. melanogaster*, we allowed one portion of a *D. yakuba* supercontig to align to one region of a chromosome and the remaining portion to align elsewhere along that chromosome. For example, four supercontigs aligned to both chromosome arms 2L and 2R. However, these 2L/2R crossovers and other interspecific non-linearities are expected given the known chromosome inversions between *D. yakuba* and *D. melanogaster*[2]. This initial ordering for 2L, 2R, 3L, 3R and X was used as the starting point for manually introducing inversions in the *D. melanogaster*-ordered *D. yakuba* supercontigs. The goal was to minimize the total number of inversions required to "rejoin" all *D. yakuba* supercontigs previously assigned to distant chromosomal regions based on *D. melanogaster* alignments (L. Hillier, unpublished). Inversions were only introduced between contigs and not within contigs. Using this process, we created the final chromosomal *D. yakuba* sequence.

***D. simulans* mosaic assembly**. We began by generating an ~2.9X WGS assembly of the *D. simulans* w501 line using PCAP. The w501contigs were initially anchored, ordered and oriented by alignment with the *D. melanogaster* genome in a manner similar to that described above for alignments between the *D. yakuba* and *D. melanogaster* genome. The assembly was then examined for places where the w501assembly suggested inversions with respect to the *D. melanogaster* assembly. One major inversion was found, confirming the documented inversion found by Lemeunier and Ashburner[2]. Six other *D. simulans* lines (c167.4, md106ts, md199s, nc48s, sim4, and sim6) were also assembled using PCAP with ~1X coverage. Using the 2.9X WGS assembly of the simulans w501 genome as a scaffold, contigs and unplaced reads from the 1X assemblies of the other individual *D. simulans* lines were used to cover gaps in the w501 assembly where possible. Thus, the resulting assembly is a mosaic containing the w501contigs as the primary scaffolding, with contigs and unplaced reads from the other lines filling gaps in the w501assembly (L. Hillier, unpublished). The *D. simulans* input read statistics from other lines are given in Supplemental Table 2.

Assembly reconciliation. Full details of the assembly reconciliation procedure are described in elsewhere[3]. A summary of the impact of reconciliation on assembly quality is presented in Supplemental Table 3.

Assembly quality assessment by synteny. Syntenic data from Synpipe was also used to pinpoint the location of probable genome assembly mis-joins. Over 95% of all *Drosophila* genes were found to be resident on the same Muller element between species[4]. This arm-level synteny conservation criterion was utilized to investigate possible mis-joins. In cases where large syntenic blocks belonging to different Muller elements were adjacent on the same scaffold without supporting evidence from other species, approximate locations of probable mis-assembly were identified (within a contig or between adjacent contigs). A number of these mis-joins, in the case of *D. sechellia* for example, were confirmed by the corresponding sequencing center (J. Chang, personal communication). Additional details have been provided in section 9.1

mtDNA sequence assembly. We used the *D. melanogaster* mtDNA sequence to probe the trace archives of the *Drosophila* species genome projects for the eight species whose mitochondrial genomes had not already been sequenced (*D. erecta, D. ananassae, D. persimilis, D. pseudoobscura, D. willistoni, D. virilis, D. mojavensis, D. grimshawi)*. For all species except *D. pseudoobscura*, we were able to assemble full mitochondrial genomes[5], excluding the A+T-rich regulatory region. The high depth of coverage for the mitochondrial genomes (15.7×–53.4×) was presumably due to the high copy number of mtDNA in cells. The seven assembled genomes were aligned to the previously sequenced *Drosophila* mtDNAs using ClustalX.

3. Annotation of repeats and transposable elements

Well-established TE/repeat libraries have been curated by the BDGP and Repbase only for *D. melanogaster*, and thus *de novo* repeat libraries were developed using PILER-DF[6] and ReAS[7]. In addition to these novel repeat libraries, we also used the BDGP TE library (http://www.fruitfly.org/p_disrupt/datasets/ASHBURNER/D_mel_transposon_sequence_set.fasta), a Dipteran PILER repeat library constructed from the 12 *Drosophila* genomes and *A. gambiae*[8], and an unbiased library of TEs from RepBase Update 11.06 that excludes all *Drosophila* entries (RepBase-noDros). We used four TE/repeat detection methods (RepeatMasker, BLASTER-TX, RepeatRunner and CompTE) to identify repetitive elements in non-*melanogaster* species. To control for the influences of different sequencing and assembly strategies, we focused our analysis of TE/repeat content on large scaffolds (>200 kb) that are likely to be representative of the euchromatic portions of each genome. We ran BLASTER-TX with three repeat libraries (BDGP, Dipteran PILER, and RepBase-noDros). We also ran RepeatMasker using the ReAS library, RepeatRunner with the PILER library, and CompTE, which does not employ a repeat library. Details of these methods are presented below.

To assess the accuracy of each method, we calibrated the different annotation methods using the estimated 5.5% TE content in *D. melanogaster* (Release 4)[9]. The BLASTER-TX method yields lower estimates of TE/repeat content relative to the *D. melanogaster* annotation and appears to be sensitive to the representation of *Drosophila*-specific repeats in each library (Repbase-NoDros: 1.9% < BDGP: 4.0% < PILER: 4.5%). RepeatMasker+ReAS (6.2%), RepeatRunner+PILER (7.2%), and CompTE (8.9%) give

higher estimates of TE/repeat content relative to the current *D. melanogaster* annotation (Supplemental Figure 1). We also investigated whether the rank order abundance of different classes of TE annotated in *D. melanogaster* (LTR retrotransposons > LINE-like retrotransposons > Terminal Inverted Repeat (TIR) DNA-based transposons[9]) was consistent across the phylogeny using an unbiased, conservative approach based on the BLASTER-tx+Repbase-NoDros analysis, which uses a library of TEs from outgroup species. The rank order abundance of major TE classes obtained by this method when applied to *D. melanogaster* is consistent with previous high-resolution TE annotations[9] (LTR: 60% > LINE-like: 23% > TIR: 9%, OTHER: 8%; Supplemental Figure 6), suggesting that this method can accurately assess TE class abundance. Applying this method to non-*melanogaster* species revealed that the rank-order abundances of TE classes are generally conserved (Supplemental Figure 6) and are not biased by uneven representation in the species-specific TE libraries.

3.1 ReAs

An improved version of ReAS[7] program was used to detect and assemble repeat consensus sequences for each of the 12 *Drosophila* genomes. For each species, raw whole-genome-shotgun reads were downloaded from NCBI TraceDB, vector sequences were screened with Cross_match (http://www.phrap.org/) and reads shorter than 100bp were removed. Candidate repeat-containing reads were identified as those having *k*-mers that occur at a frequency higher than expected based on the whole genome shotgun coverage. Reads sharing high-depth *k*-mers were aligned to each other using Cross_match with the mat70 similarity matrix, dust[10] was used to filter simple-sequence alignments, joining information between each pair of repeat segments was determined by refining pairwise alignment, and complete joining information among all repeat segments was used to form a connection network. Finally, consensus sequences were created through searching the paths in the connection network using MUSCLE[11, 12] as the multi-alignment engine. The parameters for the repeat assemblies in ReAS were: 1) *k*-mer size, K=17; 2) depth threshold, D=30 for *D. sechellia* and *D. persimilis*, D=50 for *D. melanogaster*, and D=40 for all other genomes; 3) Identity threshold of pairwise alignment hits, 70%. RepeatMasker v3.1.6 with WU-BLAST v2.2.6 as the search engine was used to annotate repeats in each species using the resulting species-specific ReAS repeat libraries.

3.2 CompTE

CompTE is a comparative method to detect repetitive, inserted elements based on the phylogenetic signature of large insertions in multiple genome alignments of related species[13]. This method identifies genomic regions enriched for transposable element (TE) sequences, and can also estimate the phylogenetic branch where the insertion occurred. Candidate repeat elements 'Insertion Regions' (IRs) were extracted from MAVID/MERCATOR whole genome alignments of the 12 *Drosophila* genomes that displayed the signature of large insertions. To confirm that IRs represented repetitive elements, a series of subsequent filters was applied to ensure sequence similarity to other such IRs in the genome, and to break long blocks of IRs into individual repeat units. Filtered IRs are called 'Repeat Insertion Regions' (RIRs) and are the output of the CompTE method. Since CompTE contains information about the branch on which a putative TE insertion has occurred, this method can also be used to identify ancestral

repeats that are inferred to have transposed on the branch leading to common ancestor of species sharing the repeat, and therefore can identify insertion events that support alternative phylogenetic relationships among species.

3.3 RepeatRunner

RepeatRunner annotations for CAF1 assemblies and construction of the Dipteran PILER libraries and have been described previously[8]. Briefly, RepeatRunner (http://www.yandell-lab.org/repeat_runner/index.html) annotates repeats based on the combined output of RepeatMasker using nucleotide sequences in the Repbase Update 10.07 *Drosophila* library, plus output of WU-BLASTX 2.0MP using a database of 37,972 TE proteins compiled from GenBank (GB-TE). For the RepeatRunner+PILER analyses reported here, RepeatMasker was run using Repbase Update 10.07 *Drosophila* library supplemented with 892 Dipteran-specific genomic repeats discovered with PILER-DF[6] using the genome sequences of all 12 *Drosophila* species plus *Anopheles gambiae*.

3.4 BLASTER-TX

A TBLASTX method using the BLASTER toolkit[14, 15] was used to annotate CAF1 genomic scaffolds >10 kb. To save computer time and reduce software memory requirements, large scaffolds were segmented into chunks of 200 kb overlapping by 10 kb. 200 kb chunks were then masked for simple repeats by RepeatMasker (6-Mar-2004) using sensitive (-s -noint -no_is) parameters and then by TRF 3.21 using Match=2, Mismatch=3, Delta=5, PM=80, PI=10, Minscore=20, MaxPeriod=15 parameters[16]. The resulting masked chunks were compared to 3 different TE libraries using BLASTER with WU-TBLASTX v2.0 (compiled 10-May-2005) with default parameters. The first library is the BDGP TE reference set v9.4.1 (BDGP-TE), which contains all known *D. melanogaster* TEs, plus a smaller number from other species in the genus. The second library is the Dipteran PILER library described above[8]. The third is derived from Repbase Update 11.06 in which we retained only TEs not belonging to the genus *Drosophila* (RU-noDros). Because most *Drosophila* TEs in Repbase are biased toward *D. melanogaster*, removing all *Drosophila* TEs allows us to remove any species bias in the TE library in estimating the relative TE content across species. For each library, we filtered for overlapping hits on the genomic sequence using the MATCHER program toolkit[14, 15], by keeping the one with the best alignment score and truncating the other so that only non-overlapping regions remain. All matches with $E > 1 \times 10^{-10}$ or length <=20 are eliminated. Redundant annotations in the overlaps between chunks were identified and merged prior converting chunk coordinates back to scaffolds sequence coordinates.

4. Protein-coding gene models

4.1 GLEAN set production and reconciliation

Annotation sets were built independently by multiple groups from the research community. Submissions were coordinated via a wiki at http://rana.lbl.gov/drosophila/wiki/index.php/Main_Page, and included gene models built using: SNAP with and without homology guidance (Don Gilbert), GeneMapper (Sourav Chatterji and Lior Pachter), Exonerate (Andreas Heger and Chris Ponting), GeneWise, Exonerate and GeneMapper (Venky Iyer and Michael Eisen), Gnomon (Kim Pruitt),

CONTRAST (Samuel Gross and Serafim Batzoglou), N-SCAN (Randall Brown and Michael Brent), and GeneID (Francisco Camara and Roderic Guigo). Predictions were standardized and filtered to retain annotations for coding exons only. These initial sets were combined into a single consensus set using GLEAN, a gene model combiner that chooses the most probable combination of start, stop, donor and acceptor sites from the input predictions[17, 18]. Non-independent annotation sets (multiple predictions using the same algorithm; GeneWise and Exonerate; SNAP +/- homology guidance) were grouped in GLEAN runs. A second consensus set of gene predictions was also built using JIGSAW by Jonathan Allen and Steven Salzberg.

Comparison of GLEAN models with the well-annotated gene features in *D. melanogaster* revealed numerous instances of gene models corresponding to well-supported *D. melanogaster* genes inappropriately joined with other genes or split into multiple genes in non-*melanogaster* species, presumably because of incorrect *de novo* predictions. Since homology-based annotations are less prone to this problem, we built a new annotation set for each species, using GLEAN to combine gene predictions from homology-based methods (GeneWise, Exonerate, GeneMapper and homology-supported Gnomon) into a "strict homology set" (GLEAN-SH). We filtered the *de novo* sets (GeneID, SNAP, N-SCAN, CONTRAST, and *de novo* Gnomon) to remove gene models that overlapped more than one GLEAN-SH model, and used these filtered sets to build a second consensus annotation set. This second annotation set was filtered further to bias the retention of homology-supported exons, while still permitting the inclusion of additional exons from *de novo* prediction sets, to produce the "filtered plus homology set" (GLEAN-FPH). Finally, we reconciled the GLEAN-SH and GLEAN-FPH to produce the "reconciled consensus set" (GLEAN-R). To do this, we removed any GLEAN-FPH models that merged multiple GLEAN-SH models and any GLEAN-FPH models that were missing one or more GLEAN-SH exons, and added back any missing GLEAN-SH models; this process yielded the reconciled consensus set GLEAN-R.

Importantly, the GLEAN-R set does not predict alternative splice forms in non-*melanogaster* species: overlapping transcript models were statistically "collapsed" to generate a consensus gene model. However, homology-based alternative-transcript models based on the *D. melanogaster* transcript annotations are available and have been used in some of our analyses (where noted).

4.2 Gene model quality

A series of Nimblegen oligonucleotide microarrays were designed to match the predicted protein-coding genes from six of the *Drosophila* species and to serve as quality control tests of predicted gene models. Array elements were designed according to preliminary annotations of draft assemblies, with 17-22k gene predictions per species and ~10 probes per gene. Genes were considered expressed if signal intensity for the probes targeting a gene model was significantly higher than the signal intensity of 2,517 negative controls that target *Arabidopsis* genes. To assess significance, we compared probe intensities for each gene to negative controls using a Mann-Whitney U test: any GLEAN-R model with probe intensities significantly above the negative controls at a false discovery rate[19] of 0.1% was considered expressed. For further details see ref. 20. Gene models for which we do not detect transcription may represent transcribed genes that are simply not

expressed at detectable levels in adult flies under laboratory conditions. Furthermore, mere presence of a transcript does not guarantee that the GLEAN-R model represents a protein-coding gene, and our design cannot assess the predicted gene structure. However, we believe that screening for transcriptional activity of GLEAN-R models represents a reasonable first step towards assessing the overall quality of our predicted gene sets.

We flagged gene models as putatively TE-contaminated using RepeatMasker with *de novo* ReAS libraries and PFAM structural annotations of the GLEAN-R gene set. We flagged gene models as false positives if ≥90% of the CDS region was masked by ReAS repeats, or if they contained "parasitic" domains characteristically found in TEs and viruses. We assessed the reliability of the two approaches used to detect TE-contaminated gene models by applying these procedures to the well-characterized *D. melanogaster* genome; consistent with previous results[21] gene models in this species are largely uncontaminated by TE sequences (Fig. 2, Supplemental Table 7). In contrast, ReAS overlaps and PFAM annotations indicate that 2.6–31.1% and 4.1–18.1%, respectively, of gene models in non-*melanogaster* species reflect TE-contamination.

Several observations support the inference that these gene models are indeed false positives derived from transposable element sequences. First, a large proportion of putative false positives are independently flagged by both ReAS overlaps and PFAM annotations. Second, the proportions of gene models that are flagged by each approach are highly correlated (Spearman's rho = 0.88, $P = 0.00017$). Third, the vast majority of genes flagged using ReAS overlaps are present in a single lineage only. Lastly, the proportion of putative TE-contaminated genes is correlated with genomic repeat content (Spearman's rho = 0.71, $P = 0.013$). We thus emphasize that gene models flagged as potentially TE-contaminated, especially those found only in single lineages, should be treated with caution, and these gene models have been removed from the final gene prediction set used in subsequent analyses.

5. Homology

5.1 Homology assignment by fuzzy reciprocal BLAST clustering

All-by-all BLASTP searches (each translation against every genome translation set) were performed. BLASTP hits were binned such that each bin was separated by a log10(E-value) jump of at least 2, with the reasoning that ordering hits by E-values is imperfect (and based on inspection of the distribution of E-value jumps between the hits). A graph consisting of nodes corresponding to all the translations from the 12 genomes was constructed, and BLASTP queries were connected by one-way edges to each of the hits from the first E-value bin. All non-reciprocal edges were removed, and the connected components of this graph, representing homology clusters, were discovered using a recursive algorithm. Homology clusters were parsed by species and by gene (combining alternative translations for *D. melanogaster*), merging clusters when necessary. Clusters were classified by the number of members from a species: multiple members indicated potential paralogy in that species. Comparisons to pairwise INPARANOID[22] runs showed that the FRB method produced very similar homology predictions (results not

shown), but took much less time and computed homologies simultaneously for all the species under consideration.

5.2 Homology assignment with Synpipe pipeline

Synpipe uses an annotated peptide set from a reference species (in this case, *D. melanogaster*), a genome assembly from the target species, and an initial set of similarity inferences based on a tool such as TBLASTN[23]. It employs a graph-based algorithm to infer blocks of synteny and to refine homology assignments with the objective of maximizing synteny based on user-defined thresholds for micro-syntenic scrambling. Homologous locations for genes, in the presence of alternate paralogous placements, are chosen with respect to the synteny maximization criteria. An initial vertex set is derived from preliminary homology hits based on input from a similarity inference tool. Graph edges are added to link neighbouring vertices that are in the same order as in the reference species. Accommodating for missing genes and scrambling thresholds, these graphs are extended into synteny chains. Singletons, or genes in paralogous locations, are moved to alternate locations to incorporate them into synteny chains, wherever possible. Collisions, overlapping hits of genes with similar coding domains, are resolved as far as possible based on increasing synteny. Synpipe does not determine gene models in the candidate assembly but assesses orthologous locations for synteny analysis. It accommodates contig and scaffold gaps in the target genome assembly by identifying homologous elements that might either fall in unsequenced assembly gaps, lie on the edges of sequenced segments, or on small assembly fragments. This is important in the context of shotgun assemblies used in this analysis, ensuring that missing elements do not disrupt synteny. Synpipe was used to analyze the set of *Drosophila* genome assemblies relative to the *D. melanogaster* gene order and the resulting syntenic dataset has been used for breakpoint analysis, a comparative study of chromosomal rearrangements between species, multi-species alignment and orthology refinement, and for mapping and orienting scaffolds along chromosome arms[24].

5.3 Final homology calls

To generate a set of homology calls, we merged the FRB homology calls with the Synpipe homology calls. Pairwise Synpipe calls (between each species and *D. melanogaster*) were mapped to GLEAN-R models, filtered to retain only 1:1 relationships, and added to the FRB calls when they did not conflict and were non-redundant. This reconciled FRB+Synpipe set of homology calls forms the basis of our subsequent analyses. Two versions are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/): one in which potentially TE-contaminated gene models are filtered prior to clustering, and one in which TE-contaminated gene models are retained.

5.4 Validation of homology calls with GeneWise

Sequence homology based pipelines for assessing the presence or absence of genes, especially is the face of assemblies with regions of low quality sequence, can incorrectly call rapidly evolving genes, or genes in assembly gaps, as missing. Indeed, a surprisingly large number of lethal-mutable genes in *D. melanogaster* were identified as absent in at

least one non-*melanogaster* species (Supplementary Fig. 2), suggesting that a number of homologs of *D. melanogaster* genes may have been missed by these homology calls. We used a GeneWise pipeline to assess the validity of gene absences inferred by FRB and Synpipe homology calls. We began with genes that were designated as "missing" in at least one species based on FRB homology assignment, and if the gene was inferred to be present based on Synpipe, the procedure terminated and we designated this gene as "not missing." However, if the gene was inferred to be missing based on Synpipe homology calls, we retrieved the first gene 5' and 3' that was present in that species that had a single orthologue in *D. melanogaster* (or a synteny-resolved orthologue that mapped to the corresponding place in *D. melanogaster*). If the neighbouring genes were not on the same contigs, this procedure terminated and we designated this gene as "not assessed." If the neighbours were on the same contigs, we checked the distance between the inferred absent gene and its neighbours. If this distance was greater than 400 kb, we terminated this process and designated this gene as "not assessed" largely because of computational difficulties associated with running GeneWise on a fragment of this size. If this distance was less than 400 kb, we ran GeneWise using the peptide from *D. melanogaster* by default. For species outside of the *melanogaster* group, we looked for a peptide from a more closely related species and ran GeneWise using that peptide if available. We first ran GeneWise on one strand, and if the score was less than 100, we ran GeneWise on the other strand, keeping track of the highest score. From the GeneWise run, we extracted the Wise score and the alignments. We also noted the proportion of ambiguous base calls (N's) in the sequence. Genes were designated with "ambiguous homology" if the syntenic region contained more than 1% N's, or if the genes were designated as "not assessed." Genes were designated as "absent" if the GeneWise score was < 35; otherwise, genes were considered "present." To be conservative, we considered genes with ambiguous homology to be present, as they likely represent cases where assembly gaps or low quality sequence has led to the absence of a gene model. Of the 13,733 *D. melanogaster* protein coding genes we analyzed, 11,644 (84.8%) could be assigned a homology pattern unambiguously and were not flagged as potentially non-protein-coding[25]. Proteins with uncertainties, either because of orthology ambiguities or because they were flagged by ref. 25 are not randomly distributed among chromosome arms ($P < 1.0 \times 10^{-63}$; see Supplemental Table 19). Revised homology patterns for each *D. melanogaster* gene are available from FlyBase (ftp://ftp.flybase.net/12_species_analysis/).

6. Alignment masking

An initial set of amino acid alignments was generated using the TCOFFEE multiple alignment tool[26]. To preserve codon phase, corresponding nucleotide alignments were generated by threading CDS sequences through amino acid alignments. Manual inspection of these alignments revealed several types of errors likely introduced by variation in gene annotations among species or by sequencing or assembly errors. The 5' and 3' ends of genes appeared most problematic: in addition to truncated gene models, we also discovered multiple cases where first and/or last exons were missed or incorrectly annotated. This necessitated the development of a computational screen, which we used to remove these problematic regions from the multi-species alignments.

This approach was largely made possible by our confidence in the gene models from *D. melanogaster*, as we assume that any alignment issues are caused by non-*melanogaster* species. Pairwise subalignments between *D. melanogaster* and each non-*melanogaster* species were screened using a sliding window, and any alignment window with divergence above a species-specific cutoff was masked in that target species (Supplemental Table 20). Species-specific cutoffs were determined by a combination of examination of the empirical distributions of divergences for each species pair, and simulations of random sequence. Notably, the alignment masking procedure reveals potential problems with low-coverage assemblies: *D. persimilis*, *D. sechellia*, and *D. simulans* all have a higher fraction of masked bases than their nearest sister taxa (*D. pseudoobscura*, and the *D. yakuba/D. erecta* clade; Supplemental Figure 3).

7. Annotation of non-coding RNAs

Non-coding RNA genes were predicted using a variety of *de novo* and homology search methods. Except where specifically noted, All BLAST analysis used WU-BLASTN 2.0 (http://blast.wustl.edu/), with a word size of 3 and sum statistics turned off (-kap option).

7.1 tRNAs

A combined set of *de novo* transfer RNA (tRNA) gene predictions were obtained from the union of tRNAscan-SE 1.23[27] using options -H -y and Aragorn 1.1[28] using options -w -t -i116 -l -d and subsequently parsed by TFAM 0.02 classifier[29] using -c TRNA2-eu.cm -m sprinzl_euk_cy.coevam to confirm tRNA identities and predict initiator tRNAs.

7.2 miRNAs

A non-redundant set of 78 *D. melanogaster* microRNA genes (miRNAs) was collected from the miRBase database v8.1[30]. Homologs of the *D. melanogaster* miRNAs were identified using a semi-automated process. Putative precursor and mature sequences were identified using BLAST (E-value $\leq 1x10^{-2}$). Two or fewer mismatches in the mature miRNA sequence were accepted and correct folding of the putative precursor to a miRNA hairpin with good folding energy ($\leq$ -20 kcal/mol) was confirmed using RNAfold[31]. Lower scoring matches were included by inspection of predicted folds, phylogenetic trees, and conservation in multiple sequence alignments, and conservation of synteny. We note that following miRBase convention, gene names in non-*melanogaster* species were assigned based on homology to the mature miRNA, rather than strict orthology.

7.3 snoRNAs

D. melanogaster small nucleolar RNA genes (snoRNAs) were collated from the EMBL and RefSeq Genome databases (Sept 2006) to supplement the 63 snoRNAs annotated in FlyBase (v4.3). Duplicated entries from multiple submissions of the same snoRNA and partial snoRNA fragments were manually removed, and the remaining 276 sequences were ultimately mapped to the *D. melanogaster* assembly. When two snoRNAs overlapped on genome coordinates, hits were manually inspected and the longer of the two hits was retained. Each snoRNA prediction was extended to the full length of the

snoRNA query sequence to produce the final set of 250 *D. melanogaster* snoRNA annotations. Homologs of all 250 *D. melanogaster* snoRNAs (plus 2 snoRNAs that could not be mapped to the *D. melanogaster* genome, AJ809564 and AJ784386) were identified in the other genomes using BLAST ($E \leq 1x10^{-6}$). Additional matches with conservation of synteny in MAVID/MERCATOR CAF1 v1 whole genome alignments were accepted with E≤1e-2. Only coordinates of the HSP alignment were used for snoRNA annotations in non-*melanogaster* species, and in many instances will not represent complete snoRNAs.

7.4 snRNAs

Sequences of all *D. melanogaster* spliceosomal, small nuclear RNA genes (snRNAs: U1, U2, U4, U4atac, U5, U6, U6atac, U11, and U12) were used as queries against the CAF1 *Drosophila* assemblies in NCBI BLAST v2.2.1 searches with following parameters: -r 5 -q -4 -G 10 -E 6 -W 7 -FF -X 150 -y 100 -Z 250 -e 0.1. In case of U11 genes, other metazoan sequences (including *D. pseudoobscura*) were used as queries in order to improve or confirm original BLAST hits. In the next step, the nucleotide sequences around each hit were extracted and used as input for INFERNAL v0.6[32] to refine the location of snRNA gene predictions for candidate genes that didn't produce global alignment with a query sequence. Functionality of the determined genes was assessed by the integrity of the gene and by the presence of the proximal sequence element A (PSEA) motif at the expected distance from the transcription start site of a gene (see [33] for details).

7.5 Other ncRNAs

D. melanogaster 18S rRNA, 28S rRNAs, *roX1* and *roX2* sequences were obtained from the public databases (GenBank entries M21017 (1..1995 bp), M2017 (3288..7232 bp), U85980, U85981). These sequences were mapped to the *D. melanogaster* genome using BLAST (requiring ≥98% identity across the full length of the query sequence). Homologs of these ncRNAs were identified in the other *Drosophila* assemblies using BLAST ($e \leq 1x10^{-6}$) and manual inspection. rRNA fragment matches were assembled into contiguous regions and inspected manually[34]. All other RNAs were predicted using the RFAM automated annotation pipeline comprising the Rfam 7.0 library of covariance models[35], INFERNAL 0.55 software[32], and NCBI BLAST 2.2.6 (http://www.ncbi.nlm.nih.gov/ftp/). Where redundant annotations were made by the RFAM pipeline and ncRNA genes listed above, the curated sets of ncRNA genes took precedence in the final genome annotation. The RFAM pipeline also identifies several classes of regulatory elements in mRNAs, which are included in the genome annotation, although these additional features are excluded from ncRNA gene counts.

8. cis-regulatory sequences

8.1 Annotation and alignment

The cis-regulatory modules (CRMs)[36] used in this study were downloaded from the REDfly database (http://redfly.ccr.buffalo.edu/) in GFF3 format, time stamped June 11th 2007. Because the 628 regulatory-region features contained many redundancies,

including large reporter constructs that encompassed numerous individual CRMs, we filtered out the large encompassing features and then collapsed the remaining overlapping features. This produced a non-redundant set of 333 CRMs for downstream analysis. The DNase I footprinted binding sites[37] used in this study were downloaded from the Drosophila DNase I Footprint Database v2.0 (http://www.flyreg.org/) in GFF2 format. The optimal position weight matrix matches for TFBS in the Drosophila DNase I Footprint Database v2.0 footprints are available here in GFF3 format. The 30 position weight matrices (PWMs) used in this study were downloaded from http://www.bioinf.manchester.ac.uk/bergman/data/motifs/optimal.xms. Optimal PWM matches within footprints were estimated using a likelihood ratio between the PWM model and genome-wide nucleotide frequencies (60/40 AT/GC) on each footprint plus 10 bp of flanking sequence. Optimal PWM matches within footprints were estimated using each footprint plus 10 bp of flanking sequence, allowing for inaccuracies in mapping of footprints to current genomic coordinates. Alignments of all features analyzed in "Evolution of cis-Regulatory DNAs" were performed as described in Halligan and Keightley[38].

We identified orthologous CRSs in the *D. simulans* genome by the following reciprocal-best-hit BLAST approach. For shorter elements (including all footprints and CRMs <500bp) we BLASTed the element itself along with 30bp flanking DNA either side to increase the chances of a best-hit. For long elements (>500bp) we chose instead to BLAST 100bp from each end of the element (to help identify the whole orthologous element in *D. simulans*) and checked that each end blasted to the same contig in the correct orientation. The orthologous elements were aligned using MCALIGN2[39], to be consistent with the alignments produced by Halligan and Keightley[38].

8.2 Estimation of constraint

Selective constraint (C; the fraction of mutations removed by natural selection) was calculated by comparing the expected number of substitutions (E), calculated using a putatively neutrally evolving standard, to the number observed (O). It has previously been shown that subsections (base pairs 8–30) of short introns (<80bp in length) in *Drosophila* represent a good candidate for a class of neutrally evolving sequence and have been termed the fastest evolving intron (FEI) sites[38]. We took FEI site alignments between *D. melanogaster* and *D. simulans* produced by Halligan and Keightley[38] and concatenated sites from ~1Mb sections of the genome together to create a local neutral standard sequence for each section. For each CRS, we estimated four different pairwise substitution rates (k_i, i=1...4; A«T, C«G, A«C/T«G, A«G/T«C) using the local neutral standard sequence from the appropriate section of the genome and calculated the expected number of substitutions as $\sum_i k_i.M_i$, where M_i is the number of noncoding sites in the CRS at which a substitution of type i can occur. Constraint in the CRS was then calculated as $C = 1 - O / E$. This method attempts to account for differences in the base composition of the putatively unconstrained and test sequences, but assumes that substitution rates of each possible type are equal in both the putatively unconstrained and test sequence.

9. Conservation of genomic context of protein-coding sequences

9.1 Synteny maps and application

Our synteny maps were based on TBLASTN hits of annotated *D. melanogaster* genes, employing an algorithm that optimizes the assignment of orthologous gene pairs to maximize global estimates of synteny between each of the species and *D. melanogaster*[24].

Synteny data were used to assign assembly scaffolds to Muller elements and infer their order and orientation along chromosome arms, supplementing experimental analysis using known markers (S. Schaeffer, personal communication). To provide higher-order organization to assemblies, we mapped most of the major scaffolds to the chromosome arms (which are referred to as Muller elements and designated by the letters A-F). For the species of the *melanogaster* subgroup, we aligned the scaffolds to the *D. melanogaster* genome, and adjusted the location to account for the known chromosomal rearrangements[2, 40, 41]. For remaining species, we used the known locations of a few genes to assign scaffolds to the proper Muller element. However, direct determination of their order and orientation was not possible because the large number of overlapping inversions reordered loci within the arms. Synteny data helped in such cases by making use of comparative evidence from evolutionarily close species to infer scaffold-joins. This was done using Synpipe homology calls on scaffold edges and synteny maximization criteria (with closely related species) while allowing for missing elements in scaffold gaps, single species breakpoints and localized rearrangements (scrambling)(S. Schaeffer, personal communication).

Analysis of the disruption in gene order in various species allowed the annotation of fixed two-break inversions between species and breakpoint reuse within species (A. Bhutkar, S. Russo, T. F. Smith and W. Gelbart, personal communication). In a number of instances, *D. melanogaster* heterochromatin genes[42] were found to have homologous placements in parts of the sequenced genomes. This suggests that either parts of the heterochromatin have been sequenced for these species, or that there could be some level of movement of genes between heterochromatin and euchromatin. In support of the first possibility, a number of assembled scaffolds were assigned to heterochromatic regions of the genome based on a majority of *D. melanogaster* heterochromatin genes being localized there. Paracentric, and in some cases pericentric, inversions are thought to play a role in the movement of heterochromatic genes into euchromatic regions of the genome.

In order to analyze fine-scale rearrangement of gene-order, we employed the NGP algorithm[43]. This allows for the inclusion of micro-syntenic changes to gene order and orientation, involving a small number of genes or even single genes, in addition to macro-syntenic changes called by Synpipe. It records adjacent pairs of homologous genes across species and uses a comparative approach to analyze the conservation and disruption of these pairs. The resulting dataset is used to infer the number of rearrangement breaks (macro and micro-syntenic) and infer ancestral gene order. Analysis of *Drosophila* species shows a wide range of such rearrangement breaks across species, from the root of the genus *Drosophila* (few hundreds to over 700)[43].

9.2 Hox identification and analysis

All previously annotated *Hox* genes (from *D. melanogaster*, *D. pseudoobscura* and *D. buzzatii*) were used as queries in BLAST searches against all other *Drosophila* genomes and *Anopheles gambiae*. Identified contigs were analysed for position and orientation of *Hox* genes and presence of non-*Hox* genes between and next to *Hox* genes. *Hox* regions were aligned with VISTA[44] to confirm correspondence of non-coding sequences and identify the exact position of splits and reorganisations. Comparison of *Hox* gene organisations along the phylogeny allowed the reconstruction the evolutionary history of these regions. See ref. 45 for details on *Hox* gene organisations and analysis of functional constraints.

10. Gene family dynamics

10.1 Gene family expansion/contraction

To estimate the average gene gain/loss rate and to identify gene families that have undergone significant size changes, we applied the probabilistic framework developed by ref. 46. By using a stochastic birth and death model for the gene gain and loss across species, and a probabilistic graphical model for the dependence relationship between branches of the phylogeny, this framework can infer the rate and direction of the change in gene family size. A total of 11,434 families including 148,326 genes were analyzed; all families that were not inferred by parsimony to have been present in at least one copy in the most recent common ancestor of all 12 species were excluded, as were all genes with evidence for TE contamination.

This likelihood approach also offers a null hypothesis against which we can compare the rate of evolution of individual gene families. Using the maximum likelihood parameters inferred from the whole dataset, we ran Monte Carlo simulations to test for significant rate accelerations in all 11,434 families. Using $P<0.0001$ we expect there to be approximately one significant result by chance; the observation of 342 families with lower *P*-values implies a false discovery rate of 0.003%. To identify the branch of the *Drosophila* tree with the most unlikely amount of change for these 342 families, we calculated the exact *P*-values for transitions over every branch (the "Viterbi" method in ref. 46). We called individual branches significant at $P<0.005$.

10.2 Lineage-specific genes

In order to produce a high-confidence set of lineage-restricted *D. melanogaster* genes, we relied on the revised homology classifications generated by the GeneWise pipeline described above; genes restricted to the *melanogaster* group or other clades were extracted from the revised homology calls produced. To test for differences in expression pattern in lineage-restricted genes compared to ancestrally present genes, we downloaded tissue expression data from FlyAtlas (www.flyatlas.org). To assign tissue specificity, we called a given gene specific for a tissue if it was expressed above background on at least 3 (out of 4) microarrays in one tissue and only 0 or 1 microarrays for all other tissues. While we believe that these genes represent lineage-restricted genes (and not rapidly diverged paralogues), we cannot exclude the possibility that some of these genes are in

fact rapidly diverging homologs of ancestrally present genes that were missed by our syntenic pipeline because of translocations or movement from heterochromatin to euchromatin.

11. Evolution of protein-coding sequences

11.1 PAML analysis of protein-coding genes

All analyses were performed on the masked set of either the guide-tree alignments of genes with a single orthologue in all 12 *Drosophila* species or just the 6 species in the *melanogaster* group (*D. melanogaster*, *D. sechellia*, *D. simulans*, *D. erecta*, *D. ananassae*). Estimation of rates of evolution and tests for positive selection were performed using codon substitution models and likelihood ratio tests implemented in the program PAML version 3.15[47] and are described in more detail in ref. 48 . For each gene within the *melanogaster* group, we ran PAML models M0, M7, and M8, with branch lengths as free parameters and codon frequencies estimated by F3x4. PAML model M0 estimates a single ω that is fixed across the phylogeny for each alignment. Unless otherwise noted, when we refer to the ω of a gene it is this estimate to which we are referring. Because the topology of the (yak,ere) clade relative to the (mel(sim,sec)) clade is uncertain[49], we ran PAML on all three possible topologies and used the data from the run with the best likelihood. Using only the best supported topology overall – (mel(sim,sec),(yak,ere)) – does not change our results. In order to avoid convergence problems, we ran each analysis three times with different initial values of ω, and used the run with the best likelihood.

To test for positive selection, we used a likelihood ratio test that compares the fit of the data to one model (M7) with ω following a beta (0,1] distribution (so codons with $\omega > 1$ are excluded) to a second model (M8) that has an additional class of sites where $\omega > 1$ [50]. A significant P-value for this test indicates that there is support for a subset of codons within a gene that are under positive selection ($\omega > 1$). P-values for the likelihood ratio test were determined by simulating 12,000 alignments under model M7 (simulated in PAML evolverNSsites), using nucleotide frequency parameters and branch lengths estimated from the empirical data, and then estimating the likelihood of M7 and M8 to generate a null distribution of likelihood ratios. P-values for the test of positive selection were corrected for multiple testing using the qvalue package in R[51]. In order to assess the extent to which we may be biased towards detecting genes with low d_S in our test for positive selection, we compared the median d_S of genes with evidence for positive selection at a 10% false discovery rate to those without. There is no significant difference between these sets of genes (Mann-Whitney U test P=0.30), suggesting that our test for positive selection does not preferentially detect genes with low d_S.

11.2 Positive selection and selective constraints

To determine whether gene function predicts patterns of evolutionary constraint and positive selection, we downloaded the Gene Ontology annotations (http://www.geneontology.org/) assigned to each gene from Flybase (http://flybase.bio.indiana.edu/). These assignments were mapped genes onto a customized ontology of 115 functional categories representing 36 terms describing a

molecular function, 15 terms describing a cellular component and 64 terms describing a biological process (see Supplemental Table 12). Low confidence evidence codes were excluded from the analysis (associations where the evidence was "inferred by electronic annotation," or any gene mapped directly to either biological process, cellular component or molecular function where the evidence was "no data"). For each gene, the relevant parameter value (either ω, d_N, amino acid divergence, or the probability of positive selection) was obtained from the PAML results described above (and described in more detail in [48]). Permutation tests were performed by shuffling the parameter values of interest for all genes with a GO annotation. The genes with their newly assigned parameter value were then re-mapped back to their assigned GO category and the median of the permuted parameters were then calculated for each GO category. P-values were assigned by comparing the true observed median parameter value for each GO term to its permuted distribution (10,000 iterations) using the ecdf function in R.

Since an elevated ω can be caused not only by an increase in the fixation rate of nonsynonymous substitutions (d_N) but also by a decrease in d_S, we cross-checked the ω results with d_N and found qualitatively similar results for most GO categories (see Supplemental Table 12). We also used a resampling approach to test whether genes with similar levels of codon bias and d_S have similar ω, since weak selection on synonymous sites (codon bias) may cause a decrease in the synonymous substitution rate, inflating ω and giving the appearance of a rapidly evolving gene. For each gene belonging to a rapidly evolving GO category (see Supplemental Table 12), we matched it to another gene outside that GO category with a similar d_S and FOP (frequency of optimal codon). We did this ranking genes by root-mean-squared difference ($d = \sqrt{(A_{i(FOP)} - B_{j(FOP)})^2 + (A_{i(dS)} - B_{j(dS)})^2}$) in FOP and d_S, where A_i and B_j correspond to individual genes belonging to a GO category with an elevated ω and a gene outside that category, respectively. We found that ω for the resampled genes were significantly different (MWU, P=0.0492), indicating that in general, the elevated ω of genes belonging to these GO categories are not necessarily due to variation in d_S and codon bias.

To determine whether genes belonging to the same functional group have similar values of ω, the permutation tests were performed as described above and instead of calculating the median ω, the standard error of ω was calculated and compared to the observed true standard error for each GO term. To determine whether there was significantly less variance in the distribution of ω than probabilities of positive selection within each GO term, we calculated the median standard error in both parameters within each term and then applied a one-tailed Mann-Whitney U test. False discovery rate correction was applied using the Benjamini and Hochberg[19] FDR procedure implemented in the p.adjust function in R and using the qvalue package to estimate the fraction of true positives. To test for clustering of positively selected codons within genes, we defined as positively selected any codon with a Bayesian posterior probability of positive selection greater than 95% (although using 90% or 75% cutoffs does not change our results). We adapted a version of the test for clustering proposed by Tang and Lewontin[52], using custom Perl scripts. InterPro domain assignments were downloaded from the InterPro website, and mapped to the masked alignments used in our analysis. InterPro assignments were filtered to retain only “domains” before use.

11.3 Factors affecting the rate of protein evolution

Principal component ANCOVAs using principal components constructed from eight potentially confounding variables as covariates (mRNA expression level, expression breadth, number of protein-protein interactions, recombination rate, protein length, average intron length, and number of introns) and gene dispensability type (essential genes described as those with lethal or sterile alleles in FlyBase; viable genes described as genes with a non-lethal/sterile effect on phenotype) as a categorical variable were performed as described in ref. 48, in order to test for an effect of dispensability class on protein evolution. Partial correlations were used to investigate the independent effect of each of the seven listed continuous variables on rates of protein evolution, also as described in ref. 48.

11.3 Chemoreceptors

Detailed methods are outlined in refs. 53 and 54. Briefly, all *Or* and *Gr* genes in *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, and *D. ananassae* in the CAF1 genome assemblies were annotated using two independent pipelines of TBLASTN searches (*melanogaster* protein sequences were used as queries) and GeneWise predictions. All putative lack-of-function mutations were verified by direct re-sequencing from the genome strains (and an additional outbred strain for *sechellia* and *erecta*). Orthologues were defined as unique reciprocal best hits that shared at least one adjacent upstream or downstream neighbour (*i.e.* were microsyntenic). To estimate the rates of evolution and the level of constraint on *Or/Gr* genes, we ran two different PAML models on each set of orthologues from the 5 *melanogaster* subgroup species; the first model gave each lineage on the five species tree its own unique while the second model assigned one ω ratio to the specialist branch (*D. sechellia* and *D. erecta*) and one to the generalist branches (rest of the lineages; ref. 53).

11.4 Immunity

Homologs of 245 *D. melanogaster* genes with immune-related functions were identified in all 12 species using the FRB homology calls as a guideline, and manually annotating and correcting GLEANR models as necessary. The 245 *D. melanogaster* genes and their homologs were divided into functional classes (recognition, signalling, effector) and homology patterns were assigned based on manually revised homology annotations. Tests for positive selection and estimates of ω were obtained in PAML as described above and elsewhere[48, 55]. Detailed methods can be found in ref. 55.

11.5 Reproduction

Sex and reproduction related (SRR) genes were determined by a reciprocal BLAST approach (described in ref. 56) for genes that either were associated with sex-related GO terms (spermatogenesis, spermiogenesis or oogenesis) or had a significant BLAST hit to EST from either testes or ovary but not the head[56]. Estimates of ω were obtained as described in ref. 48.

12. ncRNA analysis

12.1 Reconstructing ancestral miRNA sequences

For each of the 78 miRNA genes, we inferred the ancestral sequences of the ancestral nodes based on the known phylogeny of the 9 *Drosophila* species (*D.melanogaster, D. simulans, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. virilis, D. mojavensis, and D. grimshawi*) using PAML[47] and GASP[57]. The miRNA secondary structure of each node was predicted using RNAfold[31] and mfold[58]. Mutations in mature miRNAs were inferred using Fitch Parsimony.

12.2 miRNA likelihood analysis

We identified putative orthologous regions of 78 annotated *Drosophila* pre-miRNAs using the CAF1 miRNA annotations. Eighteen miRNAs were excluded for the following reasons: five (*mir-283, mir-284, mir-288, mir-31a,* and *mir-31b*) had an unusual structure of two hairpin loops, seven (*mir-2b-1, mir-311, mir-312, mir-310, mir-313, mir-289, mir-303*) did not have confident orthologue assignments, five (*mir-11, mir-263b, mir-309, mir-310, mir-313, mir-6-1* and *mir-3*) failed to be structurally aligned with their orthologues, and one (*mir-iab-4*) was annotated with two mature regions in *D. melanogaster*. As the boundaries of pre-miRNAs are experimentally uncertain, we used the operational definition that a pre-miRNA must be a dsRNA containing at least the mature miRNA and its complement. Our structural alignment procedure is as follows and used the Vienna package v1.6[31] with BioPerl v1.4[59]: 1) fold the *D. melanogaster* pre-miRNA with RNAfold (-T 25 -noLP) to identify the region complementary to the miRNA; 2) align orthologous regions using CLUSTALW (default nucleotide parameters)[60]; 3) trim the alignment to include the smallest foldable hairpin containing the miRNA and its complement; 4) structurally align using RNAfold (-T 25 -noLP) and PMMULTI v1.1[61] 5) compute the consensus structure with RNAalifold (-T 25 -noLP); and 6) re-trim the final structural alignment. We then evaluated all alignments by eye and altered one alignment (*mir-287*) manually. Alignments of individual pre-miRNA were concatenated together and sites were partitioned according to whether in the RNAalifold consensus structure they were paired or unpaired, loop, and inside, outside or complement to the miRNA, for a total of six site-classes. Gaps (ambiguous characters) were added to alignments where pre-miRNA orthologues in a species were unidentified. A bootstrap analysis (100 replicates) was conducted on maximum likelihood estimates of evolutionary rates made with OPTIMIZER in the PHASE v2.0 package[62], using the RNA6B substitution model for paired states and HKY85 for unpaired states, and the following topology
(((((((Dsim,Dsec),Dmel),(Dyak,Dere)),Dana),(Dpse,Dper)),Dwil),((Dmoj,Dvir),Dgri)).

12.3 ncRNA stem-loop rate analyses

We used cmalign in the INFERNAL package[32] to produce structurally informed multiple alignments of the predicted ncRNAs. Substitution rate estimation was performed using the general phylo-grammar engine XRate[63] to train a grammar with a PFOLD-like structure[63] on approximately half of RFAM, giving substitution rates "typical" of the stem and loop regions of the training data. We introduced two substitution rate scaling parameters, one for stem and one for loop regions, and used the EM procedure

implemented in XRate to estimate these scaling factors for the alignments of each family. In general, the absolute rate of evolution for a given multiple alignment can only be estimated to within an arbitrary scaling factor, unless the branch lengths of the underlying gene tree can be ascertained independently. Since these branch lengths are unknown in most of these ncRNA gene families, we report L/S ratios rather than absolute values for L and S.

13. Compensatory evolution in predicted intronic RNA structures

Intron sequences of six *Drosophila* species (*D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. virilis* and *D. mojavensis*) were obtained from the UCSC server (http://genome.ucsc.edu/) and aligned using tools of the Genome Browser[64] and TCoffee v2.0[26]. The sequence alignments were searched for RNA secondary structural elements (termed helices or stems) with Piranah v1.1[65] that is based on a likelihood ratio test (LRT) by [66]. In addition, we required that each predicted helix contain at least one covariation. Two other alignment-based programs – Alidot v2.0.5[67] and Alifold v1.6alpha[31] – were used to distinguish between conflicting helix predictions. Altogether 160 introns of length > 200 bp were sufficiently conserved to be alignable in all six species. The search of RNA secondary structure in these introns yielded 27 helices with an LRT > 15. Given the phylogenetic distances considered here, helices with LRT values > 15 are likely to be real[65].

14. References

1. Kent, W. J. BLAT--the BLAST-like alignment tool. *Genome Research* **12**, 656-664 (2002).
2. Lemeunier, F. & Ashburner, M. A. Relationships within the melanogaster species subgroup of the genus Drosophila (Sophophora). II. Phylogenetic relationships between six species based upon polytene chromosome banding sequences. *Proceedings of the Royal Society of London Series B: Biological Sciences* **193**, 275-294 (1976).
3. Zimin, A. V., Smith, D. R., Sutton, G. & Yorke, J. A. Assembly Reconciliation. *Bioinformatics*, submitted.
4. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome Scale Analysis of Positionally Relocated Genes. *Genome Research*, in the press.
5. Montooth, K. L., Abt, D. N., Hoffman, J. & Rand, D. M. Evolution of the mitochondrial DNA across twelve species of Drosophila. *Mol Biol Evol*, submitted.
6. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152-i158 (2005).
7. Li, Q. et al. ReAS: Recovery of ancestral sequences for transposable elements from the unassembled reads of a whole shotgun. *PloS Computational Biology* **1**, e43 (2005).
8. Smith, C. D. et al. Improved repeat identification and masking in Dipterans. *Gene* **389**, 1-9 (2007).
9. Bergman, C. M., Quesneville, H., Anxolabehere, D. & Ashburner, M. Recurrent

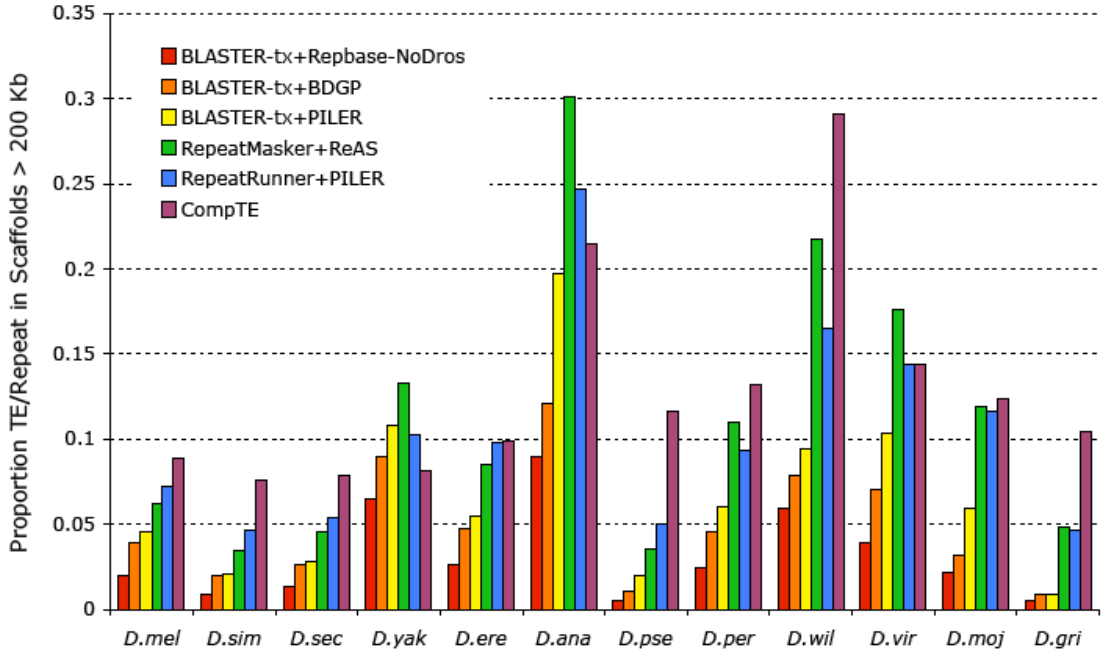
insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. *Genome Biology* **7**, R112 (2006).
10. Morgulis, A., Gertz, E. M., Schaffer, A. A. & Agarwala, R. A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology* **13**, 1028-1040 (2006).
11. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *Bmc Bioinformatics* **5**, 1-19 (2004).
12. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
13. Caspi, A. & Pachter, L. Identification of transposable elements using multiple alignments of related genomes. *Genome Research* **16**, 260-270 (2006).
14. Quesneville, H., Nouaud, D. & Anxolabehere, D. Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes. *Journal of Molecular Evolution* **57**, S50-S59 (2003).
15. Quesneville, H., Nouaud, D. & Anxolabehere, D. Recurrent recruitment of the THAP DNA-Binding domain and molecular domestication of the p-transposable element. *Molecular Biology and Evolution* **22**, 741-746 (2005).
16. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).
17. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera. *Nature* **443**, 931-49 (2006).
18. Elsik, C. G. et al. Creating a honey bee consensus gene set. *Genome Biol* **8**, R13 (2007).
19. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate- a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* **57**, 289-300 (1995).
20. Zhang, Y., D. Sturgill, M. Parisi, S. Kumar, B. Oliver. Constraint and turnover in sex-biased gene expression in the genus Drosophila. *Nature* doi:10.1038/nature06323, (in this issue).
21. Lipatov, M., Lenkov, K., Petrov, D. A. & Bergman, C. M. Paucity of chimeric transposable element transcripts in the Drosophila melanogaster genome. *BioMed Cetral Biology* **3**, 24 (2005).
22. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res* **33** (2005).
23. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
24. Bhutkar, A., Russo, S., Smith, T. F. & Gelbart, W. M. Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations. *Genome Informatics* **17** (2006).
25. Stark *et al*. Comprehensive discovery of functional elements in 12 fly genomes using evolutionary signatures. *Nature*, in the press.
26. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology* **302**, 205-217 (2000).
27. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**, 955-964 (1997).
28. Laslett, D. & Canback, B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Research* **32**, 11-16 (2004).
29. Ardell, D. H. & Andersson, S. G. E. TFAM detects co-evolution of tRNA identity rules with lateral transfer of histidyl-tRNA synthetase. *Nucleic Acids Research*

34, 893-904 (2006).
30. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research* **34**, D140-D144 (2006).
31. Hofacker, I. L. et al. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly* **125**, 167-188 (1994).
32. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**, 18 (2002).
33. Mount, S. M., Gotea, V., Lin, C. F., Hernandez, K. & Makalowski, W. Spliceosomal small nuclear RNA genes in 11 insect genomes. *Rna* **13**, 5-14 (2007).
34. Stage, D. E. & Eickbush, T. H. Sequence variation within the rRNA gene loci of 12 Drosophila species. *Genome Research*, in the press.
35. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research* **33**, D121-D124 (2005).
36. Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. REDfly: a Regulatory Element Database for Drosophila. *Bioinformatics* **22**, 381-3 (2006).
37. Bergman, C. M., Carlson, J. W. & Celniker, S. E. Drosophila DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, Drosophila melanogaster. *Bioinformatics* **21**, 1747-9 (2005).
38. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research* **16**, 875-884 (2006).
39. Wang, J., Keightley, P. D. & Johnson, T. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. *BMC Bioinformatics* **7**, 292 (2006).
40. Lemeunier, F., David, J. R. & Tsacas, L. in The Genetics and Biology of Drosophila (eds. Ashburner, M., Carson, H. I. & Thompson, J. K.) 147-188 (Academic Press, London, 1986).
41. Ashburner, M., Golic, K. G. & Hawley, R. S. Drosophila: a laboratory handbook (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y., 2003).
42. Smith, C. D., Shu, S. Q., Mungall, C. J. & Karpen, G. H. The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science* **316**, 1586-1591 (2007).
43. Bhutkar, A., Gelbart, W. M. & Smith, T. F. Inferring genome-scale rearrangement phylogeny and ancestral gene order: A Drosophila case study. *Genome Biology*, in the press.
44. Mayor, C. et al. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046-7 (2000).
45. Negre, B. & Ruiz, A. HOM-C evolution in Drosophila: is there a need for Hox gene clustering? *Trends Genet* **23**, 55-9 (2007).
46. Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C. & Cristianini, N. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research* **15**, 1153-1160 (2005).
47. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**, 555-6 (1997).
48. Larracuente, A. M. et al. Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics*, submitted.
49. Pollard, D. A., Iyer, V. N., Moses, A. M. & Eisen, M. B. Widespread discordance of gene trees with species tree in Drosophila: evidence for

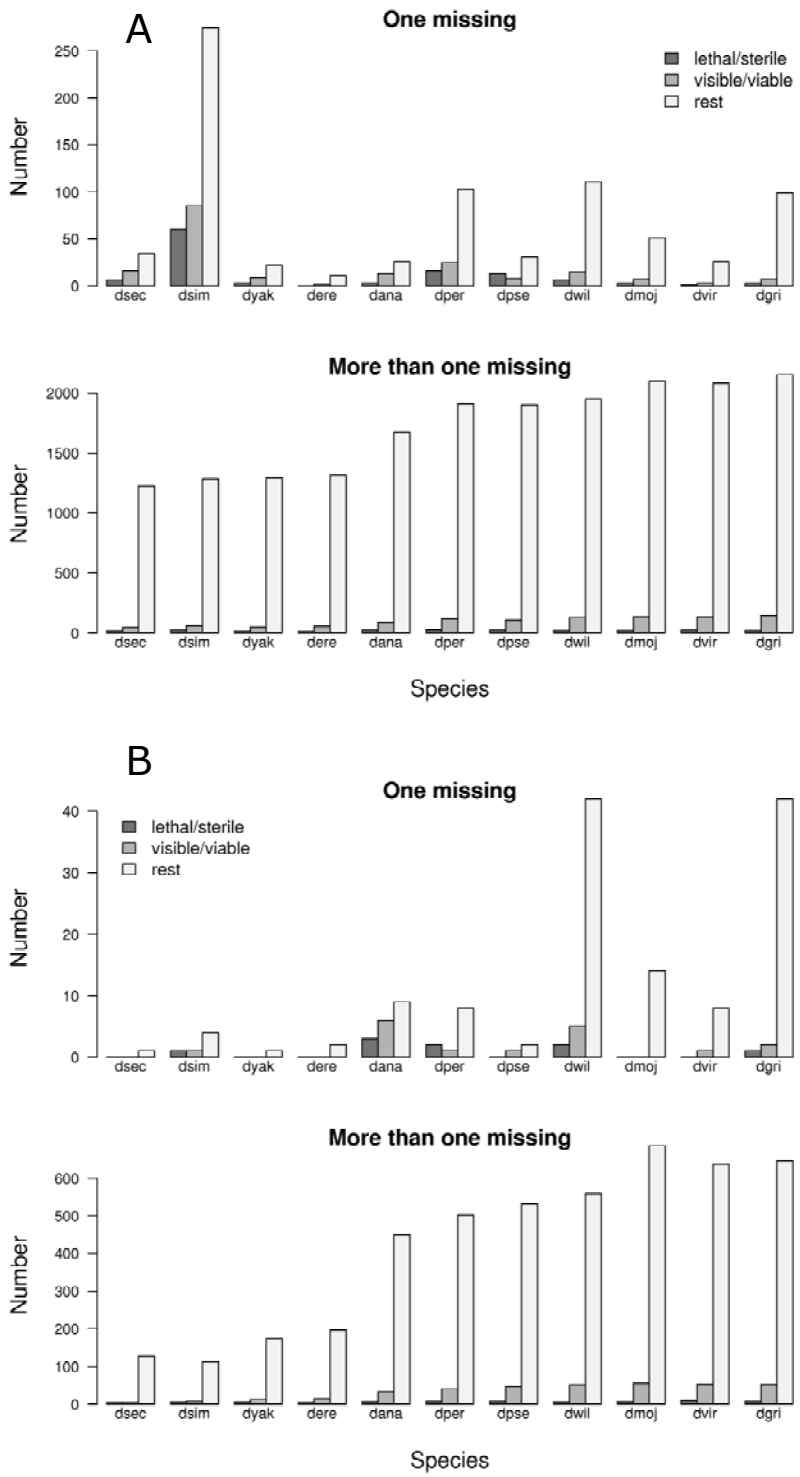
incomplete lineage sorting. *PLoS genetics* **2**, e173 (2006).
50. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**, 908-17 (2002).
51. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 479-498 (2002).
52. Tang, H. & Lewontin, R. C. Locating regions of differential variability in DNA and protein sequences. *Genetics* **153**, 485-95 (1999).
53. McBride, C. S. & Arguello, J. R. Five Drosophila genomes reveal non-neutral evolution and the signature of host specialization in the chemoreceptor superfamily. *Genetics*, in the press.
54. Vieira, F. G., Sanchez-Gracia, A. & Rozas, J. Comparative genomic analysis of the Odorant-Binding Protein family in 12 Drosophila genomes: Purifying selection and birth-and-death evolution. *Genome Biology*, in the press.
55. Sackton, T. B. et al. The evolution of the innate immune system across Drosophila. *Nature Genetics*, submitted.
56. Haerty, W. et al. Evolution in the fast lane: rapidly evolving sex-and reproduction-related genes in Drosophila species. *Genetics*, submitted.
57. Edwards, R. J. & Shields, D. C. GASP: Gapped Ancestral Sequence Prediction for proteins. *BMC Bioinformatics* **5**, 123 (2004).
58. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-15 (2003).
59. Stajich, J. E. et al. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res* **12**, 1611-8 (2002).
60. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
61. Hofacker, I. L., Bernhart, S. H. & Stadler, P. F. Alignment of RNA base pairing probability matrices. *Bioinformatics* **20**, 2222-7 (2004).
62. Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M. & Higgs, P. G. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol Phylogenet Evol* **28**, 241-52 (2003).
63. Klosterman, P. S. et al. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics* **7**, 428 (2006).
64. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).
65. Parsch, J., Braverman, J. M. & Stephan, W. Comparative sequence analysis and patterns of covariation in RNA secondary structures. *Genetics* **154**, 909-21 (2000).
66. Muse, S. V. Evolutionary analyses of DNA sequences subject to constraints on secondary structure. *Genetics*, 1429-1439 (1995).
67. Hofacker, I. L. et al. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res* **26**, 3825-36 (1998).
68. Bassett, M. H., McCarthy, J. L., Waterman, M. R. & Sliter, T. J. Sequence and developmental expression of Cyp18, a member of a new cytochrome P450 family from Drosophila. *Molecular and Cellular Endocrinology* **131**, 39-49 (1997).
69. Petryk, A. et al. Shade is the Drosophila P450 enzyme that mediates the hydroxylation of ecdysone to the steroid insect molting hormone 20-hydroxyecdysone. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 13773-13778 (2003).

70. Gutierrez, E., Wiggins, D., Fielding, B. & Gould, A. P. Specialized hepatocyte-like cells regulate Drosophila lipid metabolism. *Nature* **445**, 275-280 (2007).
71. Willingham, A. T. & Keil, T. A tissue specific cytochrome P450 required for the structure and function of Drosophila sensory organs. *Mechanisms of Development* **121**, 1289-1297 (2004).
72. Warren, J. T. et al. Molecular and biochemical characterization of two P450 enzymes in the ecdysteroidogenic pathway of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11043-11048 (2002).
73. Sztal, T. et al. Two independent duplications forming the Cyp307a genes in Drosophila. *Insect Biochem. and Mol. Biol.*, submitted.
74. Feyereisen, R. Evolution of insect P450. *Biochem Soc Trans* **34**, 1252-5 (2006).
75. Brandt, A. et al. Differential expression and induction of two Drosophila cytochrome P450 genes near the Rst(2)DDT locus. *Insect Molecular Biology* **11**, 337-341 (2002).
76. Daborn, P. J. et al. A single P450 allele associated with insecticide resistance in Drosophila. *Science* **297**, 2253-2256 (2002).
77. Amichot, M. et al. Point mutations associated with insecticide resistance in the Drosophila cytochrome P450 Cyp6a2 enable DDT metabolism. *European Journal of Biochemistry* **271**, 1250-1257 (2004).
78. Bogwitz, M. R. et al. Cyp12a4 confers lufenuron resistance in a natural population of Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 12807-12812 (2005).
79. Board, P. G. et al. Clarification of the role of key active site residues of glutathione transferase zeta/maleylacetoacetate isomerase by a new spectrophotometric technique. *Biochem J* **374**, 731-7 (2003).
80. Agianian, B. et al. Structure of a Drosophila sigma class glutathione S-transferase reveals a novel active site topography suited for lipid peroxidation products. *Journal of Molecular Biology* **326**, 151-165 (2003).
81. Singh, S. P., Coronella, J. A., Benes, H., Cochrane, B. J. & Zimniak, P. Catalytic function of Drosophila melanogaster glutathione S-transferase DmGSTS1-1 (GST-2) in conjugation of lipid peroxidation end products. *European Journal of Biochemistry* **268**, 2912-2923 (2001).
82. Kim, J. et al. Identification and characteristics of the structural gene for the Drosophila eye colour mutant sepia, encoding PDA synthase, a member of the Omega class glutathione S-transferases. *Biochemical Journal* **398**, 451-460 (2006).
83. Provost, E. & Shearn, A. The Suppressor of Killer of prune, a unique glutathione S-transferase. *Journal of Bioenergetics and Biomembranes* **38**, 189-195 (2006).
84. Sawicki, R., Singh, S. P., Mondal, A. K., Benes, H. & Zimniak, P. Cloning, expression and biochemical characterization of one Epsilon-class (GST-3) and ten Delta-class (GST-1) glutathione S-transferases from Drosophila melanogaster, and identification of additional nine members of the Epsilon class. *Biochemical Journal* **370**, 661-669 (2003).
85. Tang, A. H. & Tu, C. P. Biochemical characterization of Drosophila glutathione S-transferases D1 and D21. *J Biol Chem* **269**, 27876-84 (1994).

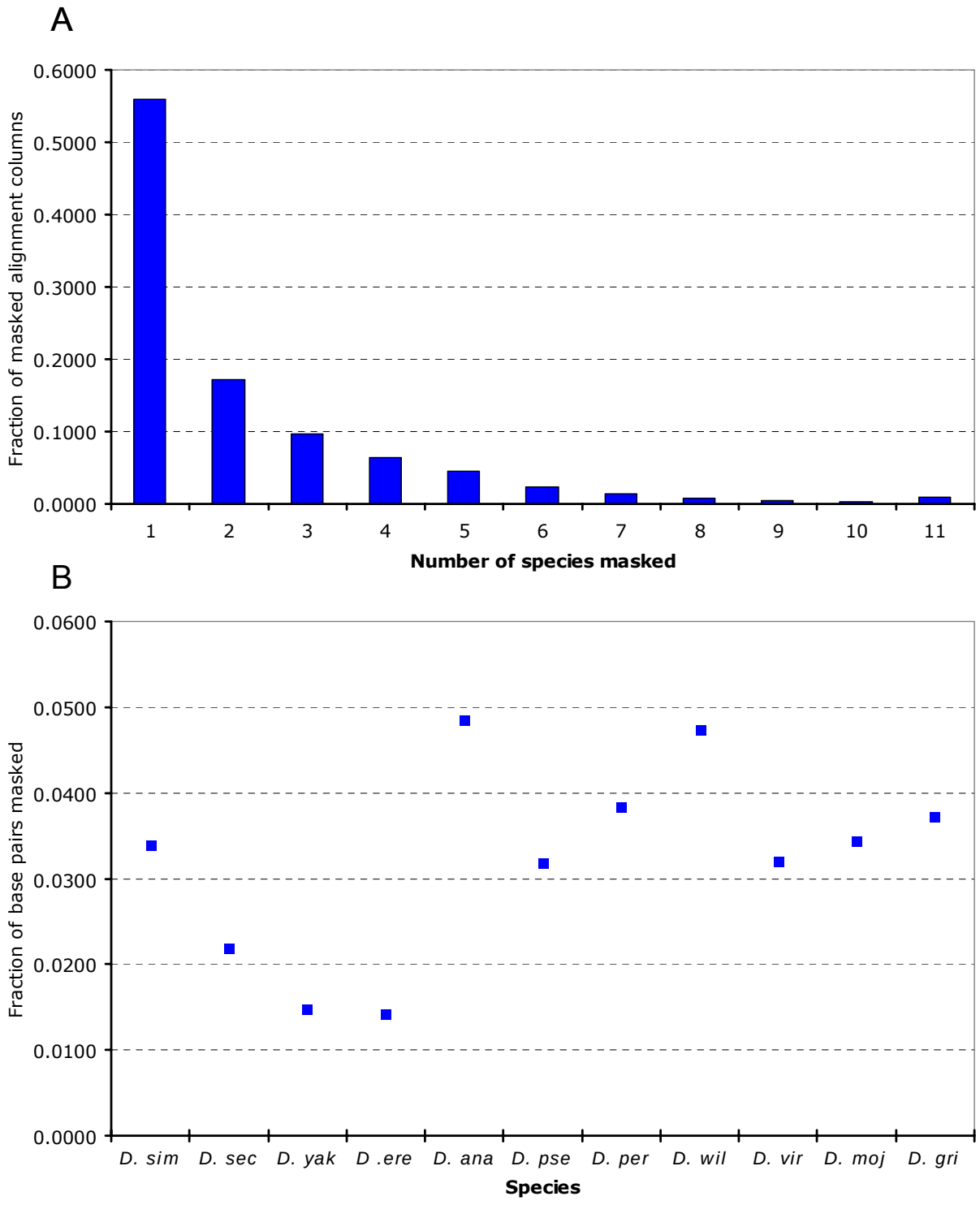
15. Figures



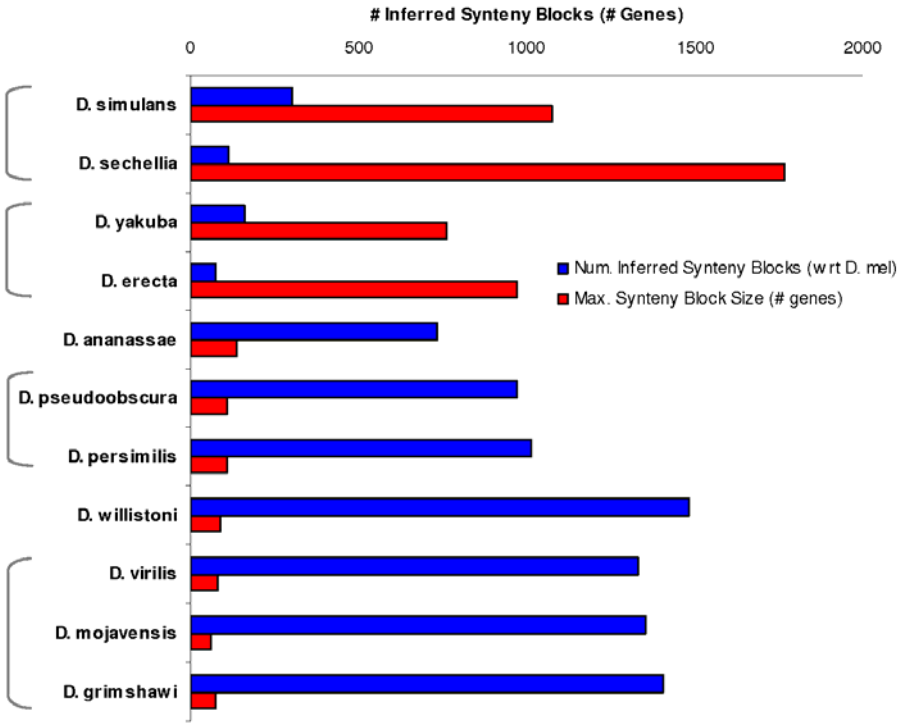
Supplemental Figure 1. **Repeat and TE content of the 12 *Drosophila* genomes.** Fraction of each genome covered by repeats based on different methods of repeat and TE annotation.



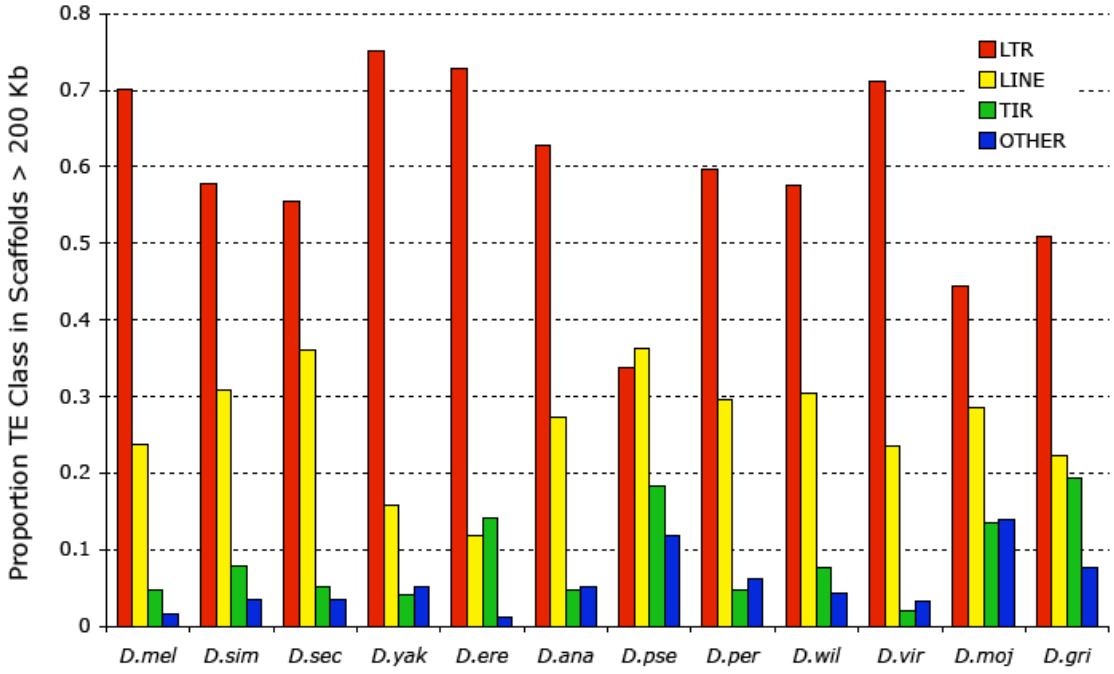
Supplemental Figure 2. **Number of missing genes in different viability classes.** Number of genes called absent in different viability classes (based on mutations in FlyBase) for each species based on FRB homology tables (A) and after GeneWise correction (B). In each panel, the upper plot shows genes missing in just one species, and the lower plot shows genes missing in any number of species.



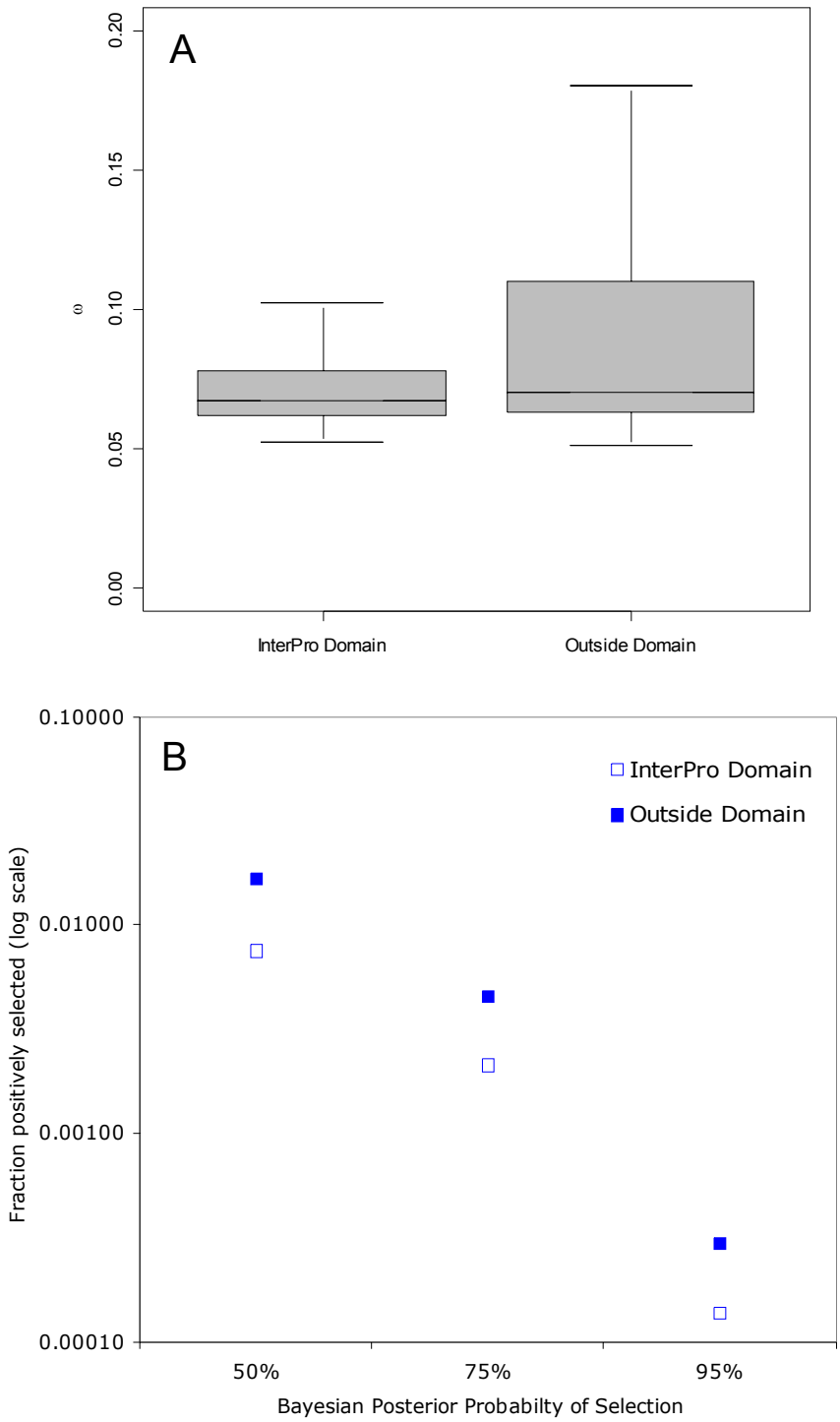
Supplemental Figure 3. **Effect of masking procedure on sequence alignments.** A) Histogram of the total number of species masked for every column in the alignment where there is at least one species masked. B) Total fraction of aligned bases masked in each of the 11 non-*melanogaster* species.



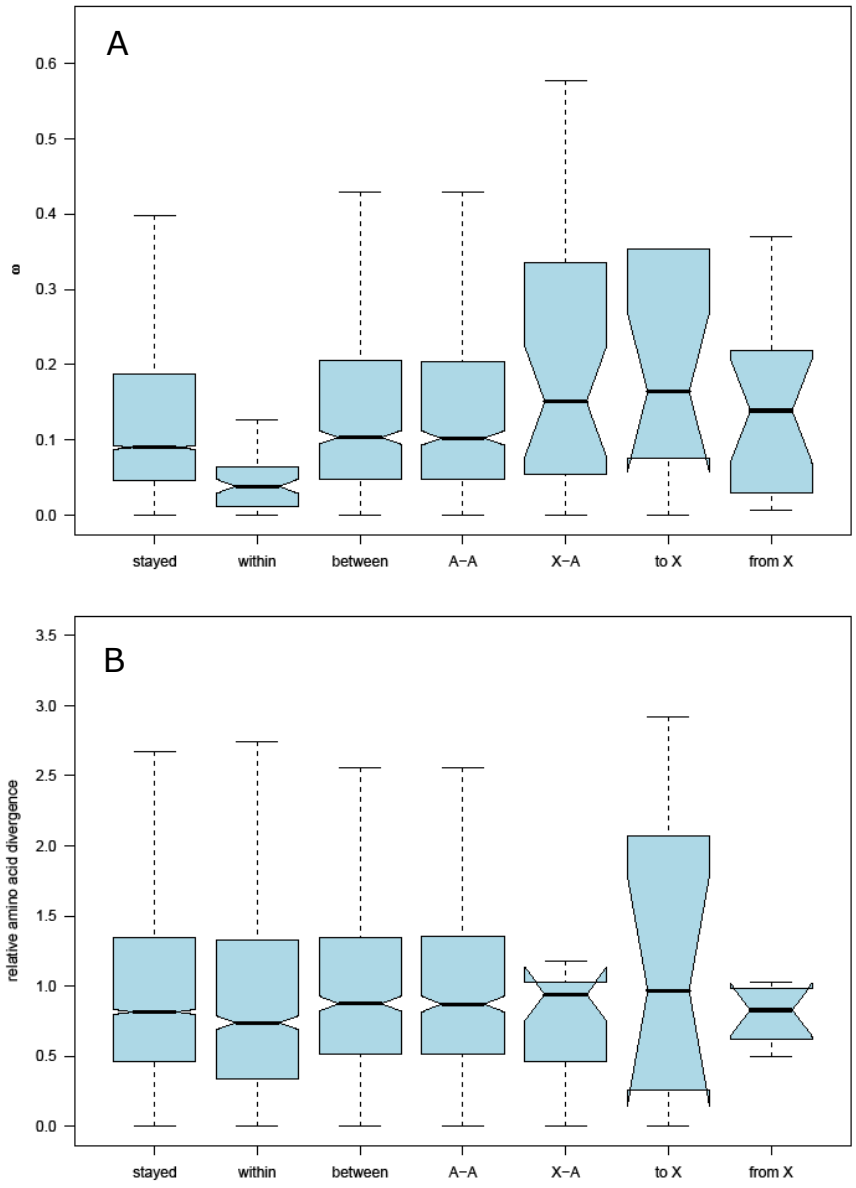
Supplemental Figure 4. **Number of inferred synteny blocks and maximum block size with respect to *D. melanogaster* gene order.** Species are listed (from top to bottom) according to increasing evolutionary divergence from *D. melanogaster*. Bracketed species are equidistant from *D. melanogaster*. Synteny blocks were inferred using Synpipe[24] where micro-syntenic scrambling within a threshold of 10 genes was allowed within syntenic blocks. Using conservative criteria, synteny blocks were terminated at assembly scaffold edges. The number of inferred blocks is generally increases with increasing divergence from the reference species (*D. melanogaster*) and the size of the largest block (by number of genes) decreases.



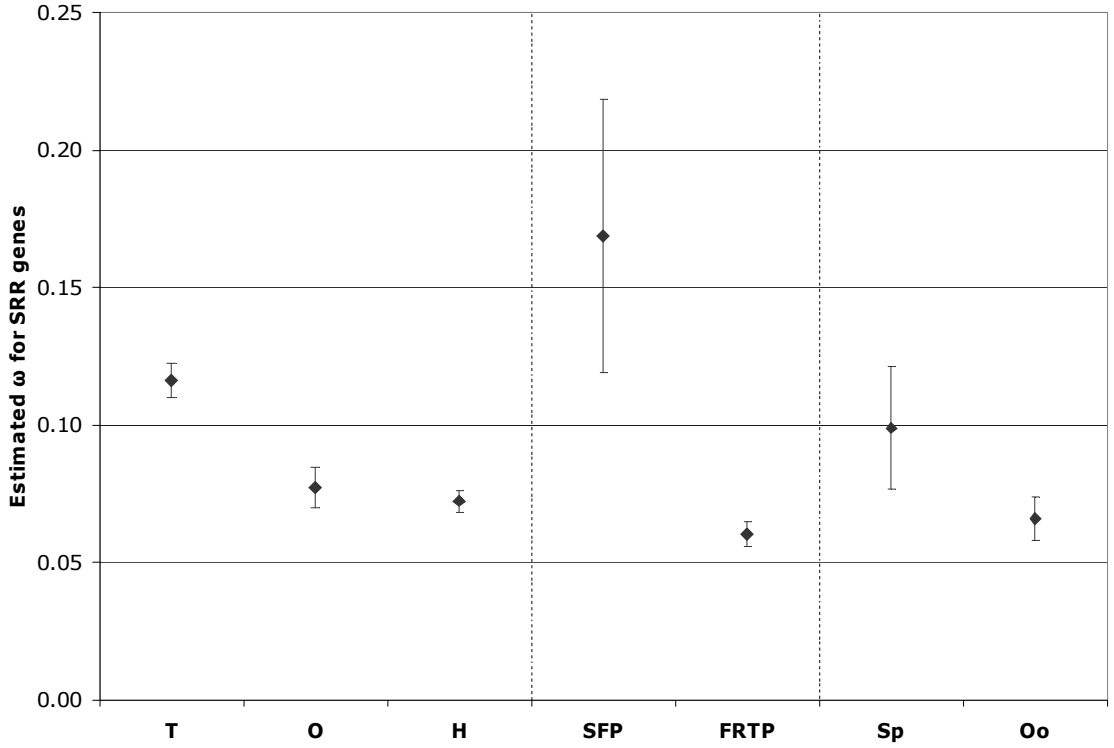
Supplemental Figure 5. **Proportion of TEs belonging to LTR, LINE-like TIR, and other element classes in each species.**



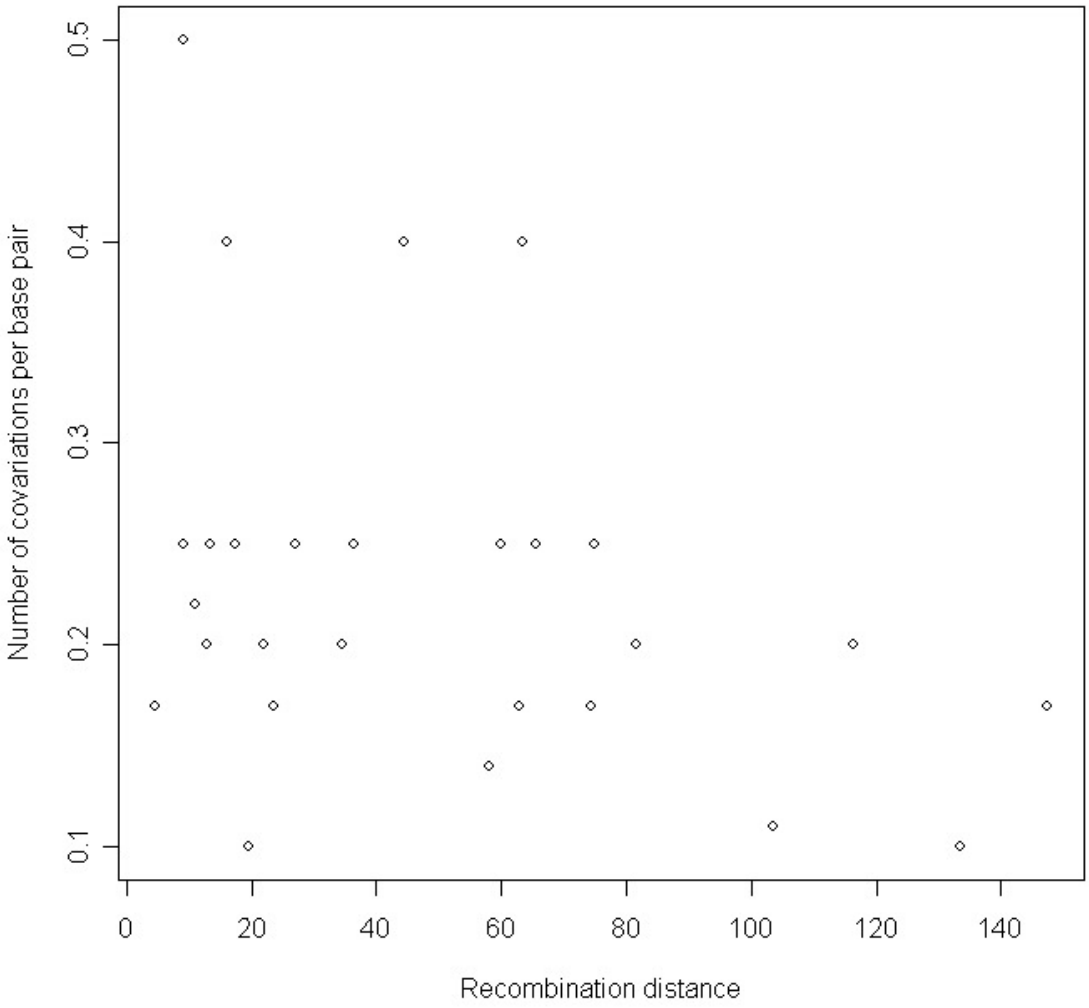
Supplemental Figure 6. **Distributions of ω and probabilities of positive selection inside and outside InterPro domains. A)** Boxplot of per-codon ω estimated for codons inside and outside InterPro domains. **B)** Fraction of codons with evidence for positive selection at three probability cut-offs (x-axis), inside and outside InterPro domains.



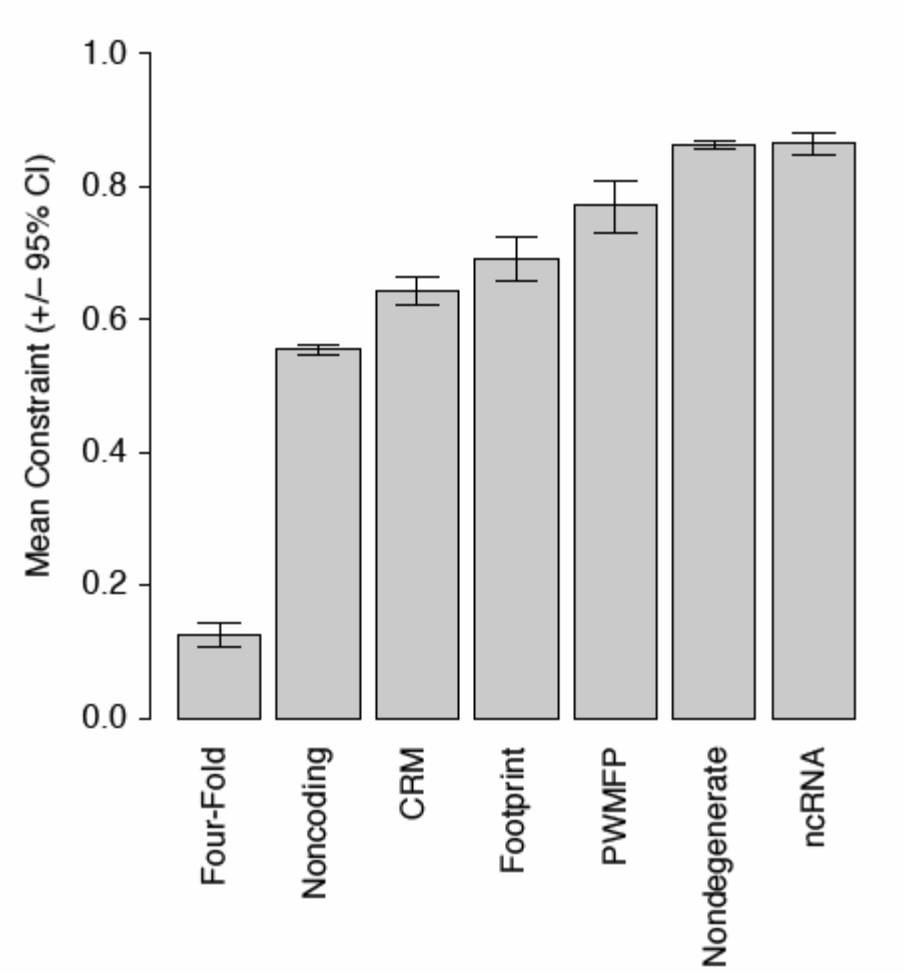
Supplemental Figure 7. **ω and amino acid divergence for different categories of genes based on gene movement.** (**A**) The distribution of ω (for the branch with the inferred gene movement) for species within the *melanogaster* group, and (**B**) the relative amino acid divergence (relative to the mean amino acid divergence across all genes for a given lineage) for all species for the following classes of genes: genes that maintain their genomic location (stayed; ω and relative divergence are averaged across the 6 and 12 species phylogeny, respectively); genes that move within a Muller element (within); genes that move between Muller elements (between); the subset of 'between' genes that moved between autosomes (A-A); the subset of 'between' genes that moved between X chromosome and autosomes (X-A); the subset of 'X-A' where the inferred movement was to the X chromosome (to X) and from the X chromosome (from X).



Supplemental Figure 8. **Patterns of constraint in sex and reproduction-related genes**. Estimated ω from genes expressed in head, ovary, testis, expressed in seminal fluid and female reproductive tract, involved in spermatogenesis and oogenesis. Error bars represent 95% confidence intervals. H: head specific (for comparison), O: ovary specific, T: testis specific, SFP: Seminal Fluid Protein, FRTP: Female Reproductive Tract Protein, Sp: spermatogenesis, Oo: oogenesis.



Supplemental Figure 9. **Number of covariations per base pair *versus* recombination distance for 27 predicted helices.** The number of covariations per base pair within a helix is the number of independently occurring covariations divided by helix length. Recombination distance is the average physical distance between covarying nucleotides scaled by 1/2 for autosomal and 2/3 for X-linked helices (because of the lack of recombination in *Drosophila* males).



Supplemental Figure 10. **Patterns of constraint on different sequence classes.** Mean constraint (+/– 95% CI) for four-fold degenerate, noncoding, CRM (*cis*-regulatory module), footprint, PWMFP (position weight matrix matches within DNase I footprints), nondegenerate and ncRNA sites. 95% confidence intervals were obtained by bootstrapping by element.

16. Tables

Supplemental Table 1. **Numbers of reads of different kinds.**

Sequence data	3-4 kb plasmid		8-10 kb plasmid		12-19 kb plasmid		37-40 kb fosmid		135-150 kb BAC		EST	finishing	assembled
	reads	pairs	reads	pairs	reads	pairs	reads	pairs	reads	pairs	reads	reads	reads
D. simulans	532	227	-	-	-	-	105	41	-	-	-	-	577
D. sechellia	1,009	486	164	80	-	-	110	53	-	-	-	-	1,177
D. yakuba	2,240	835	-	-	-	-	150	40	-	-	-	52	2,026
D. erecta	2,392	1,114	-	-	-	-	322	142	11	4	25	-	2,333
D. ananassae	2,834	1,277	-	-	-	-	511	161	9	4	25	-	2,869
D. persimilis	1,078	495	198	90	-	-	122	57	-	-	-	-	1,171
D. willistoni	1,493	716	160	78	529	255	70	33	37	17	27	-	1,984
D. virilis	2,613	1,333	-	-	-	-	676	299	20	9	25	-	2,394
D. mojavensis	2,265	1,043	-	-	-	-	427	196	25	11	21	-	2,314
D. grimshawi	2,157	1,007	-	-	-	-	405	186	21	9	25	-	2,063

(reads in thousands)

Supplemental Table 2. **Effect of reconciliation on assembly quality.**

Reconciled Assemblies	Secondary Assembler	BEFORE			AFTER				
		sum of contigs	contig N50	CE count	sum of contigs	contig N50		CE count	
D. erecta	Celera 7	145,196,048	365,805	798	145,084,019	448,166	(+23%)	645	(-19%)
D. ananassae	Celera 7	214,454,490	83,194	2,206	213,918,817	93,382	(+12%)	1,903	(-14%)
D. willistoni	ARCHANE4.5	223,295,762	144,657	1,058	224,519,948	165,230	(+14%)	893	(-16%)
D. virilis	Celera 7	189,914,823	101,385	1,566	189,205,863	118,126	(+17%)	1,094	(-30%)
D. mojavensis	Celera 7	180,519,631	100,418	1,045	180,207,831	121,352	(+21%)	841	(-20%)
D. grimshawi	Celera 7	186,365,390	78,418	1,010	186,090,669	91,175	(+16%)	936	(-7%)

Supplemental Table 3. **Assemblies of other *D. simulans* strains**.
contig data is for all contigs (not just ≥ 2 kb)

D. simulans 1X assemblies	Unplaced reads	Placed reads	Contigs >1kb	N50 length (kb)	N50 number	Supercontigs >1kb	N50 length	N50 number
c167.4	142508	224598	31805	1.5	11675	23504	2.8	5914
md106ts	51220	256980	37724	2.2	11541	18913	7.8	2877
md199s	27819	282697	35446	2.4	10376	13655	11.9	1935
nc48s	65296	246766	36503	1.9	12101	25757	3.6	5569
sim4	76043	235349	45627	1.8	15731	29839	3.7	6294
sim6	43602	290772	44461	2.4	13014	22105	8.4	3401

Supplemental Table 4. **Genbank accession numbers for genome sequences used in this paper.**

Organism	WGS	GPID	Accession Numbers of Sequences*
D. melanogaster	-	13207	AE014134, AE013599, AE014296, AE014297,AE014135, AE014298
D. simulans	'mosaic'	18237	CM000361-CM000366, CH981541-CH991539
D. sechellia	AAKO	12711	CH480815-CH482372, CH676463-CH689634
D. yakuba	AAEU	12366	CM000157-CM000162, CH891577-CH899677, CH902559-CH902573
D. erecta	AAPQ	12661	CH954177-CH959300
D. ananassae	AAPP	12651	CH902617-CH916365
D. pseudoobscura	AADE	10626	CM000070-CM000071, CH379058-CH379070, CH475397-CH476252, CH672438-CH676462
D. persimilis	AAIZ	12705	CH479179-CH480814, CH689635-CH700836
D. willistoni	AAQB	12664	CH959366-CH974203
D. virilis	AANI	12688	CH940647-CH954176
D. mojavensis	AAPU	12682	CH933806-CH940646
D. grimshawi	AAPT	12678	CH916366-CH933805

*The 'key' to the accession number prefix is that "CM" numbers are for chromosome and linkage group CON records (scaffolds built from contigs), "CH" and "DS" are for subchromosomal scaffold CON records, and AE are for complete chromosomes. Sometimes there are multiple ranges of CON records, just because of the order in which they were created. This is because for the first few projects, we only made CON records for multi-component scaffolds, before we knew that they would be annotated, so we had to go back and make CON records for the singleton contigs. All the scaffolds are available from the appropriate WGS master (the *D. simulans* records are available from the white501 strain, AAGH), and also from the Genome Project page, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=genomeprj

Supplemental Table 5. **Accession numbers for *Wolbachia* and mtDNA assemblies extracted from sequencing traces.**

Description	Genbank accession
Wolbachia endosymbiont of *D. ananassae*	NZ_AAGB00000000
Wolbachia endosymbiont of *D. simulans*	NZ_AAGC00000000
Wolbachia endosymbiont of *D. willistoni*	NZ_AAQP00000000
mtDNA assembly of *D. erecta*	BK006335
mtDNA assembly of *D. ananassae*	BK006336
mtDNA assembly of *D. persimilis*	BK006337
mtDNA assembly of *D. willistoni*	BK006338
mtDNA assembly of *D. mojavensis*	BK006339
mtDNA assembly of *D. virilis*	BK006340
mtDNA assembly of *D. grimshawi*	BK006341

Supplemental Table 6. **Expression of GLEAN-R models in six species of *Drosophila*.**

	Species					
Expression Code	*D.sim*	*D.yak*	*D.ana*	*D.pse*	*D.vir*	*D.moj*
+	10,634	12,072	11,378	11,845	11,325	11,018
0*	2,303	2,539	3,312	858	1,336	1,459
not mapped	4,112	4,205	7,861	4,625	5,018	5,261
% (of remapped) expressed	82.2%	82.6%	77.5%	93.2%	89.4%	88.3%

*Expression below background does not imply that the gene model is not expressed; it only means that we did not obtain evidence of expression in our samples. We only assayed expression on adult flies. Also, genes with few matching probes, but high expression, often do not meet the significance threshold; genes with more expression data points are of higher confidence.

Expression code key: '+' indicates expression significantly higher than background ($P \leq 0.001$), '0' indicates expression not significantly higher than background ($P > 0.001$), and 'not mapped' indicates that < 2 perfect matching array probes map to gene and so expression was not measured.

Supplemental Table 7. **Number of gene models with potential TE contamination based on two different methods**.

	# Genes	# Genes with 90% CDS masked by ReAs	% genes w/ReAS repeat	# Genes with PFAM domain	# Genes with parasitic PFAM domain	% Genes with parastic PFAM domain	Total # with evidence for TE contamination	% Gene with TE contamination
dmel	13,733	16	0.12%	10,610	34	0.32%	50	0.36%
dsim	17,049	970	5.69%	11,461	473	4.13%	1,066	6.25%
dsec	21,332	4,275	20.04%	13,938	1,384	9.93%	4,448	20.85%
dyak	18,816	2,182	11.60%	12,779	965	7.55%	2,393	12.72%
dere	16,880	1,398	8.28%	12,103	1,071	8.85%	1,556	9.22%
dana	22,551	7,104	31.50%	15,202	2,753	18.11%	7,275	32.26%
dpse	17,328	464	2.68%	12,207	641	5.25%	965	5.57%
dper	23,029	5,540	24.06%	14,963	2,140	14.30%	5,704	24.77%
dwil	20,211	3,891	19.25%	13,535	1,476	10.91%	4,395	21.75%
dvir	17,679	2,871	16.24%	12,227	973	7.96%	2,999	16.96%
dmoj	17,738	2,647	14.92%	12,427	1,331	10.71%	2,889	16.29%
dgri	16,901	1,429	8.46%	12,138	706	5.82%	1,631	9.65%

Supplemental Table 8. **Available alignment sets.**

Homology Set	Masked version available?
single copy orthologues in the *melanogaster* group	yes
single copy orthologues in all 12 species	yes
All *D. melanogaster* genes with single copy orthologues or Synpipe resolved orthologues in any species	yes
All clusters with only single copy orthologues in any species	no

All alignment sets were based on FRB+Synpipe resolved homology calls, unless otherwise noted. Versions produced with or without a guide tree, as well as versions using either the longest dmel translation or all dmel translations are available for all alignment sets.

Supplemental Table 9. **Number and percentage of gene models that fall into each homology class.**

	# Genes	# with a dmel homolog	% with dmel homolog	# present in all species	% present in all species	# single copy orthologue in all species	% single copy orthologues
dmel	13,733	13,733	100.00%	10,614	77.29%	6,698	48.77%
dsim	17,049	13,533	79.38%	11,499	67.45%	6,698	39.29%
dsec	21,332	16,338	76.59%	13,822	64.79%	6,698	31.40%
dyak	18,816	15,136	80.44%	12,615	67.04%	6,698	35.60%
dere	16,880	14,328	84.88%	11,911	70.56%	6,698	39.68%
dana	22,551	16,933	75.09%	14,959	66.33%	6,698	29.70%
dpse	17,328	13,520	78.02%	11,782	67.99%	6,698	38.65%
dper	23,029	16,050	69.69%	14,420	62.62%	6,698	29.09%
dwil	20,211	14,783	73.14%	13,180	65.21%	6,698	33.14%
dvir	17,679	13,770	77.89%	12,246	69.27%	6,698	37.89%
dmoj	17,738	14,074	79.34%	12,650	71.32%	6,698	37.76%
dgri	16,901	13,384	79.19%	12,012	71.07%	6,698	39.63%

Supplemental Table 10. **Identities of gene families with significantly elevated rates of turnover along the *D. melanogaster* lineage.**

Cluster ID*	Annotation#	FBgns	*P*-value&	Mel. species group ancestral copy number&	D. mel copy number	Change
223	FLYWCH zinc finger domain	NA	0	4	0	-4
2548	*longitudinals lacking*	FBgn0005630	0	5	1	-4
322	FLYWCH zinc finger domain	NA	0.000001	3	0	-3
3206	*pipe*	FBgn0003089	0.000003	4	1	-3
711		NA	0.000068	2	0	-2
2743		NA	0.000068	2	0	-2
2939		NA	0.000068	2	0	-2
2956	Zinc finger, C2H2 type	NA	0.000068	2	0	-2
3344	FLYWCH zinc finger domain	NA	0.000068	2	0	-2
5072		NA	0.000068	2	0	-2
6325	FLYWCH zinc finger domain	NA	0.000068	2	0	-2
6367	Kunitz/Bovine trypsin inhibitor	NA	0.000068	2	0	-2
8248	S-adenosylmethionine synthetase	NA	0.000068	2	0	-2
12554	EF hand	NA	0.000068	2	0	-2
879	*broad*	FBgn0000210	0.000198	3	1	-2
1049	*Ecdysone-induced protein 75B*	FBgn0000568	0.000198	3	1	-2
1784	*sallimus*	FBgn0003432	0.000198	3	1	-2
3726	*Stretchin-Mlck*	FBgn0013988	0.000198	3	1	-2
250	α–εTrypsin family	FBgn0003863 FBgn0010357 FBgn0010358 FBgn0010359 FBgn0010425 FBgn0050025 FBgn0050031	0.002064	5	7	+2
1703	Jonah family	FBgn0001285 FBgn0003356 FBgn0003357 FBgn0020906 FBgn0031653 FBgn0035886 FBgn0035887 FBgn0039777 FBgn0039778	0.003991	7	9	+2
6175	Sdic family	FBgn0003654 FBgn0052823 FBgn0053497 FBgn0053499 FBgn0067861	0.000003	2	5	+3
2187	Stellate family	FBgn0031809 FBgn0044817 FBgn0053236 FBgn0053237 FBgn0053238 FBgn0053239 FBgn0053240 FBgn0053241 FBgn0053242 FBgn0053243 FBgn0053244 FBgn0053245 FBgn0053246 FBgn0053247	0	4	14	+10

*Cluster ids are identifiers from the FRB+Synpipe homology assignments
#Annotation is based on PFAM domain assignment (for genes with no paralogue in *D. melanogaster*), or the identities of the *D. melanogaster* genes in the cluster.
& P-value is for test of accelerated rate of gene turnover along the *melanogaster* branch from CAFE[46]; ancestral copy number for the *melanogaster* species group is the maximum likelihood estimate from the same program.

Supplemental Table 11: **List of 44 lineage-specific genes arising in the *melanogaster* group or some subset of the *melanogaster* group phylogeny.** Columns correspond to: gene ID (FBgn); gene name; CDS length; number of introns; chromosome position in melanogaster; whether the gene occurs in the intron of another gene (yes or no) and the orientation of the novel gene with respect to the gene it occurs in; the specific tissue the gene is expressed in (testes/accessory gland/head/tubule/not specific/no data); lineage the gene arose in. One of the novel genes overlaps with another gene, and its orientation is opposite to the gene it overlaps with. (available as a separate dataset at www.nature.com/nature)

Supplemental Table 12. **Median value of ω, the negative log of the P-value from the test of positive selection and d_N for each of the 115 GO categories.** Right and left-tail P-values and corresponding false discovery rates (FDR) are from permutation tests. Numbers of genes (out of the set of genes with a single orthologue in the *melanogaster* group) annotated to each term are indicated. The unknown class includes any genes lacking a GO annotation or any genes directly annotated to Biological process (BIO; GO:0008150), Cellular Component (CELL; GO:0005575) or Molecular Function (MOL; GO:0003674). (available as a separate dataset at www.nature.com/nature)

Supplemental Table 13. **Number of times P450 and GST genes duplicated in the *Drosophila* radiation.**

	Gene name	substrate	duplications in *Drosophila*
P450s	Cyp18a1[68]		0
	Cyp314a1[69]	ecdysone precursor	0
	Cyp4g1[70]	omega-hydroxylase	0
	Cyp303a1[71]		0
	Cyp302a1[72]	ecdysone precursor	0
	Cyp315a1[72]	ecdysone precursor	0
	Cyp306a1[72]	ecdysone precursor	0
	Cyp307a2[73]	ecdysone precursor	1
	Cyp6a8[74]	lauric acid, DDT	2
	Cyp12d1[75]	DDT	2
	Cyp6g1[76]	DDT, lufenuron, nitenpyrum	4
	Cyp6a2[77]	DDT	6
	Cyp12a4[78]	lufenuron	7
GSTs	CG9363[79]	putative Maleyl acetoacetate	0
	GstS1[80, 81]	4HNE	0
	CG6781[82]	6-PTP	0
	CG10065[83]		0
	Gst D1[84, 85]	DDT, 4HNE, cumene hydroperoxide, H_2O_2	1

Shading indicates genes that have been associated with the detoxification of insecticides in the literature. Rows that are not shaded are for genes in P450/GST families that have an association with non-detoxification function in the literature.

Supplemental Table 14. **Mutations in the mature miRNA component of 60 conserved pre-miRNA sequences with reliable alignments were inferred using Fitch parsimony.**

miRNA	Type	Mutation	Clade	miRNA pos.
mir-9b	5′	G → C	*D. grimshawi*	19
mir-100	5′	A → U	*obscura* group	10
mir-282	5′	A → G	*D. ananassae*	1
mir-305	5′	ΔG	*D. simulans*	15
mir-316	5′	G → A	*D. mojavensis*	20
mir-274	5′	G → A	*mel./obscura* groups	25
mir-274	5′	A → U	*D. ananassae*	25
mir-274	5′	U·C ↔ G:C	*Sophophora*	1
mir-287	3′	C → U	*D. yakuba*	21
mir-277	3′	G:C → G·U	*D. simulans*	9
mir-277	3′	G:C → G·U	*D. simulans*	13
mir-317	3′	C:G → U:A	*obscura* group	23
mir-317	3′	R·U → G·A	*D. mojavensis*	24

Ambiguous ancestral states are indicated by IUPAC codes, unclear polarities indicated with a bidirectional arrow, and the single deletion denoted by Δ. Base pairs are written with the 5′-base on the left. Mutations are inferred to occur in the ancestor of the species in the clade column. miRNA pos. is the position where the mutation occurs in the miRNA counting in the alignment from the *D. melanogaster* 5' end.

Supplemental Table 15. **The ratio L/S of substitution rates in loop (L) and stem (S) regions of the predicted ncRNAs was estimated with the EM procedure implemented in XRate (Klosterman *et al*. 2006).**

RFAM Accession	Family name	Description	# seq	Stem nts.	Loop nts.	L/S
RF00009	RNaseP_nuc	Nuclear RNase P	2	88	322	2.57
RF00017	SRP_euk_arch	Eukaryotic type signal recognition particle RNA	2	172	150	1.28
RF00004	U2	U2 spliceosomal RNA	87	90	122	1.18
RF00002	5_8S_rRNA	5.8S ribosomal RNA	638	52	402	1.18
RF00031	SECIS	Selenocysteine insertion sequence	14	44	29	1.17
RF00028	Intron_gpI	Group I catalytic intron	6	106	657	1.03
RF00485	K_chan_RES	Potassium channel RNA editing signal	67	48	75	0.91
RF00020	U5	U5 spliceosomal RNA	83	60	101	0.88
RF00003	U1	U1 spliceosomal RNA	95	80	145	0.73
RF00015	U4	U4 spliceosomal RNA	36	62	109	0.64
RF00001	5S_rRNA	5S ribosomal RNA	601	68	110	0.56

Rate estimations were made for the 11 families for which there was enough data such that the maximum expected error in estimating any rate is less than 15%.

Supplemental Table 16. **Estimated constraint for different sequence classes on each chromosome.**

Site Type	Mean Constraint [95 % CI]		
	Chr 2 and 3	Chr X	Chr 4
ncRNA	0.862 [0.858 - 0.867]	0.766 [0.683 - 0.837]	-
Nondegenerate	0.862 [0.858 - 0.867]	0.850 [0.835 - 0.865]	0.848 [0.771 - 0.899]
PWMFP[1]	0.794 [0.752 - 0.832]	0.549 [0.338 - 0.729]	-
Footprints	0.706 [0.670 – 0.739]	0.540 [0.397 - 0.667]	-
CRM[2]	0.656 [0.634 - 0.681]	0.584 [0.532 - 0.633]	0.477
Noncoding	0.561 [0.554 - 0.568]	0.506 [0.484 - 0.526]	0.0958 [0.0104 - 0.185]
Four-fold	0.127 [0.116 - 0.138]	0.178 [0.143 - 0.213]	0.274 [0.118 - 0.399]

1 Position weight matrix matches within footprints. 2 *cis*-Regulatory module. 95% confidence intervals were obtained by bootstrapping by element. Mean constraint could not be estimated on chromosome 4 for ncRNA, PWMFP and FP sites due to a lack of any data and confidence limits could not be obtained for CRMs by bootstrapping due to limited data (n = 4).

Supplemental Table 17. **Statistics of the new *D. pseudoobscura* assembly.**

Draft *D. pseudoobscura* assembly (ARACNHE)	Count	Total length	N50
Contigs	5,986	168,284,336	96,690
Supercontigs (gapped)	2,663	170,161,593	2,099,360
Supercontigs (ungapped)	2,663	168,284,336	1,953,110

Supplemental Table 18. **Total number of GLEAN models for each species for each of three different GLEAN sets, as described in supplemental materials.**

Species	GLEAN-SH	GLEAN-FPH	GLEAN-R
D. simulans	13,942	18,892	18,273
D. sechellia	13,917	21,913	21,332
D. yakuba	14,110	20,075	19,430
D. erecta	13,429	17,281	16,881
D. ananassae	13,016	22,485	22,551
D. pseudoobscura	13,024	17,660	17,328
D. persimilis	12,972	23,629	23,029
D. willistoni	12,567	19,847	20,257
D. virilis	12,350	18,230	17,684
D. mojavensis	12,253	18,278	17,739
D. grimshawi	12,714	17,385	16,901

Supplemental Table 19. **Counts of potentially problematic and high-confidence genes for each chromosome arm in *D. melanogaster*.**

Chromosome Arm	High Confidence Protein-Coding Genes	Flagged Genes
2L	2,253	337
2R	2,397	323
3L	2,301	382
3R	2,982	420
4	62	26
X	1,649	600

"High Confidence" genes correspond to those protein-coding genes not flagged by Stark et al.[25], with homology calls and no ambiguities from the GeneWise pipeline. "Flagged" genes are those designated as 'problematic' by Stark *et al*.[25] as well as genes with ambiguities in the homology calls from the GeneWise pipeline.

Supplemental Table 20. **Criteria used to determine if a given alignment region should be masked.**

Alignment Pair	Masking Criteria
dmel-dsim	>11 nucleotide differences in a 30 bp window
dmel-dsec	>11 nucleotide differences in a 30 bp window
dmel-dyak	>15 nucleotide differences in a 30 bp window
dmel-dere	>15 nucleotide differences in a 30 bp window
dmel-dana	>18 nucleotide differences in a 30 bp window
dmel-dpse	>16 amino acid differences in a 20 aa window
dmel-dper	>16 amino acid differences in a 20 aa window
dmel-dwil	>16 amino acid differences in a 20 aa window
dmel-dmoj	>17 amino acid differences in a 20 aa window
dmel-dvir	>17 amino acid differences in a 20 aa window
dmel-dgri	>17 amino acid differences in a 20 aa window