Supplemental information

Detailed Methods

Transcript prediction pipeline

The principal problem in prediction by homology is the selection of a template sequence. Alternative transcripts and paralogous genes in both *D. melanogaster* and or in the target species result in cases in which a single *D. melanogaster* transcript can be aligned to several candidate regions in the target species and, vice versa, a candidate region might be matched to several transcripts. The pipeline solves these problems by (1) selecting a representative transcript from each *D. melanogaster* gene and (2) pre-selecting likely pairings of orthologous transcripts and candidate regions and processing these first. Paralogous transcripts and variant transcripts are then predicted in subsequent steps.

The pipeline runs on GNU/Linux (http://www.gnu.org, http://www.kernel.org) and is controlled by a set of python (http://www.python.org) scripts. Results are stored in a relational database (http://www.posgresql.org).

The transcript prediction pipeline falls into four steps corresponding to the section subheadings below: (1) Pre-processing of template sequences, (2) identification of transcript containing loci, (3) prediction of transcripts in such loci, and (4) post-processing and classification of transcript predictions.

Step 1: Pre-processing

We cluster into genes those *D. melanogaster* transcripts that share at least one exon with identical exon boundaries. For each gene, the longest unspliced transcript is chosen as its representative; all others are designated as variants. The set of *D. melanogaster* representative transcripts is required for steps 2 and 3.

A set of 163 sequences templates were removed from the input list as they appear to be associated with transposable elements (repeatmasker and manual annotation). The genes involved were: CG10598, CG10598, CG11231, CG11458, CG12454, CG12462, CG12683, CG12749, CG12852, , CG12912, CG1294, CG13033, CG13235, CG13291, CG13591, CG13673, CG13785, CG14148, CG14191, CG14410, CG14460, CG14494, CG14503, CG14519, CG14560, CG14578, CG14975, CG15036, CG15166, CG15381, CG15494, CG15690, CG16901, CG16937, CG17506, CG17618, CG17702, CG17742, CG1840, CG1903, CG2042, CG2175, CG2175, CG30126, CG31483, CG32015. CG32231, CG32606, CG32616, CG32710, CG32711, CG32781, CG32788, CG32821, CG33236, CG33237, CG33238, CG33239, CG33240, CG33241, CG33242, CG33243, CG33244, CG33245, CG33246, CG33247, CG33267, CG33269, CG33270, CG33271, CG33272, CG33500, CG3894, CG40032, CG40043, CG40096, CG40103, CG40119, CG40122, CG40128, CG40138, CG40154, CG40156, CG40166, CG40167, CG40177, CG40182, CG40192, CG40221, CG40224, CG40241, CG40251, CG40264, CG40267, CG40315, CG40317, CG40322, CG40341, CG40342, CG40343, CG40346, CG40367, CG40373, CG40376, CG4038, CG40383, CG40396, CG40400, CG40409, CG40410, CG40442, CG40473, CG40484, CG40496, CG40497, CG41040, CG41065, CG41098, CG41104, CG41114, CG41118, CG41119, CG41122, CG41124, CG41132, CG41135, CG41141, CG4268, CG4345, CG5079, CG5172, CG5386, CG5812, CG6900, CG7105, CG7552, CG9106, CG9106, CG9983, CG40090, CG11260, CG9843, CG40265, CG40236, CG41059, CG40295,

Step 2: Genome scan

The second step produces candidate transcript-containing regions from within the eleven newly sequenced genome assemblies. The amino acid sequences encoded in the *D. melanogaster* representative

transcripts are masked for low complexity regions using SEG (Wootton 1994, default options). These masked peptide sequences are then aligned against the unmasked genomes using Exonerate in heuristic mode with sensitive search options (options: "model=p2g, forcegtag=TRUE, bestn=200, maxintron=50000, proteinwordthreshold=3, proteinhspdropoff=5, proteinwordlen=5, score cutoff=50"). These settings produce a multitude of matches between template amino acid sequences and genomic regions. The level of sensitivity for identifying matches was found to be comparable to that of TBLASN (see section on Benchmarking below). Adjacent matches within 50kB associated with the same template are merged and their scores added. Matches of aggregate scores of less than 80 are discarded as representing likely spurious alignments. The result of this step is a list of pairs of transcripts and matching genomic regions.

Step 3: Transcript prediction

The third step produces transcript predictions by aligning the template amino acid sequence to its paired genomic region using the Genewise mode of Exonerate. When the genomic region to be searched is large, this is computationally expensive. At the same time, paralogous *D. melanogaster* sequences in the input set and in the target genome will give rise to a multitude of redundant predictions. Consequently, the pipeline tries to minimize the numbers of alignments that are performed by first processing pairs between orthologous pairs. Only then it will process the remaining paralogous pairs to properly account for lineage specific duplications. Finally, variant transcripts as opposed to representative transcripts are aligned to genomic segments, where its representative transcript matched previously.

Pairs of matching template transcripts and genomic regions are predicted in three passes. Each pass selects a subset of pairs of transcripts and genomic regions to predict with exonerate (see next section).

The first pass processes likely ortholog matches by discarding all low scoring matches. For a given template these are defined as having an alignment coverage, score or percent identity to the query that is less than half of the best coverage, score or percent identity obtained for any predictions from this particular *D. melanogaster* sequence.

The second pass excludes all pairs of of templates and genomic regions that overlap with the genomic locations of transcripts predicted in the first pass. The purpose of this step is to predict paralogous sequences.

The third pass predicts transcripts which are variants to those predicted in passes one and two. For each prediction in pass one and two, variant as opposed to representative template transcripts (see step 1) are aligned to the same genomic region as the representative template transcript. This pass does not make use of priority lists.

Prediction of transcripts from a list of pairs of aligned templates with genomic regions

The resulting collection of pairs between templates and various genomic regions is transformed into a collection of priority lists. The priority lists are used to solve the problem of paralogous *D. melanogaster* queries that will align to the same region containing a homologous gene in the target genome. The priority list ensures, that the likely ortholog sequence is selected for transcript prediction before attempting prediction with paralogous sequences.

Priority lists are built from matches encompassed within the same genomic region. All templates in each priority list are first grouped into clusters on the basis of sequence identity (at least 60%) and alignment coverage (at least 50%). Each cluster thus groups homologous template transcripts that align to a particular genomic region. Each cluster is sorted by decreasing alignment score between the template and genomic prediction. The templates are thus arranged in a priority list, such that more similar templates are used before less similar ones to prediction transcript. As soon as a satisfactory prediction is returned, less similar template sequences are skipped. At most, five templates are in each priority list.

Templates are aligned in order of their priority. A prediction is defined as being satisfactory, if its exon structure corresponds to the exon structure of the template. To this end, exon boundaries of the prediction are mapped onto the template sequence. The exon structure of multi-exon genes is considered to be conserved if at most one exon is missing and/or at most one exon boundary has shifted. Single exon predictions are considered to be conserved if the template is a single exon protein and the prediction aligns to at least 80% of this template.

In order to reduce computation time, the size of the genomic region presented to Exonerate is determined first by another search with Exonerate in heuristic mode, but using more sensitive settings than used in the initial genomic scan (options: proteinwordthreshold=5, proteinhspdropoff=5, proteinwordlen=3). The region is extended with 50kb at either end, unless the template's ends are aligned, in which case only 0.3kb are added.

Transcripts are predicted from template peptide sequences in the final alignment using Exonerate's genewise mode (options: "exhaustive, subopt=FALSE, forcegtag=TRUE, subopt=FALSE"). Even though Exonerate permits the detection of suboptimal matches, this capability was switched off as it led to excessive running time. Instead, if a match is found, adjacent regions of 100kB on either side are checked for local duplications. The local search for paralogs is repeated, until no more matches are found.

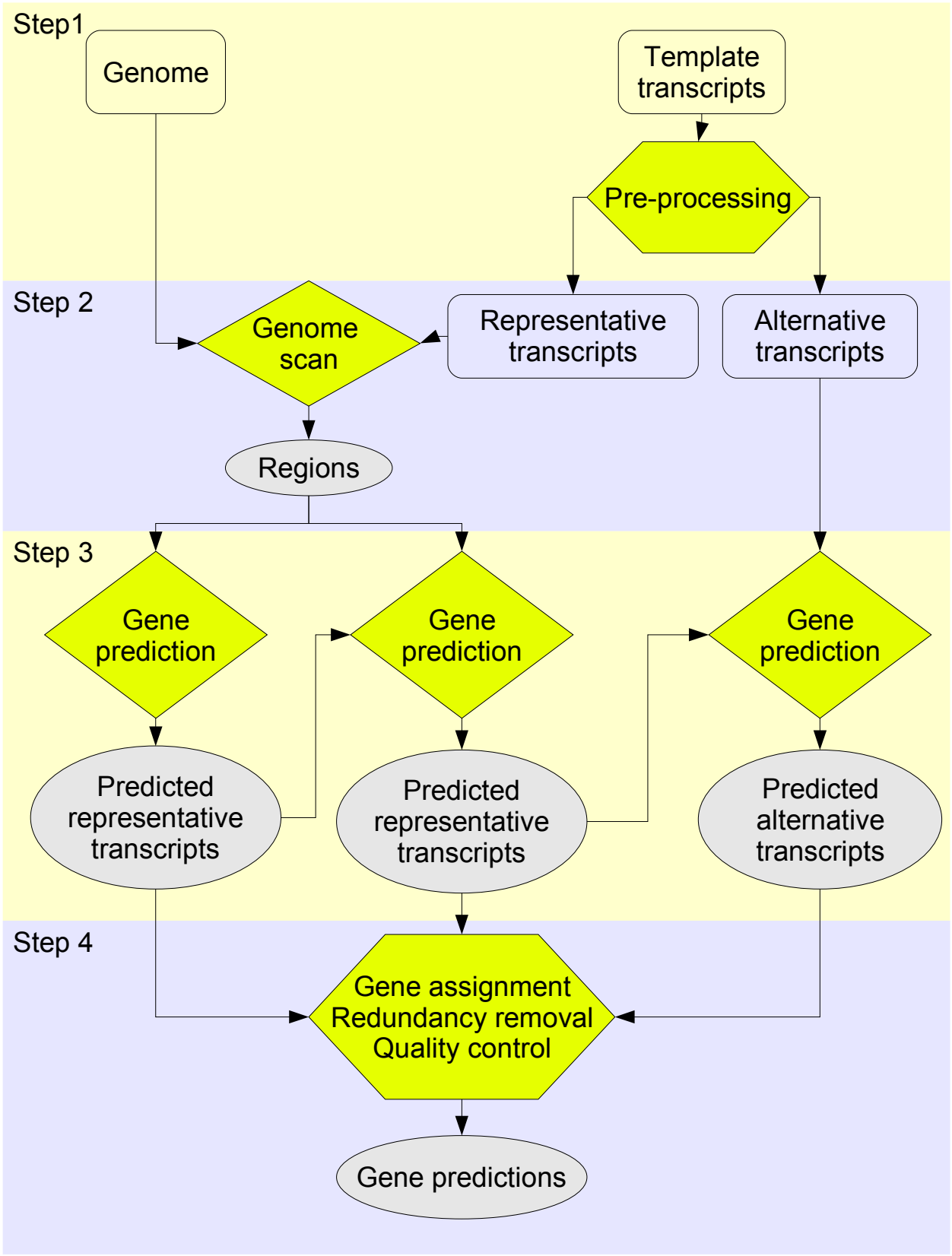

Figure S1: The four steps of the transcript prediction pipeline. See text for further details.

Post-processing

The pipeline attempts to predict transcripts in a genomic region until it identifies a satisfactory prediction (as defined using a set of heuristics, see below). In multi-gene families several equally suitable templates result in identical or highly similar predictions. The resultant set of predictions thus needed to be filtered to reduce redundancy. Similarly, inappropriate templates might have been used for prediction. Such spurious predictions have to be removed since they are most often invalidated by a better matching, orthologous prediction. To this end, an index is assigned to each prediction that is intended to reflect whether it is orthologous, for all exons, to the template.

This quality index aggregates the following attributes: alignment coverage of the template, presence of frameshifts and/or stop-codons, and conservation of exon structure when compared to its template. Indices are ranked to present an approximate indicator of the prediction quality. Briefly, full length predictions are preferred over incomplete predictions (namely, less than 80% alignment coverage of the template); predictions without in-frame stop-codons and frameshifts are preferred over those containing frame disruptions; and finally, predictions with matching exon boundaries between template and prediction are preferred over those with shifted exon boundaries or those with inserted or deleted introns.

Redundant and spurious predictions are removed by applying a set of heuristic rules. Briefly, predictions are sorted by quality index, length and alignment score. Once a prediction is accepted, other overlapping predictions satisfying one or more of the following rules are removed: (1) the predicted peptide sequence is identical to the template sequence or a part thereof; (2) the prediction is of a lower quality index and the sequences are essentially identical (better than 98% sequence identity, fewer than 20 gap positions); (3) the prediction is a fragment and its sequence is more than 80% identical to the accepted prediction; and (4) the prediction has a lower quality index and spans the better ranked prediction on either strand. These rules permit alternative transcripts per locus but disallow truncated transcripts. An additional filter removes two other types of predictions, those sharing exons between adjacent duplications, and those predictions without conserved gene structure that span other predictions with conserved gene structure, but do not contain overlapping exons.

These rules remove dubious predictions that are invalidated by a better prediction. Nevertheless, those predictions of low confidence but showing no clear favorite are all retained, which reflects the difficulty of predicting that particular transcript.

A particular problem arises from genes that are duplicated locally. These can result in spurious transcripts containing exons from adjacent genes. In most cases, exon boundaries between duplicates are conserved and thus all transcripts will possess the same quality index. We found no heuristic to reliably winnow out these spurious transcripts while keeping true alternative transcripts. Instead, we defer identification of these transcripts until after the orthology assignment.

Transcript predictions are combined into genes if they share at least one identical exon. Partial overlap is allowed for single exon genes and terminal exons, although at least 80% nucleotide overlap is required. The quality index of a gene is assigned as the best quality index of its transcripts.

Quality indices for transcripts

Quality index	fragment	pseudogene	conserved	partially conserved	Prediction multi-exon	Query multi-exon
Genes						
CG: conserved gene	N	N	Y	Y	Y	Y
SG: single exon gene	N	N	?	?	N	N
PG: partially conserved gene	N	N	N	Y	Y	Y
RG: retrotransposed gene	N	N	?	?	N	Y
UG: not conserved gene	N	N	?	?	?	?
Pseudogenes						
CP: duplicated pseudogene	N	Y	Y	?	Y	?
SP: single exon pseudogene	N	Y	?	?	N	N
PP: partially conserved pseudo	N	Y	N	Y	Y	?
RP: processed pseudogene	N	Y	?	?	N	Y
UP: not conserved pseudogene	N	Y	?	?	?	?
Fragments						
SF: single exon fragment	Y	N	?	?	N	N
CF: conserved fragment	Y	N	Y	?	?	?
PF: partially conserved fragment	Y	N	N	Y	?	?
UF: not conserved fragment	Y	N	?	?	?	?
BF: pseudogenic fragment	Y	Y	?	?	?	?

Table S1: Classes of predictions. The quality index of a gene is determined by a set of flags. The flags are set to true or false according to the following criteria:

- *Fragment: True, if alignment between template and prediction covers less than 80% of the template.*
- *Pseudogene: True, if the prediction contains frame-shifts and disruptions. Frame-shifts and disruptions are allowed in dubious exons, which exceptionally low sequence similarity to the template when compared to other exons in the transcript.*
- *Conserved: True, if all exon boundaries are conserved*
- *Partially conserved: True, if at least one conserved exon is present.*
- *Prediction/Query multi-exon: the prediction/query has multiple exons.*

The priority of quality indices is: CG > PG > SG > RG > CP > PP > SP > RP > CF > PF > SF > UG > UP > UF > BF

Orthology assignment

Orthology assignment was performed in two stages. First, pairwise orthology was computed in pairwise species comparisons using PhyOP (Goodstadt and Ponting 2006). Then, multiple orthology assignments involving more than two species were inferred from clusters derived from the graph of pairwise orthology relationships.

In our terminology, we follow the definitions of (Fitch 1970, Remm et al. 2001): orthologs are related through speciation, while paralogs are related through an intra-genome duplication event. In-paralogs duplicated after the speciation of the two species being compared, whereas out-paralogs duplicated before speciation. Strict orthologous pairs are those with only a single copy in each genome, while degenerate orthologous pairs contain in-paralogs for at least one species.

Pairwise orthology assignment

For each pair of genomes, amino acid sequences of full length predictions (>80% coverage of template) are collated. Each amino acid sequence is aligned against every other sequence using BLASTP (Altschul et al. 1997). Alignments with an E-value of more than 10^{-5} or covering less than 75% of the smaller sequence are removed. The remaining alignments are weighted according to a normalized bit score

$s_{ij} = 1 - \frac{max(s'_{ij}, s'_{ji})}{min(s'_{ii}, s'_{jj})}$, where s'_{ij} is the bit score for a BLASTP alignment between sequence *i* and sequence *j*.

Briefly, the PhyOP pipeline then constructs an initial transcript phylogeny based on the UPGMA method and then optimizes branches in the tree using KITSCH (30 iterations, down weighting large distances by a power of 3) modified from the PHYLIP package (Felsenstein 1989). Orthologs are assigned to members of the smallest subtree containing leaves of both species. Nodes closer to the root than a node joining branches of orthologs are considered to be orphans. Depending on the tree topology, several predictions in one species (in-paralogs) can be assigned to the same ortholog in another species.

Distances based on amino acid substitutions can be quickly derived from BLASTP alignments, but in some circumstances they can be an unreliable guide for orthology assignment (Goodstadt and Ponting 2006). We thus verified pairwise orthology relationships using the numbers of synonymous substitutions per synonymous site d_S calculated with CodeML from the PAML package (Yang 1997) based on the pairwise BLASTP alignments calculated previously. Only orthologous and in-paralogous pairs were considered, thereby reducing computational time. The resultant set of alignments was submitted to another iteration of PhyOP orthology assignment.

Orthologous genes were defined based on orthologous transcripts. The d_S value between genes was defined as the minimum d_S value estimated between transcripts.

Multiple orthology assignment

In theory, orthologs across a multitude of species should be easily grouped through pairwise orthology relationships. In practice, however, inaccurate estimation of the phylogeny causes errors in pairwise orthology assignment, which leads to the grouping of out-paralogs. Here we use a graph clustering approach to define orthologs across multiple species (see also Tatusov et al. 1997 and Alexeyenko et al. 2006).

The set of pairwise orthology relationships defines an undirected graph with transcripts as vertices and orthology and in-paralogy relationships as edges. We assume that spurious orthology assignments will be identifiable as inconsistencies in the graph.

Both inconsistent edges and inconsistent vertices are removed. Let N_a denote the set of direct neighbors

of vertex *a*. Inconsistent edges link vertices *a* and *b*, where the proportion of shared neighbors between vertices *a* and *b* $\frac{|N_a \cap N_b|}{|N_a \cup N_b|}$ is less than 0.3. Inconsistent vertices are those with a clustering coefficient C_a, $C_a = \frac{|E_a|}{|N_a| * (|N_a| - 1)}$ of less than 0.5, where the set E_a contains all edges between vertices in the neighborhood of *a*. These thresholds were chosen heuristically following consideration of a few benchmark cases. Orthologous groups are given by the connected components of the filtered graph.

The filtering procedure introduces a potential bias towards removing ancient duplication events. Ancient duplication events result in two clusters of densely connected components with few edges in between the clusters. The between-cluster edges might be wrongly identified as inconsistent. This occurs, given the chosen thresholds, if there are more than four in-groups than out-groups for the ancient duplication event. Given the phylogeny of the Drosophila species, this should only occur on the two branches leading to the speciation points of *D. willistoni* and *D. pseudoobscura/D. pseudoobscura* and will not affect our analysis of paralogs in the *D. melanogaster* group.

Multiple alignment

Evolutionary rate estimates were inferred from multiple alignments of orthologous and paralogous sequences within an orthologous cluster. Multiple alignments were built for all coding sequences in each orthologous group using Dialign (Morgenstern 1999), for clusters containing fewer than 50 sequences, and Muscle (Edgar 2004) for clusters containing greater than 50 sequences. For Dialign, the nucleotide sequences were internally translated into peptide sequences. All transcripts per gene were provided for the alignment, which resulted in occasional spurious alignment of incompatible exons from alternative transcripts of the same gene. In such cases multiple alignment columns joining incompatible exons were split into compatible groups.

For codon based rate analysis, in-frame stop codons and frameshifts in transcripts from the newly sequence genomes were removed from the multiple alignment. Codons were assigned across the multiple alignment using *D. melanogaster* sequences (assumed to be pseudogene-free) as internal references.

Poorly aligned regions were pruned with Gblocks (Castresana 2000) using default parameters except that alignment columns with up to 50 percent gap characters were tolerated. This step retained the majority of codons (75% of all multiple alignments retained at least 90% of codons), while it removed more than half the codons in only 3% of multiple alignments. Pairwise d_N and d_S values of multiply aligned sequences were estimated using CodeML from the PAML package (Yang 1997).

Assignment of orthologous transcripts

Orthologous groups might contain several transcripts per gene and duplicated genes. For orthology based analyses, it is desirable to work with a set of ortholog transcripts which discount duplications and alternative transcripts. To this end, a reference set of orthologous transcripts was derived from the multiple alignment of each orthologous group. This reference multiple alignment contains at most one entry from each species.

The reference multiple alignment of orthologous transcripts was built using an approach similar to that using reciprocal best BLASTP hits (Tatusov et al. 1997), except that transcripts within a multiple alignment are clustered by components in a graph of reciprocal minimum d_S values. We use the reciprocity of minimum dS values as a criterion for orthology inference, assuming that problems due to misalignment and paralogy will increase d_S values and thus will have the least effect on our approach. If several transcripts from the same species and gene are in the same cluster, the transcript with the highest quality index is retained and all others removed.

Rate computation among orthologs

Evolutionary rates were computed for orthologous sequences based on multiple alignments of orthologous transcripts. All analyses are based on 6,375 multiple alignments of orthologous transcripts (see above) where each of the 12 species is represented. Synonymous and non-synonymous substitution rates were estimated using CodonML from the PAML package (Yang and Nielsen 2002). In all measurements, codon frequencies were estimated from nucleotide frequencies at each codon position (model F3x4). No correlation among sites was assumed and the transition/transversion ratio was allowed to vary.

Reconciliation of gene trees with species tree

We briefly checked the congruence between the topology of gene trees and the species tree. Pairwise synonymous and non-synonymous substitution rates were estimated using CodonML and phylogenetic trees were built using FITCH (Felsenstein 1989) and the number of gene trees supporting a particular branch in the species phylogeny were recorded. In later analyses, we set the topology of gene trees to the species phylogeny, regarded it as constant and only estimated branch lengths.

Branch specific d_N/d_S values

Species tree branch specific d_N/d_S values were calculated using the CodonML method from the PAML package (Yang and Nielsen 2002). 200 randomly chosen multiple alignments of orthologous transcripts were concatenated and submitted for evolutionary rate estimation. Further parameters were: (1) no rate heterogeneity over sites; and (2) codon frequencies given by nucleotide frequencies at each codon position. The log likelihood value obtained was compared to the log likelihood estimated assuming the same d_N/d_S ratio for all branches. The experiment was replicated 20 times. The assumption of homogeneous d_N/d_S across all branches was rejected in a likelihood ratio test with 14 degrees of freedom in all instances (Felsenstein 1981). The species *D. willistoni, D. persimilis* and *D. sechellia* were excluded from the analysis, the former because of its exceptional G+C content compared to the other species (A. Heger & C. P. Ponting, submitted) and the latter two because of their genes' high similarities to those from *D. pseudoobscura* and *D. simulans*, respectively.

Identification of sites under positive selection

Positively selected sites were predicted using the program SLR (Massingham and Goldman 2005). The analysis was restricted to the five species *D. melanogaster, D. simulans, D. yakuba, D. erecta* and *D. ananassae,* whose divergences are substantially less than those prone to substitution saturation effects.

Multiple alignments were rigorously pruned of regions that are, or could be, badly aligned. In compositionally biased regions, the signal for aligning such positions on the translated sequences is effectively lost, while it might still be retained in the untranslated codons. As a result, non-homologous characters might be aligned which affect synonymous and non-synonymous substitution rate estimates. We removed compositionally biased regions rigorously by first masking each individual amino acid sequence with seg (Wootton 1994, default parameters). Masked residues in a single sequence were then propagated to all residues in an alignment column.

Sites under positive selection are commonly defined by a significant excess of non-synonymous substitutions over synonymous substitutions giving a d_N/d_S ratio, ω, of larger than 1. Neutral sites have an ω of about 1, while sites under purifying selection have an ω of less than 1. Parsimony (Suzuki and Gojobori 1999) and maximum likelihood models (Nielsen and Yang 1998) exist to predict site-specific synonymous and non-synonymous substitution rates. However predictions of sites under positive selection are often contentious (Suzuki and Nei 2004, Wong et al. 2004, Yang 2006), especially when they are embedded in neutrally evolving sequence. Here we have adopted a practical approach.

SLR was run with default parameters. Positive sites were collected at a 95% significance threshold. Sites at indel positions were removed. The multiple testing correction by SLR was not applied. Instead we filtered proteins with positive sites based on their over-representation by such sites. Our null hypothesis is that all sites identified to be under positive selection are instead false positives and thus that they are distributed uniformly over all possible sites. We only report proteins whose numbers of putative sites under positive selection are larger than expected by chance from a binomial distribution with a significance threshold of 0.01.

Analysis of gene duplications

Duplication and speciation events can be inferred from the reconciliation of a gene tree with a known species tree. This process relies on a gene tree whose topology is determined accurately. This becomes increasingly difficult to estimate for diverse species sets owing to gene absences, and problems of reliably estimating divergence rates over long distances due to saturation of substitutions. Here we adopt a protocol that takes account of these issues by constructing a tree incrementally.

Multiple alignments were masked for compositionally biased regions. Note, here we used the multiple alignments involving all sequences in each orthologous cluster. The topology of the tree was derived from the consensus tree based on 100 bootstrap samples. Trees were built using the Neighbor-joining method employing the F84 distance (NEIGHBOR, Felsenstein 1989) and the consensus tree was chosen using the extended majority rule (CONSENSE, Felsenstein 1989).

Branch lengths of the tree were estimated based on pairwise synonymous substitution rates (d_S; estimated using CodeML). Saturated pairwise distance estimates larger than $d_S = 5.0$ were discarded. In order to minimize the influence of long-range distance estimates that tend to be unreliable due to saturation, branch lengths of the tree were estimated in three stages. First, branch lengths for the whole tree were determined from all pairwise distances using the program FITCH from the PHYLIP package. This allows for unequal evolutionary rates among branches. Secondly, sub-trees with a maximum distance to leaf of 1.5 were fitted separately to optimize branch lengths within the *D. melanogaster* subgroup, where codon usage bias is homogeneous and neutral rates estimates can be estimated reliably (A. Heger & C. P. Ponting, submitted). Thirdly, sub-trees with a maximum distance of node to leaf of 0.5 were fitted; in this step, the topology was re-estimated to resolve alternative transcripts. Missing values were treated in all steps as replicates of 0. Trees were rooted at the mid-point.

The species tree was reconciled with the gene tree using the RIO method (Zmasek and Eddy 2002), which assigns speciation and duplication events to nodes in the gene tree. Node types in RIO were extended to include lineage-specific duplications and alternative transcripts. Deletions in speciation and duplication events were permitted, if the deletion was confined to only one branch. If a duplication event encompassed all species, it was recorded as an out-paralogous duplication. All other events were recorded as inconsistencies.

Dating duplication events rests on the assumption of a molecular clock (Zuckerkandl and Pauling L 1962) such that synonymous substitutions on all branches occur at a constant rate. However, codon usage bias has varied since the split between the *D. melanogaster* subgroup and the *D. virilis, D. mojavensis* and *D. grimshawi* lineages (A. Heger & C. P. Ponting, submitted). Thus neutral substitution rates are expected to vary both between gene families and within gene families. We circumvented these problems by (1) restricting our analysis to the *D. melanogaster* subgroup, in which codon usage bias has remained largely unchanged (A. Heger & C. P. Ponting, submitted), and by (2) normalizing duplication events between trees.

D. pseudoobscura and *D. persimilis* sequences were used as an outgroup to root gene trees and all in-groups in the melanogaster subgroup were submitted to branch length estimation by maximum likelihood using CodeML from the PAML package. A model allowing no site variation and assuming constant d_N/d_S

across all branches was chosen to maintain a small number of parameters.

We employed two methods to normalize the date of duplication events. Firstly, the largest distance of a duplication event to any of its children was divided by the total height of the tree, which was given by the median distance to root of all leaves excluding pseudogene and outgroup sequences. The resultant date of a duplication event was expressed in percentage distance from the tip with the most recent duplications having a rate of 0%. Secondly, we dated a duplication event by its relative position on a particular branch on the species tree. Here, duplications are either close (0%) or distant (100%) from the most recent speciation or extant species in terminal lineage-specific branches.

Counting gene duplications has many pitfalls due to the species asymmetry in the gene prediction methodology and the unfinished status of these genomes. Duplications in the pair of *D. yakuba* and *D. melanogaster* serve as an illustration of the latter. We found 169 duplication events for genes of *D. melanogaster* with respect to *D. yakuba* and 575 duplication events for *D. yakuba* with respect to *D. melanogaster*. Most of these duplications were, however, due to unplaced contigs in the newly sequenced *D. yakuba* genome assembly. If we removed all entries that are located on chromosome assemblies containing the suffixes "random" or "U", the numbers reduce to 83 and 96, respectively. The majority of the remaining duplications were specific to a chromosome arm (82 and 94, respectively) and most were within 100kb of one another (77 and 93, respectively). The same reduction was observed for longer evolutionary distances (*D. melanogaster* versus *D. pseudoobscura* (**Table S2**), but no data are available for the extant clade of *D. mojavensis, D. grimshawi* and *D. virilis*. An alignment based measure for detecting spurious duplications failed, as sequencing or assembly errors produce gene copies that are indistinguishable from recent duplications. In our analysis we removed all predictions located on unplaced contigs in the genomes, where this information was available, namely for *D. melanogaster, D. yakuba, D. simulans* and *D. pseudoobscura.*

Gene Ontology (GO) Analysis

GO assignments (Ashburner et al. 2000) for *D. melanogaster* genes were obtained from ENSEMBL. GO categories were mapped onto generic GO slim categories (version from 6/12/2005). For a given list of genes from a set of background genes, P-values of over- or under-representation were calculated using a hypergeometric distribution controlling for a false discovery rate of 5% (Boyle et al. 2004).

Additional results

This section contains further analyses, figures and tables that are referred to in the main paper.

Conservation of gene structure

A gene structure of a prediction different to that of its template can be due to evolutionary processes or to artefacts. To examine these alternatives, we selected pairs of orthologous transcripts from two different genomes (see below for orthology assignment) that were predicted from the same template and where one was predicted to have a fully conserved exon structure in a species distantly related to *D. melanogaster* and the other a partially conserved gene structure in a species more closely related to *D. melanogaster*. We parsimoniously assume such differences to be due to sequence or method artefacts.

We tested this assumption by examining 40 predicted genes in *D. yakuba* with degraded gene structure, whose orthologs in *D. willistoni* show full conserved gene structure. Of these 40 genes, 21 genes encompass sequence gaps and can readily be explained as artefacts. Of the remaining 19 genes, three genes have one inserted intron, four have a deleted intron, eight have one or more deleted terminal exons, three have one exon of conserved boundaries, but low sequence identity, and in one gene, the exon boundary has moved 28 nucleotides.

The lengths of inserted introns all are of a multiple of three bases, suggesting that they are artefacts arising either from a deletion in the *D. melanogaster* template or insertion of coding sequence in *D. yakuba*. Missing terminal exons are due to gaps, contig boundaries or over-predicted exons in *D. melanogaster* transcripts. For example, we find that the non-predicted exon of CG3250 aligns upstream to the predicted ortholog in *D. yakuba*, albeit without a starting methionine codon and at lower identity than the other exons.

We repeated this analysis with all genomes that are more distantly related to *D. melanogaster* than *D. yakuba.* Combining the results, we estimate that at least 290 of 1038 genes (28%) with changed exon structure are due to sequence artefacts. This is a lower bound estimate, as it relies on gene structure conservation in species distantly related to *D. melanogaster,* whereas gene structure does evolve over time, albeit slowly. The proportion of artefacts in other species are similar (*D. simulans:* 49%*, D. sechellia:* 28%, *D. ananassae:* 29%, *D. pseudoobscura*: 26%, and *D. persimilis:* 25%) except for a comparatively low value for *D. erecta* (15%) and a very high value for *D. simulans* (49%). Both these rates can be explained by the exceptionally low and high number of gaps in the two genomes, respectively (**Table 1**).

Coverage and percent identity of predictions

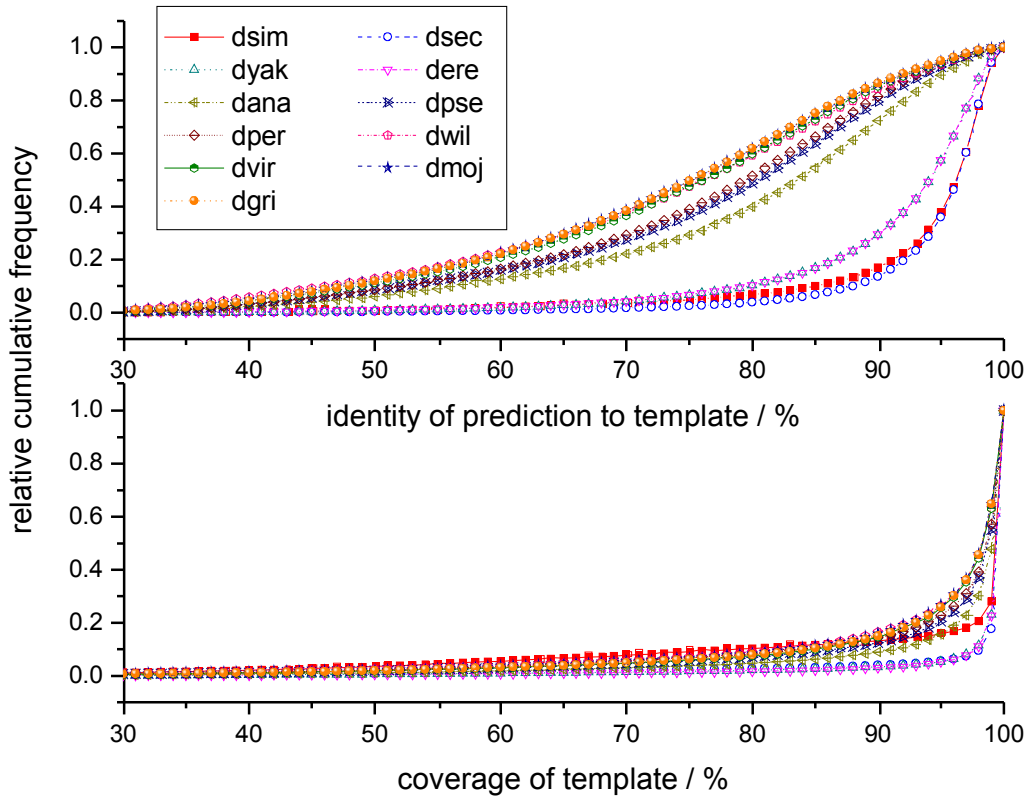

Figure S2: Cumulative histograms of coverage and percent identity between template transcript and their prediction. Only the prediction of highest similarity to the template is shown. Most templates have a prediction that aligns from end to end. The similarity of templates and predictions drops for the further divergent species, but sequence similarity for most templates is in general above 80% , thus showing that prediction by homology in this set of genomes is feasible.

Separation of orthologs and paralogs in pairwise normalized bitscore distances

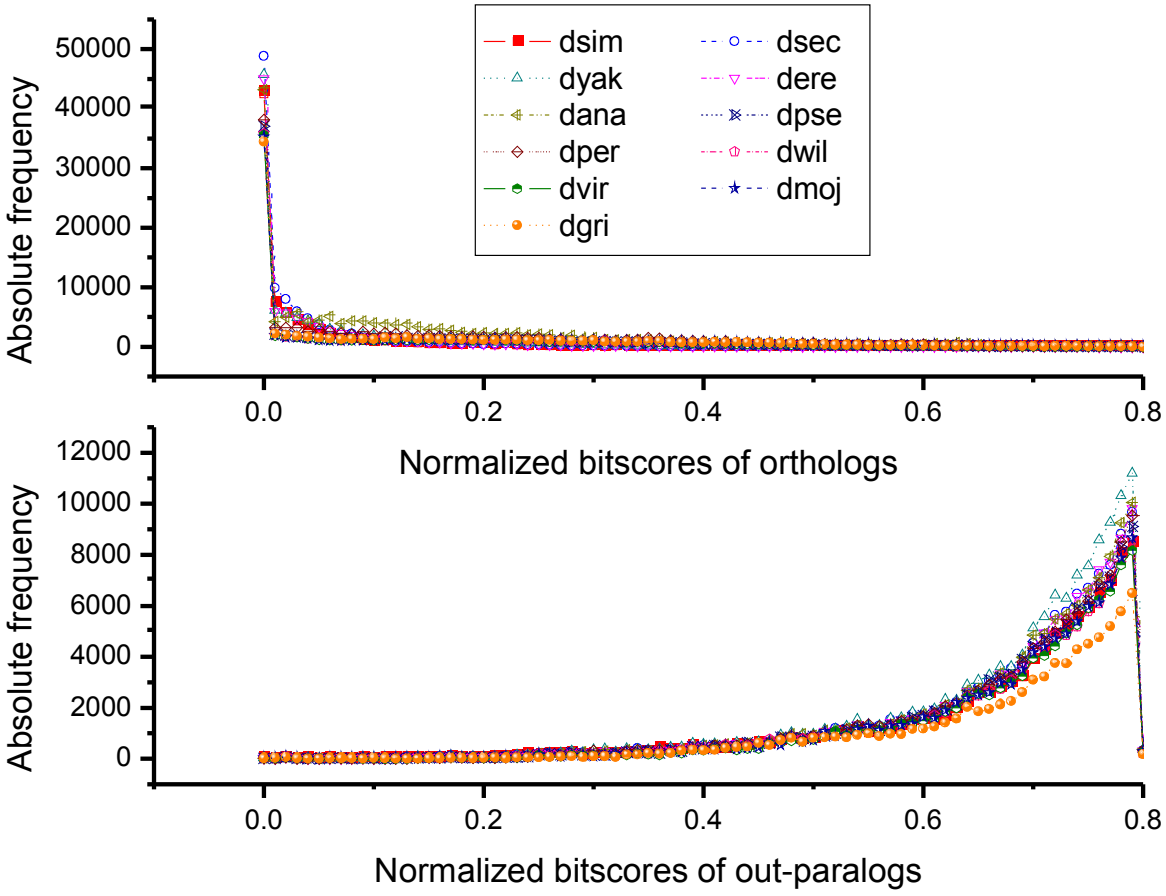

Figure S3: Bitscores of orthologs and paralogs do not overlap, indicating that the orthology assignment procedure is consistent with the similarity measure used.

Length of synteny blocks

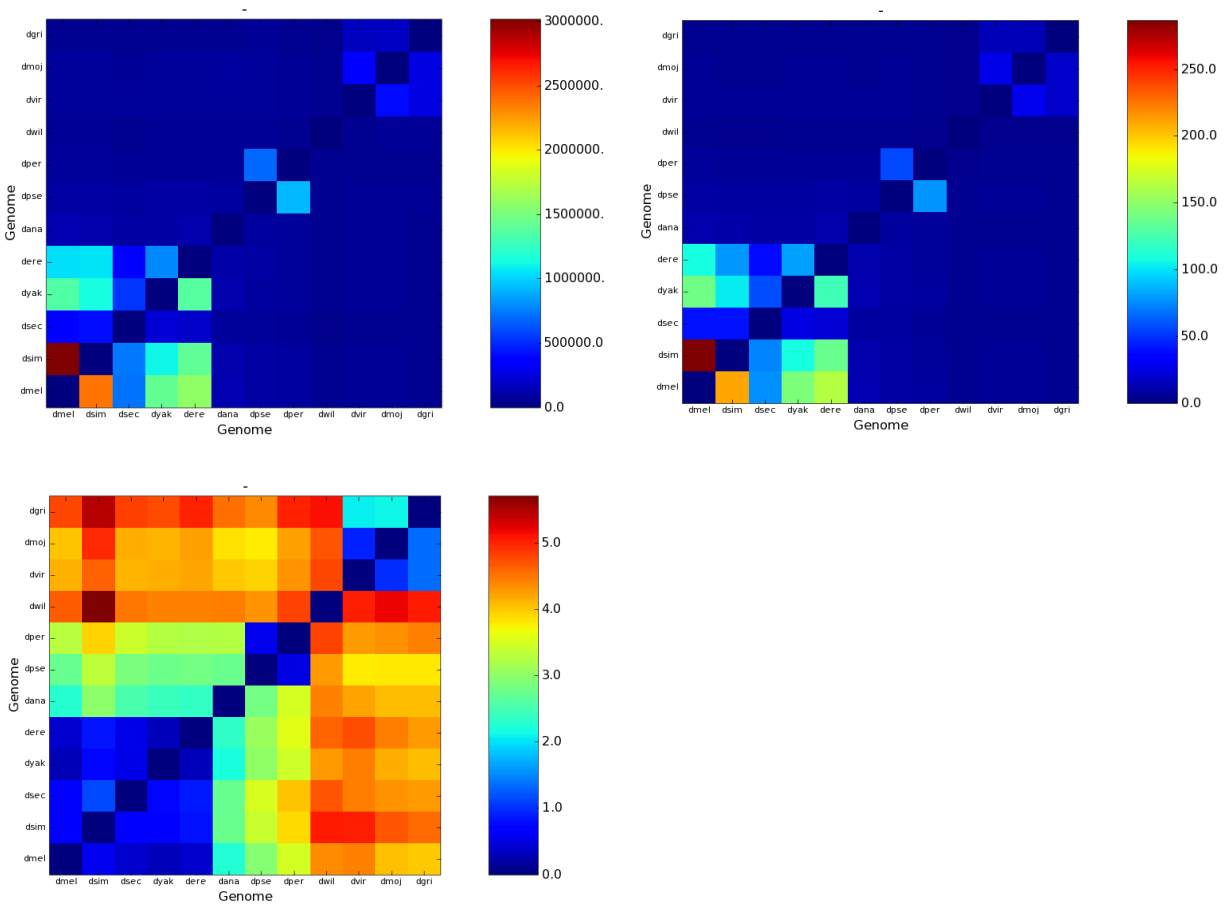

Figure 4: Top row: Average length of synteny blocks (left) and average number of genes (left) per synteny block in each pairwise comparison between the 12 fly species. Bottom row: Percent of orthologous gene pairs that conflict with synteny.

Synonymous substitution rates

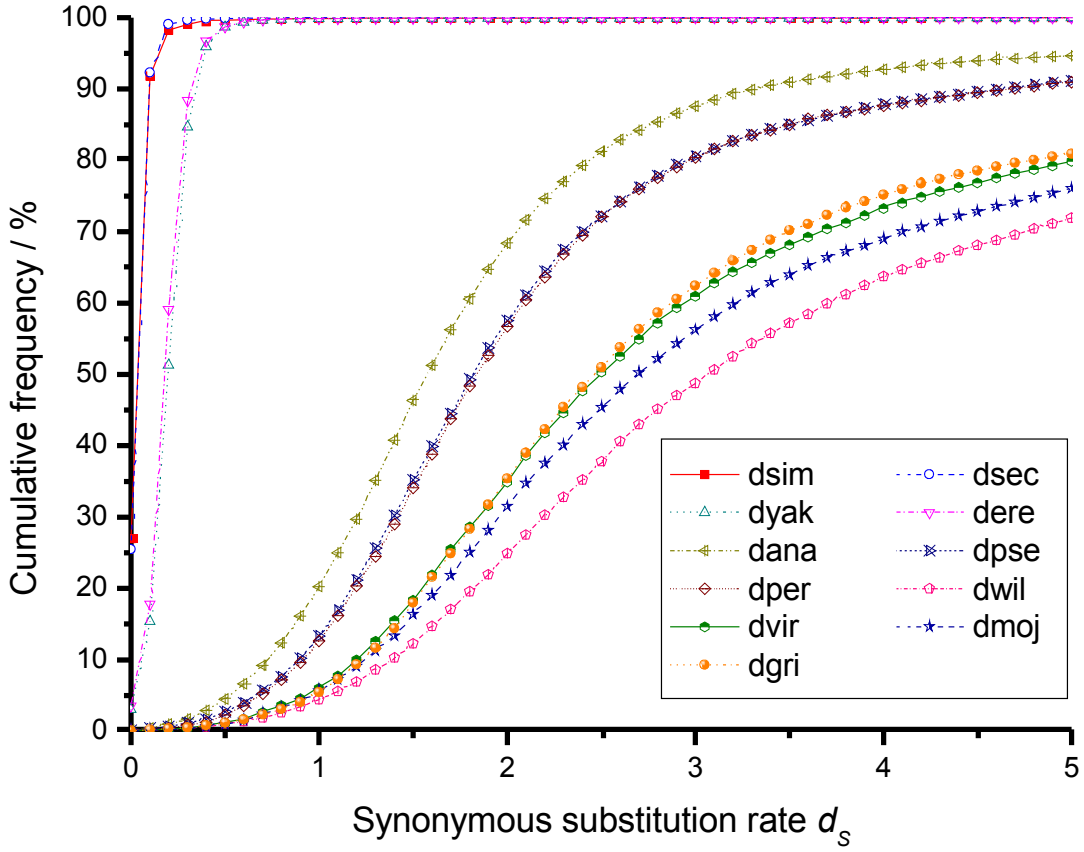

Figure S5: Synonymous rates can be computed between dmel and all other species. Shown here are cumulative frequencies for orthologs in each genome when compared to D. melanogaster. We argue that even the divergent species of D. mojavensis, D. virilis, D. grimshawi, and D. willistoni can be used for synonymous rate computation.

Trees of paralogous sequences

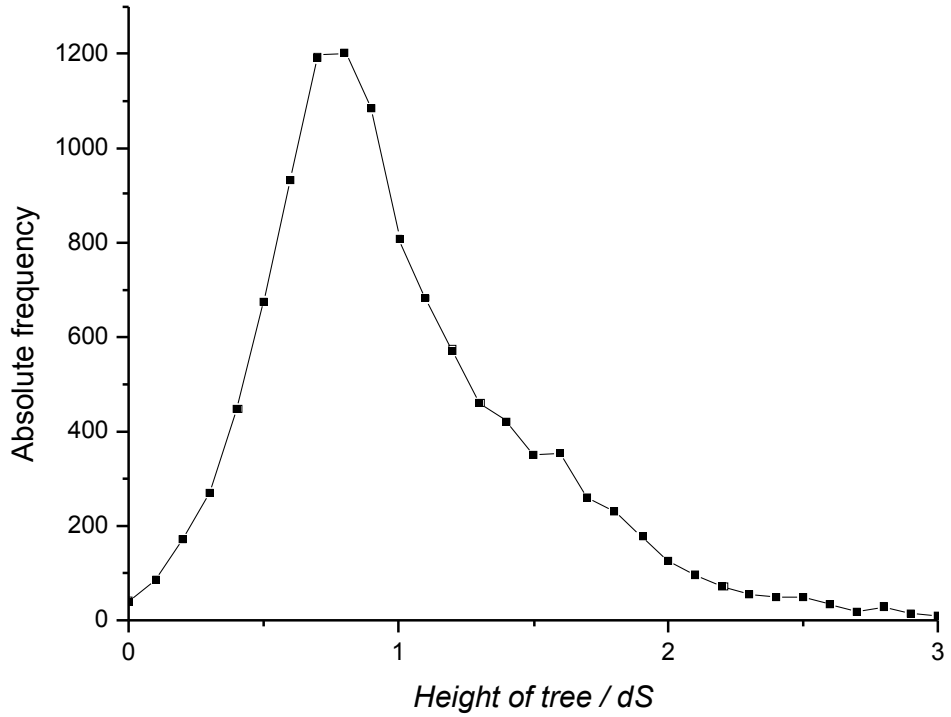

Figure S6: Heights of trees vary considerably between orthologous groups. Shown here are the heights of trees in the D. melanogaster subgroup rooted with D.pseudoobscura/D.persimilis. The height of the tree is given as the median distance of leaf to root. Only genes are considered for the determination of the tree height, excluding pseudogenes.

Branch length variation across gene trees

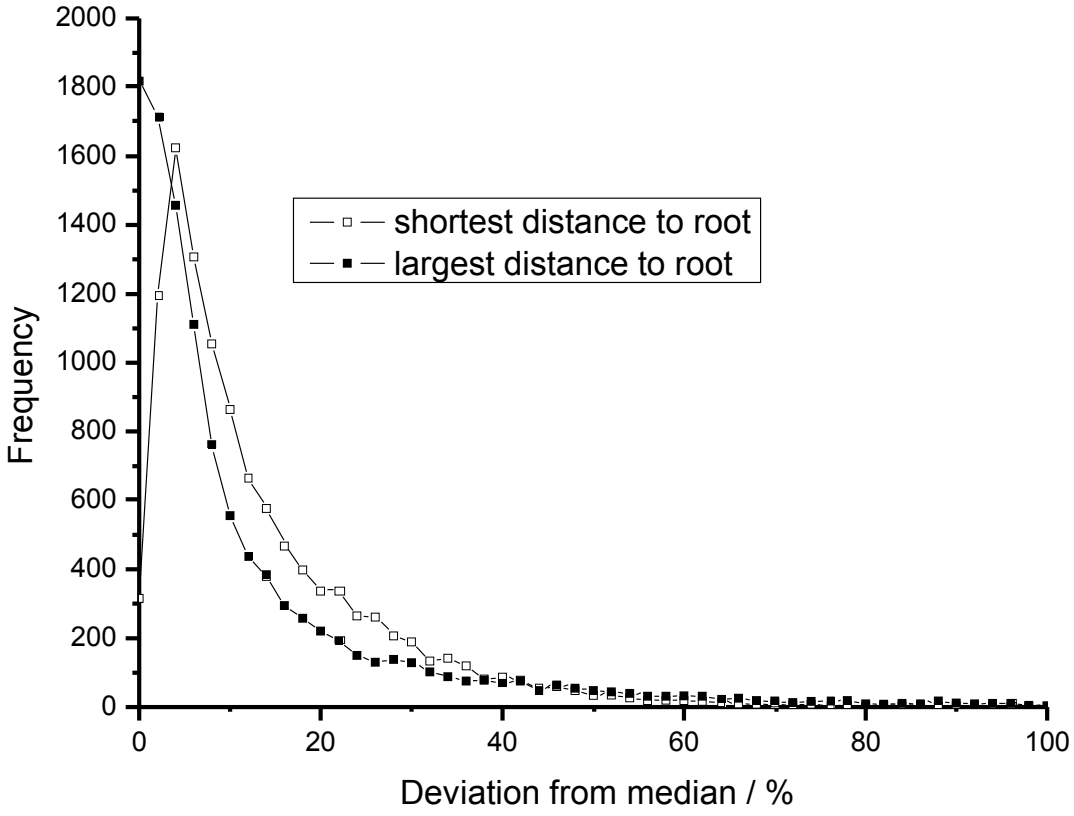

Figure S7: Distribution of the minimum/maximum distance to root of branches. Distances are normalized by the tree height (median distance of leaves to root). Most trees contain largely contemporaneous tips.

Branch length variation across species

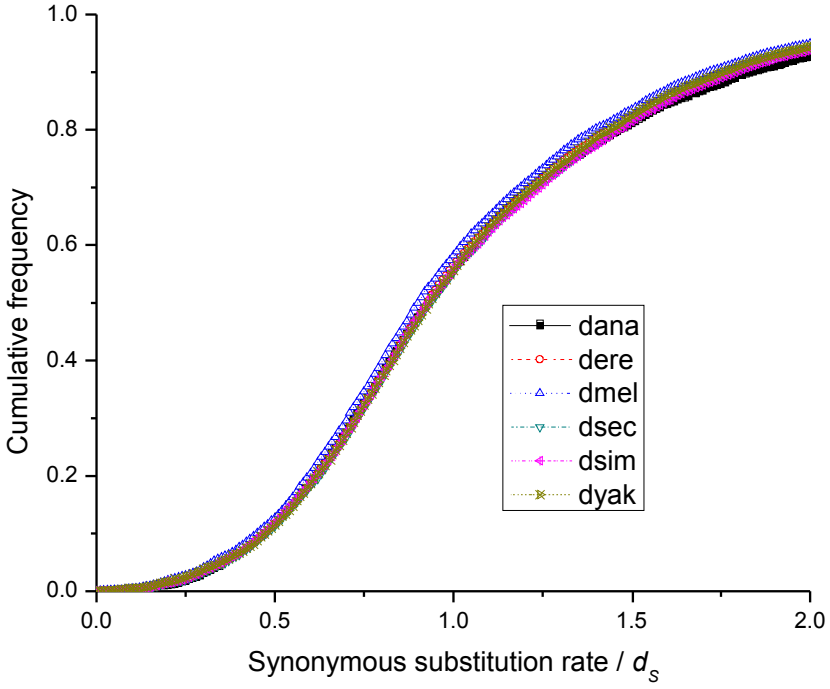

Figure S8: Individual species show no particular trend towards long or short branches. Shown here is the distance of leaves to root for each species in d_S for each species over all gene trees.

Normalized duplication rate for terminal branches

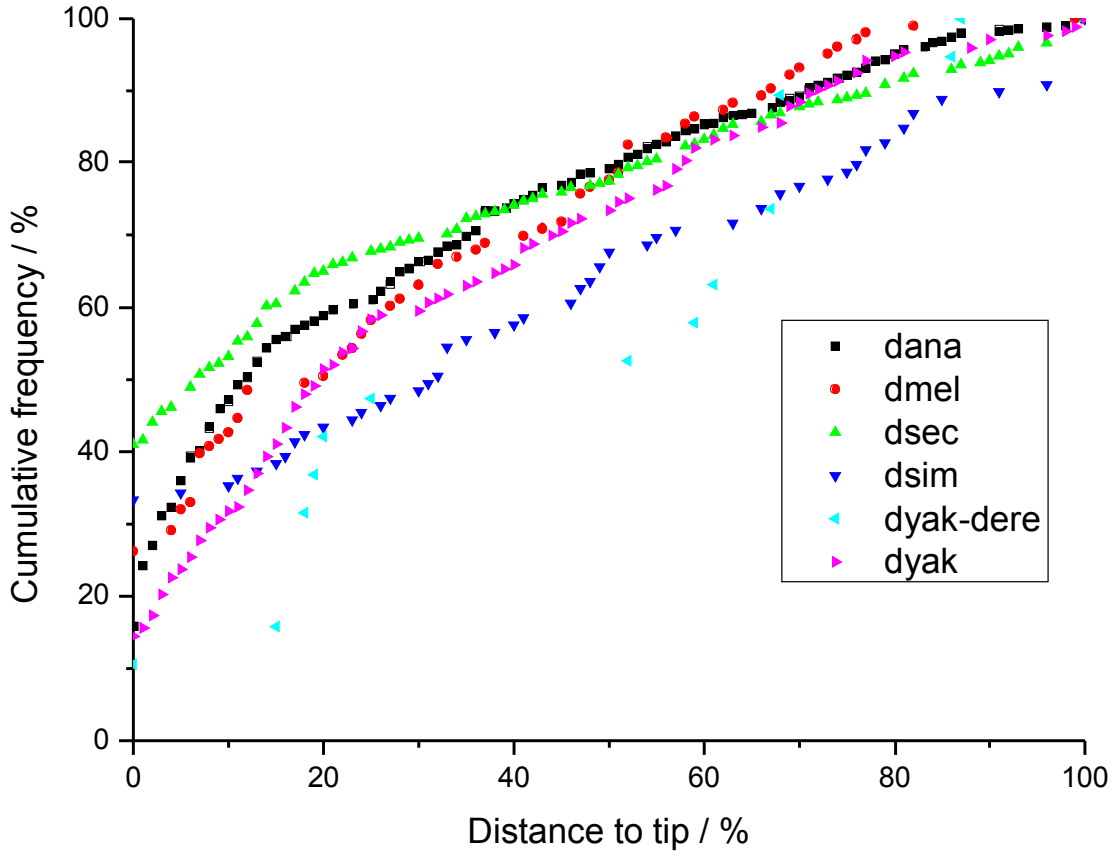

Figure S9: Lineage specific duplications in the D. melanogaster subgroup have often been near-contemporaneous or, previously, at a relatively uniform rate.

Normalized duplication rates for internal branches

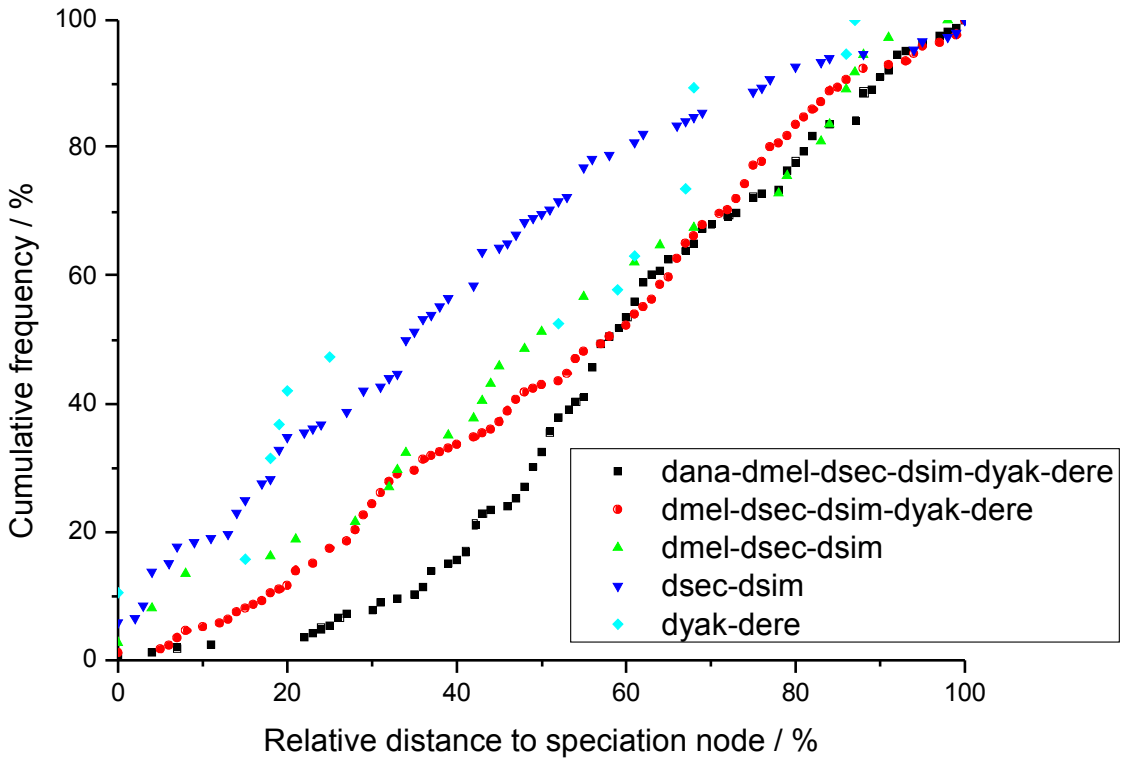

Figure S10: Gene duplications in internal branches in the D. melanogaster subgroup have occurred at uniform rates. Distance are based on d_S values and normalized by the branch of the reconciled species where a particular duplication has occurred.

Normalized duplication rates for genes and pseudogenes

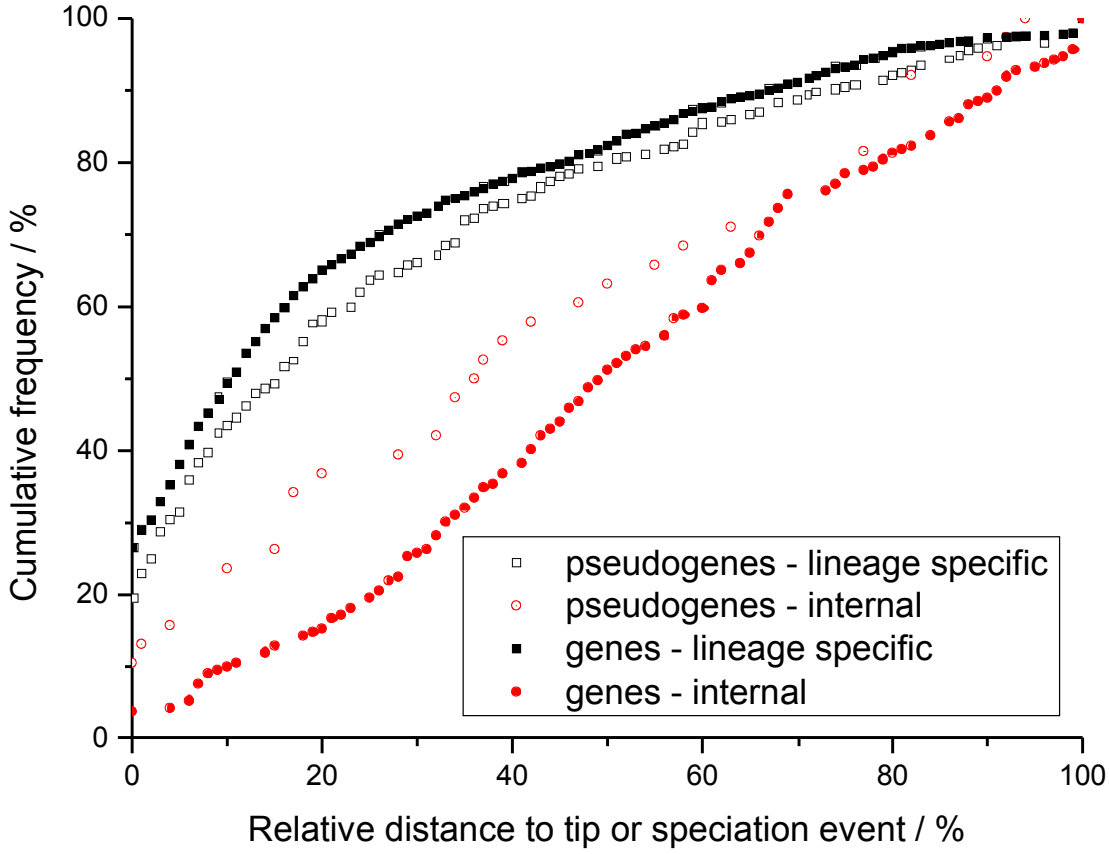

Figure S11: Duplication rates normalized by branch length within the D. melanogaster subgroup (see above). Duplication rates are shown for nodes in the tree producing exclusively genes and exclusively pseudogenes (filled and open symbols, respectively). Duplications in internal species-tree branches leading to pseudogenes in all extant species are rare and restricted to the branches leading to the sibling species D. erecta/D. yakuba and D. simulans/D. sechellia

The effect of unplaced contigs on duplication counts

Species pair	*All*	*Placed*	*Cis*	*Local*
D. melanogaster/D.yakuba	146/811	120/119	118/112	87/109
D. melanogaster/D.pseudoobscura	341/323	146/169	137/110	105/100
D.yakuba/D.pseudoobscura	435/248	76/129	68/85	64/74

Table S2: Unplaced contigs contribute strongly to counts of duplications. Placed duplications involve genes which are not on unplaced contigs. Cis duplicated genes are on the same chromosomal arm, local duplications are within 100kb of one another. The numbers represent duplication events in one genome compared to the other.

The benefit of multiple orthology assignment

Species	*Pairwise*	*Multiple*	*Increase*
D. simulans	11566	11593	100.2%
D. sechellia	12809	12848	100.3%
D. yakuba	12920	12971	100.4%
D. erecta	12883	12974	100.7%
D. ananassae	11681	11802	101.0%
D. pseudoobscura	10829	10998	101.6%
D. persimilis	10474	10637	101.6%
D. willistonis	10035	10162	101.3%
D. virilis	10399	10590	101.8%
D. mojavensis	10163	10340	101.7%
D. grimshawi	10148	10307	101.6%

Table S3: The number of D. melanogaster genes in orthology relationships increases after multiple orthology assignment. Shown are the number of D. melanogaster genes with orthologs in the other 11 species after pairwise and multiple orthology assignment.

Lineage specific duplications in *Drosophila melanogaster*

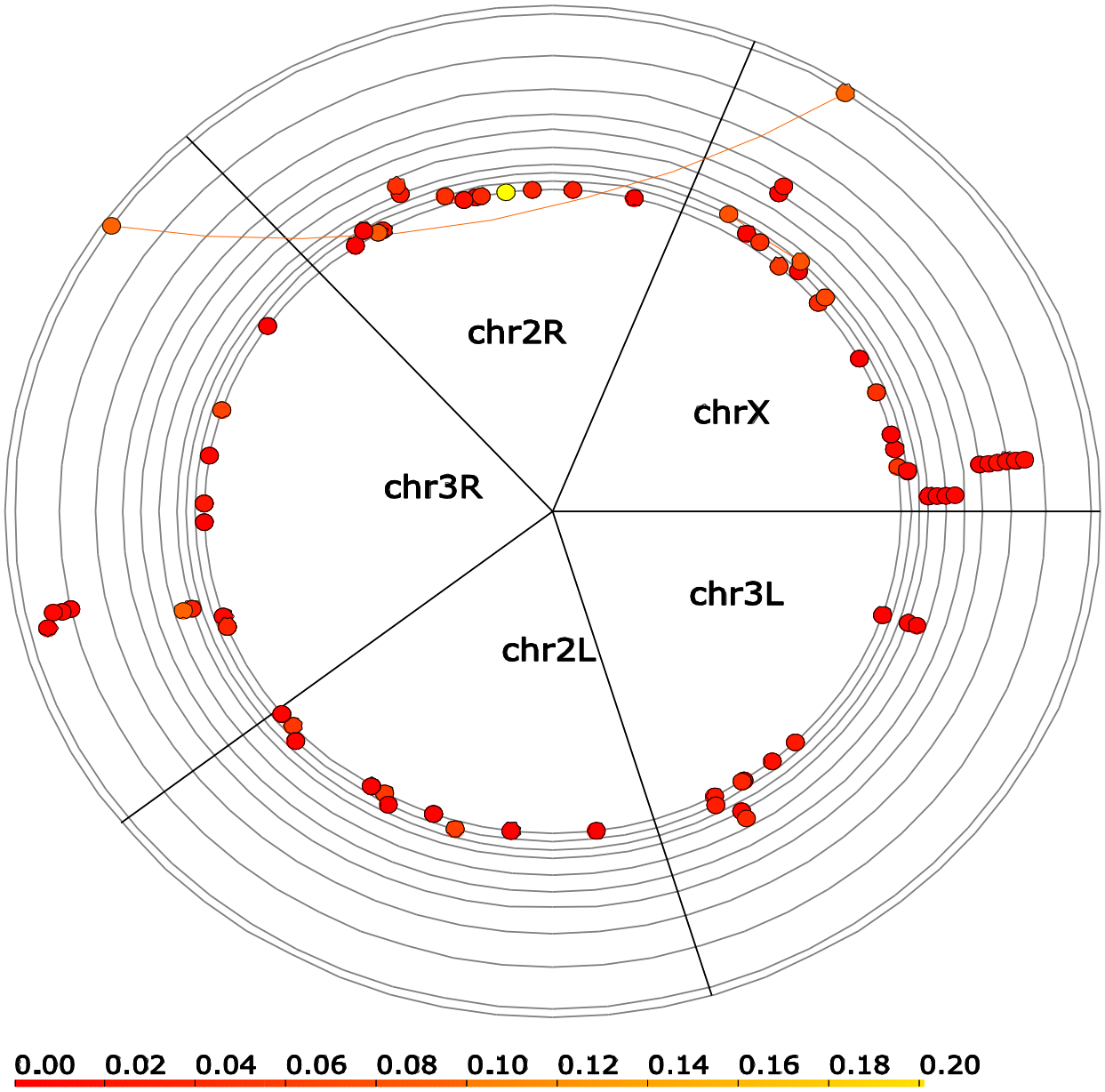

Figure S12: Shown here are lineage-specific duplications in D. melanogaster for the four large chromosomal arms. Each duplication is represented by two dots connected by an arc. These are colored by their divergence (d_S value, see scale). Pseudogenes are colored gray. Genes are placed on the chromosomal arms according to their physical location. Most duplications are local such that only a single dot is visible. Overlapping or very close duplications are stacked on top of each other. Multiple duplications within the same gene family are stacked on top of each other in the outer rings whose increased radius reflects the family size. Each member of a multi-gene family is connected to all other members resulting in a connected path of arcs within a family.

Lineage specific duplications in D. simulans

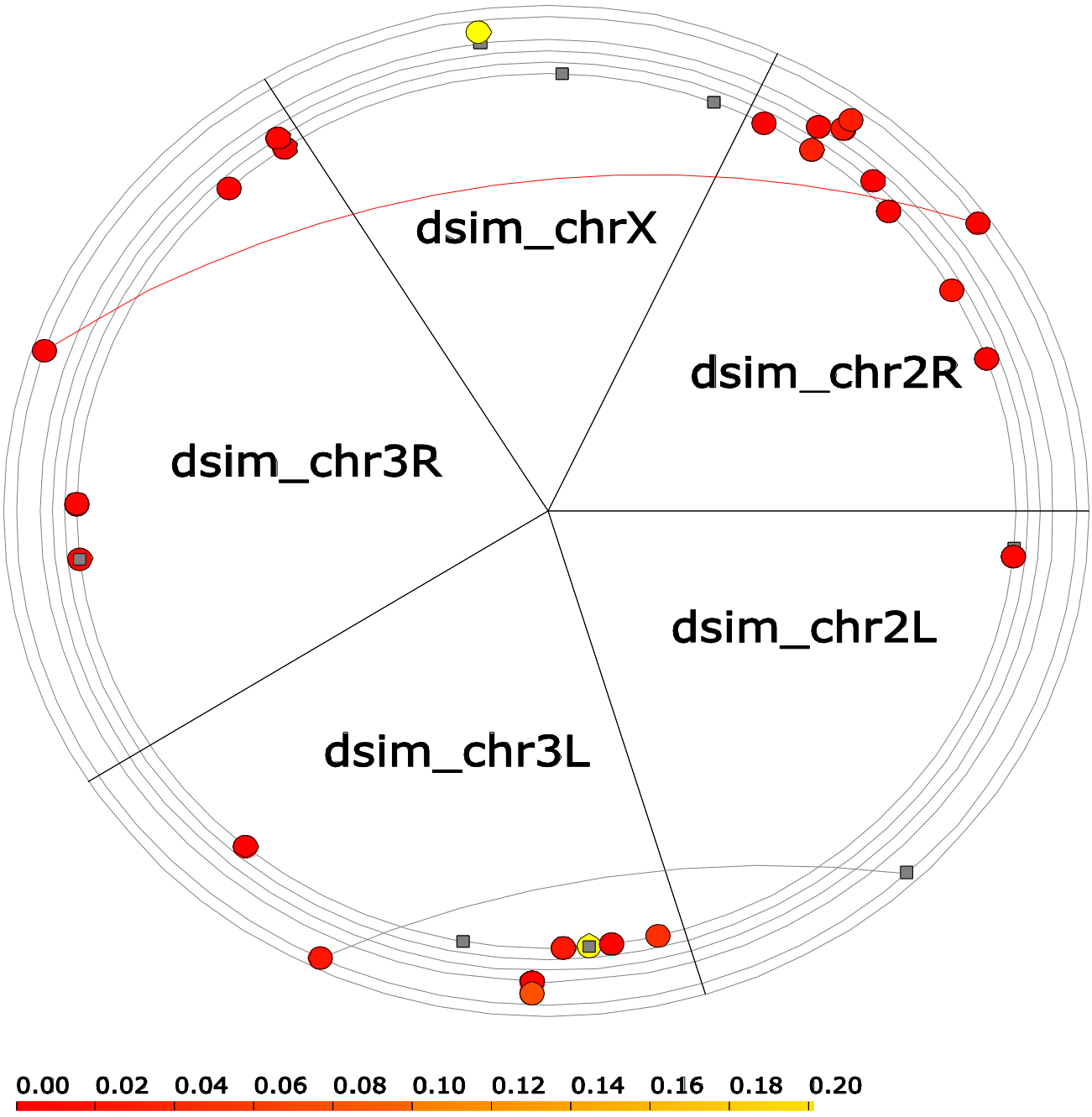

Figure S13: Shown here are lineage-specific duplications in D. simulans for the five large chromosomal arms. See legend of Figure S12 for an explanation of the plot layout.

Proteins with sites predicted to by subject to positive selection:

Transcript	*N*	*P-Value*	*Sites*	*Interpro*		*PFAM*	
CG11538-RA	43	1.37E-71	G638,H659,P636,G640,I433				
CG7052-RB	26	2.80E-23	E637,G638,P647,Y649,W643	IPR000215	Proteinase inhibitor I4, serpin	PF00207	Alpha-2-macroglobulin family
				IPR001599	Alpha-2-macroglobulin	PF01835	Alpha-2-macroglobulin family N-terminal region
				IPR002890	Alpha-2-macroglobulin, N-terminal	PF07677	A-macroglobulin receptor
				IPR011625	Alpha-2-macroglobulin, N-terminal 2	PF07678	A-macroglobulin complement component
				IPR011626	A-macroglobulin complement component	PF07703	Alpha-2-macroglobulin family N-terminal region
				IPR011627	A-macroglobulin receptor		
CG17054-RA	22	5.13E-21	L1133,L1143,K1139,M1118,D1119				
CG32092-RB	18	3.56E-10	H1858,E24,Q227,A1798,Y1856	IPR000859	CUB	PF00431	CUB domain
CG10363-RA	14	1.72E-09	H672,D670,A678,T375,F674	IPR001599	Alpha-2-macroglobulin	PF00207	Alpha-2-macroglobulin family
				IPR002890	Alpha-2-macroglobulin, N-terminal	PF01835	Alpha-2-macroglobulin family N-terminal region
				IPR011625	Alpha-2-macroglobulin, N-terminal 2	PF07677	A-macroglobulin receptor
				IPR011626	A-macroglobulin complement component	PF07678	A-macroglobulin complement component
				IPR011627	A-macroglobulin receptor	PF07703	Alpha-2-macroglobulin family N-terminal region
CG6947-RA	14	4.61E-12	L693,H434,R989,T1074,L859	IPR001687	ATP/GTP-binding site motif A (P-loop)	PF01607	Chitin binding Peritrophin-A domain
				IPR002557	Chitin binding Peritrophin-A		
CG2233-RA	13	3.06E-16	I81,Y74,N79,R51,E26				
CG7298-RA	13	6.55E-17	L136,A195,V127,S194,S76	IPR002557	Chitin binding Peritrophin-A	PF01607	Chitin binding Peritrophin-A domain
CG11321-RA	11	1.51E-05	G1941,S2390,Q2386,V2388,P1574	IPR001841	Zinc finger, RING-type	PF01485	IBR domain

				IPR002867	Zinc finger, C6HC-type		
CG11856-RA	11	1.20E-05	T1818,A2359,G1275,L2021,T3	IPR000156	RanBP1	PF00638	RanBP1 domain
				IPR000694	Proline-rich region	PF00641	Zn-finger in Ran binding protein and others
				IPR000697	EVH1		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR001876	Zinc finger, RanBP2-type		
				IPR013026	Tetratricopeptide region		
CG8201-RA	11	6.55E-13	A185,S182,G179,S186,T180	IPR000104	Antifreeze protein, type I	PF00069	Protein kinase domain
				IPR000449	Ubiquitin-associated	PF00627	UBA/TS-N domain
				IPR000719	Protein kinase	PF02149	Kinase associated domain 1
				IPR001245	Tyrosine protein kinase	PF07714	Protein tyrosine kinase
				IPR001772	Kinase-associated, C-terminal		
				IPR008271	Serine/threonine protein kinase, active site		
CG17838-RA	10	1.58E-12	K489,I497,N484,S494,G495	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
				IPR001472	Bipartite nuclear localization signal		
CG3158-RA	10	3.92E-06	K1038,A854,P616,S39,P706	IPR001410	DEAD/DEAH box helicase	PF00270	DEAD/DEAH box helicase
				IPR001650	Helicase, C-terminal	PF00271	Helicase conserved C-terminal domain
				IPR001687	ATP/GTP-binding site motif A (P-loop)	PF00567	Tudor domain
				IPR002999	Tudor	PF04408	Helicase associated domain (HA2)
				IPR007087	Zinc finger, C2H2-type		
				IPR007502	Helicase-associated region		
				IPR008191	Maternal tudor protein		
				IPR011545	DEAD/DEAH box helicase, N-terminal		
CG11098-RA	9	2.13E-07	V306,F753,Q372,A701,P436	IPR000694	Proline-rich region		
				IPR001472	Bipartite nuclear localization signal		
CG6483-RA	9	1.42E-11	Q194,S234,L142,T268,T	IPR001254	Peptidase S1 and S6,	PF00089	Trypsin

			76			chymotrypsin/Hap	
				IPR001314	Peptidase S1A, chymotrypsin		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR008256	Peptidase S1B, glutamyl endopeptidase I		
CG6493-RA	9	1.05E-04	W433,L261,G603,T489, Y510	IPR000999	Ribonuclease III	PF00035	Double-stranded RNA binding motif
				IPR001159	Double-stranded RNA binding	PF00270	DEAD/DEAH box helicase
				IPR001410	DEAD/DEAH box helicase	PF00271	Helicase conserved C-terminal domain
				IPR001650	Helicase, C-terminal	PF00636	RNase3 domain
				IPR003100	Argonaute and Dicer protein, PAZ	PF02170	PAZ domain
				IPR005034	Protein of unknown function DUF283	PF03368	Domain of unknown function
				IPR006935	Type III restriction enzyme, res subunit	PF04851	Type III restriction enzyme, res subunit
				IPR011545	DEAD/DEAH box helicase, N-terminal		
CG9125-RA	9	4.40E-10	L346,S104,P257,T210,K 229				
CG14303-RA	8	1.69E-04	S480,D461,P1251,P117 9,C1585	IPR002999	Tudor	PF00567	Tudor domain
				IPR008191	Maternal tudor protein		
CG15287-RA	8	1.36E-04	P1075,R1080,S889,H10 77,P611	IPR001472	Bipartite nuclear localization signal		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG2941-RA	8	4.95E-06	S159,D798,D67,A675,K 545				
CG11328-RB	7	8.49E-06	V370,R369,G170,P167, K366	IPR001463	Sodium:alanine symporter	PF00999	Sodium/hydrogen exchanger family
				IPR002090	Na+/H+ exchanger, isoform 6 (NHE6)		
				IPR004709	Sodium/hydrogen exchanger subfamily		
				IPR006153	Sodium/hydrogen exchanger		
CG13742-RA	7	5.52E-06	G576,G441,E105,M28,				

			V291				
			L193,N197,P185,V500,				PDZ domain (Also known as DHR
CG14168-RA	7	3.18E-06	N496	IPR001478	PDZ/DHR/GLGF	PF00595	or GLGF)
			G218,A220,S459,G366,				
CG2791-RA	7	3.88E-06	E355				
CG32469-RA	7	4.57E-11	S56,Y39,G37,S51,A112				
			G1047,R526,P481,S104		RNA-binding region RNP-1		RNA recognition motif. (a.k.a.
CG9809-RA	7	2.48E-05	8,S408	IPR000504	(RNA recognition motif)	PF00076	RRM, RBD, or RNP domain)
					Bipartite nuclear localization		
				IPR001472	signal		
					ATP/GTP-binding site motif		
				IPR001687	A (P-loop)		
CG14247-RA	6	1.43E-08	T112,N135,I115,G26,N6				
			S432,A586,I242,P285,S				
CG14685-RA	6	6.79E-05	459				
			G121,S120,W122,G270,		ATP/GTP-binding site motif		
CG15214-RA	6	1.38E-07	L271	IPR001687	A (P-loop)	PF00047	Immunoglobulin domain
				IPR013098	Immunoglobulin I-set	PF07679	Immunoglobulin I-set domain
				IPR013106	Immunoglobulin V-set	PF07686	Immunoglobulin V-set domain
				IPR013151	Immunoglobulin		
			S91,A92,N93,Y262,Y45				
CG17930-RA	6	2.69E-06	6	IPR000694	Proline-rich region	PF00083	Sugar (and other) transporter
				IPR005828	General substrate transporter		
					Sugar transporter		
				IPR005829	superfamily		
			N14,P81,T186,G48,G26				Cytochrome b5-like Heme/Steroid
CG30502-RA	6	8.87E-07	7	IPR001199	Cytochrome b5	PF00173	binding domain
			T302,D87,S308,S368,T3				
CG3568-RA	6	1.44E-05	01				
			T675,H776,P190,V375,		Bipartite nuclear localization		
CG4616-RB	6	5.14E-06	A786	IPR001472	signal		
			A303,I238,V306,S314,I				
CG5630-RA	6	9.78E-06	200				
			L978,Q754,A720,S109,		ABC transporter,		
CG8523-RA	6	2.87E-03	L96	IPR001140	transmembrane region	PF00005	ABC transporter
					ATP/GTP-binding site motif		ABC transporter transmembrane
				IPR001687	A (P-loop)	PF00664	region
				IPR003439	ABC transporter related		
			S1185,I609,V1187,V489		Bipartite nuclear localization		
CG8590-RA	6	7.47E-04	,Q1127	IPR001472	signal	PF00225	Kinesin motor domain

				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR001752	Kinesin, motor region		
CG9508-RA	6	2.29E-04	W294,S291,R477,S875,I504	IPR000718	Peptidase M13, neprilysin	PF00047	Immunoglobulin domain
				IPR006025	Peptidase M, neutral zinc metallopeptidases, zinc-binding site	PF01431	Peptidase family M13
				IPR013098	Immunoglobulin I-set	PF07679	Immunoglobulin I-set domain
				IPR013106	Immunoglobulin V-set	PF07686	Immunoglobulin V-set domain
				IPR013151	Immunoglobulin		
CG10095-RA	5	1.28E-04	L441,A25,T442,T608,K610	IPR000104	Antifreeze protein, type I	PF07686	Immunoglobulin V-set domain
				IPR013106	Immunoglobulin V-set		
CG10586-RA	5	1.22E-05	G147,P38,A269,I193,S153	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
CG1082-RA	5	4.61E-05	A179,A113,L454,W350,S76	IPR000379	Esterase/lipase/thioesterase	PF00135	Carboxylesterase
				IPR002018	Carboxylesterase, type B		
CG12437-RB	5	1.84E-03	A282,A69,G490,S729,A64	IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG12713-RA	5	2.17E-05	K26,S70,D125,F67,G234				
CG12885-RA	5	1.26E-04	P328,P332,A331,G334,T621				
CG13190-RA	5	5.16E-05	N207,S58,P220,L21,N187				
CG13400-RA	5	1.48E-03	K206,I380,K387,I685,G386	IPR005033	YEATS	PF03366	YEATS family
CG13594-RA	5	1.47E-05	G171,P464,V410,A70,G166	IPR000104	Antifreeze protein, type I		
				IPR001472	Bipartite nuclear localization signal		
				IPR004051	Kv1.4 voltage-gated K+ channel		
CG13793-RA	5	3.64E-05	W79,L108,A77,T75,N109	IPR000175	Sodium:neurotransmitter symporter	PF00209	Sodium:neurotransmitter symporter family
CG14499-RA	5	2.82E-07	S16,K39,Q15,L19,Q47				

CG15427-RD	5	2.02E-03	A898,Q100,R900,T902, E505	IPR003961	Fibronectin, type III	PF00041	Fibronectin type III domain
				IPR003962	Fibronectin, type III subdomain	PF00047	Immunoglobulin domain
				IPR013098	Immunoglobulin I-set	PF07679	Immunoglobulin I-set domain
				IPR013106	Immunoglobulin V-set	PF07686	Immunoglobulin V-set domain
				IPR013151	Immunoglobulin		
CG18596-RA	5	7.44E-03	T640,V820,S46,E675,E 853	IPR001472	Bipartite nuclear localization signal	PF00588	SpoU rRNA Methylase family
				IPR001537	tRNA/rRNA methyltransferase, SpoU		
CG2221-RA	5	2.60E-04	M149,Q148,D142,H541, F268	IPR002165	Plexin	PF01437	Plexin repeat
CG30421-RA	5	3.97E-03	K1042,A631,V1043,V9 72,G1025	IPR001394	Peptidase C19, ubiquitin carboxyl-terminal hydrolase 2	PF00443	Ubiquitin carboxyl-terminal hydrolase
CG3126-RA	5	3.27E-03	A1101,E805,S874,Q890, S853	IPR000651	Guanine nucleotide exchange factor for Ras-like GTPases, N-terminal	PF00617	RasGEF domain
				IPR000694	Proline-rich region	PF00618	Guanine nucleotide exchange factor for Ras-like GTPases
				IPR001472	Bipartite nuclear localization signal		N-terminal motif
				IPR001895	Guanine-nucleotide dissociation stimulator CDC25		
				IPR006077	Vinculin/alpha-catenin		
CG33349-RA	5	1.37E-04	T43,K259,S377,S250,F2 31	IPR001873	Na+ channel, amiloride-sensitive	PF00858	Amiloride-sensitive sodium channel
CG4613-RA	5	5.69E-05	S156,S149,Q269,P152,P 49	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
CG4898-RE	5	1.88E-07	N307,H304,P308,T312, N306	IPR000533	Tropomyosin	PF00261	Tropomyosin
CG6204-RA	5	2.76E-03	I831,G230,M428,M749, T213	IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG6467-RA	5	5.97E-06	T75,S143,K146,K216,S 231	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		

CG6967-RA	5	8.49E-04	L179,H181,N604,E44,P651	IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG7413-RA	5	2.67E-03	T800,K808,E820,M572,E558	IPR002719	Retinoblastoma-associated protein, B-box	PF01857	Retinoblastoma-associated protein B domain
				IPR002720	Retinoblastoma-associated protein, A-box	PF01858	Retinoblastoma-associated protein A domain
CG7795-RA	5	2.50E-04	S77,V79,S312,N482,D163				
CG7869-RA	5	1.76E-03	L513,S499,Y648,L461,S250	IPR000330	SNF2-related	PF00176	SNF2 family N-terminal domain
				IPR001472	Bipartite nuclear localization signal		
CG9925-RA	5	1.64E-03	I827,I856,K864,C53,R857	IPR000694	Proline-rich region	PF00567	Tudor domain
				IPR002893	Zinc finger, MYND-type	PF01753	MYND finger
				IPR002999	Tudor		
				IPR008191	Maternal tudor protein		
CG10101-RA	4	5.04E-03	R269,L263,P277,N265	IPR001320	Ionotropic glutamate receptor	PF00060	Ligand-gated ion channel
CG10700-RA	4	3.13E-03	E518,L534,E79,S122	IPR000103	Pyridine nucleotide-disulphide oxidoreductase, class-II	PF00070	Pyridine nucleotide-disulphide oxidoreductase
				IPR000759	Adrenodoxin reductase	PF00355	Rieske [2Fe-2S] domain
				IPR001100	Pyridine nucleotide-disulphide oxidoreductase, class I	PF07992	Pyridine nucleotide-disulphide oxidoreductase
				IPR001327	Pyridine nucleotide-disulphide oxidoreductase, NAD-binding region		
				IPR005806	Rieske [2Fe-2S] region		
				IPR013027	FAD-dependent pyridine nucleotide-disulphide oxidoreductase		
CG10901-RA	4	1.47E-03	S20,R139,P138,L375				
CG11912-RA	4	1.36E-04	N195,S225,T151,S168	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
				IPR008256	Peptidase S1B, glutamyl endopeptidase I		
CG11983-RA	4	9.55E-04	S417,I134,V126,V111				

CG12096-RA	4	1.91E-03	Y259,L205,P165,L155	IPR001472	Bipartite nuclear localization signal		
CG12945-RA	4	9.21E-05	K208,P207,P233,D216	IPR001472	Bipartite nuclear localization signal		
CG12951-RA	4	1.78E-04	T80,D79,L11,D108	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
				IPR001316	Peptidase S1E, alpha-lytic endopeptidase		
CG13075-RA	4	5.35E-05	L70,A115,I134,Q29				
CG13329-RA	4	6.24E-05	D73,R72,Y87,G222	IPR000164	Histone H3	PF00125	Core histone H2A/H2B/H3/H4
				IPR001472	Bipartite nuclear localization signal		
				IPR002185	Dopamine D4 receptor		
				IPR007124	Histone-fold/TFIID-TAF/NF-Y		
				IPR007125	Histone core		
CG13353-RA	4	1.13E-04	R77,A76,W48,G94				
CG14355-RA	4	5.47E-03	S56,A242,T438,T136				
CG14369-RA	4	8.78E-07	N44,G61,L46,L45				
CG14893-RA	4	2.17E-03	T326,L154,N337,A82	IPR004262	Male sterility	PF03015	Male sterility protein
				IPR013120	Male sterility C-terminal	PF07993	Male sterility protein
CG15098-RA	4	2.14E-05	Q121,V67,D138,N125				
CG15279-RC	4	6.37E-03	L11,S558,N403,D36	IPR000175	Sodium:neurotransmitter symporter	PF00209	Sodium:neurotransmitter symporter family
CG15415-RA	4	7.10E-03	P488,T312,A164,V508	IPR001472	Bipartite nuclear localization signal		
CG15707-RA	4	4.64E-04	E327,G559,P625,F552	IPR000571	Zinc finger, CCCH-type	PF00567	Tudor domain
				IPR002999	Tudor	PF00642	Zinc finger C-x8-C-x5-C-x3-H type (and similar)
				IPR008191	Maternal tudor protein		
CG16801-RB	4	1.46E-03	P119,S118,R112,L121	IPR000324	Vitamin D receptor	PF00104	Ligand-binding domain of nuclear hormone receptor
				IPR000536	Nuclear hormone receptor, ligand-binding	PF00105	Zinc finger, C4 type (two domains)
				IPR000694	Proline-rich region		
				IPR001628	Nuclear hormone receptor, DNA-binding		

				IPR001723	Steroid hormone receptor		
CG16965-RA	4	8.16E-03	D223,S660,H656,Q644	IPR005195	Glycoside hydrolase family 65, central catalytic	PF03632	Glycosyl hydrolase family 65 central catalytic domain
CG18389-RA	4	9.74E-03	A350,S353,N519,I37	IPR001472	Bipartite nuclear localization signal	PF05225	helix-turn-helix, Psq domain
				IPR007889	Helix-turn-helix, Psq		
CG18550-RA	4	1.36E-03	S68,S69,G209,S67	IPR003534	Major royal jelly protein	PF03022	Major royal jelly protein
CG1863-RB	4	6.66E-03	L104,R437,S677,V62	IPR000454	Eubacterial/plasma membrane H+-transporting two-sector ATPase, C subunit		
CG2681-RA	4	2.15E-03	M69,K77,I161,T78	IPR001841	Zinc finger, RING-type	PF03145	Seven in absentia protein family
				IPR004162	Seven in absentia protein		
CG3024-RA	4	3.88E-04	R293,N158,L290,N151	IPR001270	Chaperonin clpA/B	PF06309	Torsin
				IPR010448	Torsin		
CG31301-RA	4	1.23E-03	C276,A176,T396,P198	IPR000467	D111/G-patch	PF00035	Double-stranded RNA binding motif
				IPR001159	Double-stranded RNA binding	PF01424	R3H domain
				IPR001374	Single-stranded nucleic acid binding R3H	PF01585	G-patch domain
CG31322-RA	4	4.06E-03	T199,S551,S522,R194	IPR002300	Aminoacyl-tRNA synthetase, class Ia	PF00133	tRNA synthetases class I (I, L, M and V)
				IPR002304	Methionyl-tRNA synthetase, class Ia		
CG31453-RA	4	1.33E-03	H63,A36,L21,A56	IPR000767	Disease resistance protein	PF00004	ATPase family associated with various cellular activities (AAA)
				IPR001270	Chaperonin clpA/B		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR003959	AAA ATPase, central region		
CG31519-RA	4	7.77E-04	R110,A103,R220,S104	IPR004117	Olfactory receptor, Drosophila	PF02949	7tm Odorant receptor
CG31753-RA	4	7.06E-03	T691,R789,G742,T759	IPR004051	Kv1.4 voltage-gated K+ channel	PF00096	Zinc finger, C2H2 type
				IPR007086	Zinc finger, C2H2-subtype		
				IPR007087	Zinc finger, C2H2-type		
CG31928-RA	4	7.60E-04	H22,V71,T188,D189	IPR001461	Peptidase A1, pepsin	PF00026	Eukaryotic aspartyl protease
CG32221-RA	4	1.48E-03	N166,Y167,C316,A393	IPR001810	Cyclin-like F-box	PF00646	F-box domain

CG32364-RA	4	6.80E-05	T19,C107,E113,E186	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)		
CG32438-RD	4	8.33E-03	S731,M697,A287,I713	IPR000533	Tropomyosin	PF02463	RecF/RecN/SMC N terminal domain
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR003395	SMC protein, N-terminal		
CG33261-RA	4	7.44E-04	G451,G447,V445,S446	IPR000210	BTB	PF00651	BTB/POZ domain
				IPR000354	Involucrin repeat	PF00904	Involucrin repeat
				IPR007087	Zinc finger, C2H2-type		
				IPR013069	BTB/POZ		
CG3829-RA	4	3.66E-03	T239,A236,N251,W240	IPR002159	CD36 antigen	PF01130	CD36 family
CG4435-RA	4	1.50E-03	N97,K164,R318,S170	IPR001503	Glycosyl transferase, family 10	PF00852	Glycosyltransferase family 10 (fucosyltransferase)
CG5250-RA	4	3.13E-04	Y79,I142,I112,N110	IPR000402	Na+/K+ ATPase, beta subunit	PF00287	Sodium / potassium ATPase beta chain
				IPR001472	Bipartite nuclear localization signal		
CG6176-RA	4	5.13E-03	G98,K581,S238,N140	IPR001719	AP endonuclease, family 2	PF04130	Spc97 / Spc98 family
				IPR007259	Spc97/Spc98		
CG7643-RA	4	3.45E-03	E320,P318,V153,P322	IPR000719	Protein kinase	PF00069	Protein kinase domain
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
				IPR008271	Serine/threonine protein kinase, active site		
CG8295-RD	4	1.20E-04	K1,K3,K5,Q2				
CG8589-RA	4	1.00E-03	S385,F131,C91,P132	IPR008191	Maternal tudor protein	PF00567	Tudor domain
CG8712-RB	4	2.18E-04	L219,T355,T110,P141				
CG9109-RA	4	2.12E-03	V191,S42,P184,E267	IPR003378	Fringe-like	PF02434	Fringe-like
CG9606-RA	4	7.77E-04	R311,P71,S369,L297	IPR001247	Exoribonuclease	PF01138	3' exoribonuclease family, domain 1
				IPR001472	Bipartite nuclear localization signal		
CG9698-RA	4	2.32E-04	G80,K82,L84,E81	IPR005123	2OG-Fe(II) oxygenase	PF03171	2OG-Fe(II) oxygenase superfamily
						PF08336	Prolyl 4-Hydroxylase alpha-subunit, N-terminal region
CG10090-RA	3	1.66E-03	A142,T168,S204	IPR003397	Mitochondrial import inner membrane translocase, subunit Tim17/22	PF02466	Tim17/Tim22/Tim23 family

CG10938-RA	3	2.95E-03	R188,D194,L186	IPR000426	Proteasome alpha-subunit	PF00227	Proteasome A-type and B-type
				IPR001353	20S proteasome, A and B subunits		
CG11347-RD	3	1.43E-03	E124,N169,L2	IPR001472	Bipartite nuclear localization signal		
CG11356-RA	3	1.32E-03	I50,S45,S96	IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG11470-RA	3	1.45E-03	L118,P122,H86	IPR001778	Pollen allergen Poa pIX/Phl pVI, C-terminal	PF05267	Protein of unknown function (DUF725)
				IPR007931	Protein of unknown function DUF725		
CG11723-RA	3	3.53E-03	T228,L233,L216	IPR004210	BESS motif	PF02944	BESS motif
CG11843-RA	3	4.79E-03	V26,V25,G27	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
CG12108-RA	3	4.69E-03	K165,A186,E164	IPR002472	Palmitoyl protein thioesterase	PF02089	Palmitoyl protein thioesterase
CG13732-RA	3	7.75E-04	A119,E46,L103				
CG13788-RA	3	5.50E-03	Q159,T162,A157			PF08395	7tm Chemosensory receptor
CG14102-RA	3	5.91E-03	A204,P246,V279				
CG14151-RA	3	7.32E-04	T138,P141,A137	IPR001472	Bipartite nuclear localization signal		
CG14379-RA	3	8.83E-04	Y65,N69,T24	IPR010512	Protein of unknown function DUF1091	PF06477	Protein of unknown function (DUF1091)
CG14669-RA	3	5.26E-04	Q83,F85,T84	IPR001230	Prenyl group, CAAX box, attachment site		
				IPR001806	Ras GTPase		
				IPR010916	TonB box, N-terminal		
CG14971-RA	3	9.81E-03	G292,Q110,T39	IPR004853	Protein of unknown function DUF250	PF03151	Triose-phosphate Transporter family
						PF08449	UAA transporter family
CG15905-RA	3	8.05E-04	L159,T160,T140				
CG17610-RA	3	1.54E-03	S97,G99,D124	IPR006209	EGF-like	PF00008	EGF-like domain
				IPR013032	EGF-like region		
CG18094-RA	3	8.15E-03	R81,C343,R318	IPR000086	NUDIX hydrolase	PF00293	NUDIX domain
CG18143-RA	3	9.67E-03	H364,F361,S216	IPR002048	Calcium-binding EF-hand	PF01979	Amidohydrolase family
				IPR006680	Amidohydrolase 1		
CG1830-RB	3	4.60E-03	G298,R300,R301	IPR000719	Protein kinase	PF00069	Protein kinase domain

				IPR001245	Tyrosine protein kinase		
				IPR001472	Bipartite nuclear localization signal		
				IPR002291	Phosphorylase kinase, gamma catalytic subunit		
				IPR008271	Serine/threonine protein kinase, active site		
CG2050-RA	3	5.45E-03	V537,A378,T180	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
				IPR001472	Bipartite nuclear localization signal		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG30190-RA	3	2.35E-03	R194,L199,Y192	IPR000694	Proline-rich region		
CG30327-RA	3	5.65E-03	P279,R25,F106	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)
CG31267-RA	3	1.80E-03	S74,W80,Q182	IPR001230	Prenyl group, CAAX box, attachment site	PF00089	Trypsin
				IPR001254	Peptidase S1 and S6, chymotrypsin/Hap		
				IPR001314	Peptidase S1A, chymotrypsin		
CG31605-RF	3	7.61E-04	V201,G204,A160	IPR013098	Immunoglobulin I-set	PF00047	Immunoglobulin domain
				IPR013106	Immunoglobulin V-set	PF07679	Immunoglobulin I-set domain
				IPR013151	Immunoglobulin	PF07686	Immunoglobulin V-set domain
CG3186-RA	3	8.36E-04	T17,S107,L127	IPR001884	Eukaryotic initiation factor 5A hypusine (eIF-5A)	PF01287	Eukaryotic initiation factor 5A hypusine, DNA-binding OB fold
CG32510-RA	3	1.83E-05	S63,K58,I93				
CG33005-RB	3	3.02E-03	L12,E244,S7				
CG33479-RA	3	4.74E-03	E13,V47,N147				
CG4413-RA	3	6.89E-03	T256,V261,P252	IPR007086	Zinc finger, C2H2-subtype	PF00096	Zinc finger, C2H2 type
				IPR007087	Zinc finger, C2H2-type	PF07776	Zinc-finger associated domain (zf-AD)
				IPR012934	Zinc finger, AD-type		
CG4688-RA	3	2.12E-03	S224,V226,I142	IPR004045	Glutathione S-transferase, N-terminal	PF00043	Glutathione S-transferase, C-terminal domain
				IPR004046	Glutathione S-transferase, C-terminal	PF02798	Glutathione S-transferase, N-terminal domain
CG5326-RA	3	3.81E-04	A116,R117,V119	IPR002076	GNS1/SUR4 membrane protein	PF01151	GNS1/SUR4 family

CG5783-RA	3	3.38E-03	S148,G278,S250	IPR000182	GCN5-related N-acetyltransferase	PF00583	Acetyltransferase (GNAT) family
						PF08445	FR47-like protein
CG6678-RA	3	8.86E-03	A287,T354,A191	IPR000408	Regulator of chromosome condensation, RCC1	PF00415	Regulator of chromosome condensation (RCC1)
CG7138-RA	3	4.17E-03	K236,P28,S80	IPR001159	Double-stranded RNA binding	PF00035	Double-stranded RNA binding motif
CG8667-RA	3	2.53E-03	S118,S117,A116	IPR001092	Basic helix-loop-helix dimerisation region bHLH	PF00010	Helix-loop-helix DNA-binding domain
CG9631-RA	3	9.06E-03	P144,S297,I76	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap	PF00089	Trypsin
				IPR001314	Peptidase S1A, chymotrypsin		
				IPR001687	ATP/GTP-binding site motif A (P-loop)		
CG11072-RA	2	1.51E-03	V36,T41	IPR013069	BTB/POZ	PF00651	BTB/POZ domain
CG11076-RA	2	8.69E-03	G168,E127				
CG11769-RA	2	7.22E-03	E30,Q188	IPR000104	Antifreeze protein, type I		
CG18779-RA	2	3.77E-03	Q82,N84	IPR000618	Insect cuticle protein	PF00379	Insect cuticle protein
				IPR001419	HMW glutenin		
CG30441-RA	2	7.87E-03	T58,V63				
CG31469-RA	2	9.98E-03	S99,T101	IPR000106	Low molecular weight phosphotyrosine protein phosphatase	PF01451	Low molecular weight phosphotyrosine protein phosphatase
				IPR002115	Mammalian LMW phosphotyrosine protein phosphatase		
CG9415-RB	2	6.72E-03	A176,N21	IPR001472	Bipartite nuclear localization signal	PF00170	bZIP transcription factor
				IPR004827	Basic-leucine zipper (bZIP) transcription factor	PF07716	Basic region leucine zipper
				IPR011616	bZIP transcription factor, bZIP_1		
				IPR011700	Basic leucine zipper		
CG9568-RA	2	5.76E-03	E52,A108				

Table S4: Transcript in D. melanogaster with positively selected sites. N gives the number of sites; only the first five residues with highest significance are shown. Interpro (Mulder et al. 2005) and PFAM annotations are shown (Finn et al. 2006).

Genes recently duplicated in *D. melanogaster lineage*

Cluster	*Distance to tips (min/max)*		*Genes*	*Identifiers*	*Locations*	*Interpro*	*Description*
7572	0.000	0.000	2	CG31953	chr2L:3162514:3163144	IPR010347	Tyrosyl-DNA phosphodiesterase
				CG8825	chr2L:3161912:3160148		
4864	0.025	0.026	2	CG13041	chr3L:7487251:7487191	IPR007614	Retinin-like protein
				CG13060	chr3L:16285256:16285196		
1993	0.000	0.000	2	CG9650	chrX:7082960:7063124	IPR007087	Zinc finger, C2H2-type
				CG9650	chrX:7082351:7080140		
9025	0.036	0.036	2	CG4216	chr3L:5213027:5214311	IPR007087	Zinc finger, C2H2-type
				CG7271	chr3L:18554117:18555398		
6396	0.040	0.087	2	CG10146	chr2R:10262580:10262516	IPR005520	Attacin, N-terminal region
				CG18372	chr2R:10264435:10264371	IPR005521	Attacin, C-terminal region
6133	0.003	0.012	2	CG18859	chrX:2248569:2248002	IPR004117	Olfactory receptor, Drosophila
				CG32825	chrX:19929638:19929071		
6891	0.000	0.007	2	CG18858	chr2L:1962839:1961744	IPR003386	Lecithin:cholesterol acyltransferase
				CG31683	chr2L:1971956:1970861		
4303	0.022	0.049	3	CG32214	chr3L:19432729:19432670	IPR003072	Orphan nuclear receptor, NOR1 type
				CG32208	chr3L:4334532:4334473		
				CG32213	chr3L:19441219:19441160		
6833	0.000	0.036	2	CG14746	chr2R:4221575:4222130	IPR002502	N-acetylmuramoyl-L-alanine amidase, family 2
				CG8577	chr2R:4225288:4225843		
7327	0.036	0.063	2	CG1524	chrX:7779565:7779506	IPR001971	Ribosomal protein S11
				CG1527	chrX:7780734:7780675		
3729	0.026	0.032	2	CG17637	chr3L:20438026:20439538	IPR001958	Tetracycline resistance protein
				CG18281	chr3L:20433712:20435338	IPR011701	Major facilitator superfamily MFS_1
8456	0.000	0.000	2	CG8974-RE	chrX:6596252:6595957	IPR001841	Zinc finger, RING-type
				CG32581	chrX:6592850:6592765		
7734	0.000	0.075	2	CG31352	chr3R:22384108:22374387	IPR001781	LIM, zinc-binding
				CG31332-RD	chr3R:22391694:22387692	IPR003128	Villin headpiece
5742	0.032	0.078	2	CG31624	chr2L:1220304:1220838	IPR001781	LIM, zinc-binding
				CG31988	chr2L:1217989:1218523		
8910	0.034	0.041	2	CG17876	chr2R:12639662:12641144	IPR001687	ATP/GTP-binding site motif A (P-loop)
				CG18730	chr2R:8131652:8133134	IPR006046	Glycoside hydrolase family 13
						IPR006047	Alpha amylase, catalytic region

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
						IPR006048	Alpha amylase, C-terminal all-beta
8653	0.003	0.018	2	CG31809	chr2L:5558043:5557019	IPR001687	ATP/GTP-binding site motif A (P-loop)
				CG31810	chr2L:5562644:5559751	IPR002198	Short-chain dehydrogenase/reductase SDR
						IPR002347	Glucose/ribitol dehydrogenase
						IPR003560	2,3-dihydro-2,3-dihydroxybenzoate dehydrogenase
4033	0.000	0.000	2	CG32495	chrX:17739511:17734765	IPR001680	WD-40 repeat
				CG6835-RC	chrX:17732089:17724086	IPR004887	Eukaryotic glutathione synthase
						IPR005615	Eukaryotic glutathione synthase, ATP-binding
8825	0.000	0.000	2	CG4329	chr2R:2647625:2646183	IPR001680	WD-40 repeat
				CG4329	chr2R:2647808:2645883		
9530	0.000	0.000	2	CG32640	chrX:13012038:13012434	IPR001623	Heat shock protein DnaJ, N-terminal
				CG32641	chrX:13018511:13018907	IPR003095	Heat shock protein DnaJ
8865	0.048	0.062	2	CG4478	chr2L:7529648:7529599	IPR001472	Bipartite nuclear localization signal
				CG4479	chr2L:7532717:7532668	IPR010477	Protein of unknown function DUF1074
1549	0.085	0.115	2	CG33213	chr3R:4160048:4161071	IPR001472	Bipartite nuclear localization signal
				CG33221	chrX:18580258:18581569	IPR007087	Zinc finger, C2H2-type
8654	0.096	0.413	4	CG18157	chrX:8075041:8075632	IPR001472	Bipartite nuclear localization signal
				CG32601	chrX:14212913:14213699	IPR002715	Nascent polypeptide-associated complex NAC
				CG32598	chrX:8073474:8074260		
				CG18313	chrX:14214443:14215232		
7699	0.004	0.004	2	CG31702	chr2L:22060200:22060151	IPR001472	Bipartite nuclear localization signal
				CG31703	chr2L:317319:317270	IPR001687	ATP/GTP-binding site motif A (P-loop)
						IPR004343	Plus-3
2321	0.001	0.026	4	CG32823	chrX:2199321:2196680	IPR001472	Bipartite nuclear localization signal
				CG18000	chrX:2184595:2178159	IPR001680	WD-40 repeat
				CG33499	chrX:2192469:2188820		
				CG33497	chrX:2208028:2204379		
4631	0.017	0.079	2	CG12608	chrX:7046951:7046721	IPR001472	Bipartite nuclear localization signal
				CG9123	chrX:7049462:7049232	IPR001680	WD-40 repeat
5014	0.000	0.012	2	CG12819	chr3R:6163229:6159035	IPR001472	Bipartite nuclear localization signal
				CG12592	chr3R:6165832:6164137		
8297	0.000	0.018	5	CG9579	chrX:2219060:2217340	IPR001464	Annexin
				CG33491	chrX:2211897:2211237	IPR002388	Annexin, type I

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
				CG33498	chrX:2188474:2187818	IPR002389	Annexin, type II
				CG33496	chrX:2204033:2203365	IPR002390	Annexin, type III
				CG33487	chrX:2196334:2195678	IPR002391	Annexin, type IV
						IPR002392	Annexin, type V
						IPR002393	Annexin, type VI
6575	0.021	0.025	2	CG17556	chr3R:15084272:15083406	IPR001440	Tetratricopeptide TPR_1
				CG3678	chr3R:15088493:15087598	IPR013026	Tetratricopeptide region
8938	0.000	0.005	2	CG2947	chrX:3636713:3635573	IPR001440	Tetratricopeptide TPR_1
				CG32789	chrX:3645796:3644656	IPR001472	Bipartite nuclear localization signal
						IPR013026	Tetratricopeptide region
						IPR013105	Tetratricopeptide TPR_2
10950	0.038	0.057	2	CG8821	chr2R:12742491:12740554	IPR001356	Homeobox
				CG8819-RC	chr2R:12738689:12736749	IPR001472	Bipartite nuclear localization signal
6325	0.019	0.161	3	CG17210	chr3R:20838519:20838466	IPR001283	Allergen V5/Tpx-1 related
				CG5106	chr3R:7051424:7051371	IPR003438	Glial cell line-derived neurotrophic factor receptor
				CG5207	chr3R:7067102:7067037		
5815	0.000	0.029	2	CG31034	chr3R:2154744:2155539	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
				CG31362	chr3R:25747963:25748758	IPR001314	Peptidase S1A, chymotrypsin
						IPR008256	Peptidase S1B, glutamyl endopeptidase I
3343	0.924	0.993	2	CG33461	chr2R:11210925:11210475	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
				CG33462	chr2R:11212304:11211915	IPR001314	Peptidase S1A, chymotrypsin
5439	0.000	0.000	2	CG18478	chr2L:15666351:15666251	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
				CG31827	chr2L:15701348:15701248	IPR001314	Peptidase S1A, chymotrypsin
5819	0.000	0.000	2	CG18477	chr2L:15657432:15657202	IPR001254	Peptidase S1 and S6, chymotrypsin/Hap
				CG31780	chr2L:15692429:15692199	IPR001314	Peptidase S1A, chymotrypsin
8621	0.046	0.061	2	CG11941	chrX:19650074:19650548	IPR001232	SKP1 component
				CG12700	chrX:19646737:19647211	IPR001687	ATP/GTP-binding site motif A (P-loop)
7022	0.000	0.005	2	CG30489	chr2R:14129753:14129132	IPR001128	Cytochrome P450
				CG33503	chr2R:14125987:14125366	IPR002397	B-class P450
						IPR002401	E-class P450, group I
						IPR002403	E-class P450, group IV
6080	0.058	0.962	2	CG17176	chr2L:12924758:12921215	IPR001054	Adenylyl cyclase class-3/4/guanylyl cyclase
				CG17174	chr2L:12920365:12916999	IPR002086	Aldehyde dehydrogenase

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
6506	0.076	0.133	6	CG6489	chr3R:8335039:8336962	IPR001023	Heat shock protein Hsp70
				CG31359	chr3R:8331756:8333679	IPR001472	Bipartite nuclear localization signal
				CG5834	chr3R:8328473:8330396	IPR013126	Heat shock protein 70
				CG31449	chr3R:19611774:19613697		
				CG31366	chr3R:20123042:20124968		
				CG18743	chr3R:7784510:7786436		
9375	0.056	0.191	5	CG2885	chrX:10262716:10263301	IPR001019	Guanine nucleotide binding protein (G-protein), alpha subunit
				CG9807	chrX:11756014:11756605	IPR001687	ATP/GTP-binding site motif A (P-loop)
				CG32678	chrX:10434399:10434990	IPR001806	Ras GTPase
				CG32671	chrX:11428209:11428800	IPR002078	Sigma-54 factor, interaction region
				CG32673	chrX:11717414:11718005	IPR006689	ARF/SAR superfamily
2209	0.032	0.105	2	CG1179	chr3L:1190925:1191345	IPR000974	Glycoside hydrolase, family 22, lysozyme
				CG9118	chr3L:22577188:22577608	IPR001916	Glycoside hydrolase, family 22
6726	0.012	0.017	2	CG18278	chr2R:11842565:11841064	IPR000917	Sulfatase
				CG30059	chr2R:11837838:11836337		
6497	0.005	0.005	2	CG11659	chr3R:12302874:12302245	IPR000873	AMP-dependent synthetase and ligase
				CG6300	chr3R:12306932:12306300	IPR001589	Actin-binding, actinin-type
8422	0.029	0.070	3	CG12405	chr2R:5937066:5936427	IPR000866	Alkyl hydroperoxide reductase/ Thiol specific antioxidant/ Mal allergen
				CG12896	chr2R:14827875:14827236		
				CG11765	chr2R:5931339:5930700		
1665	0.025	0.069	2	CG12179	chrX:17747191:17745276	IPR000694	Proline-rich region
				CG12184	chrX:17751412:17749414	IPR001472	Bipartite nuclear localization signal
7592	0.000	0.000	3	CG32500	chrX:791496:790113	IPR000629	ATP-dependent helicase, DEAD-box
				CG32857	chrX:797588:796221	IPR001075	Nitrogen-fixing NifU, C-terminal
				CG33502	chrX:803680:802313		
5058	0.006	0.017	2	CG32713	chrX:13921594:13922146	IPR000626	Ubiquitin
				CG33223	chrX:13912842:13913394		
5937	0.000	0.009	3	CG10530	chr3L:17662917:17662856	IPR000618	Insect cuticle protein
				CG18779	chr3L:17667201:17667140	IPR001419	HMW glutenin
				CG10534	chr3L:17664598:17664537		
4787	0.000	0.175	2	CG11650	chr2R:16823847:16823783	IPR000618	Insect cuticle protein
				CG8697	chr2R:16820513:16820451		
5784	0.000	0.030	2	CG1252	chr3R:25376600:25376524	IPR000618	Insect cuticle protein

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
				CG2360	chr3R:2529935:2529880		
5962	0.000	0.000	2	CG18773	chr3L:6120668:6120980	IPR000618	Insect cuticle protein
				CG32400	chr3L:6123539:6123851		
5823	0.000	0.000	2	CG18787	chr2L:10408738:10408683	IPR000571	Zinc finger, CCCH-type
				CG18789	chr2L:10412649:10412594		
6715	0.000	0.021	2	CG18259	chrX:3608596:3607158	IPR000504	RNA-binding region RNP-1 (RNA recognition motif)
				CG6961	chrX:3603911:3602473	IPR001472	Bipartite nuclear localization signal
3670	0.005	0.005	3	CG32820	chrX:800987:799091	IPR000435	Tektin
				CG32819	chrX:794895:794024		
				CG17450	chrX:807082:806208		
9481	0.000	0.000	2	CG18495	chr2R:17427090:17426563	IPR000426	Proteasome alpha-subunit
				CG30382	chr2R:17423980:17423453	IPR001353	20S proteasome, A and B subunits
1545	0.024	0.091	2	CG32745	chrX:16098715:16098666	IPR000380	DNA topoisomerase I
				CG3458	chrX:16094264:16091514	IPR001412	Aminoacyl-tRNA synthetase, class I
						IPR001472	Bipartite nuclear localization signal
						IPR006171	TOPRIM
7579	0.010	0.021	2	CG32164	chr3L:16572686:16568536	IPR000357	HEAT
				CG32165	chr3L:16564999:16560858	IPR001494	Importin-beta, N-terminal
3310	0.049	0.185	2	CG18412	chrX:20240674:20232218	IPR000354	Involucrin repeat
				CG3895	chrX:20250617:20246548	IPR001660	Sterile alpha motif SAM
						IPR001778	Pollen allergen Poa pIX/Phl pVI, C-terminal
						IPR011510	Sterile alpha motif homology 2
9341	0.000	0.000	2	CG18816	chr2R:2511295:2510994	IPR000301	CD9/CD37/CD63 antigen
				CG30160	chr2R:2517583:2517282		
6922	0.010	0.036	2	CG6289	chr3L:20277750:20276901	IPR000215	Proteinase inhibitor I4, serpin
				CG6663	chr3L:20255891:20255042		
6349	0.000	0.000	2	CG7216	chr2L:14666477:14666359	IPR000175	Sodium:neurotransmitter symporter
				CG31904	chr2L:14672628:14671497		
8028	0.067	0.068	2	CG7045	chr3R:18317648:18318026	IPR000135	High mobility group proteins HMG1 and HMG2
				CG7046	chr3R:18318984:18319383	IPR000910	HMG1/2 (high mobility group) box
6907	0.000	0.000	2	CG4220	chr2L:8015835:8000240	IPR000104	Antifreeze protein, type I
				CG4220-RC	chr2L:8015895:7999467	IPR006162	Phosphopantetheine attachment site
						IPR007087	Zinc finger, C2H2-type

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
2262	0.000	0.074	2	CG10102	chr2R:10890134:10889985	IPR000104	Antifreeze protein, type I
				CG12505	chr2R:10891722:10892484	IPR005162	Retrotransposon gag protein
1969	0.000	0.165	2	CG4575	chrX:6796016:6795797	IPR000104	Antifreeze protein, type I
				CG7952	chrX:19939509:19939434	IPR001472	Bipartite nuclear localization signal
						IPR004827	Basic-leucine zipper (bZIP) transcription factor
						IPR011700	Basic leucine zipper
5912	0.000	0.000	2	CG8118-RC	chr2R:9574675:9554330	IPR000104	Antifreeze protein, type I
				CG8118	chr2R:9573168:9527088	IPR000817	Prion protein
						IPR001419	HMW glutenin
						IPR001472	Bipartite nuclear localization signal
						IPR002952	Eggshell protein
5067	0.069	0.075	2	CG17438	chrX:13118766:13119882	IPR000104	Antifreeze protein, type I
				CG32703	chrX:13116157:13113988	IPR000719	Protein kinase
						IPR001472	Bipartite nuclear localization signal
						IPR001859	Ribosomal protein P2
						IPR003527	MAP kinase
						IPR008271	Serine/threonine protein kinase, active site
6983	0.052	0.073	2	CG18284	chr2L:10645369:10644150	IPR000073	Alpha/beta hydrolase fold-1
				CG31872	chr2L:10642997:10642093	IPR000379	Esterase/lipase/thioesterase
						IPR006693	AB-hydrolase associated lipase region
						IPR008262	Lipase, active site
2172	0.000	0.000	2	CG31292	chr3R:11732308:11732249		
				CG3303	chr3R:11735365:11734441		
399	0.016	0.033	2	CG12487	chr3L:8860570:8860804		
				CG13465	chr3L:8862346:8862580		
6188	0.016	0.067	2	CG15797	chrX:9101663:9101598		
				CG15910	chrX:9105917:9105716		
6613	0.000	0.000	2	CG33470	chr2R:8928632:8927728		
				CG18279	chr2R:8923905:8923001		
6817	0.000	0.000	2	CG18542	chr3R:5511968:5511693		
				CG32939	chr3R:5519288:5519013		
7435	0.000	0.000	2	CG18064	chr3L:5337040:5337193		
				CG32197	chr3L:5334274:5334427		
7528	0.000	0.000	2	CG31131	chr3R:4116178:4117288		

Cluster	Distance to tips (min/max)		Genes	Identifiers	Locations	Interpro	Description
				CG31253	chr3R:4118797:4119934		
7764	0.000	0.000	2	CG31825	chr2L:6753287:6753815		
				CG33311	chr2L:6719690:6720218		
7785	0.000	0.005	2	CG31865	chr2L:11997767:11997703		
				CG31866	chr2L:11993856:11993792		
8224	0.000	0.012	2	CG3176	chrX:339339:339255		
				CG32817	chrX:336851:336767		
8539	0.028	0.029	2	CG10476	chr2R:5966703:5967369		
				CG18606	chr2R:14802884:14803511		
8937	0.011	0.026	3	CG32783	chrX:3792885:3790610		
				CG32786	chrX:3788689:3786414		
				CG2941	chrX:3784451:3782176		
8950	0.000	0.000	2	CG31447	chr3R:12815240:12814690		
				CG31274	chr3R:12819452:12819388		

Table S5: Genes recently duplicated in the D. melanogaster lineage. If there are several lineage specific duplications within one family, all members are shown. The distance to tip refers to the minimum and maximum distance (in d_S) to any tip from the earliest duplication event. The last two columns show Interpro annotations (Mulder et al. 2005).

Benchmarking against *D. pseudoobscura* gene set

Benchmarking tests were performed against the predicted gene set from release 1.04 of *D. pseudoobscura* (Richards et al. 2005). Briefly, this set was derived computationally by applying TBLASTN, Genewise (Birney et al. 2004), Genscan (Burge and Karlin 1997) and Twinscan (Korf et al. 2001). Genes were accepted if they were detected by TBLASTN and at least one of the other methods and template and prediction were reciprocal best matches.

After removing genes on unplaced contigs the reference gene set contained 9,946 genes with 39,580 exons. Sensitivity and selectivity on the nucleotide, exon and gene level were calculated as in Burset and Guigo 1996 (Burset and Guigo 1996) with a relaxed criterion for matching exons: two overlapping exons between the sets were declared to match if they shared at least one exon boundary within nine nucleotides. Variant exons and transcripts for a single gene were compared individually to the reference, but counted as a single entry.

Representative template transcripts were compared against the unmasked genome using TBLASTN and an E-Value upper threshold of 10^{-5}.

The pipeline predicts 83% of transcripts of the published prediction set from *D. pseudoobscura r1.04*, 72% of these predictions are identical. Only 2% of transcripts in the reference set do not overlap with any transcript in the predicted set. The amount of overlap between the two gene prediction sets is expected given that both are derived from different computational methods.

5% of exons in the reference set are missing. Of the missing exons, 28% are in missed genes and 55% are terminal exons at either end of a gene. In most cases these extra exons are not part of the template used for the prediction and are likely spurious exons from ab-initio predictions.

We compared the performance of exonerate in the initial scanning step and compared it to TBLASTN using the annotations from *D. pseudoobscura* release 1.04. Exonerate needs to be run at low levels of nucleotide specificity in order to obtain a level of sensitivity comparable to TBLASTN. The gene model in exonerate allows it to predict exon boundaries in scanning mode with high sensitivity (92%). TBLASTN lacks such a model and only predicts 60% of exons correct. Again, Exonerate's selectivity is low. Exonerate can replace TBLASTN as a method to locate exons and genes, but its result set needs to be filtered for spurious matches.

Benchmarking showed Exonerate to be 1.7 times slower than TBLASTN, but execution speed is highly dependent on parameter choices and can be increased using the default parameters.

	Nucleotides		*Exons*				*Genes*			
Category	*sp*	*sn*	*sp*	*sn*	*me*	*we*	*sp*	*sn*	*mg*	*wg*
Exonerate50	0.21	0.94	0.06	0.95	0.02	0.87	-	-	-	-
Exonerate60	0.64	0.88	0.41	0.93	0.06	0.59	-	-	-	-
Exonerate70	0.73	0.87	0.64	0.92	0.06	0.37	-	-	-	-
Exonerate80	0.74	0.87	0.66	0.91	0.07	0.34	-	-	-	-
Tblastn	0.67	0.94	0.69	0.80	0.10	0.35	-	-	-	-
Pipeline (all genes)	0.73	0.95	0.69	0.92	0.05	0.30	0.48	0.83	0.02	0.52
Pipeline (CG,PG,SG)	0.78	0.83	0.75	0.81	0.17	0.25	0.64	0.75	0.14	0.36

Table T6: Gene prediction quality indices for predictions in D. pseudoobscura.

References

Alexeyenko A., Tamas I., Liu G., Sonnhammer E.L.L. 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**: e9-15.

Birney E., Clamp M., Durbin R. (2004) Genewise and genomewise. *Genome Res* **14**: 988-995.

Boyle E.I., Weng S., Gollub J., Jin H., Botstein D., Cherry J.M., Sherlock G. 2004. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**: 3710-3715.

Burge C., Karlin S. (1997) Prediction of complete gene structures in human genomic dna. *J Mol Biol* **268**: 78-94.

Burset M., Guigo R. (1996) Evaluation of gene structure prediction programs. *Genomics* **34**: 353-367.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540-552.

Chen R., Meisel R.P. et al. (2005) Comparative genome sequencing of drosophila pseudoobscura: chromosomal, gene, and cis-element evolution. *Genome Res* **15**: 1-18.

Finn R.D., Mistry J., Schuster-Bockler B., Griffiths-Jones S., Hollich V., Lassmann T., Moxon S., Marshall M., Khanna A., Durbin R. et al. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res* **34**: D247-51.

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

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**: 368-376.

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

Fitch W.M. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* **19**: 99-113.

Korf I., Flicek P., Duan D., Brent M.R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**: S140-8.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**: 211-218.

Mulder N.J., Apweiler R., Attwood T.K., Bairoch A., Bateman A., Binns D., Bradley P., Bork P., Bucher P., Cerutti L. et al. (2005) Interpro, progress and status in 2005. *Nucleic Acids Res* **33**: D201-5.

Remm M., Storm C.E., Sonnhammer E.L. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol* **314**: 1041-1052.

Richards S., Liu Y., Bettencourt B.R., Hradecky P., Letovsky S., Nielsen R., Thornton K., Hubisz M.J., Nielsen R., Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929-936.

Tatusov R.L., Koonin E.V., Lipman D.J. 1997. A genomic perspective on protein families. *Science* **278**: 631-637.

Suzuki Y., Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**: 1315-1328.

Suzuki Y., Nei M. 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. *Mol Biol Evol* **21**: 914-921.

Wong W.S.W., Yang Z., Goldman N., Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041-1051.

Wootton J.C. 1994. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* **18**: 269-285.

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

Zmasek C.M., Eddy S.R. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* **3**: 14.

Zuckerkandl E., Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In: *Horizons in biochemistry* (ed. M. Kasha, B. Pullman), pp. 189–225. Academic Press, New York.