Supplemental material

Age-dependent chromosomal distribution of male-biased genes in *Drosophila*

Yong E. Zhang, Maria D. Vibranovski, Benjamin H. Krinsky and Manyuan Long

Correspondence and requests for materials should be addressed to M.L. (mlong@uchicago.edu).

This file includes:
Materials and methods
References
Legends for supplemental table 1 to 9
Supplemental figures 1 to 4

Other Supplemental material for this manuscript includes the following:

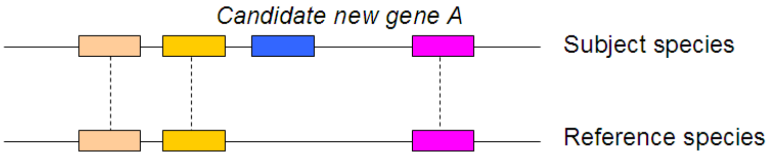
Supplemental tables 1 to 9

Materials and methods

We downloaded FlyBase release V5.3 (September, 2007) from UCSC. We used MySQL V5.0.45 to organize the data, BioPerl (Stajich *et al.* 2002) and BioEnsembl (Stabenau *et al.* 2004) to fold the pipeline, and R V2.7.1 (Team 2007) to perform several statistical tests. Notably, for the two by two contingency table test, we prefer the Chi-square test. However, if one cell had not more than five samples, the later was used.

1 Dating *D. melanogaster* protein-coding genes on the *Drosophila* genus phylogenetic tree

For a *D. melanogaster* gene A, we deduced its origination time by inferring its ortholog distribution in different species. In other words, we used gene A presence or absence in an outgroup species such as *D. simulans*. More specifically, we addressed this question by investigating whether gene A exists in the syntenic chain between *D. melanogaster* and another species (namely the subject and reference species, respectively).

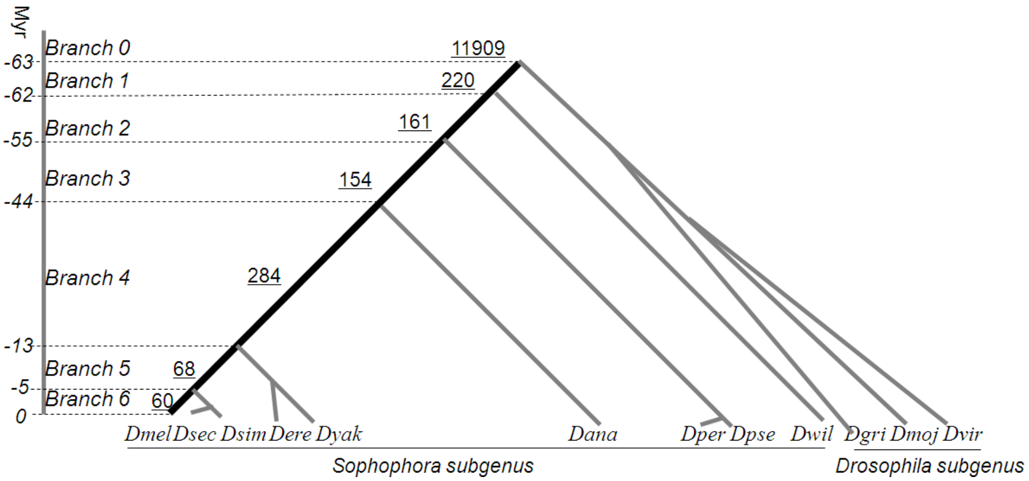


The two lines indicate a syntenic chain between species. Boxes marked in the same color indicate orthologous regions. The blue box represents a candidate new gene because it is absent in the reference species. Since the syntenic chain represents a linear combination of all conserved anchors (such as exons, regulatory elements and other sequences under constraint), this method is more effective in inferring the true orthology relative to traditional gene-based comparison methods.

We developed this synteny-based strategy by following Zhou *et al.* (Zhou *et al.* 2008). They mapped annotated *D. melanogaster* genes to *D. simulans* and *D. yakuba* using BLAST and then constructed syntenic chains requiring at least two consecutive genes. We further improved this method by using a genome-alignment based strategy. Specifically, we began with a netted genomic alignment provided by UCSC, which is based on a complicated post-processing genome alignment result generated by BLASTZ (Schwartz *et al.* 2003; Kuhn *et al.* 2007). Unlike gene-based synteny, this mapping profited from all alignment sequences, both genes and intergenic elements, and is thus more robust and reliable. Moreover, due to its higher sensitivity compared to BLAST, the BLASTZ-based pipeline is capable of building syntenic mapping between two evolutionarily distant species, like *D. melanogaster* and *D. virilis*. In addition, the UCSC pipeline does not depend on some parameters such as the number of genes used to define syntenic mappings, and therefore is tolerant to single gene translocations.

However, a disadvantage of the UCSC netted track and of any other similarity-based strategy is that several regions in *D. melanogaster* might map to the same region of the reference species. In other words, if duplication has occurred in *D. melanogaster*, two different segments will present the same best hit in the reference genome, since the UCSC pipeline identifies the best hit, rather than the reciprocal best hit. In order to solve this problem, for exons of gene A in *D. melanogaster,* we first scanned the netted alignment table for a corresponding region mapped in the reference species. Then, we checked if the mapped region in the reference species had gene A as the reciprocal best hit in *D. melanogaster*. Therefore, if gene A overlapped with the reciprocal best genomic mapping, we assigned “gene presence” in the reference genome.

Given presence and absence information, we dated gene A on the *Drosophila* genus phylogenetic tree containing 12 species.



For instance, branch 0 indicates the *Drosophila* genus and genes assigned to this branch are therefore shared by all 12 species. Branch 1 corresponds to the Sophophora-lineage and so forth. To be conservative, we attempted to use information from two outgroup levels when assigning gene presence an absence. For example, gene A was considered to be a *D. melanogaster*-specific gene only if it was not present in *D. simulans- D. sechellia* (first outgroup) and in *D. yakuba-D. erecta* (second outgroup).

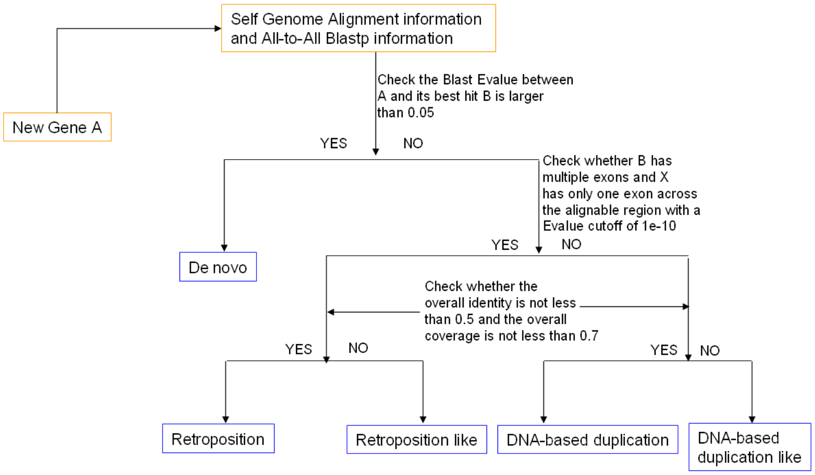
FlyBase protein annotation v5.3 was downloaded from UCSC (Kuhn *et al.* 2007) and processed by this pipeline. We estimated the performance of our method by counting the proportion of new genes in *D. melanogaster* assigned to branches consistent with previous studies. We compared our dataset to three different studies including retrogenes and DNA-based duplicated genes: i) Bai and collaborators analyzed the gene-based synteny, reconstructing the gene tree to date 188 genes (97 retrogenes and 91 parental genes) (Bai *et al.* 2007); ii) Yang and collaborators dated five genes by manually checking BLAST hits across different species (Yang *et al.* 2008); iii) Zhou and collaborators dated 73 *D. melanogaster* specific genes by inferring the micro-synteny of at least two genes (Zhou *et al.* 2008). Comparison results between our dataset and these compiled datasets are shown below:

97 Retrogenes			91 Parental Genes			5 multi-exon Genes	73 D. melanogaster specific genes		
Consistent	Conflict	Absent	Consistent	Conflict	Absent	Consistent	Consistent	Conflict	Absent
84	6 (5*)	7	88	2	1	5	51	9 (7*)	13 (3 removed from FlyBase);
									Specificity:93% Sensitivity: 93%

Our pipeline was able to date 93%of all 266 non-overlapping cases (a measure of sensitivity), and 93% are consistent with our dataset (a measure of specificity). For those cases in conflict with previous studies, we usually assigned genes to older branches due to the conservative nature of our method. In addition, "*" stands for the number of genes assigned to older branches in our dataset.

2 Inference of gene origination mechanism

Following the strategy of previous studies (Levine *et al.* 2006; Bai *et al.* 2007; Zhou *et al.* 2008), we classified all young genes into three categories: DNA-based duplication, retroposition and *de novo* gene origination. Briefly, as the following Figure shows, for a young gene A in *D. melanogaster*, we investigated whether there is one paralog B based on all against all protein alignments across 12 species. For the negative case, gene A emerged through a *de novo* origination. For the positive case, we checked whether all introns of gene B have been lost in their alignable gene A region. Young genes that emerged via a retroposition mechanism do not contain introns, whereas DNA-based duplication generates multiple-exon new genes. We also checked the alignment quality between A and B and only used those high quality duplicates (identity>=0.5 and coverage>=0.7) in the analysis.


Self Genome Alignment information and All-to-All Blastp information
New Gene A
Check the Blast Evalue between A and its best hit B is larger than 0.05
YES
NO
De novo
Check whether B has multiple exons and X has only one exon across the alignable region with a Evalue cutoff of 1e-10
YES
NO
Check whether the overall identity is not less than 0.5 and the overall coverage is not less than 0.7
YES
NO
YES
NO
Retroposition
Retroposition like
DNA-based duplication
DNA-based duplication like

We made an improvement compared to previous efforts regarding the assignment of parental and child genes (Zhou *et al.* 2008). Briefly, we used the information of genomic alignment, profiting from neighboring regions. As is done for gene dating, genome alignment (self genome alignment) information was used to infer the most likely parent-child gene mappings and then the probable origination mechanism. Specifically, we implemented the ChainSelf pipeline of UCSC (Kent *et al.* 2003; Schwartz *et al.* 2003) and generated the alignment between the *D. melanogaster* genome (UCSC Release, dm3) and itself. After a series of processing steps like Chainning and Netting, we found the best hit for any genomic region. Compared to traditional gene-based methods, this pipeline considered the downstream, upstream, exonic and intronic sequences and is capable of identifying the most probable parental genes. Moreover, it automatically identifies the duplication block borders without additional procedures. Guided by this Chainself information, we further checked all-against-all BLASTP information to retrieve alignment information such as identity, coverage and Evalue (Levine *et al.* 2006; Bai *et al.* 2007; Zhou *et al.* 2008). For genes accidentally not covered by ChainSelf mappings, we followed the traditional pipeline, *i.e.*, beginning directly with the all-against-all BLASTP information.

After the inference of parent-child relationship, we divided duplication including DNA-based duplication and retroposition into three sub categories. If the parental gene and child gene were encoded on different chromosomes, such a case was defined as a “movement duplicate”. If there was at least one gene localized between the duplication block (inferred from ChainSelf mappings), it was defined as “dispersed duplicate”; otherwise, it was defined as “tandem duplicate”. Moreover, since we assigned new and old protein-coding genes to phylogenetic branches, we filtered the parent-child matches using relative origination timing, *i.e.*, only parent-child mapping with parent genes assigned to an older branch were retained. Such filtered mapping was used in Table 1 and Figure 5 (main text).

Generally, our classification result is highly consistent with previous work. For example, out of 25 young retrogenes identified by Bai *et al* (Bai *et al.* 2007), we classified 23 (92%) entries as retrogenes. For those two cases in conflict, we identified them as DNA level duplicates since our pipeline found more similar single exon paralogs.

3 Expression profiling of *D. melanogaster* protein-coding genes

Since gene annotation and probe annotation are constantly changing, a refined probe mapping file is essential for the interpretation of microarray experiments (Dai *et al.* 2005). This means that the original Affymetrix probe set definitions might be inaccurate, and the results from previous GeneChip analyses may need to be reviewed (Dai *et al.* 2005). Thus, we re-analyzed the high-quality microarray datasets of FlyAtlas (Chintapalli *et al.* 2007) as the basic source to identify sex-biased genes, which covers the following 12 samples as of October 2008: hind gut, mid gut, accessory gland, brain, crop, larval fat, head, larvae tubules, ovary, testes, whole fly and salivary gland.

Specifically, we used the customized probe mapping file (gene-level mapping, Drosophila2_Dm_ENTREZG on June 2008), which filtered low-quality probes, like those mapping to the removed gene models, or those mapping to multiple genomic locations (Dai *et al.* 2005). Then, we performed all the subsequent analyses based on Bioconductor software (Gentleman *et al.* 2004). We used the GCRMA package (V2.12.1) to adjust the background intensity, normalize and summarize the expression value and MAS5 function of the Affy package (V1.18.2) to call the presence or absence of each gene. After that, we implemented the linear models of the Limma package (V2.14.6) to assess whether genes show differential expression between testis and ovary (Smyth 2004). Limma can analyze comparisons between many RNA targets simultaneously. It can also make the analyses stable even for experiments with a small number of arrays by borrowing information across genes. More importantly, with an empirical Bayes fitting strategy, it does not necessarily specify an arbitrary value like two-fold as a cutoff for differential expression. In our analysis, we defined genes with a False Discovery Rate (FDR) of 0.05 as a threshold for candidate sex-biased genes. Genes with high expression in testis, which were also called as “present” in testis across all four duplicates by MAS5, were defined as male-biased genes. The genes with the opposite pattern were defined as female-biased genes. The remaining genes were classified as unbiased genes.

4 Evolutionary analysis

For within-species divergence analysis, that is, to compare parental genes and children genes, we used the bl2seq program of the BLAST package (Altschul *et al.* 1997) to construct protein-level alignments and the PAL2NAL script (V12) to convert protein-level alignments to codon-level alignments (Suyama *et al.* 2006). We used the DNAStatistics module of BioPerl to calculate *dN* and *dS* (Stajich *et al.* 2002) for Figure 5.

In order to integrate both between-species divergence data and within-species polymorphism data, we needed to infer the sequence of one outgroup, *i.e.*, *D. simulans*. While dating all *D. melanogaster* protein-coding genes, we already generated the genome-level ortholog blocks. Using gene sequences of *D. melanogaster* as the reference, we predicted the corresponding protein sequence and coding sequence in *D. simulans* using Genewise (V2-2-0) (Birney *et al.* 2004). The *Drosophila* Population Genomics Project (DPGP, http://www.dpgp.org/) released 7MB of re-sequencing data from up to 50 *D. melanogaster* strains, which consists of about 3MB of chromosome 2L data and 3MB of chromosome X sequences.

Concurrently, DPGP also released the whole-genome population data based on the syntenic alignment of six strains of *D. simulans* (Begun *et al.* 2007). Unfortunately, both datasets are anchored to the old *D. melanogaster* assembly (UCSC dm2). In order to deal with possible sequence updates between different assemblies, we developed a pipeline.

Specifically, we mapped the 7MB data to the latest genomic coordinates of dm3 using the liftOver tool of UCSC. Then, we extracted all non *D. melanogaster* specific gene sequences annotated within this range and aligned them with 50 strains using BLAT (Kent 2002). Gene models generated by BLAT were further refined using Genewise (Birney *et al*. 2004). Finally, proteins from 50 strains together with the orthologs of *D. simulans* were aligned using the linsi program of the MAFFT package (v6.603b) (Katoh and Toh 2008), which might be the most accurate sequence alignment program. The alignment was further filtered: entries with "N" or sequencing gaps contributing to more than half of the alignable region were purged. Then, similarly, the protein-level alignment was converted back to a codon-level alignment using PAL2NAL. Given this alignment, we finally counted the number of synonymous and non-synonymous substitutions using the PGEToolbox (Cai 2008). The "rmcodongaps" function was used to remove codons with gaps and "mktestcmd" was used to calculate statistics.

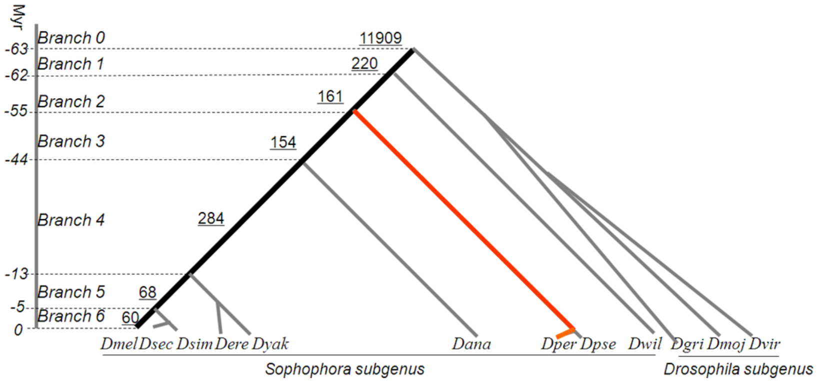
The pipeline used to handle the six strains of *D. simulans* data was almost identical except for the following one point: we mapped all non *D. melanogaster* specific gene models to six genomes and required the chromosomal location of the best hit to overlap with the location of this gene in dm3 or its flanking 500KB region. Considering that the sequence difference between dm2 (*D. simulans* is anchored to Dm2) and dm3 is not so large, 500KB tolerance should have been able to ensure *D. melanogaster* genes mapped to their real orthologous regions. Since *D. simulans* polymorphism data only cover six strains, it is possible that some alignments included only one or two strains after the alignment purge. So, we only used the alignment including at least four strains and at least one polymorphism for the final statistics

(Table S4B).

Based on the number of mutations, we implemented a multi-locus MK-test implemented in DoFE (Distribution of Fitness Effects) (Bierne and Eyre-Walker 2004), which provided the maximum-likelihood based estimation of α, the proportion of substitutions fixed by adaptive mutation. We used the LikeLihood-Ratio test (LLR) to measure whether one group of genes such as male-biased genes has different α compared to other groups of genes such as female-biased genes. Specifically, we first ran the analysis on the total set of genes together and recorded the log likelihood (namely as LLt). Next, we performed the analysis on each group of genes separately and recorded the log likelihoods (namely, LL1 and LL2). We then performed a standard likelihood ratio test, namely 2((LL1+LL2) - LLt) being chi-square distributed with one degree of freedom under the null hypothesis that alpha was the same in the two groups of genes.

5 Dating *D. pseudoobscura* protein-coding genes on the *Drosophila* genus phylogenetic tree

We downloaded FlyBase annotation V2.2 information regarding gene structure and orthology mapping, which consisted of numerous updates compared to releases of 2007 (Clark *et al.* 2007). As shown below, we dated genes that originated in the obscura group branch and in *D. pseudoosbcura* specific branch (marked in orange).



Since UCSC only provides *D. melanogaster* centric synteny tracks and BLASTZ-based genome alignment is time-consuming, we implemented a lightweight dating pipeline in *D. pseudoobscura*. Specifically, for genes with an ortholog in *D.*

melanogaster, we directly used the assignment information in *D. melanogaster*. As for the remaining genes, we used the orthology annotation of FlyBase. Herein, for one *D. pseudoobscura* gene A, if it had no annotated orthology in the reference species, we checked whether one reciprocal best hit existed. In such a way, we recovered several “Presence” rather than “Absence” cases.

Given this information, we identified new young genes that originated after the divergence of *D. melanogaster* and *D. pseudoobscura*.

6 Expression profiling of *D. pseudoobscura* protein-coding genes

Due to the limit of public raw microarray data for *D. pseudoobscura*, we directly used a pre-computed dataset of sex-biased genes (Sturgill *et al.* 2007), which was generated from a whole-body comparison of two sexes on the NimbleGen double-channel platform. Using GleanR ID as a reference (Clark *et al.* 2007), we mapped this dataset to *D. pseudoobscura* gene annotation V2.2. Notably, although the original authors tried to cover as many gene models as possible, this dataset is biased toward old genes. For the old genes shared by all 12 species, the coverage of probes is as high as 91%. By contrast, the coverage for *D. pseudoobscura* specific genes is as low as 15%. Although such a decrease of sample size affects our statistical power, this dataset was sufficient for most of all the analyses done.

Thus, it is noteworthy that our conclusions obtained for both *D. melanogaster* and *D. pseudoobscura* were robust regardless the Microarray platform analyzed (Affymetrix or NimbleGen platform).

7 Analysis of microRNA (miRNA)

151 miRNAs in *D. melanogaster* with genomic coordinates were collected from miRBase V10.0 (Griffiths-Jones *et al.* 2008). Mature miRNA sequences were defined as regions associated with the most abundant reads after mapping 12 small RNA deep sequencing samples (Ruby *et al.* 2007; Czech *et al.* 2008) onto the genome. Four miRNA entries including dme-mir-280, dme-mir-287, dme-mir-288 and dme-mir-289 were excluded from further analyses since they have no support of any

reads across all samples.

The orthologous sequences in the other 11 *Drosophila* species (Consortium 2007) were parsed out from the whole genome sequence alignments containing 12 *Drosophila* species provided by UCSC (Kuhn *et al.* 2007). The multiple-species alignment of miRNA's "seed" region (2-8nt of mature sequence) was constructed using those orthologous sequences. Since miRNA is too short and the alignment between remote species tends to be unreliable, we limited the phylogenetic analysis to the following species: *D. melanogaster, D. simulans* and *D. sechellia*, *D. yakuba* and *D. erecta*, *D. ananassae*, and *D. persimilis* or *D. pseudoobscura*. We defined the new miRNA as miRNA absent both in *D. pseudoobscura* and *D. persimilis*: the putative new miRNA had to have no orthologous locus in these species, or the seed region included mismatches in both *D. ananassae* and *D. pseudoobscura* or *D. persimilis*. According to this definition, out of 147 miRNAs, we found 29 new miRNAs. Notably, we do not think this number indicates that 20% (29/147) of *D. melanogaster* miRNAs originated in less than 40 million years. As reported before, it seems that hairpin structures are generated at a high frequency in *Drosophila* genomes (Lu *et al.* 2008). Thus, it is likely that some of these young miRNAs might not be functional, but instead are transcriptional noise. Nevertheless, such a miRNA dataset provides an opportunity to investigate the origination of this type of noncoding gene.

We used two distinct strategies to call sex-biased expression. First, we implemented a likelihood ratio statistical scoring framework (Herbert *et al.* 2008) to investigate whether one miRNA of interest has differential expression between testis and ovary. We followed (Herbert *et al.* 2008) and considered a FDR-controlled *p* (q-value) of 0.01 as significant. The result is tabulated in Table S7. Second, we identified expression bias as sex-specific by defining testis-limited genes as those with at least 50 reads in testis but no reads in ovary. The result (Table S8) reproduces what we know for protein-coding genes (Figure S1D), which again suggests miRNAs follow a similar pattern.

8 Mechanism analysis regarding how young male-biased genes move out of X

chromosome

As stated in the main text, various mechanisms contribute to the depletion of X-linked male-biased new genes, like inter-chromosomal movement or selective gene loss (Betran *et al.* 2002; Sturgill *et al.* 2007; Vibranovski *et al.* 2009). We analyzed three aspects: whether young genes of *D. melanogaster* have different chromosomal linkage in other species, whether inter-chromosomal duplication already occurred for these new genes, and whether *D. melanogaster* young genes have already become pseudogenes in other species. One reason that we did not concentrate on non-melanogaster gene models is that the automatic annotation in the 11 genomes is actually error-prone, especially for those species and lineage specific models. For example, we can find some genes which are absent from *D. melanogaster* but present in the other species. However, it is very possible that these gene models are just annotation error. Another reason is high-quality expression data only exists for *D. melanogaster*.

First, we downloaded orthology mapping from FlyBase and filtered this mapping based on our branch alignment. Only the concordant mapping was retained for subsequent analysis. Taking advantage of recently mapped chromosomal linkage between *D. melanogaster* and the other 11 Drosophila genomes (Schaeffer *et al.* 2008), we then screened out genes linked to different chromosomes in different species. We discarded those ambiguous cases where we cannot ensure the translocation direction. For example, a gene X is located on 2L in *D. melanogaster*, while it is on 3R in *D. simulans* and *D. sechellia*. Without any other groups, it is impossible to figure out the original movement pattern. We also filtered those species-specific translocations, which might only indicate occasional assembly error. Finally, we identified three reliable cases, one of which is from X chromosome to autosomes (Table S9).

Secondly, based on our parental gene and children gene inference, we identified inter-chromosomal duplication events (Table S9). In this case, the parental gene is one young gene (assigned to branches 1 to 5), while the child gene is even younger. Here, we only found two cases and one of them is an out-of-X duplicate.

Finally, we checked how many *D. melanogaster* young genes have become pseudogenes in other species. Specifically, we implemented Genewise (Birney *et al.* 2004) and conceptually translated orthologous loci previously inferred by genomic alignment using *D. melanogaster* proteins as the reference. From the raw output of Genewise, we counted the pseudogenization events (frameshifts or premature stop codons). Only a locus with at least one such event and without FlyBase annotation was marked as a pseudogene. Considering the small life span of pseudogenes in *D. melanogaster* (Petrov and Hartl 1998; Harrison *et al.* 2003), we only investigated the recent pseudogenization events by comparing the *D. simulans*, *D. sechellia* and *D. melanogaster* complex. We identified two cases in which the gene is already pseudogenic in either *D. simulans* or *D. sechellia* (Table S9). One of them is X-linked.

9 Data analysis note

Our main dataset consists of 947 evolutionarily young genes (that have originated since *Drosophila* and *Sophophora* subgenus split). However, in order to perform some accurate analyses for specific problems, we used two different filters, generating smaller datasets. Although we have already briefly explained inside table or figure legends (such as Table S2), here we present more details.

As mentioned in the main text, we classified genes into 102 (11%) retrogenes, 741 (78%) DNA-level duplicates, and 104 (12%) de novo genes. Notably, for some duplicates, we could not ensure that the parental gene inference was reliable. First, alignment coverage or identity might be very low. Second, the parental gene can be assigned to the same branch as the child gene or an evolutionarily younger one. After requiring both age relationship and sequence alignment quality, we screened out only 38 DNA-level inter-chromosomal duplicates (Table S2B).

The other filter was applied to require probe uniqueness. We used the customized probe mapping file (gene-level mapping, Drosophila2_Dm_ENTREZG on June 2008), which filtered low-quality probes, like those mapping to retired gene models, or those mapping to multiple genomic locations (Dai *et al*. 2005). Since child genes are often quite similar to parental genes, this filter is necessary and therefore we retained 716 genes with unique probes out of all 845 non-retroposed entries.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17): 3390.

Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome Biol.* **8**(1): R11.

Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, Nista PM, Jones CD, Kern AD, Dewey CN et al. 2007. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biol.* **5**(11): e310.

Betran E, Thornton K, Long M. 2002. Retroposed New Genes Out of the X in Drosophila. *Genome Research*: 6049.

Bierne N, Eyre-Walker A. 2004. The Genomic Rate of Adaptive Amino Acid Substitution in Drosophila. *Mol. Biol. Evol.* **21**(7): 1350-1360.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res.* **14**(5): 988.

Cai JJ. 2008. PGEToolbox: A Matlab toolbox for population genetics and evolution. *J. Hered.* **99**(4): 438-440.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nat. Genet.* **39**(6): 715.

Clark AG Eisen MB Smith DR Bergman CM Oliver B Markow TA Kaufman TC Kellis M Gelbart W Iyer VN et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167): 203-218.

Consortium DCGSaA. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450**(7167): 203-218.

Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, Perrimon N, Kellis M, Wohlschlegel JA, Sachidanandam R. 2008. An endogenous small interfering RNA pathway in Drosophila. *Nature* **453**(7196): 798.

Dai M, Wang P, Boyd AD, Kostov G, Athey B, Jones EG, Bunney WE, Myers RM, Speed TP, Akil H et al. 2005. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**(20): e175.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**(10): R80.

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**(Database issue): D154-158.

Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. 2003. Identification of pseudogenes in the Drosophila melanogaster genome. *Nucleic Acids Res.* **31**(3): 1033-1037.

Herbert JM, Stekel D, Sanderson S, Heath VL, Bicknell R. 2008. A novel method of differential gene expression analysis using multiple cDNA libraries applied to the identification of tumour endothelial genes. *BMC Genomics* **9**: 153.

Katoh K, Toh H. 2008. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**(4): 286-298.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res.* **12**(4): 656-664.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. U. S. A.* **100**(20): 11484-11489.

Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A et al. 2007. The UCSC genome browser database: update 2007. *Nucleic Acids Res.* **35**(Database issue): D668-673.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proc. Natl. Acad. Sci. U. S. A.* **103**(26): 9935-9939.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in Drosophila. *Nat. Genet.* **40**(3): 351-355.

Petrov DA, Hartl DL. 1998. High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol. Biol. Evol.* **15**(3): 293-302.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.* **17**(12): 1850-1864.

Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C, Valente VLS, Aguade M, Anderson WW. 2008. Polytene Chromosomal Maps of 11 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and Physical Maps. *Genetics* **179**(3): 1601.

Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W. 2003. Human-Mouse Alignments with BLASTZ. *Genome Res.* **13**(1): 103-107.

Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**: Article3.

Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E. 2004. The Ensembl Core Software Libraries. *Genome Res.* **14**(5): 929.

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H et al. 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* **12**(10): 1611-1618.

Sturgill D, Zhang Y, Parisi M, Oliver B. 2007. Demasculinization of X chromosomes in the Drosophila genus. *Nature* **450**(7167): 238.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**(Web Server issue): W609-612.

Team RDC. 2007. R: A Language and Environment for Statistical Computing. http://www.R-project.org.

Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the Drosophila genus. *Genome Res.* **19**(5): 897-903.

Yang S, Arguello JR, Li X, Ding Y, Zhou Q, Chen Y, Zhang Y, Zhao R, Brunet F, Peng L. 2008. Repetitive Element-Mediated Recombination as a Mechanism for New Gene Origination in Drosophila. *PLoS Genet.* **4**(1): e3.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in Drosophila. *Genome Res* **18**(9): 1446-1455.

Supplemental Tables

Table 1: ***D. melanogaster* gene branch assignment and expression profile.** Branch assignment follows the convention of Figure 1. "tissue_number" indicates the number of samples out of 12 FlyAtlas samples, where the gene of interest is transcribed. "testis_value" and "ovary_value" indicate the expression intensity in testis and ovary based on FlyAtlas data, respectively. "adj_pvalue" indicates the FDR controlled pvalue generated by LIMMA while comparing testis and ovary expression value. Since all probes mapping to multiple genomic locations were filtered, some genes might have no probes and thus had no expression value. In this case, the expression is shown as "NA".

Table 2A: **Classification of new gene origination for 947 young genes.** “child_id” and “parent_id” indicate the young gene and potential parental gene, respectively. “gene_type” includes five categories, “A”, “D”, “Dl”, “R” and “RL”, which respectively corresponds to “De novo”, “DNA-based duplication”, “DNA-based duplication like”, “Retroposition” and “Retroposition like” (for details, please refer to Supplementary methods). “m_type” could be “D”, “T” and “M”, which indicate “within-chromosome dispersed”, “within-chromosome tandem” and “between-chromosome movement”, respectively. The “note” column summarizes the alignment information between the parental gene and the child gene. For example, “1:7:1:7;0:chr3R:8062716-8064928:8060468-8062716;0.591:1:1e-177;Self” is coded as follows: the first four numbers indicate the alignment region ranging from the first to the seventh exon of both child gene and parental genes; the fifth number (for within-chromosome duplicates) shows how many genes exist between the parental gene and the child gene; the next section, “chr3R:8062716-8064928:8060468-8062716” shows the chromosomal coordinates for both child gene and parental gene; “0.591:1:1e-177” gives the alignment identity, alignment coverage for the child gene and the alignment Evalue; finally, “Self” indicates self-chained genome alignment covering the block between this parent gene

and child gene. For *de novo* genes or “A” category, the last three columns, “gene_type”, “m_type” and “note” are blank.

Table 2B: **High quality DNA-level movement.** Table 2B is a subset of Table 2A. In this table, only DNA-level movements fitting the following criteria are shown: 1) Parental gene originated from an older branch compared to child gene; 2) The pairwise alignment between parental gene and child gene is of high quality including overall protein identity not less than 50% and coverage not less than 70% for the parental gene by following (Bai *et al.* 2007). In other words, for Table 2A, we identified the most similar paralog to the candidate parental gene for DNA-level or retroposition-level young genes. However, this relationship might not be that robust for many cases.

Table 3: **163 inter-chromosomal DNA-level gene duplication statistics.** The convention follows Table 1 in the main text.

Table 4A: **MK-table based on *D. melanogaster* 7MB data.** "seq_number" corresponds to the total number of sequences in the alignment with *D. melanogaster* individuals and one *D. simulans* outgroup sequence. The maximum is 51. It is possible that the seq_number is lower than 51 since some low quality sequences were purged. "seq_length" is the length of the open reading frame. "ds", "ps", "dn" and "pn" indicate the synonymous divergence, synonymous polymorphism, non-synonymous divergence and non-synonymous polymorphism, respectively. "ln" and "ls" indicate the total number of non-synonymous sites and synonymous sites. "fet_pvalue" and "g_pvalue" indicate p for both the Fisher Exact Test and G-test. "alpha" indicates the proportion of positive selection for "neutrality" defined as 1-alpha. NaN indicates not applicable. "name" is the accession for the representative transcript (the transcript with the longest coding sequence).

Table 4B: **MK-table based on *D. simulans* six-strain data.** "seq_number" is the total

number of sequences in the alignment with *D. melanogaster* individuals and one *D. simulans* outgroup sequence. The maximum should be 7. All the other columns are the labeled the same way as Table S2.

Table 5: **Analysis based on DoFE (Distribution of Fitness Effects) package.**

S5a shows the maximum-likelihood estimation of α for different groups of genes. The highest α is marked in red for both polymorphism datasets.

[1] DoFE gives a Maximum-likelihood estimation of α.

[2] Gene number covered by this 7MB data or *D. simulans* data in different categories. No new X-linked female-biased genes were covered by *D. melanogaster* 7MB data.

[3] LikeLihood Ratio (LLR) test shows that the estimated α is significantly different compared to the neutral estimation, *i.e.*, α of 0.

[4] Ns, not significant.

[5] For six strain data, the α does not change a lot after removing *de novo* young genes.

S5B shows LLR tests between groups. The comparison between male-biased genes of X chromosome and female-biased genes of 2L was skipped due to the small number of female-biased genes of 2L.

Herein, we used two polymorphism datasets: one generated by DPGP, 7MB array-reseqencing data of X chromosome and chromosome 2L across 50 *D. melanogaster* strains; and the second consisting of the whole genome sequencing data across six *D. simulans* strains. These datasets are complementary: the former covers many more strains but many fewer genes (for example only three X-linked young genes), while the later covers many fewer strains but many more genes. Our conclusion is robust since it is consistent across the two datasets.

Table 6: ***D. pseudoobscura* gene branch assignment and expression bias.** “id” and

“chrom” indicate the FlyBase ID and the chromosome. The expression “bias” information is extracted from (Sturgill *et al.* 2007) . “Branch” categorizes three groups, *D. pseudoobscura* specific genes, obscura group genes and the other old genes. “GleanR ID” corresponds to the original gene prediction ID (Clark *et al.* 2007).

Table 7: **miRNA expression profiling based on short-sequencing data.** "seed-conservation" is described as (*D. simulans*, *D. sechellia*)-(*D. yakuba*, *D. erecta*)-(*D. ananassae*)-(*D. persimilis*, *D. pseudoobscura*). The number of each column in conservation code is minimized pairwise nucleotide difference between species in each major branch and *D. melanogaster*. "na" means the miRNA does not exist in that species. For each miRNA, the count of supporting reads is presented for all 14 samples.

Table 8: **Sex specific expression of miRNAs.** “Old” and “new” stand for miRNAs that originated before or after the divergence between *D. melanogaster* and *D. pseudoobscura*, respectively. “Fisher-test *p*” was given for contingency tables tests comparing male-specific and non male-specific gene proportions between new and old genes.

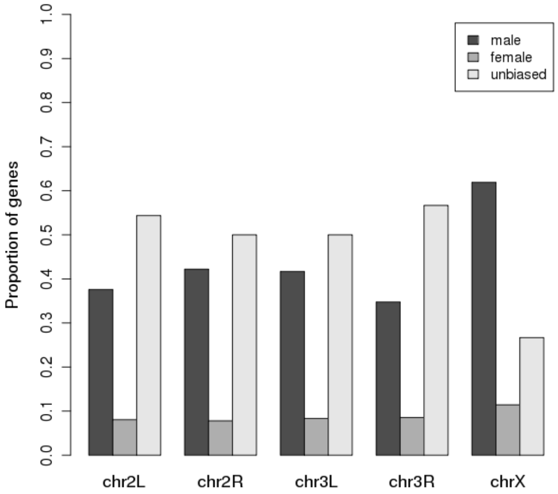
Table 9: **Pseudogenization, translocation and duplication involving young parental genes.**

A. *D. simulans/D. sechellia/D. melanogaster* specific genes (branch 5), which already show some degeneration in either *D. simulans* or *D. sechellia*. “Indel+Stop” indicates the number of disabled genes, while “annotation” indicates whether FlyBase annotates one ortholog for the gene of interest. The conceptual alignment generated by Genewise based on proteins in *D. melanogaster* is shown on the left with "!" indicating stop codons or frame-shifts indels.
B. Translocations for *D. melanogaster* new genes. The initial two letters indicate

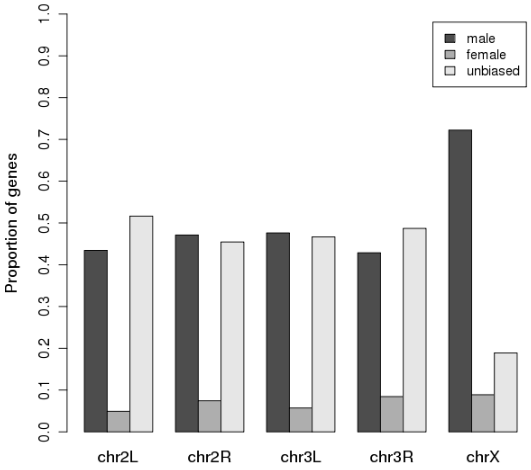
the species. For example, “me” indicates “melanogaster”. “?” means the corresponding contig has not been assigned to chromosomal arms. “LA” indicates loss of annotation where Flybase does not annotate a gene in this species. In case of CG11262, it occurred in the ancestor of *D. melanogaster* and *D. ananassae*. “Bias” indicates the expression bias of a *D. melanogaster* gene.

C. Two recent duplicates. We calculated evalue, identity and coverage by summarizing and parsing BLAST’s local alignment (Altschul *et al.* 1997) with BioPerl chained-BLAST module (Stajich *et al*. 2002).

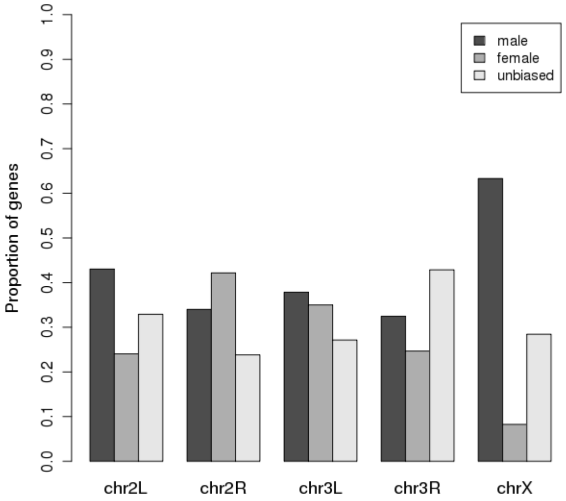
Figure 1: Chromosomal distribution of genes with respect to their expression bias.



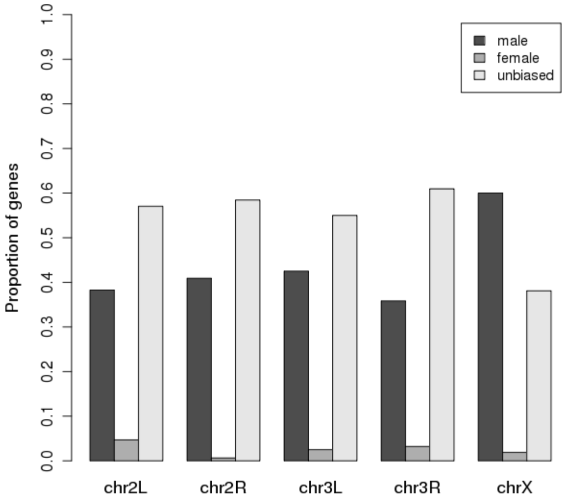
1A: **Proportion of young sex-biased genes (branches 1 to 6) on chromosome arms (chr).**



1B: **Proportion of young sex-biased genes (branches 2 to 5) on chromosome arms (chr).**



1C: **Proportion of young sex-biased genes (branches 1 to 6) with peptide evidence on chromosome arms (chr).** Peptides mapped to multiple genomic locations were discarded.



1D: **Proportion of young sex-biased genes (branches 1 to 6) on chromosome arms (chr).** Expression bias was defined as sex-specific expression. Specifically, we identified new genes with exclusive presence call in all four microarray replicates of testis or ovary as male-specific or female-specific genes, respectively.

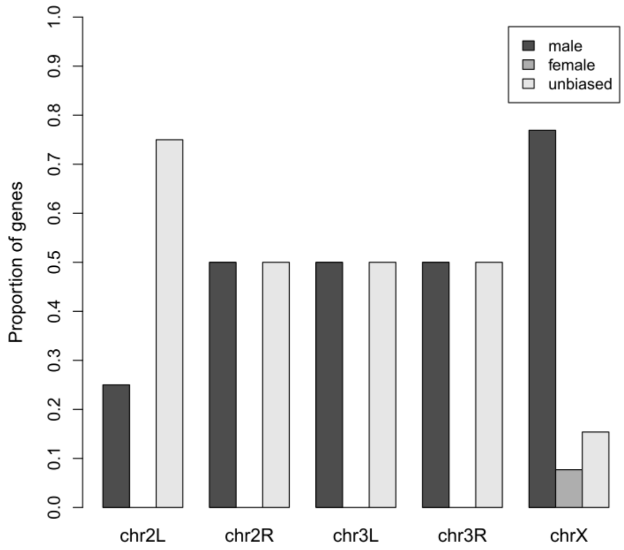


Figure 2: **Proportion of young DNA-level duplicated sex-biased genes (branches 5 to 6) on chromosome arms (chr).** This figure depicts the distribution of the DNA-level duplicates, and complements Figure 2B that includes both the DNA level duplicates and the de novo genes.

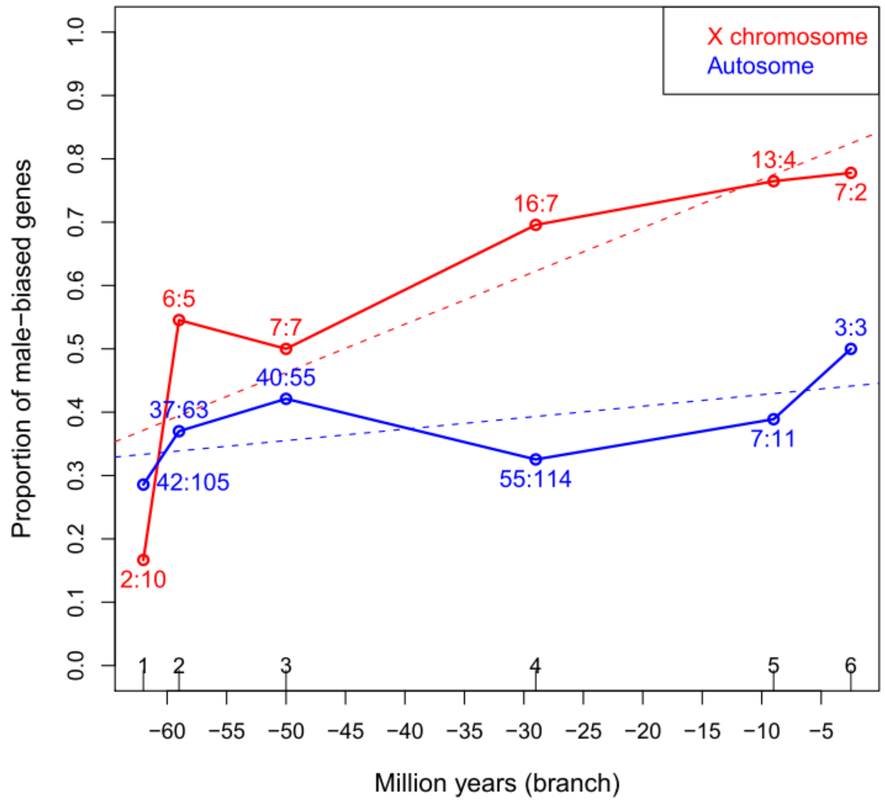


Figure 3: **The distribution of the ages and proportion of the male-biased genes that originated via DNA-level duplication only.** This is a supplement to Figure 3 of the main text that includes the genes that originated from both DNA-level duplication and *de novo* origination.

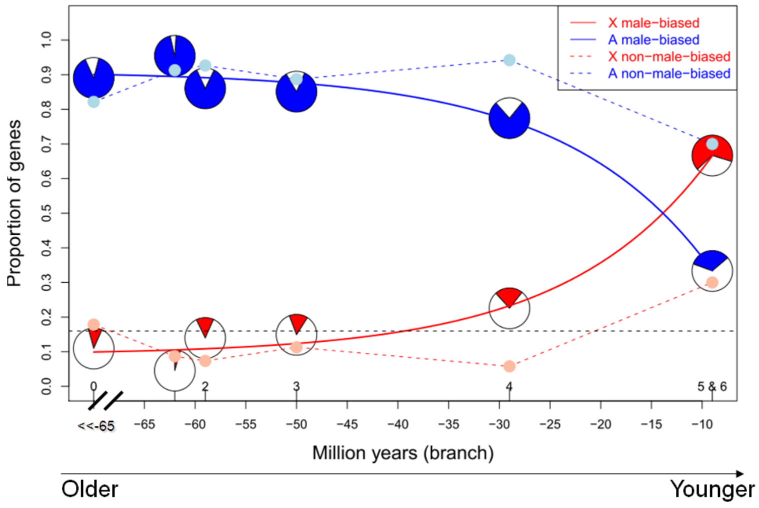


Figure 4: **The distribution of the ages and the proportion of the genes that originated via DNA-level duplication only.** This is a supplement to Figure 4 of the main text that includes the genes that originated from both DNA-level duplication and *de novo* origination.