Supplemental Data S1.
The detailed comparison between our identified new genes and other works' predictions (Bai et al. 2007; Bhutkar et al. 2007; Clark et al. 2007; Hahn et al. 2007; Heger and Ponting 2007)

We compared our results with other recent 5 genome-wide studies of *Drosophila* paralog/orthologs (Bai et al. 2007; Bhutkar et al. 2007; Clark et al. 2007; Hahn et al. 2007; Heger and Ponting 2007). Basically, the major differences among the results are due to different analytical pipelines and database versions. We have provided the comparison results in a gene-by-gene fashion. We also explain the differences, if any, as follows:

1)Heger and Ponting, Evolutionary rate analysis of orthologs and paralogs from 12 *Drosophila* genomes *Genome Res.* 2007 Dec;17(12):1837-49.
This paper studied evolutionary patterns of orthologs and paralogs across 12 *Drosophila* species. The authors found gene duplications in terminal lineages are strongly skewed toward very recent events and proposed a rapid-birth and infrequent-fix model for recent duplications. Consistently, we found more new genes in *D. melanogaster* or *D. yakuba* ('terminal lineage'), compared with new gene number in the *D. melanogaster* species complex. And lots of *D. melanogaster*-specific new genes show copy number polymorphisms according to our PCR assays. Thus our major conclusions are compatible. To be specific, the authors didn't describe exact lineage specific duplication numbers in *D. melanogaster* and the *D. melanogaster* species complex while they mentioned that they identified 200 *D. yakuba*-specific gene duplications (Table 2 in that paper). Similarly, we identified 177 *D. yakuba*-specific new genes in our work, although we could not compare gene by gene because they didn't provide gene list. The difference is probably caused by a more stringent criteria when we assigning orthologous relationship between species. We mapped a putative ortholog by both sequences similarity and syntenic relationship, while Heger and Ponting mapped orthologs first by sequence similarities and then validated them by syntenic relationships. It appears that they didn't remove non-syntenic orthologs at last (p.1842 in that paper: …nonsyntenic ortholog assignments dropped from 629 to 66 for the pair of *D. melanogaster* and *D. simulans* assemblies). Another difference is that they identified 94 'orphan' genes compared to our 2 *D. melanogaster* specific '*de novo*' genes. As mentioned before in the Materials and Methods , some of these 'orphan' genes actually result from gene duplication followed by fast evolution. Such cases can be syntenicly mapped by BLASTZ rather than BLAST. Also, a newer version of gene annotation used by Heger and Ponting (v.4.2.1) (Grumbling and Strelets 2006) and our work (v.4.2, Sep. 2005, http://genome.ucsc.edu/) should also have effect. Finally, the author found vast majority of gene duplications occur within single Muller elements, i.e., X chromosome in *D. melanogaster* and autosome 3 in *D. yakuba* (Figure 5 and Figure S12 of its Supplemental File in that paper). This result is consistent with our analysis on chromosomal distribution of new genes.

2)Hahn et al. Gene family evolution across 12 *Drosophila* species *PLoS Genet*. 2007. 3(11):e197

This paper used two approaches to study gene family expansion/contraction across 12 *Drosophila* species, i.e., maximum likelihood approach and nonparametric gene tree/species tree reconciliation approach. As discussed in the text, our estimation of new gene origination rate is consistent with but slightly lower than Hahn et al.'s estimate using maximum likelihood approach (0.000391-0.000925 vs. 0.0010 per gene per million years). This slight difference results from our different methods for estimating the rate, which is more close to their second approach. Consistently, they inferred a minimum estimate of 77 *D. melanogaster*-specific duplications with the gene/species tree reconciliation approach, while we have identified 72 *D. melanogaster*-specific new genes Again, this difference may be caused by more stringent criteria of requiring for syntenic relationship of each gene in our work.

3) Bhutkar et al. Genome-scale analysis of positionally relocated genes Genome Res. 2007. 17(12):1880-7.

This paper studied relocated genes between different chromosome arms to address the genomic changes in *Drosophila* species. Such relocation, i.e., break of syntenic relationship can be caused either by origination of a new gene on a different chromosome arm or simply a relocation of a preexisting gene. To be specific, we compared our results in detail with relocated gene identified in this work (please see Figure 2A and Supplemental File of that paper):

a. There are three *D. melanogaster*-specific relocated genes pairs: *CG2033-CG12324* and *CG33213-CG33221* are contained in our dataset, they are created by lineage-specific duplication/retroposition. *CG9140-CG11423*, this pair is not contained in our result and they located on Muller element B and C in *D. melanogaster*. Their curated ortholog *GA21571-GA10997* also located on Muller element B and C in *D. pseudoobscura* according to flybase annotation (www.flybase.org). So this pair maybe a questionable relocated gene pair specific to *D. melanogaster*.
b. There are three *D. yakuba*-specific relocated genes: *CG13902, CG13888* and *CG13889*. All of them have curated orthologs (www.flybase.org) in other *Drosophila* species, so they are not considered as new genes in our work.
c. There are two *D. melanogaster* species complex-specific relocated genes: *CG30354* and *CG9068*. They also have curated orthologs in other species and so not included in our study either.

We do find some false negative results of Bhutkar et al.'s work. For example, the well characterized retrogene *sphinx* (*CG11091*, on Muller F) (Wang et al. 2002) originated specifically in *D. melanogaster* should be characterized as a *D. melanogaster*-specific relocated gene from Muller C and it was not included in Bhutkar et al.'s result.

4) Bai et al. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila Genome Biol.* 2007;8(1):R11.

This paper has studied retrogenes at different time nodes across the phylogeny of 12 *Drosophila* species. They identified one *D. melanogaster*-specific and five the *D. melanogaster* species complex-specific new retrogenes (Figure 1 and Additional Data File 1 in that paper). But we identified five and six new retrogenes in the corresponding

lineages (Supplemental Table S1 and S2). We would thus have a slightly different estimate of origination rate of retrogenes in the *D. melanogaster* species complex. Besides the overlapped results, we identified the additional following genes as new retrogenes:

a. *CG11091* (*sphinx*): this new retrogene specific to *D. melanogaster* has been well characterized in (Wang et al. 2002).
b. *CR12628*: this retroposed RNA gene specific to *D. melanogaster* was firstly reported in (Toba and Aigaki 2000) and then in (Betran et al. 2002). It is not included in Bai et al.'s work probably because it is termed with a 'CR-id' as a non-coding gene. Although it doesn't contain intact open reading frames, there's evidence showing this gene expresses over 10-24 hours in early developmental stages of *D. melanogaster* (Manak et al. 2006) but its flanking regions don't show this pattern. Therefore, we suggest to keep this gene in the dataset to provide interested readers for further studies.
c. *CG32733* and *CG32797*: they are both generated from the gene *CG9821*. These two genes contain no introns and no curated orthologs according to the annotations of Flybase (www.flybase.org). Blast results show that both their coding sequences have mapped to 3' part of *CG9821*'s mRNA. These data indicate they should be new retrogenes.
d. *CG11235* and *CG7804*: Both of these two new genes contain introns, that's maybe why they are removed in the above paper. However, a detailed BLAST comparison using their introns showed they are homologous to their parental genes' coding regions (data not shown). Thus it is likely that new splicing sites and introns originated after the retroposition.

5) Clark et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 450: 203-218

We manually scrutinized our results and predicted orthologs across 12 *Drosophila* species on the Gbrowse site of flybase (http://www.flybase.org/cgi-bin/gbrowse/), where independent ortholog predictions from 8 groups are provided. We took the GLEANR consensus orthology dataset (ftp://ftp.flybase.net/genomes/) published along with this 12-*Drosophila*-genome paper (Clark et al. 2007) as the reference, and found 73.8% (48 out of 65, 7 genes haven't been annotated in GLEANR) *D. melanogaster* new genes are compatible with the GLEANR predictions. 17 non-compatible genes have same or higher gene copy numbers in at least one of the other 11 *Drosophila* species compared with *D. melanogaster*. However, a detailed check on predicted gene copies in other species found all of the 16 (except those of *CG30160* with an independent new gene origination in *D. grimshawi*) don't show a conserved syntenic relationship with the new gene in *D. melanogaster*. There are several explanations for the discrepancies: First, independent gene origination events occurred in other species. For example, *CG7046* are predicted to have two copies in *D. melanogaster* and *D. pseudoobscura* and one or no copy in all other species. An origination before the split of these two species following subsequent gene losses in several *Drosophila* species plus a relocation in *D.*

pseudoobscura should be much less likely than simply an independent origination event in *D. pseudoobscura*. Second, there are relocations breaking the synteny after the origination. This could explain the same gene copy number in *D. melanogaster* and its closely-related species (*D. simulans* or *D. sechellia*) such as the case of *CG32789*. In such cases, we could not discriminate it with the fist explanation. Since all of such cases show a lower copy number in *D. yakuba* and *D. ananassae*, they are still new genes originated either in *D. melanogaster* or the ancestor of the *D. melanogaster* species complex. At last, because of the different sequence qualities of 12 *Drosophila* species, there could be assembly/annotation artifacts. For example, there are two predicted orthologs in *D. ananassae* (dana_GLEANR_18790, dana_GLEANR_18791) corresponding to the *CG12819* (parental gene)-*CG12592* (new gene) pair in *D. melanogaster*. A blast check on protein sequences of dana_GLEANR_18790 and dana_GLEANR_18791 found they actually map to different part of *CG12819*, the parental gene. Also these two copies are predicted to be one single gene from other pipelines (Eisen group's pipeline) on Gbrowse of flybase. Also, we found several predicted orthologs (e.g. orthologs of *CG32788* and *CG12592*) locate in a short scaffold (<10k) without any flanking genes. They may represent sequence assembly redundancies. In summary, these differences are generated by different criteria and pipelines adopted by us and other groups. We speculate similar phenomenon could occur in the new gene datasets in the *D. melanogaster* species complex and *D. yakuba*. If needed, more details specific to each non-compatible gene and its predicted orthologs in other *Drosophila* species could be provided.

References

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

Betran, E., Thornton, K., and Long, M. 2002. Retroposed new genes out of the X in Drosophila. *Genome Res* **12:** 1854-1859.

Bhutkar, A., Russo, S.M., Smith, T.F., and Gelbart, W.M. 2007. Genome-scale analysis of positionally relocated genes. *Genome Res* **17:** 1880-1887.

Clark, A.G. Eisen, M.B. Smith, D.R. Bergman, C.M. Oliver, B. Markow, T.A. Kaufman, T.C. Kellis, M. Gelbart, W. Iyer, V.N. et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450:** 203-218.

Grumbling, G. and Strelets, V. 2006. FlyBase: anatomical data, images and queries. *Nucleic Acids Res* **34:** D484-488.

Hahn, M.W., Han, M.V., and Han, S.-G. 2007. Gene family evolution across 12 Drosophila genomes. *PLoS Genet* **3:** e197.

Heger, A. and Ponting, C.P. 2007. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes. *Genome Res.***:** gr.6249707.

Manak, J.R., Dike, S., Sementchenko, V., Kapranov, P., Biemar, F., Long, J., Cheng, J., Bell, I., Ghosh, S., Piccolboni, A. et al. 2006. Biological function of unannotated transcription during the early development of Drosophila melanogaster. *Nat Genet* **38:** 1151-1158.

Toba, G. and Aigaki, T. 2000. Disruption of the microsomal glutathione S-transferase-like gene reduces life span of Drosophila melanogaster. *Gene* **253:** 179-187.

Wang, W., Brunet, F.G., Nevo, E., and Long, M. 2002. Origin of sphinx, a young chimeric RNA gene in Drosophila melanogaster. *Proc Natl Acad Sci U S A* **99:** 4448-4453.